LESSON BOOK

Distributed Energy Resources as Grid Resources

From DER fundamentals to a worked virtual power plant

Eighteen lessons · 308 pages

Ricardo Marquez, Ph.D.

ricardo-marquez.com

Distributed Energy Resources as Grid Resources

From device physics to virtual power plants

Track B lessons, written up as a book

Ricardo Marquez

October 5, 2026 (work in progress: one chapter per lesson, added as the lessons proceed)

How to read this book

Each chapter of this book expands one lesson of "Track B", a sequence that follows the chain

DER fundamentals → demand response → DERMS → virtual power plants → thermostats, EVs and batteries → fleet aggregation → dispatch → dynamic load shaping → grid services.

The emphasis is on the *physical and operational logic*, not on vocabulary. Where the industry uses a shorthand term, a *Jargon* box expands it once and the text then uses the expanded meaning. Every symbol is defined in the notation chapter that follows and is used with the same meaning in every chapter.

Each chapter closes with a short section that connects the lesson to the TERRA framework, a graph-based way of modeling physical systems as conservation laws on nodes with flows on edges. That framework is a convenient mental model for everything in this book, because a distributed energy resource is, at bottom, a device that exchanges a conserved quantity (electrical energy, thermal energy) across a boundary under constraints.

Contents

How to read this book **ii**

Notation and vocabulary **viii**

1 What makes a DER a grid resource? **1**
1.1 Devices and resources . . . 1
1.2 The device as a dynamical system . . . 1
1.3 Flexibility as a quantity . . . 4
1.4 Constraints and uncertainty . . . 7
1.5 From one device to a fleet . . . 9
1.6 The working definition . . . 11
1.7 Exercises . . . 12
1.8 Connection to the TERRA framework . . . 12

2 Demand response and demand flexibility **14**
2.1 Demand response as a subset of demand flexibility . . . 14
2.2 The demand-response event . . . 14
2.3 The physics of an event . . . 16
2.4 Demand flexibility . . . 17
2.5 Objectives and decisions . . . 19
2.6 From events to continuous operation . . . 20
2.7 The key distinction . . . 22
2.8 Exercises . . . 23
2.9 Connection to the TERRA framework . . . 23

3 What a DERMS does **25**
3.1 Definition and the operational chain . . . 25
3.2 Sensing the fleet . . . 25
3.3 Aggregation . . . 28
3.4 Decision and action . . . 30
3.5 The DERMS as a control system . . . 31
3.6 Architecture and neighbors . . . 32
3.7 The working definition . . . 34
3.8 Exercises . . . 34
3.9 Connection to the TERRA framework . . . 35

4 Virtual power plants **37**
4.1 Many devices and one resource . . . 37
4.2 Capacity . . . 37
4.3 The service product . . . 39
4.4 The VPP as a resource abstraction . . . 42

4.5 Commitment under uncertainty . . . 43
4.6 Operation and forecasting . . . 45
4.7 The working definition . . . 46
4.8 Exercises . . . 47
4.9 Connection to the TERRA framework . . . 47

5 Thermostatically controlled loads 49
5.1 The question . . . 49
5.2 The building and the thermostat . . . 49
5.3 What control does . . . 52
5.4 The state-space model . . . 54
5.5 Fleet behavior . . . 55
5.6 Forecasting and dispatch . . . 57
5.7 What to retain . . . 59
5.8 Exercises . . . 59
5.9 Connection to the TERRA framework . . . 60

6 Batteries as DERs 61
6.1 The question . . . 61
6.2 State and dynamics . . . 61
6.3 Constraints . . . 62
6.4 Dispatch . . . 65
6.5 Fleet and forecasting . . . 66
6.6 Comparison and control formulation . . . 69
6.7 What to retain . . . 70
6.8 Exercises . . . 71
6.9 Connection to the TERRA framework . . . 71

7 EV managed charging 72
7.1 The question . . . 72
7.2 One vehicle's night . . . 72
7.3 The model . . . 73
7.4 Scheduling . . . 74
7.5 Direction, forecasting and risk . . . 76
7.6 Vehicle-to-grid and the fleet . . . 78
7.7 What to retain . . . 81
7.8 Exercises . . . 81
7.9 Connection to the TERRA framework . . . 81

8 Baselines and verification 83
8.1 The counterfactual problem . . . 83
8.2 Baseline methods . . . 86
8.3 Verification metrics . . . 89
8.4 Attributing underperformance . . . 91
8.5 Probabilistic and experimental verification . . . 93
8.6 Verification inside the learning loop . . . 96
8.7 What to retain . . . 98
8.8 Exercises . . . 98
8.9 Connection to the TERRA framework . . . 99

9 Fleet aggregation **101**
9.1 The aggregation problem . . . 101
9.2 Representations of a fleet . . . 103
9.3 Diversity and common-mode uncertainty . . . 106
9.4 Location, service and hierarchy . . . 109
9.5 Probabilistic aggregation . . . 111
9.6 The mathematical object . . . 114
9.7 What to retain . . . 116
9.8 Exercises . . . 117
9.9 Connection to the TERRA framework . . . 117

10 Forecasting flexibility **119**
10.1 Load forecasting and flexibility forecasting . . . 119
10.2 Resource-specific forecasts . . . 122
10.3 Probabilistic forecast targets . . . 124
10.4 Features and realizable flexibility . . . 128
10.5 Endogeneity of the forecast target . . . 129
10.6 The forecast target and the interface to optimization . . . 132
10.7 What to retain . . . 133
10.8 Exercises . . . 133
10.9 Connection to the TERRA framework . . . 135

11 Dispatch and optimization **137**
11.1 The dispatch problem . . . 137
11.2 Costs and constraints . . . 139
11.3 Horizon and model predictive control . . . 142
11.4 Dispatch under uncertainty . . . 144
11.5 Coordinating heterogeneous resources . . . 148
11.6 Economics and the value of a forecast . . . 151
11.7 What to retain . . . 153
11.8 Exercises . . . 153
11.9 Connection to the TERRA framework . . . 154

12 Dynamic load shaping **156**
12.1 From events to trajectories . . . 156
12.2 Cohort rotation, rebound and pre-conditioning . . . 159
12.3 Shaping as tracking control . . . 162
12.4 Shapes and services . . . 164
12.5 Feasibility over the horizon and feedback . . . 167
12.6 Objective, trade-offs and controlled recovery . . . 169
12.7 What to retain . . . 172
12.8 Exercises . . . 172
12.9 Connection to the TERRA framework . . . 173

13 Distribution-grid constraints **175**
13.1 System-wide and local flexibility . . . 175
13.2 Equipment and voltage limits . . . 178
13.3 Network-aware dispatch . . . 182
13.4 Location in aggregation and in the merit order . . . 185
13.5 A worked example with uncertainty . . . 187
13.6 Topology, the ADMS, and planning . . . 190
13.7 What to retain . . . 192

13.8 Exercises . . . 193
13.9 Connection to the TERRA framework . . . 193

14 Grid services and markets **195**
14.1 Products, not megawatts . . . 195
14.2 Wholesale products . . . 197
14.3 Demand response, load shaping, and locational services . . . 200
14.4 Value stacking and co-optimization . . . 203
14.5 Requirements differ by service . . . 206
14.6 Commitment economics . . . 208
14.7 What to retain . . . 211
14.8 Exercises . . . 212
14.9 Connection to the TERRA framework . . . 212

15 Customer and device constraints **214**
15.1 Purpose and the intersection of constraints . . . 214
15.2 Customer constraints . . . 216
15.3 Device constraints . . . 220
15.4 Program rules and the economics of an event . . . 222
15.5 Heterogeneity, fairness and participation . . . 224
15.6 The feasible resource . . . 227
15.7 What to retain . . . 230
15.8 Exercises . . . 230
15.9 Connection to the TERRA framework . . . 231

16 DERMS and VPP architecture **232**
16.1 The layered architecture . . . 232
16.2 Sensing: device integration, telemetry and state . . . 235
16.3 Forecasting and aggregation layers . . . 236
16.4 Decision, actuation and monitoring . . . 238
16.5 Verification, learning and cadence . . . 240
16.6 Interfaces, robustness and uncertainty . . . 242
16.7 What to retain . . . 247
16.8 Exercises . . . 248
16.9 Connection to the TERRA framework . . . 248

17 A commercial platform: the architecture in practice **250**
17.1 The platform and its value chain . . . 250
17.2 Telemetry and DER-specific forecasting . . . 252
17.3 From demand response to dynamic load shaping . . . 255
17.4 Objectives, targets and resource classes . . . 258
17.5 Scale and the platform architecture . . . 260
17.6 The forecasting and optimization role . . . 261
17.7 What to retain . . . 266
17.8 Exercises . . . 266
17.9 Connection to the TERRA framework . . . 268

18 A worked case: design and operate a VPP **269**
18.1 The case and the service definition . . . 269
18.2 Resource models . . . 271
18.3 Finding a feasible plan . . . 274
18.4 Uncertainty, reserves and real-time operation . . . 278

18.5 Attribution and verification 283
18.6 Presenting the reasoning 285
18.7 What to retain 289
18.8 Exercises 290
18.9 Connection to the TERRA framework 291

Notation and vocabulary

The same symbols are used throughout. Power is in kilowatts (kW) or megawatts (MW); energy is in kilowatt-hours (kWh); temperature is in degrees Celsius; time t is a clock time or a discrete step index, and τ is a duration.

Symbols

Symbol	Unit	Meaning
$P_{\text{grid}}(t)$	kW	Net power drawn from the grid at the point where the premises meet the network. Positive means import, negative means export.
$P_{\text{load}}(t)$	kW	Ordinary, uncontrolled consumption of the premises.
$P_{\text{EV}}(t)$	kW	Electric-vehicle charging power (positive when charging).
$P_{\text{bat}}(t)$	kW	Battery power; positive when charging, negative when discharging.
$P_{\text{PV}}(t)$	kW	Solar photovoltaic generation (always ≥ 0).
$P_{\text{AC}}(t)$	kW	Electrical power drawn by an air conditioner or heat pump.
x_t	–	Internal *state* of a device at step t (what the device "remembers").
u_t	–	*Control action* applied at step t (what we are allowed to choose).
w_t	–	*Disturbance*: an external condition we do not choose (weather, occupancy, driving).
$\mathcal{U}(x_t)$	–	The set of control actions that are allowed when the state is x_t.
$f(x, u, w)$	–	State-transition map: $x_{t+1} = f(x_t, u_t, w_t)$.
$g(x, u, w)$	kW	Output map: $P_{\text{grid}} = g(x, u, w)$.
$P^{\max}$	kW	Nameplate (rated) power of a device.
$P^{\text{avail}}(x)$	kW	Power actually available right now, given the state.
E^{avail}	kWh	Energy available to be delivered (or absorbed) before a limit is hit.
P_t^{base}	kW	Baseline: the power the premises *would have* drawn without control.
P_t^{ctrl}	kW	Power actually drawn under control.
$\Delta P_t = P_t^{\text{base}} - P_t^{\text{ctrl}}$	kW	Load reduction delivered by control at step t.
$F(t, \tau)$	kW	Flexibility: the sustainable load change if a response starts at t and must last τ.

Symbol	Unit	Meaning
SOC	% or –	State of charge of a battery (fraction of usable energy stored).
$\mathrm{SOC}^{\mathrm{res}}$	%	Reserve the owner insists on keeping (backup, warranty).
$\mathrm{SOC}^{\mathrm{req}}$	%	State of charge the EV driver requires at departure.
t_{dep}	time	Departure time of an EV.
$T_{\mathrm{in}}, T_{\mathrm{out}}, T_{\mathrm{set}}$	°C	Indoor, outdoor and setpoint temperatures.
$T^{\min}, T^{\max}$	°C	Customer comfort band on T_{in}.
$P^{\mathrm{act}}(t)$	kW, MW	Actual, metered load during an event.
$P^{\mathrm{tgt}}(t)$	MW	Target trajectory the fleet is asked to follow.
$P^{\mathrm{del}}(t)$	MW	Load change actually delivered by the fleet.
$P_{\mathrm{fleet}}(t)$	MW	Net power of the whole fleet.
$\hat{P}$	MW	A forecast of a power quantity (hat denotes an estimate).
$E^{\max}$	kWh	Usable energy capacity of a battery.
η	–	Charging efficiency (fraction of drawn energy that is stored).
c_t	\$/kWh	Price (or marginal cost) of energy at step t.
J	–	Objective function minimized by a dispatch optimizer.
e_t	MW	Tracking error, $P^{\mathrm{tgt}} - P^{\mathrm{del}}$, at step t.
H	steps	Optimization horizon (number of steps looked ahead).
y_t	–	Telemetry: what the platform *observes* at step t.
$h(x)$	–	Measurement map, $y_t = h(x_t) + v_t$.
v_t	–	Measurement noise.
$F_i(t)$	kW	Flexibility of device i; $F(r, t, d)$ indexes by region r and type d.
ρ	–	Correlation coefficient between device flexibilities.
a_i	–	Availability factor of device i, between 0 and 1.
P^{name}	MW	Nameplate (installed) capacity of a fleet, $\sum_i P_i^{\max}$.
P^{commit}	MW	Capacity committed to a utility or market.
P^{enroll}	MW	Nominal flexibility of everything enrolled.
$P_{\mathrm{VPP}}(t)$	MW	Power of the virtual power plant as seen by the grid.
$C(\tau)$	MW	Deliverable capacity for a required duration τ.
p^{ovr}	–	Probability that a customer overrides an event.
$z = \phi(x)$	–	Reduced representation of the fleet state handed to the grid.
R, C	°C/kW, kWh/°C	Thermal resistance and capacitance of a building; RC is its time constant.
Q_{int}, Q_{HVAC}	kW	Internal heat gains; heat removed by the HVAC system.
s_t	–	Compressor state at step t, on (1) or off (0).
D	–	Duty cycle: fraction of time the compressor is on.
$\mathrm{SOC}^{\mathrm{th}}$	–	Thermal state of charge: position of T_{in} inside the comfort band.
E^{shift}	kWh	Energy shifted by an event, $\int \Delta P(t)\,\mathrm{d}t$.

Symbol	Unit	Meaning
η_{c}, η_{d}, η_{RT}	–	Charging, discharging and round-trip efficiency, $\eta_{\mathrm{RT}} = \eta_{\mathrm{c}}\,\eta_{\mathrm{d}}$.
$\mathrm{SOC}^{\min}$, $\mathrm{SOC}^{\max}$	–	Operating limits on state of charge (protection, degradation, reserve).
P^{ch}, P^{dis}	kW	Charging and discharging power limits (magnitudes).
$F^{\uparrow}$, $F^{\downarrow}$	kW	Upward grid support (reduce net load / inject) and downward support (increase net load).
a_t	–	Plug-in indicator of an EV at step t: 1 if connected, 0 if away.
T^{arr}, T^{dep}	time	Arrival and departure times of an EV (random in general).
E^{need}	kWh	Energy an EV must receive before departure, $(\mathrm{SOC}^{\mathrm{req}} - \mathrm{SOC}_0)\,E^{\max}$.
E^{rem}, T^{rem}	kWh, h	Energy still to deliver and time left before departure.
U_i	–	Urgency of EV i: charging time needed divided by time left.
β_k	–	Coefficients of a regression baseline or forecast model.
RMSE	MW	Root-mean-square tracking error over an event.
ATE	MW	Average treatment effect: control-group load minus treated-group load.
R (response)	MW	Verified response of the fleet; distinguished from forecast F and request P^{tgt}.
n_k	–	Number of devices in state bin k of a histogram representation.
$\mathcal{F}$	MW	A feasible set of fleet power (instantaneous) or of fleet trajectories.
$\mathbf{H}$	–	Sensitivity matrix from DER power changes to network quantities, $\Delta z = \mathbf{H}\,\Delta P$.
S, $S^{\max}$	kVA	Apparent power through a transformer and its rating.
V_i, I_ℓ	V, A	Voltage at node i; current on branch ℓ.
Q	kvar	Reactive power.
r, ℓ	–	Electrical region or location index.
α	–	Reliability level of a commitment or forecast quantile (also a derating factor in dispatch).
π_C, π_P, π_t	\$/MW, \$/MWh	Price of committed capacity, penalty for shortfall, energy price at step t.
λ_k	–	Weight of cost term k in a dispatch objective.
N^{events}	–	Remaining event budget of a demand-response program.
r_i	–	Learned reliability score of device i.

Vocabulary

DER (distributed energy resource)
A physical device connected on the customer side of the distribution network that

generates, stores or consumes electricity and can, under some conditions, change how much it exchanges with the grid. Batteries, electric vehicles, thermostats with the heating/cooling plant behind them, rooftop solar, electric water heaters and controllable commercial loads are all DERs.

DERMS (DER management system)
The software layer that monitors DERs, forecasts what they can do, decides what to ask of them and sends the commands. Chapter 3 of the lesson sequence.

VPP (virtual power plant)
A fleet of DERs operated together so that, from the grid's point of view, it behaves like a single controllable plant. Chapter 4.

Aggregator
The company that enrolls customers' devices, runs the DERMS and sells the fleet's capability to a utility or market. In the lessons the aggregator is one large residential DER aggregator, which is not named.

Demand response (DR)
Asking customers' loads to reduce (or shift) consumption for a period, usually at a time of system stress. Chapter 2.

Flexibility
The ability to change net power exchange with the grid on request, while respecting the device's own limits and the customer's requirements. The central quantity of this book.

Baseline
The counterfactual consumption that would have happened without the control request. Needed to measure what a response delivered.

Dispatch
The act of deciding and sending control actions to a set of devices for a period.

Setpoint
The target value a controller tries to hold, for example the thermostat temperature.

State of charge (SOC)
The fraction of a battery's usable energy currently stored.

Nameplate rating
The maximum power printed on the device label. Rarely the amount actually available.

Point of common coupling (PCC)
The physical point where the customer's premises connect to the utility network; P_{grid} is measured there.

Pre-cooling / pre-heating
Deliberately moving a building's temperature before an event so that more flexibility is available during the event.

Event
A bounded period during which an aggregator asks enrolled devices to change their consumption; the unit of traditional demand response.

Rebound
The increase in load above baseline after an event ends, as devices recover the state they gave up during it.

Load shifting
Moving consumption from one period to another with the same total energy delivered.

Load shedding
Turning load off, sometimes involuntarily, to protect the system.

Demand-side management (DSM)
The historical umbrella for everything done on the consumption side: efficiency, tariffs, conservation and demand response.

Measurement and verification (M&V)
The procedure that estimates what an event delivered, by comparing metered load with a baseline.

Chance constraint
A constraint required to hold with a stated probability, for example $\Pr(P^{\text{del}} \geq 100\,\text{MW}) \geq 0.95$.

Telemetry
The stream of measurements a device reports to the platform (temperatures, state of charge, power, status).

API (application programming interface)
The documented way one piece of software talks to another; each device vendor exposes its own.

State estimation
Inferring the full internal state of a system from partial, noisy measurements.

MPC (model predictive control)
A control scheme that repeatedly estimates the state, optimizes actions over a finite horizon, applies the first action, observes, and repeats.

EMS (energy management system)
The control-center software of the bulk transmission grid: SCADA, state estimation, contingency analysis, economic dispatch, automatic generation control.

ADMS (advanced distribution management system)
The utility's software for operating the distribution network itself: feeder topology, switching, voltage, outage management.

SCADA (supervisory control and data acquisition)
The measurement-and-command infrastructure of a control center.

Volt/VAR optimization
Coordinated control of voltage and reactive power on a distribution feeder; an ADMS function.

Nameplate, available, firm capacity
Three successively smaller numbers: what is installed; what the current states permit; what can be promised with a stated confidence.

Derating
Reducing a resource's credited capacity below nameplate to account for unavailability; standard for conventional generators.

Coverage ratio
Enrolled nominal flexibility divided by committed capacity; the overbooking margin.

CVaR (conditional value at risk)
The expected shortfall in the worst fraction of outcomes, for example the mean delivery in the worst 5 % of days; a tail-risk measure.

Model reduction
Replacing a high-dimensional state by a small set of quantities that suffice for a decision.

Deadband (hysteresis)
The temperature band around the setpoint inside which a thermostat leaves the compressor in its current state.

Duty cycle
Fraction of time a cycling device is on; its average power is the duty cycle times its on-power.

Thermal time constant
The product RC of a building; the time for the indoor temperature to cover about 63 % of a step change in the outdoor temperature with the HVAC off.

Synchronization
The loss of diversity that occurs when many cycling devices receive the same command at the same time and thereafter switch together.

Staggered recovery
Ending an event for groups of devices at different times to avoid a synchronized rebound.

BESS (battery energy storage system)
A stationary battery with its inverter and controls; residential (a few kWh to tens of kWh) or utility scale.

Round-trip efficiency
Fraction of the energy put into a battery that comes back out; the product of charging and discharging efficiencies.

Degradation
Loss of battery capacity and life with use; cycling has a cost even when energy is free.

Arbitrage
Buying (charging) when the price is low and selling (discharging) when it is high.

Managed charging
Scheduling when an EV charges, within its plugged-in window, so that it is ready by departure.

V2G (vehicle-to-grid)
Discharging an EV battery into the grid; bidirectional charging.

Slack Time plugged in minus time needed to charge at full power; the scheduling freedom of one EV.

Urgency
Charging time needed divided by time left before departure; above one the deadline cannot be met.

Earliest-deadline-first (EDF)
A scheduling rule that serves the job with the nearest deadline first.

Valley filling
Placing flexible load in the hours where system load is lowest.

Counterfactual
What would have happened without the intervention; never observed, always inferred.

Control group
Enrolled devices deliberately left uncontrolled during an event so their load estimates the counterfactual.

Histogram state
A fleet representation that counts devices per state bin instead of listing every device.

Virtual battery
A fleet of flexible loads approximated by one power limit, one energy state and its bounds.

Common-mode uncertainty
Error shared by every device in a fleet (for example a weather forecast error); aggregation does not average it away.

Quantile forecast
A forecast reported as values exceeded with stated probabilities rather than as one number.

Scenario optimization
Optimizing expected cost over a set of sampled futures.

Robust optimization
Optimizing against the worst case inside an uncertainty set.

Receding horizon
Re-solving a finite-horizon plan at every step and applying only its first action; the mechanism of MPC.

Reference signal
The trajectory a controller is asked to track.

Feeder
A distribution circuit leaving a substation and serving a group of transformers and customers.

Hosting capacity
How much additional DER a network can accommodate before a constraint is violated.

Optimal power flow (OPF)
Dispatching resources to minimize a cost subject to the network's power-flow equations and limits.

Regulation
A fast balancing service that follows a frequently updated signal around a baseline.

Operating reserve
Capacity held back, unused, to respond to an unexpected event.

Value stacking
: Using one device for several services at different times; forbidden to count the same capability twice for services that can call at once.

Event budget
: The number of events a program allows per season; a finite resource to allocate over time.

Vendor adapter
: The software that translates between a device vendor's interface and the platform's common model.

Cadence
: How often a module of the system runs.

Edge DERMS
: A DERMS operated by an aggregator on customer-owned devices at the grid edge, as opposed to a utility-owned system.

Enrollment
: The customer-program step of signing a customer and connecting their device; a VPP's capacity is bounded by it.

Snapback
: An operator's term for the post-event rebound.

Firm multi-hour dispatch
: Grouping thermostats and scheduling the groups so the fleet's multi-hour reduction is approximately flat.

Packetized energy management
: A coordination technique in which devices request short "packets" of energy that a coordinator grants or defers.

Receding horizon
: See model predictive control.

Chapter 1

What is a DER, and what makes it a grid resource?

1.1 Devices and resources

The first distinction of the subject is simple but is often lost.

Key idea

A *distributed energy resource* (DER) is a physical device. It is not useful to the grid merely because it exists. It becomes useful when it can *change its net power exchange with the grid when asked.*

Batteries, electric vehicles (EVs), thermostats with the heating and cooling plant behind them, rooftop solar photovoltaics (PV), electric water heaters and controllable commercial loads are all DERs. Each one sits on the customer's side of the meter, each one moves electrical energy, and each one was bought for a reason that has nothing to do with the grid: the owner wants a comfortable house, a charged car, backup power, or a lower bill.

For an aggregator the only question that matters at first is:

Can this device change its net power exchange with the grid when requested?

The rest of this chapter makes precise what "can" means in that sentence. It depends on the device's internal state, on how long the change must last, on what the customer will tolerate, and on quantities that are not known in advance. Those four dependencies, state, time, customer, and uncertainty, organize the chapter.

Jargon: DER, aggregator, flexible capacity

DER is any customer-side device that generates, stores or consumes electricity. An *aggregator* is the company that enrolls many such devices and operates them together (a commercial aggregation platform in these lessons). *Flexible capacity* is the amount of power the aggregator can add to or remove from the grid on request. This chapter makes "flexible capacity" precise.

1.2 The device as a dynamical system

This section sets up the vocabulary used for every device in the book: a power balance at the meter, and a state, a control and a disturbance behind each term.

1.2.1 Power exchange at the point of common coupling

Consider one house connected to the grid, shown in Figure 1.1. Define $P_{\text{grid}}(t)$ as the instantaneous power drawn from the grid at the point of common coupling, the physical point where the house meets the network. Positive values mean the house is importing, negative values mean it is exporting.

A simple balance at that point is

$$P_{\text{grid}}(t) = P_{\text{load}}(t) + P_{\text{EV}}(t) + P_{\text{bat}}(t) - P_{\text{PV}}(t). \tag{1.1}$$

Here P_{load} is ordinary household consumption (lights, cooking, appliances), P_{EV} is EV charging power, P_{bat} is battery power with the convention that $P_{\text{bat}} > 0$ means charging and $P_{\text{bat}} < 0$ means discharging, and P_{PV} is solar generation, which always enters with a minus sign because it offsets import.

Equation (1.1) is conservation of energy at one node: whatever the house consumes or stores must come from the grid or from its own solar generation. The same statement holds at every bus of a power network, and Section 1.8 uses it to connect the chapter to the TERRA framework.

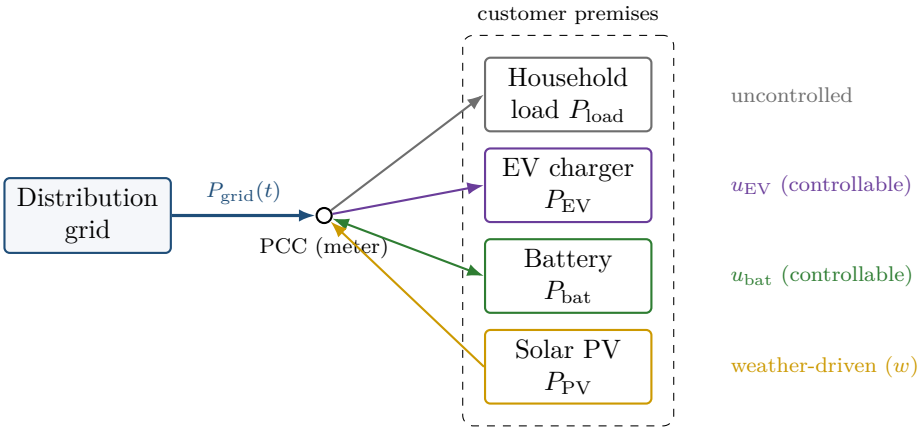


Figure 1.1: One house as a power-balance node. The net exchange P_{grid} at the point of common coupling (PCC) is the sum of the device powers, equation (1.1). Only some of the terms are controllable.

Now suppose some of the terms can be commanded. A control signal u_{EV} might set EV charging power; u_{bat} might set battery charging or discharging. Then the net exchange can be written schematically as

$$P_{\text{grid}} = g(x, u, w), \tag{1.2}$$

with three kinds of argument:

- x, the **internal device states**: what the devices "remember" from the past (how full the battery is, how warm the house is);
- u, the **control actions**: what the aggregator is allowed to choose;
- w, the **external disturbances or conditions**: what nobody chooses (outdoor temperature, sunshine, whether the car is home).

This triple (x, u, w) is the standard language of control theory, and it is worth adopting from the start because every later lesson (DERMS, dispatch, load shaping) is a statement about how to pick u given x and a forecast of w. Table 1.1 lists the three arguments for the common device types.

1.2.2 State dependence of available power

At this point a DER starts to differ from a conventional generator. Consider the question

Table 1.1: The state, control and disturbance of the common DER types.

Device	State x	Control u	Disturbance w
Battery	SOC; temperature; cycle count	charge/discharge power P_{bat}	household load P_{load}, PV output
Electric vehicle	SOC; plugged-in or not; departure time t_{dep}	charging power P_{EV}	arrival time, trips, driver overrides
Thermostat / HVAC	indoor temperature T_{in}; compressor on/off; current setpoint T_{set}	setpoint T_{set} (or direct on/off)	outdoor temperature T_{out}, solar gain, occupancy
Water heater	tank temperature; stratification	element on/off	hot-water draws
Solar PV	(none controllable in most installs)	curtailment, if any	irradiance, clouds

Can this battery give me 5 kW right now?

The answer depends on the state. If the state of charge is SOC = 90% the answer is yes. If SOC = 8% and the owner insists on keeping a 10% emergency reserve, the answer is no, even though the label on the inverter says 5 kW.

So the available power is not simply the nameplate rating, $P^{\max} = 5\,\text{kW}$. It is a function of the state,

$$P^{\text{avail}} = P^{\text{avail}}(x). \tag{1.3}$$

For the battery in the example a reasonable form is

$$P^{\text{avail}}(\text{SOC}) = \begin{cases} P^{\max}, & \text{SOC} \geq \text{SOC}^{\text{res}} + \delta, \\ P^{\max} \cdot \dfrac{\text{SOC} - \text{SOC}^{\text{res}}}{\delta}, & \text{SOC}^{\text{res}} \leq \text{SOC} < \text{SOC}^{\text{res}} + \delta, \\ 0, & \text{SOC} < \text{SOC}^{\text{res}}, \end{cases} \tag{1.4}$$

where δ is a small band over which the inverter tapers power as it approaches the reserve. Figure 1.2 draws this. The exact taper is a detail; the point is the shape: a flat region where the device is fully available, a reserve region where it is not available at all, and a transition between them.

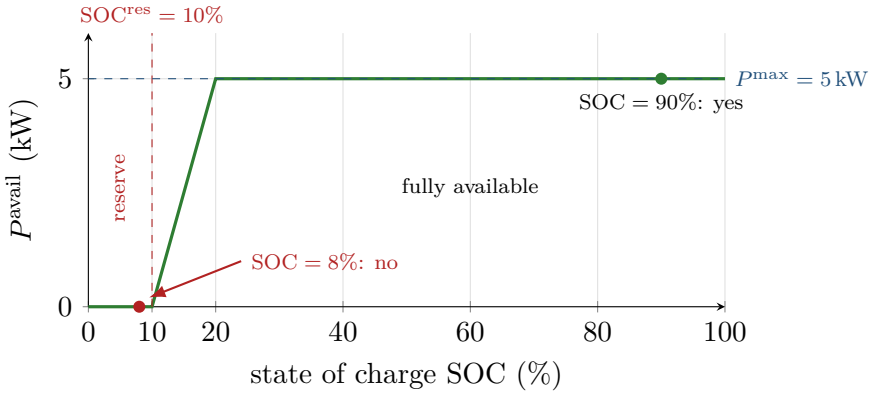


Figure 1.2: Available discharge power of a 5 kW battery as a function of state of charge, equation (1.4), with a 10 % owner reserve and a 10 % taper band. The same device answers "yes" or "no" to the same request depending only on its state.

The state that matters is different for each device type:

$$\begin{aligned} \text{battery:} &\quad x \supset \{\text{SOC}\}, \\ \text{EV:} &\quad x \supset \{\text{SOC},\ \text{plugged-in status},\ t_{\text{dep}}\}, \\ \text{thermostat:} &\quad x \supset \{T_{\text{in}},\ \text{HVAC on/off},\ T_{\text{set}}\}. \end{aligned}$$

The symbol $\supset$ reads "contains at least"; a real model may carry more (battery temperature, EV charger rating, building thermal mass), but these are the states without which no sensible answer can be given.

Principle 1.1. *DER flexibility is* state-dependent*: the power a device can deliver on request is a function $P^{\text{avail}}(x)$ of its internal state, not a fixed nameplate number.*

1.3 Flexibility as a quantity

With the vocabulary in place, flexibility can be defined as a quantity. Four properties follow in turn: it is the change caused by control, it has a duration, it is bounded by power and by energy, and it is a state-dependent feasible set.

1.3.1 Controllability versus flexibility

Suppose a house has a smart thermostat. Technically the aggregator can change the setpoint,

$$u_t = T_{\text{set}}.$$

Now imagine the house is already warm, the thermostat is satisfied, and the compressor is off. Raising the setpoint by another degree produces essentially no change in power:

$$\Delta P \approx 0.$$

The device is fully controllable and yet, at that moment, provides no flexibility. Controllability is a property of the interface; flexibility is a property of the physics at the current operating point.

A useful conceptual definition is

Flexibility = the change in feasible grid power caused by allowable control.

or, as a quantity at a single time step,

$$\Delta P_t = P_t^{\text{base}} - P_t^{\text{ctrl}}, \tag{1.5}$$

where P_t^{base} is the *baseline*, the power the premises would have drawn with no control, and P_t^{ctrl} is the power actually drawn under control. With the sign in (1.5), a positive ΔP_t is a load *reduction*, which is what a grid under stress usually asks for.

Example 1.1 (One thermostat at 5 PM)**.** If the baseline would have been $P_t^{\text{base}} = 3.5\,\text{kW}$ and control reduces consumption to $P_t^{\text{ctrl}} = 1.0\,\text{kW}$, then

$$\Delta P_t = 3.5 - 1.0 = 2.5\,\text{kW}.$$

That 2.5 kW is the quantity a virtual power plant is interested in, not the 3.5 kW of load and not the air conditioner's 4 kW nameplate.

Jargon: baseline

The baseline is a *counterfactual*: what would have happened without the request. It cannot be measured, only estimated, because once control is applied the uncontrolled outcome can no longer be observed. Baseline estimation is one of the oldest arguments in demand response and gets its own treatment in Chapter 2.

1.3.2 Duration dependence of flexibility

A second property separates a DER from a generator. Suppose the thermostat above gives 2.5 kW of load reduction. Can it sustain that for five minutes? Almost certainly. For four hours? Probably not, because the house warms up and eventually the compressor has to run again to keep the customer within their comfort band.

So flexibility is not a number; it has a time dimension. The better object is a function of two arguments,

$$F(t, \tau), \tag{1.6}$$

where t is when the service begins and τ is how long the response must be sustained. For the thermostat, numbers like

$$F(17{:}00,\ 15\text{ min}) = 2.5\,\text{kW} \qquad \text{but} \qquad F(17{:}00,\ 4\text{ h}) = 0.7\,\text{kW}$$

are entirely plausible. The second number is the *average* reduction the device can hold over four hours without violating comfort: it may deliver 2.5 kW at the start, then cycle back on, so that over the whole window the sustainable reduction is much smaller. Figure 1.3 plots this dependence for the two devices used in this chapter: the battery of Section 1.3.3 and the air-conditioned house of Section 1.5.2. The horizontal axis is the required duration τ; the vertical axis is the average reduction that can be sustained over that duration. The battery curve is flat until its energy is exhausted and then falls as $1/\tau$; the thermostat curve is flat only until the compressor must restart and then falls toward a small sustained value.

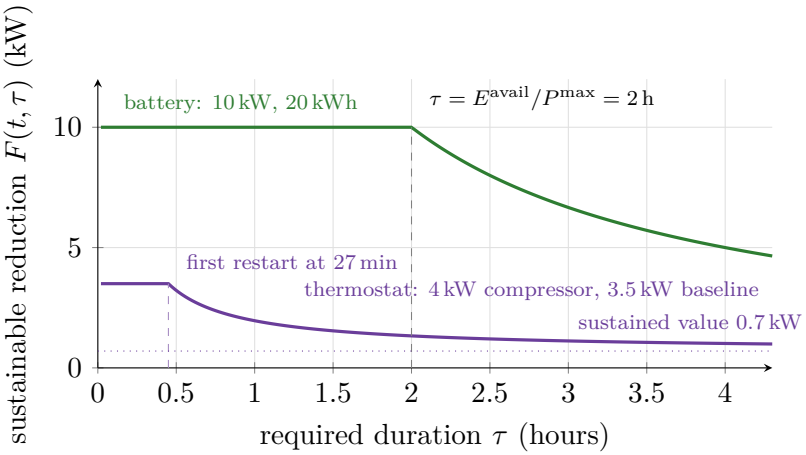


Figure 1.3: Sustainable average reduction versus required duration for the two devices of this chapter. The battery (green) holds its full 10 kW until its 20 kWh is exhausted at two hours, then the average it can sustain falls as E^{avail}/τ. The thermostat (purple) is the air-conditioned house of Section 1.5.2: a 4 kW compressor with a 3.5 kW baseline average, switched off at t. The full 3.5 kW reduction relative to baseline lasts until the compressor must restart at 27 minutes; after that the average reduction over the window, $[3.5\,t^* + 0.7\,(\tau - t^*)]/\tau$, settles toward the 0.7 kW that holding 24°C instead of 22°C sustains.

A conventional generator, by contrast, has essentially $F(t, \tau)$ independent of τ as long as there is fuel. The dependence on τ is what makes a DER a fundamentally different kind of resource.

Principle 1.2. *DER flexibility is* time-dependent*: it is a function $F(t, \tau)$ of when the response starts and how long it must last, and for most devices it decays with τ.*

1.3.3 Power and energy limits

The duration dependence has a simple closed form for a battery, which makes the battery the clearest case. Suppose a battery has a rated power of $P^{\max} = 10\,\text{kW}$ and an available energy of $E^{\text{avail}} = 20\,\text{kWh}$. It can deliver 10 kW for approximately

$$\frac{E^{\text{avail}}}{P^{\max}} = \frac{20\,\text{kWh}}{10\,\text{kW}} = 2 \text{ hours},$$

but it cannot provide 10 kW for four hours. Over four hours the most it can average is $20/4 = 5\,\text{kW}$.

So there are at least two independent quantities that describe a DER:

P	**how fast** the device can move power (kW)
E	**how long** it can keep doing so (kWh)

Figure 1.4 draws the feasible region in the (P, τ) plane: any request with $P \le P^{\max}$ and $P\,\tau \le E^{\text{avail}}$ is feasible.

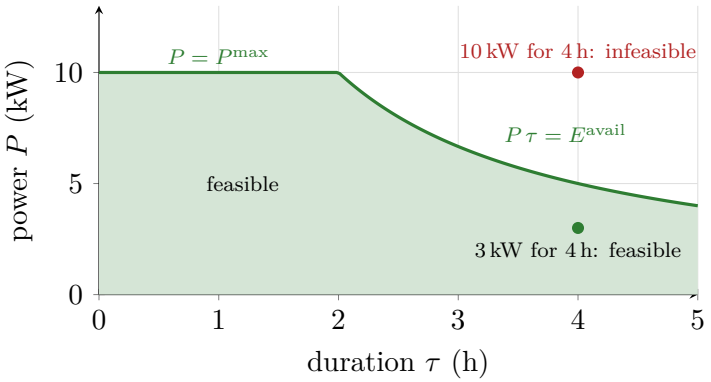


Figure 1.4: The power–duration capability of a 10 kW, 20 kWh battery. The power limit caps the height, the energy limit caps the area under a request.

For the other device types the "energy" is not stored as electricity but the same two-quantity picture applies:

- For a **thermostat**, the analog of stored energy is the building's *thermal inertia*: how much heat the walls, floors and air can absorb before the temperature leaves the comfort band.
- For an **EV**, it is the combination of the battery's SOC and how long the car remains plugged in before t_{dep}.
- For a **water heater**, it is the tank's thermal mass between the minimum acceptable and maximum safe water temperature.

This is why the state-space picture of Section 1.2.1 maps so naturally onto DERs: "how fast" is a limit on u and "how long" is a limit on how far x may travel.

1.3.4 The feasible control set and the state transition

Sections 1.2.2 to 1.3.3 together give the appropriate description of a DER. Instead of saying "this battery is a 10 kW resource" one says

$$u_t \in \mathcal{U}(x_t), \tag{1.7}$$

read as: *the allowable control actions depend on the current state.* For the battery, $\mathcal{U}(\text{SOC})$ is the interval $[-P^{\text{dis}}(\text{SOC}),\ P^{\text{chg}}(\text{SOC})]$, where the discharge limit P^{dis} is the curve of Figure 1.2

and the charge limit P^{chg} has the mirror-image shape near full charge; for the thermostat it is the set of setpoints that keep T_{in} inside the comfort band over the next step.

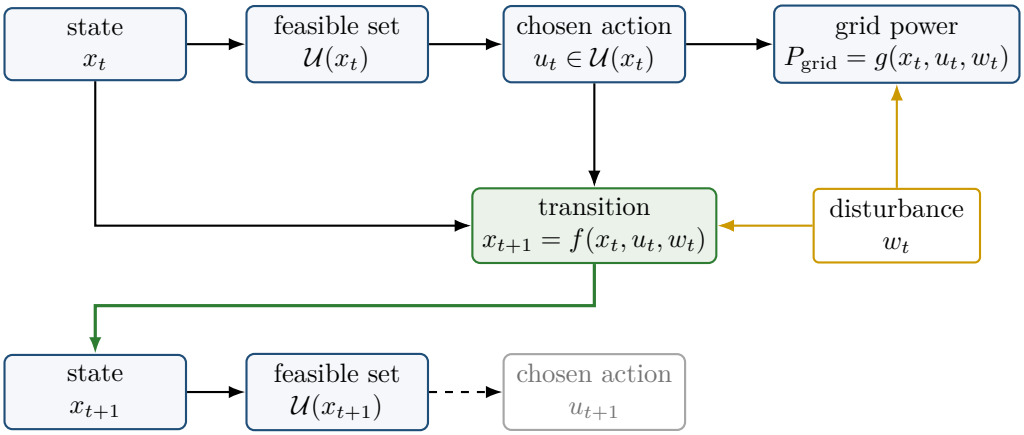
Then the device evolves:

$$\boxed{x_{t+1} = f(x_t, u_t, w_t)} \tag{1.8}$$

This equation accounts for a large part of the DERMS problem, because it states that *today's dispatch changes tomorrow's flexibility*:

- If I heavily discharge a battery now, $\text{SOC}_{t+1} \downarrow$ and less is available later.
- If I pre-cool a home now, $T_{\text{in},t+1} \downarrow$ and I may have *more* flexibility later.
- If I delay EV charging now, SOC_{t+1} may become too low to meet the driver's departure requirement.

DER control is therefore intrinsically *intertemporal*: a decision at t cannot be evaluated without its effect on $\mathcal{U}(x_{t+1})$, $\mathcal{U}(x_{t+2})$, and so on. Figure 1.5 shows the loop.



Intertemporal coupling. The action chosen at t sets the state at $t+1$, and the state at $t+1$ sets which actions are allowed at $t+1$.

Figure 1.5: The state–control loop of a single DER. The disturbance w_t (weather, occupancy, driving) enters both the output and the transition. The green arrow is the coupling that makes dispatch a planning problem rather than a switch.

Principle 1.3. *Because $x_{t+1} = f(x_t, u_t, w_t)$, every dispatch decision spends or stores future flexibility. DER control is intertemporal.*

1.4 Constraints and uncertainty

Two further properties come from outside the device: the customer's requirements and the aggregator's ignorance of the future.

1.4.1 Customer constraints

Conventional generation does not have this in the same form. A thermostat is installed because somebody wants to be comfortable. An EV exists because somebody needs to drive. A home battery may exist mainly for backup during outages. So the control problem carries constraints

that express the owner's purpose:

$$\text{thermostat:} \quad T^{\min} \le T_{\text{in}}(t) \le T^{\max} \quad \text{for all } t, \tag{1.9}$$

$$\text{EV:} \quad \text{SOC}(t_{\text{dep}}) \ge \text{SOC}^{\text{req}}, \tag{1.10}$$

$$\text{battery:} \quad \text{SOC}(t) \ge \text{SOC}^{\text{res}} \quad \text{for all } t. \tag{1.11}$$

These are not secondary conditions. They define whether a grid service is feasible at all. A 4 kW reduction that leaves the house at 27°C is not a service; it produces a complaint and, typically, withdrawal from the program. Formally, the customer constraints are part of what shapes $\mathcal{U}(x_t)$ in (1.7): the feasible set is the set of actions that respect both the device's physics *and* the owner's requirements over the horizon.

Two features make customer constraints harder than device constraints:

- They are stated over a *horizon*, not an instant. Equation (1.10) is a constraint at a future time t_{dep} that restricts actions now.
- They are *soft in practice*. A customer who is slightly uncomfortable may tolerate it once; one who is uncomfortable repeatedly will override the thermostat or leave the program. The override is itself part of the disturbance w.

Principle 1.4. *DER flexibility is* customer-constrained*: comfort bands, departure requirements and reserves are hard constraints on the control problem, and violating them removes the resource.*

1.4.2 Uncertainty

The central forecasting problem can now be stated. Suppose 10 000 EVs are enrolled, each with a 7 kW charger. Does the aggregator have

$$10\,000 \times 7\,\text{kW} = 70\,\text{MW}$$

available? It does not. At 6 PM on a given day:

- some cars are not plugged in;
- some are already fully charged, so there is no charging to defer;
- some need charging urgently to meet SOC^{req} by t_{dep};
- some customers will override the control;
- chargers have different ratings and some are on single-phase circuits that cannot reach 7 kW;
- some communications will fail, so the command never arrives or the confirmation never returns.

Figure 1.6 shows the nameplate total shrinking through these filters to the amount actually deliverable.

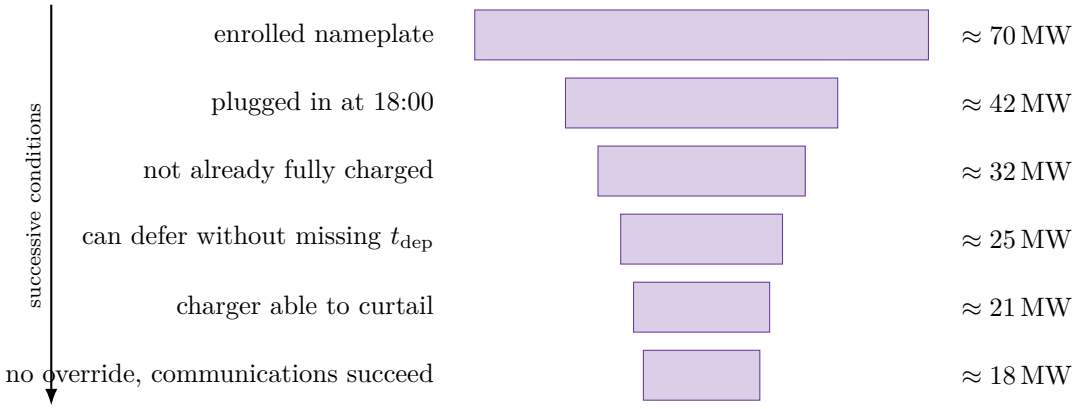
The consequence is that P^{avail} is not a number to be computed but a random variable to be forecast. What the aggregator really wants is a probability distribution,

$$p\Big(P^{\text{avail}} \,\Big|\, \text{information available at time } t\Big), \tag{1.12}$$

conditioned on everything observed so far: telemetry from the devices, weather, calendar, historical participation rates. This is the point where the two tracks of the lesson sequence meet:

Track A asks	*How much flexibility will be available?*
Track B asks	*Given that flexibility, how should we use it?*

Principle 1.5. *DER flexibility is* uncertain*: the aggregator never knows P^{avail} exactly, only a distribution conditioned on current information. Forecasting and control cannot be separated.*



The numbers are illustrative. The shape is not: the deliverable amount is a small, uncertain fraction of nameplate, and each condition is itself a random quantity on any given day.

Figure 1.6: From nameplate to deliverable: 10 000 enrolled EVs at 7 kW each are a 70 MW label but an 18 MW resource on this evening, and a different number tomorrow.

1.5 From one device to a fleet

The chapter closes its argument by placing the single device in the hierarchy the rest of the book climbs, and by working one thermostat through an event.

1.5.1 The hierarchy from device to fleet

Figure 1.7 organizes the domain. Reading left to right on the first row: a *device* has a *state*; the state determines the *feasible control*; applying a control produces a *power response*; the response, net of the baseline and over a duration, is the *flexibility*. The second row is the transition that the later lessons are about: individual flexibilities are aggregated into a *fleet* flexibility, and a fleet operated as a single controllable plant is a *virtual power plant*.

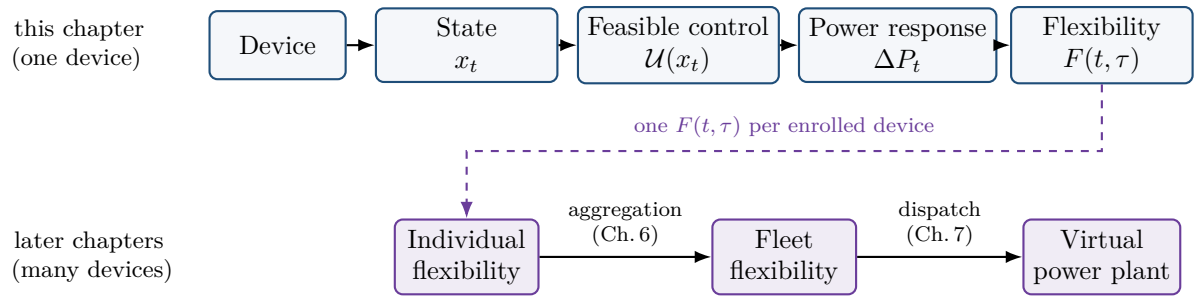


Figure 1.7: The two-level hierarchy of the subject. The top row is one device; the bottom row is what the rest of the book builds on top of it.

The step from the first row to the second, aggregation, carries most of the mathematical difficulty (fleet feasible sets are not simply sums of individual ones), and it is also where the uncertainty of Section 1.4.2 becomes manageable: ten thousand uncertain devices are far more predictable in total than any one of them.

1.5.2 Worked example: thermostat pre-cooling

All of the properties above appear in a single air-conditioned home. The parameters are listed in Table 1.2.

Table 1.2: Parameters of the thermostat example.

Air conditioner electrical power when running	$P_{\mathrm{AC}} = 4\,\mathrm{kW}$
Indoor temperature at 5 PM	$T_{\mathrm{in}} = 22°\mathrm{C}$
Customer comfort band	$21°\mathrm{C} \le T_{\mathrm{in}} \le 24°\mathrm{C}$
Outdoor temperature	$T_{\mathrm{out}} = 32°\mathrm{C}$
Current setpoint	$T_{\mathrm{set}} = 22°\mathrm{C}$, compressor running

The event. At 5 PM the aggregator raises the setpoint by one degree, $T_{\mathrm{set}} : 22°\mathrm{C} \to 23°\mathrm{C}$. The thermostat is now satisfied, the compressor turns off, and the grid load changes approximately

$$4\,\mathrm{kW} \to 0, \qquad \text{so initially } \Delta P \approx 4\,\mathrm{kW}.$$

The decay. But the house is 10 degrees cooler than outdoors, so heat leaks in and the indoor temperature starts rising,

$$\frac{\mathrm{d}T_{\mathrm{in}}}{\mathrm{d}t} > 0.$$

A simple and adequate model of a house for this purpose is a single thermal mass C (kWh per °C) behind a thermal resistance R (°C per kW) to the outdoors, with the air conditioner removing heat at rate Q (kW thermal) when it runs:

$$C\,\frac{\mathrm{d}T_{\mathrm{in}}}{\mathrm{d}t} = \frac{T_{\mathrm{out}} - T_{\mathrm{in}}}{R} - Q\,s(t), \qquad s(t) \in \{0, 1\}. \tag{1.13}$$

With the compressor off ($s = 0$) the solution is an exponential approach to T_{out} with time constant RC:

$$T_{\mathrm{in}}(t) = T_{\mathrm{out}} - \left(T_{\mathrm{out}} - T_{\mathrm{in}}(0)\right) e^{-t/RC}. \tag{1.14}$$

Eventually T_{in} reaches $T^{\max} = 24°\mathrm{C}$ and the compressor must restart to avoid violating the comfort band. With the numbers of Table 1.2 and $RC = 2\,\mathrm{h}$, equation (1.14) gives the first restart at

$$t^{*} = RC \ln \frac{T_{\mathrm{out}} - T_{\mathrm{in}}(0)}{T_{\mathrm{out}} - T^{\max}} = 120\,\mathrm{min} \times \ln \frac{10}{8} \approx 27\,\mathrm{min}.$$

From then on the compressor cycles to hold 24°C. The heat it must remove per unit time is now proportional to $T_{\mathrm{out}} - 24 = 8°\mathrm{C}$ instead of $T_{\mathrm{out}} - 22 = 10°\mathrm{C}$, so its average electrical power is 8/10 of the baseline average. If the baseline average is 3.5 kW (the 4 kW compressor running 87.5% of the time at 22°C), the sustained load is 2.8 kW and the sustained reduction is 0.7 kW.

Two reductions therefore describe the same event: 4 kW at the instant the compressor stops, measured against the instantaneous load, and 3.5 kW measured against the baseline *average*, which is the quantity a settlement uses. The distinction is the subject of Chapter 2. Either way the full reduction lasts only until t^{*} and then settles toward 0.7 kW; in the language of Section 1.3.2, $F(17{:}00, 15\,\mathrm{min})$ is the full value while $F(17{:}00, 4\,\mathrm{h})$ is about 1 kW, and Figure 1.3 plotted exactly this history.

Key idea

Available flexibility decays as the thermal state evolves, *unless you manipulate the state beforehand.*

Pre-cooling. The transition equation (1.8) says the state at the start of the event is something the aggregator can influence. Before 5 PM, run the air conditioner a little harder and bring the house down,

$$T_{\text{in}} : 22°\text{C} \to 21.5°\text{C},$$

which is still inside the comfort band. The building's thermal mass now holds extra "cold", and when the compressor is switched off at 5 PM it takes longer for T_{in} to reach 24°C. Future flexibility has been increased by spending a little energy earlier. Figure 1.8 plots both trajectories using equation (1.14) with a 2-hour time constant.

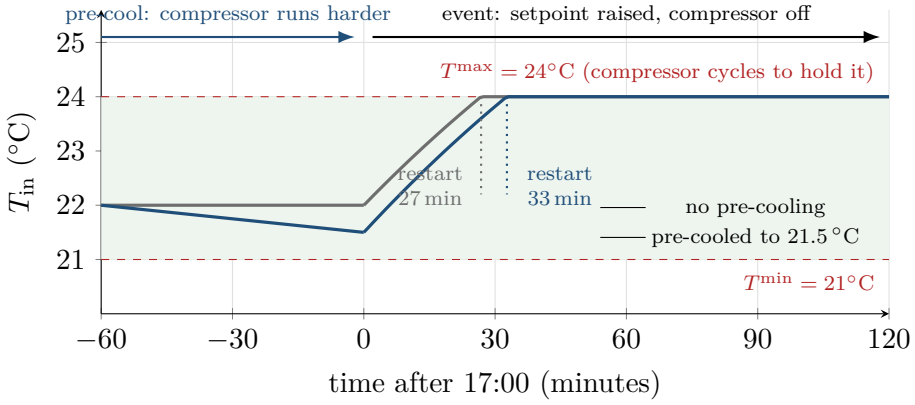


Figure 1.8: Indoor temperature around a 5 PM event. The shaded band is the customer's comfort range; the horizontal axis is minutes relative to the start of the event. Before the event the gray trajectory holds 22°C and the blue one is cooled to 21.5°C. After the event starts both rise according to equation (1.14) with $RC = 2$ h and $T_{\text{out}} = 32°\text{C}$ until they reach $T^{\max}$, where the compressor restarts and cycles. Pre-cooling by half a degree moves the first restart from 27 to 33 minutes; a deeper pre-cool, or a slower building, moves it more.

The gain in this example is modest because half a degree is a small investment in a fast building. With a 3-hour time constant and a pre-cool to 21°C, the first restart moves from 40 minutes to 57. The qualitative lesson does not depend on the numbers: pre-cooling is a genuine control problem, with a cost (the extra energy and the slightly higher load before the event) and a benefit (more flexibility during it), and the right amount depends on the forecast of T_{out}, on the building's RC, and on when the event will be called. That is exactly the shape of the dispatch problems in the later chapters.

1.6 The working definition

The chapter closes with the definition to retain, followed by Table 1.3, which lists the four properties of DER flexibility established above and the mathematical object that captures each.

Definition 1.1 (DER as a grid resource)**.** A DER becomes a grid resource when its *feasible, state-dependent power response* can be *predicted* and *controlled* while respecting *device and customer constraints*.

Each italicized phrase is one of the sections above: feasible (Section 1.3.4), state-dependent (Section 1.2.2), predicted (Section 1.4.2), controlled (Sections 1.3.1 and 1.3.2), device and customer constraints (Section 1.4.1). The definition is more useful than the expansion "distributed energy resource" alone, and it already explains why an aggregator needs forecasting, optimization and control, not merely an interface that turns devices on and off.

Table 1.3: The four properties of DER flexibility, and the object that captures each.

Property	Object	Consequence
State-dependent	$P^{\text{avail}}(x)$, $\mathcal{U}(x)$	nameplate is not availability
Time-dependent	$F(t,\tau)$, $x_{t+1} = f(x_t, u_t, w_t)$	flexibility decays, and today's use changes tomorrow's
Customer-constrained	$T^{\min} \le T_{\text{in}} \le T^{\max}$, $\text{SOC}(t_{\text{dep}}) \ge \text{SOC}^{\text{req}}$	the service is feasible only inside the owner's band
Uncertain	$p(P^{\text{avail}} \mid \text{info}_t)$	forecast a distribution, not a number

1.7 Exercises

Exercise 1.1. A home battery has $P^{\max} = 5\,\text{kW}$, a usable capacity of 13.5 kWh, is at SOC = 60%, and the owner keeps a 20% reserve. For how long can it discharge at full power? What is $F(t, 3\,\text{h})$?

Exercise 1.2. Using equation (1.14), show that the time until the compressor must restart after being switched off at $T_{\text{in}}(0)$ is $t^* = RC \ln\!\big[(T_{\text{out}} - T_{\text{in}}(0))/(T_{\text{out}} - T^{\max})\big]$. Evaluate it for $RC = 2\,\text{h}$, $T_{\text{out}} = 32°\text{C}$, $T^{\max} = 24°\text{C}$ and $T_{\text{in}}(0) \in \{22, 21.5, 21\}°\text{C}$.

Exercise 1.3. An EV arrives at 18:00 with SOC = 30% on a 60 kWh pack, charges at 7 kW, and must leave at 07:00 with SOC $\ge$ 80%. How many hours of charging are needed, and how many hours of *deferral* does the aggregator have? What happens to that deferral if the driver changes the departure to 22:00?

Exercise 1.4. Write the feasible set $\mathcal{U}(x_t)$ for the EV of the previous exercise at a time t before departure, in terms of SOC_t, $t_{\text{dep}} - t$, SOC^{req} and the charger rating.

1.8 Connection to the TERRA framework

TERRA models a physical system as a graph: *nodes* are places where a conserved quantity is balanced, *edges* carry flows of that quantity between nodes, and *components* attached to the graph contribute equations (residuals) that must all be satisfied at once. Every quantity the engine solves for is collected in one state vector, and the whole model is the statement "all residuals are zero". On top of that, the framework treats unknown parameters and unmeasured flows as quantities to be *inferred* from measurements, with a probability distribution rather than a point value. The lesson of this chapter maps onto that picture almost line by line.

- **The house is a conservation node.** Equation (1.1) is a power balance at the point of common coupling: inflow from the grid equals the sum of the device flows. In TERRA that is one node with five edges (grid, load, EV, battery, PV) and one residual, $P_{\text{grid}} - P_{\text{load}} - P_{\text{EV}} - P_{\text{bat}} + P_{\text{PV}} = 0$. A feeder with a thousand houses is a thousand such nodes joined by the network's own balance equations.
- **Each device is a component with a state.** The battery contributes its state-of-charge dynamics (an accumulator), the house contributes the thermal mass of equation (1.13) (a second conservation law, for heat, in the same graph), the EV contributes a battery plus a plugged-in schedule. The transition map f of equation (1.8) is what the engine integrates in time; the output map g is the set of edge flows the solved state implies. Multi-conservation, electrical and thermal in one model, is precisely what a thermostat DER needs.

- **Feasible sets and customer constraints are bounds on the state.** $T^{\min} \le T_{\text{in}} \le T^{\max}$ and $\text{SOC} \ge \text{SOC}^{\text{res}}$ are constraints on components of the state vector. Checking whether a dispatch is feasible is checking whether the integrated trajectory stays inside those bounds; finding the best dispatch is an optimization over the control inputs with the graph as the model.

- **Flexibility $F(t, \tau)$ is a query on the model.** "Switch this compressor off at t and integrate forward: how long until a bound is hit, and what was the average change in the grid edge flow?" is a forward simulation of the graph. The chapter's $F(t, \tau)$ is that query's answer, and the pre-cooling experiment of Section 1.5.2 is the same query run from a different initial state.

- **Uncertainty is inference, not a fudge factor.** The distribution $p(P^{\text{avail}} \mid \text{info}_t)$ of equation (1.12) is what the framework produces when device states (an EV's SOC, a building's RC) are only partially observed: measurements at the meter constrain, but do not pin down, the hidden states, and the posterior over those states propagates into a posterior over the available power. The funnel of Figure 1.6 is a chain of such partially observed conditions.

- **Aggregation is the same graph, larger.** The step from one device to a fleet (Figure 1.7) adds nodes and edges but no new kind of equation. This is the practical reason to adopt a conservation-graph view from the first lesson: the single-house model of Figure 1.1 and the virtual-power-plant model of a later chapter are the same object at different scales.

In summary, a DER is a component on a conservation graph whose state is partially known and whose control is constrained; a grid resource is such a component whose edge flow can be predicted and steered inside those constraints. The rest of the book is about doing that for many components at once.

Chapter 2

Demand response and demand flexibility

2.1 Demand response as a subset of demand flexibility

This chapter covers the transition from the older, event-based idea of *demand response* to the broader idea of continuously shaping when electricity is consumed, which this book calls *demand flexibility*. The relationship between the two is containment:

$$\text{demand response} \subset \text{demand flexibility}. \tag{2.1}$$

Demand response is usually an *event*: a bounded period in which enrolled devices are asked to reduce consumption. Demand flexibility is an *ongoing operating capability*: the ability to decide, at every hour of every day, when controllable consumption should occur. The chapter develops the event first, shows why its measurement and its physics are harder than they look, and then generalizes to the continuous problem.

Jargon: demand response (DR), event, demand flexibility

Demand response is a deliberate change in consumption in response to a request from a utility or market operator. The request is an *event*: it has a start time, an end time, and a required reduction. *Demand flexibility* is the general capability of changing the timing or magnitude of consumption, of which an event is one use.

2.2 The demand-response event

Demand response is developed first in its traditional, event-based form: the request, its measurement, and a fleet-scale example.

2.2.1 The event-based problem

Consider a utility that expects a system peak tomorrow evening between 17:00 and 19:00. Without intervention the forecast load, written $P^{\text{base}}(t)$ in the notation of Chapter 1, peaks at 5 000 MW. The utility wants to reduce that peak by 100 MW, so it calls a demand-response event. The sequence is

$$\text{forecast peak} \rightarrow \text{event} \rightarrow \text{load reduction}.$$

For an aggregator the event means temporarily adjusting many thermostats, EV chargers, batteries and other devices. The traditional procedure has five steps: forecast the load that can be shed, group the resources, define the event parameters, dispatch, and compare the metered

load with a baseline. Figure 2.1 shows the intended outcome on the system load curve: the horizontal axis is the hour of day, the vertical axis is system load, and the controlled curve sits 100 MW below the forecast during the event window.

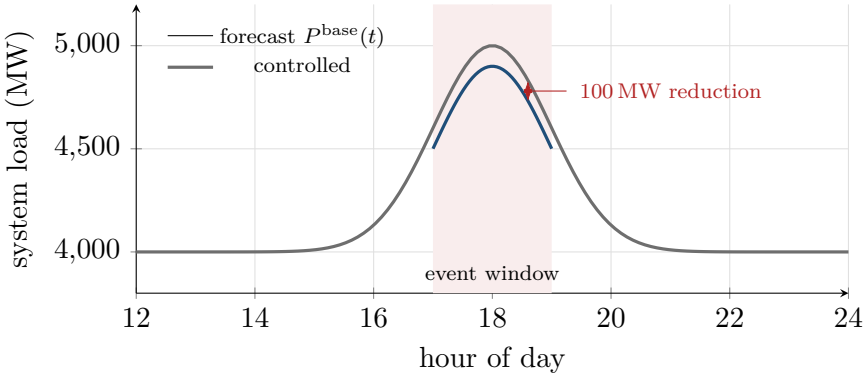


Figure 2.1: The event-based demand-response problem. The forecast load (gray) peaks at 5 000 MW at 18:00; the event asks the fleet to hold the load 100 MW below the forecast (blue) between 17:00 and 19:00. The forecast curve is illustrative.

2.2.2 Measuring the reduction: the baseline is a counterfactual

Suppose the actual, metered load during the event is $P^{\text{act}}(t)$. The quantity of interest is how much the event reduced load, but the load that *would have occurred* without control is not observed. That counterfactual is the baseline $P^{\text{base}}(t)$, and the estimated reduction is

$$\Delta P(t) = P^{\text{base}}(t) - P^{\text{act}}(t). \tag{2.2}$$

For example, with $P^{\text{base}} = 1\,000\,\text{MW}$ and $P^{\text{act}} = 930\,\text{MW}$ the fleet is said to have delivered $\Delta P = 70\,\text{MW}$.

The difficulty is immediate: there is no meter that measures "what would have happened". The baseline is *inferred*, from historical load on similar days, from weather, from the load in the hours before the event, or from a model. Its quality determines the quality of the performance estimate.

Consider the same metered load, $P^{\text{act}} = 930\,\text{MW}$, evaluated against two baseline models:

Baseline model	P^{base}	Estimated ΔP
Model A	1 000 MW	70 MW
Model B	960 MW	30 MW

Same physical system, same measured load, and a factor of more than two between the two estimates of how well the resource performed. Figure 2.2 draws the situation: the metered load is a single curve, and the two baselines are two hypotheses above it. The reduction is the gap between hypothesis and measurement, and the measurement does not say which hypothesis is right.

The conclusion is a decomposition worth keeping:

$$\text{DR performance} = \text{measurement} + \text{counterfactual estimation}. \tag{2.3}$$

The second term is easy to underestimate. It is the subject of measurement and verification, which later chapters treat in more detail.

> **Principle 2.1.** *An event's delivered reduction is never measured directly; it is the difference between a measurement and an inferred baseline, and the inference can dominate the result.*

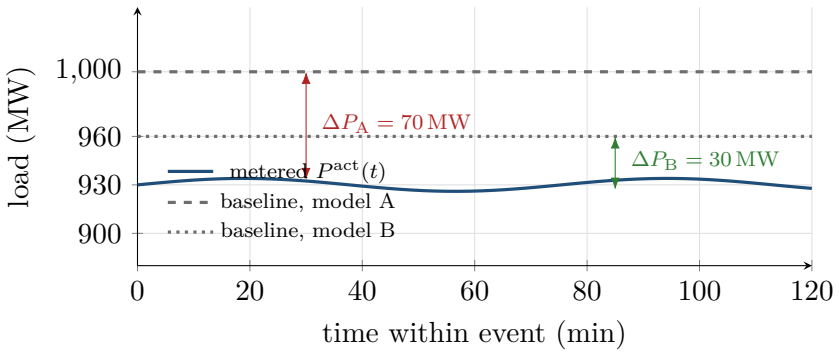


Figure 2.2: One metered load, two baseline hypotheses, two performance estimates. Nothing in the measurement distinguishes model A from model B; the difference is entirely in the counterfactual.

2.2.3 A thermostat fleet event

The second difficulty is physical rather than statistical. Consider a fleet of $N = 100\,000$ thermostats. At 17:00 perhaps 30 000 compressors are running, at an average of 3 kW each, so the instantaneous controllable load is

$$30\,000 \times 3\,\mathrm{kW} = 90\,\mathrm{MW}.$$

It is tempting to record $P^{\mathrm{avail}} = 90\,\mathrm{MW}$. That is not what the fleet delivers over the next two hours, for the reasons established in Chapter 1:

- once compressors are switched off, houses warm and some compressors restart within the comfort band;
- some customers override their thermostats;
- some of the 30 000 compressors were about to cycle off anyway, so switching them off changes nothing relative to baseline;
- weather, and therefore the rate of warming, varies by location.

Hence

$$P^{\mathrm{del}}(t) \neq P^{\text{initially controllable}}, \tag{2.4}$$

and the gap grows with the duration of the event. This is the state-dependent, time-dependent flexibility $F(t, \tau)$ of Chapter 1, now at fleet scale.

2.3 The physics of an event

Two physical facts complicate the event: the response is a dynamic process, and it leaves a debt that is repaid after the event ends.

2.3.1 Demand response is a dynamic process

A naive description of an event is a switch: $u = 1$ implies a 100 MW reduction. The accurate description is the state-space pair of Chapter 1,

$$x_{t+1} = f(x_t, u_t, w_t), \qquad P_t = g(x_t, u_t, w_t). \tag{2.5}$$

For a thermostat the relevant state is the indoor temperature, $x_t \supset T_{\mathrm{in},t}$. When the control curtails the HVAC, $u_t = \text{curtail}$, the thermal state moves: $T_{\mathrm{in},t+1} > T_{\mathrm{in},t}$ under hot outdoor conditions. That changes the future response. Dispatch now affects available flexibility later, and the event is a controls problem, not an instruction to turn 50 000 thermostats off.

2.3.2 The rebound

Suppose the event ends at 19:00. During the event many buildings warmed above their normal trajectory. At 19:00 normal control resumes and many air conditioners turn on at once, because every thermostat is now above its setpoint. The load does not simply return to baseline; it overshoots it. Figure 2.3 shows the shape: load below baseline during the event, then above baseline for a period afterward.

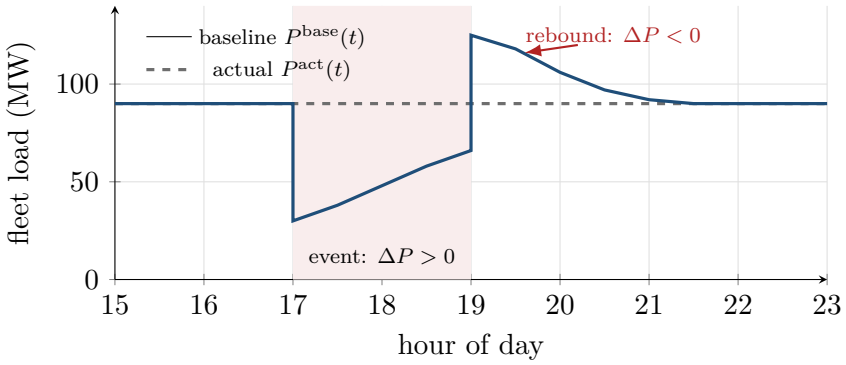


Figure 2.3: Load of a thermostat fleet through an event and its rebound. During the event the actual load (blue) falls below baseline (dashed) and then creeps upward as compressors restart; when the event ends the load overshoots the baseline while the fleet re-cools. The curve is illustrative.

In the sign convention of equation (2.2),

$$\Delta P(t) > 0 \text{ during curtailment}, \qquad \Delta P(t) < 0 \text{ afterward},$$

because for a while $P^{\text{act}} > P^{\text{base}}$. Consumption was borrowed from one period and pushed into another. This is why *load shifting* is often a more accurate description than *load reduction*: the energy not consumed during the event is largely consumed later. The rebound is itself a grid event, and an aggregator that ignores it can create a second, smaller peak immediately after the one it was asked to cut.

2.4 Demand flexibility

This section generalizes from the event to the continuous question of when consumption should occur.

2.4.1 From events to demand flexibility

Traditional demand response asks for a reduction between 17:00 and 19:00. Demand flexibility asks a more general question:

When should these devices consume electricity, given grid conditions and customer constraints?

The lesson plan gives several instances: increase EV charging while renewable output is high; reduce or shift charging while prices are high; reshape local load when a distribution transformer approaches its rating; coordinate thermostats and batteries through an evening system peak. None of these is an “event” in the traditional sense; each is a continuous decision about timing.

2.4.2 Load as a shape to be chosen

The clearest instance is EV charging. Uncontrolled, every car is plugged in when its driver arrives home, and the fleet produces a sharp evening peak. But the cars only need to be charged by 07:00. The aggregator can spread the same energy over the night, or concentrate it at 02:00 when power is cheapest. The driver's requirement, $\mathrm{SOC}(t_{\mathrm{dep}}) \geq \mathrm{SOC}^{\mathrm{req}}$, is met in all three cases; only the timing of consumption changes. Figure 2.4 shows the three profiles. Each panel plots fleet charging power against hour; the area under each curve, the energy delivered, is the same.

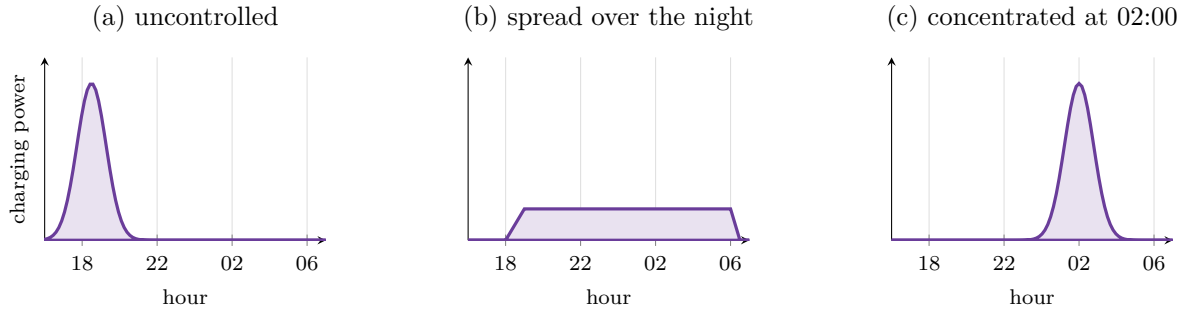


Figure 2.4: Three charging profiles for the same EV fleet with the same total energy. (a) Uncontrolled: every car charges on arrival, producing an evening peak. (b) Spread: the same energy delivered at low power over the whole night. (c) Concentrated: charging moved to the cheapest hours. All three satisfy $\mathrm{SOC}(t_{\mathrm{dep}}) \geq \mathrm{SOC}^{\mathrm{req}}$. Profiles are illustrative.

Choosing among such shapes, continuously, for every device, is demand flexibility.

2.4.3 The optimization viewpoint

The choice of shape can be made mathematical. Let the control be the charging power, $u_t = P_{\mathrm{EV},t}$, over steps $t = 1, \ldots, T$ of length Δt. Minimizing the cost of energy at price c_t is

$$\min_{u_1,\ldots,u_T} \sum_{t=1}^{T} c_t \, u_t \, \Delta t \tag{2.6}$$

subject to the charger rating,

$$0 \leq u_t \leq P^{\max}, \tag{2.7}$$

the battery dynamics,

$$\mathrm{SOC}_{t+1} = \mathrm{SOC}_t + \frac{\eta \, u_t \, \Delta t}{E^{\max}}, \tag{2.8}$$

and the driver's requirement,

$$\mathrm{SOC}(t_{\mathrm{dep}}) \geq \mathrm{SOC}^{\mathrm{req}}. \tag{2.9}$$

Here η is the charging efficiency and $E^{\max}$ the usable pack capacity. Equation (2.8) is the transition map $f(x, u, w)$ of Chapter 1 written out for one device, and equation (2.9) is the customer constraint.

The solution of this problem answers "when should the car charge?" rather than "should this charger be off right now?". With a flat price every feasible schedule costs the same; with a price that is low at 02:00 the solution is profile (c) of Figure 2.4. That change of question, from a yes/no at one instant to a schedule over a horizon, is the conceptual step from demand response to demand flexibility.

2.4.4 Flexibility in both directions

DERs are not useful only because they can reduce load. At times the grid wants *more* load. At noon on a sunny day wholesale prices can fall to zero or below because solar generation exceeds demand; rather than curtail the solar, the system can increase P_{EV} or charge batteries. In the convention of equation (2.2) this is

$$P^{\mathrm{ctrl}} > P^{\mathrm{base}} \quad \Longrightarrow \quad \Delta P < 0,$$

and a negative ΔP is not a failure. It may be exactly what the grid asked for. A reader used to demand response has to unlearn the idea that "down" is the only useful direction.

2.5 Objectives and decisions

Once timing is a choice, the choice needs an objective, and the objective must respect the future and the uncertainty of the forecast.

2.5.1 Tracking a target trajectory

Once both directions are allowed, the natural target is no longer "reduce load" but a desired trajectory $P^{\mathrm{tgt}}(t)$ that the fleet is asked to follow,

$$P_{\mathrm{fleet}}(t) \approx P^{\mathrm{tgt}}(t). \tag{2.10}$$

One objective that expresses this is the squared tracking error summed over the horizon,

$$J = \sum_{t} \big(P_{\mathrm{fleet}}(t) - P^{\mathrm{tgt}}(t)\big)^2, \tag{2.11}$$

minimized subject to every device and customer constraint. This is the form that the dynamic load shaping of a later chapter takes. The objective says nothing about direction; a target above baseline is followed as readily as one below.

2.5.2 Different grid objectives, different target shapes

The same fleet can serve different purposes, and each purpose is a different objective function. Table 2.1 lists the common ones. The device model, the constraints and the fleet are identical across the rows; only the objective changes, and with it the shape of the optimal consumption.

Table 2.1: Grid objectives and the optimization each implies. P_{tr} is the loading of a distribution transformer.

Objective	Mathematical form	Resulting shape
Peak shaving	min $\max_t P_{\mathrm{grid}}(t)$	flattened top
Price arbitrage	min $\sum_t c_t\, P_{\mathrm{grid}}(t)$	consumption moved to cheap hours
Renewable absorption	increase load when renewable output is high	midday bump
Transformer constraint	enforce $P_{\mathrm{tr}}(t) \le P_{\mathrm{tr}}^{\max}$	local cap
Scheduled grid service	$P_{\mathrm{fleet}}(t) \approx P^{\mathrm{tgt}}(t)$	whatever was requested

2.5.3 Opportunity cost

Demand flexibility introduces a problem that an event does not have. Suppose a battery holds 20 kWh of usable energy. It could discharge between 16:00 and 17:00. But if the larger system peak occurs at 18:00, using the battery now reduces what it can do then. The question is no longer "is this DER available?" but "is this the best time to use it?". That is a harder question, because its answer depends on a forecast of the rest of the day.

Formally this is the intertemporal coupling of Chapter 1: u_t changes x_{t+1}, which changes $\mathcal{U}(x_{t+1})$, which restricts every later action. Dispatch changes future flexibility, and a dispatcher that does not look ahead will spend the resource at the first opportunity rather than the best one.

2.5.4 Uncertainty and chance constraints

Forecast uncertainty makes the problem harder still. Suppose tomorrow's requirement is 100 MW from 17:00 to 19:00 and the forecast says $\hat{P}^{\text{avail}} = 120\,\text{MW}$. That looks safe. But if the forecast has a standard deviation of 30 MW, so that

$$P^{\text{avail}} \sim \mathcal{N}(120, 30^2)\ \text{MW}, \tag{2.12}$$

then the probability of having less than 100 MW is

$$\Pr(P^{\text{avail}} < 100) = \Phi\left(\frac{100 - 120}{30}\right) = \Phi(-0.67) \approx 0.25,$$

where Φ is the standard normal distribution function. One event in four would fall short. Figure 2.5 draws the forecast distribution with the shortfall region shaded.

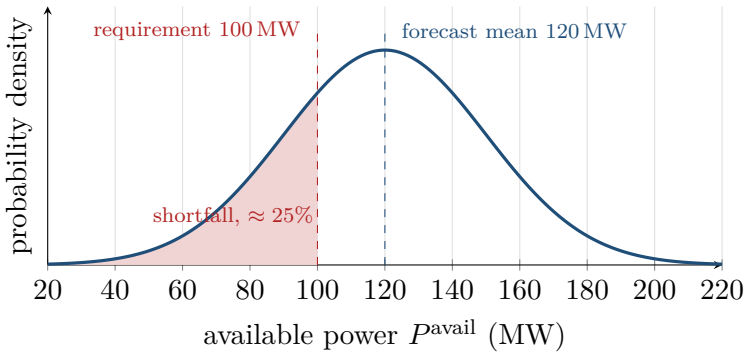


Figure 2.5: Forecast distribution of available power, equation (2.12). The mean exceeds the 100 MW requirement, yet a quarter of the probability mass lies below it.

A control problem that takes this seriously replaces the deterministic requirement with a *chance constraint*,

$$\Pr(P^{\text{del}} \geq 100\,\text{MW}) \geq 0.95. \tag{2.13}$$

Under the normal model this means committing resources whose forecast mean exceeds the requirement by $1.645\,\sigma$, that is, forecasting about 150 MW to promise 100 with 95 % confidence. The dispatch is no longer deterministic; it combines

$$\text{forecasting} + \text{uncertainty} + \text{optimization}.$$

2.6 From events to continuous operation

The last part of the chapter fixes the vocabulary, contrasts the two modes of operation, and works a two-hour service through a three-resource portfolio.

2.6.1 Terminology

Several phrases in this area are easy to confuse. Table 2.2 fixes them for the rest of the book. The relationship to retain is that demand response is one operational use of demand flexibility, and both sit inside the older umbrella of demand-side management.

Table 2.2: Terms that are easy to confuse, from narrowest to broadest.

Term	Meaning
Load shedding	Turning load off, sometimes involuntarily, to protect the system.
Demand response	A deliberate change in consumption in response to a request or event.
Load shifting	Moving consumption from one period to another with the same total energy.
Demand flexibility	The general ability to change the timing or magnitude of demand, in either direction.
Demand-side management	The historical umbrella: efficiency, tariffs, conservation, and demand response.

2.6.2 Open-loop events and closed-loop operation

Traditional demand response was easier than modern fleet operation because it was open-loop: forecast the peak, select customers, call the event, reduce load, measure the result. Each step happened once, in order. Modern flexibility is closed-loop: weather, prices, grid conditions, device telemetry and customer behavior feed a forecast; the forecast feeds an optimization; the optimization produces a dispatch; the realized state is measured; and the cycle repeats. Figure 2.6 shows both pipelines side by side.

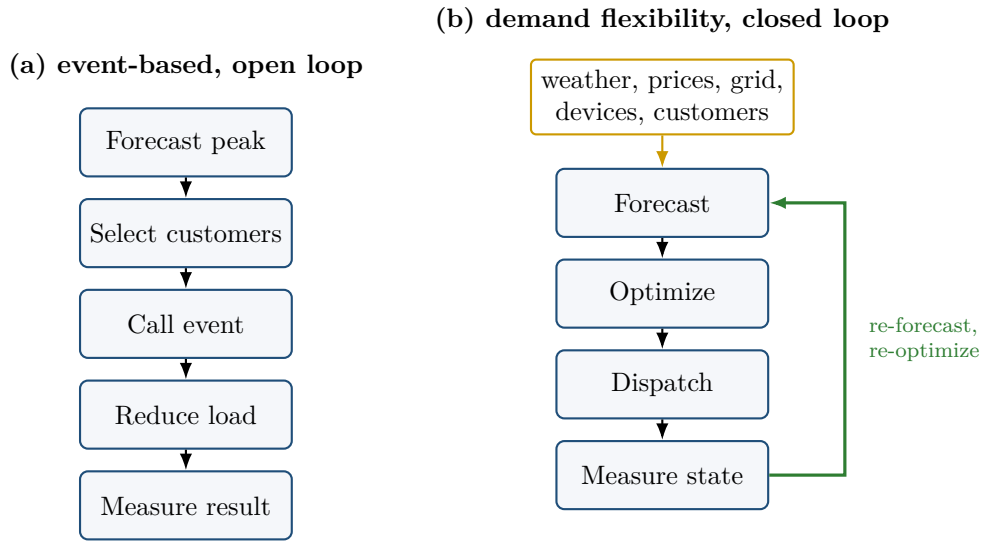


Figure 2.6: (a) The open-loop pipeline of a traditional demand-response event: five steps executed once. (b) The closed-loop pipeline of continuous demand flexibility: measurements of the realized state feed back into the next forecast and optimization.

The closed loop is the sequence *state estimation, forecasting, optimization, control* that the rest of the book develops; Chapter 3 describes the software system that runs it.

2.6.3 Worked example: 20 MW for two hours from three resources

A utility needs 20 MW of reduction from 18:00 to 20:00. The aggregator has three resource classes, listed in Table 2.3 with the power each can provide and for roughly how long.

Table 2.3: Resources available for the worked example.

Resource	Available power	Sustainable duration
Thermostats	15 MW	about 1 h
EV charging	10 MW	about 3 h
Batteries	8 MW	about 2 h

The sum, $15 + 10 + 8 = 33$ MW, suggests the request is easy. But the thermostats cannot sustain 15 MW for two hours: by the second hour their flexibility has decayed, as Section 2.2.3 explained. One schedule that works is shown in Table 2.4 and drawn in Figure 2.7: the thermostats carry most of the first hour, and the EVs and batteries take over as the thermostats fade.

Table 2.4: A feasible schedule for the worked example (MW).

Hour	Thermostats	EV charging	Batteries	Total
18:00–19:00	10	5	5	20
19:00–20:00	4	8	8	20

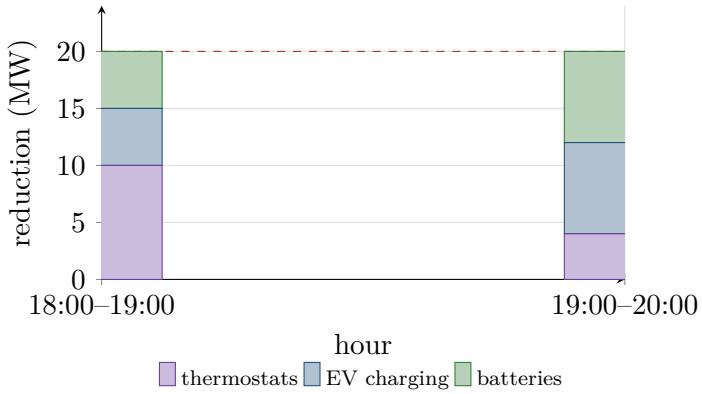


Figure 2.7: The schedule of Table 2.4 as stacked bars. The dashed line is the 20 MW requirement. The thermostat contribution shrinks in the second hour and the storage-backed resources grow to compensate.

The question answered here is not “how many megawatts are available?” but

What combination of resources can produce the required trajectory over time?

That is the virtual-power-plant operating problem in miniature, and it is already a scheduling problem over resources with different power–duration characteristics, exactly the capability regions of Chapter 1.

2.7 The key distinction

Definition 2.1 (Demand response versus demand flexibility)**.** *Demand response* asks whether load can respond to an event. *Demand flexibility* asks how controllable consumption should

evolve over time.

The second question requires a state, a forecast, constraints and an optimization, which is why a modern fleet-operating system is much richer than a traditional demand-response platform. Table 2.5 contrasts the two along the dimensions this chapter developed.

Table 2.5: Demand response and demand flexibility compared.

Dimension	Demand response	Demand flexibility
Unit of operation	an event	a continuous schedule
Direction	reduce	reduce or increase
Target	a megawatt number	a trajectory $P^{\mathrm{tgt}}(t)$
Time structure	open loop, once	closed loop, repeated
Decision	which devices now	which devices, when, how much
Uncertainty	handled by over-enrollment	handled by chance constraints
Needs	baseline, dispatch	state, forecast, constraints, optimization

2.8 Exercises

Exercise 2.1. In Section 2.2.2 the metered load is 930 MW. A third baseline model estimates 985 MW. Compute the delivered reduction under all three models and the ratio of the largest to the smallest estimate.

Exercise 2.2. For the EV of Section 2.4.3, take $E^{\max} = 60\,\mathrm{kWh}$, $\eta = 0.92$, $P^{\max} = 7\,\mathrm{kW}$, $\Delta t = 1\,\mathrm{h}$, arrival at 18:00 with SOC $= 0.30$, departure at 07:00 with $\mathrm{SOC}^{\mathrm{req}} = 0.80$. How many charging hours are needed? If the price is \$0.30/kWh from 18:00 to 22:00 and \$0.10/kWh afterward, write down the cost-minimizing schedule and its cost.

Exercise 2.3. Using the normal model of equation (2.12), what forecast mean is required so that $\Pr(P^{\mathrm{avail}} \geq 100\,\mathrm{MW}) \geq 0.99$ with $\sigma = 30\,\mathrm{MW}$? Repeat for $\sigma = 15\,\mathrm{MW}$ and comment on the value of a better forecast.

Exercise 2.4. Redo the worked example of Section 2.6.3 for a three-hour event (18:00–21:00) with the same resources. Is 20 MW feasible in the third hour? If not, what is the largest constant requirement that is?

2.9 Connection to the TERRA framework

Chapter 1 mapped a single DER onto a conservation graph: a node where power balances, edges carrying device flows, components holding state. This chapter adds three objects, and each has a natural place in that picture.

- **The baseline is a counterfactual simulation.** The quantity $P^{\mathrm{base}}(t)$ of Section 2.2.2 is the grid-edge flow the model predicts when the control inputs are held at their uncontrolled values. In a graph model the baseline and the controlled trajectory are two forward integrations from the same estimated initial state, differing only in u. The ambiguity of Figure 2.2 is the uncertainty in that initial state and in the model parameters, and a probabilistic framework reports it as a distribution over P^{base} rather than choosing model A or model B by convention.

- **The rebound is a conservation statement.** The thermal energy not removed from the buildings during the event must be removed afterward; the rebound of Figure 2.3 is the heat-balance node of each building settling its account. A model that carries both the electrical and the thermal conservation law predicts the rebound automatically; a model that carries only the electrical one cannot.
- **The objective and the chance constraint are queries on the posterior.** Tracking a target, equation (2.11), is an optimization over control inputs with the graph as the model. The chance constraint of equation (2.13) asks for a quantile of the predicted edge flow, which is exactly what a posterior over the fleet state provides. The 25 % shortfall of Figure 2.5 is read off that posterior rather than assumed.
- **Both directions are one model.** Increasing load at noon and decreasing it at 18:00 are the same forward simulation with different inputs; nothing in the conservation-graph view privileges one sign of ΔP.

In summary, an event is a short, open-loop use of the graph; demand flexibility is the same graph run continuously in a loop with its state re-estimated at every step. The next chapter is about the system that runs that loop.

Chapter 3

What does a DERMS actually do?

3.1 Definition and the operational chain

This section fixes what a DERMS is and introduces the eight-stage chain that organizes the rest of the chapter.

3.1.1 Definition and position

A distributed energy resource management system (DERMS) is the software and control system that integrates DERs and manages their telemetry, forecasting, aggregation, optimization, dispatch, monitoring and measurement and verification. A compact statement of its role is

DERMS = the operational intelligence and control layer between DERs and grid objectives.

It is not the DER itself, and it is not the virtual power plant itself. Those two boundaries are the subject of Section 3.6.1; the body of the chapter describes what sits between them.

Jargon: DERMS

Pronounced "derms". The term is used both for a product class (software a utility or aggregator buys) and for the function (the loop described in this chapter). The function is what this book means; which vendor's product performs it is immaterial.

3.1.2 The operational chain

A DERMS sits in a loop of eight stages,

$$\text{connect} \rightarrow \text{observe} \rightarrow \text{estimate} \rightarrow \text{forecast} \rightarrow \text{aggregate} \rightarrow \text{optimize} \rightarrow \text{dispatch} \rightarrow \text{verify}, \tag{3.1}$$

after which the loop returns to "observe". Figure 3.1 draws the chain with the object each stage produces. The remainder of the chapter takes the stages in order; this decomposition is far more useful than the one-line description "software that controls distributed resources".

3.2 Sensing the fleet

The first four stages take the platform from raw device interfaces to a forecast of what each resource can do.

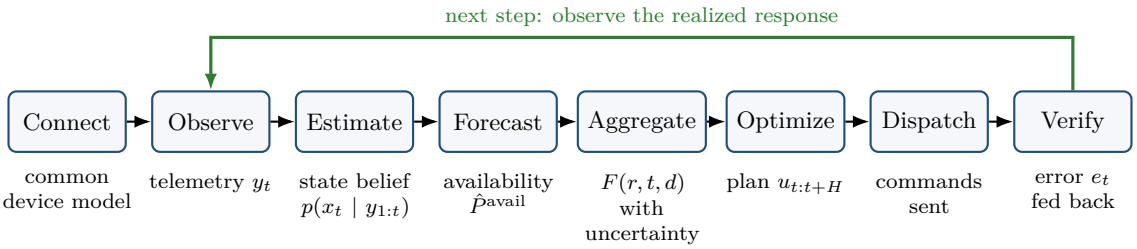


Figure 3.1: The eight-stage operational chain of a DERMS, equation (3.1), with the object each stage produces. The green return arrow closes the loop.

3.2.1 Connect

Before anything can be controlled the platform must communicate with the devices. A real fleet mixes thermostats, home batteries, vehicles from several manufacturers, public and private chargers, and commercial HVAC controllers, each exposing its own API and its own data model. The same physical quantity arrives under different names and units: Table 3.1 shows three ways the same battery state might be reported.

Table 3.1: The same quantity, three device vocabularies.

Device vendor	Field reported	Meaning
A	`SOC`	fraction of usable capacity, 0–1
B	`battery_percentage`	percent, 0–100
C	`energy_remaining`	kWh stored

The DERMS normalizes these into one internal representation, a *common DER model*, so that every downstream stage reasons about "a battery with state of charge SOC" and never about a vendor's field name. Figure 3.2 shows the integration layer. Without it, every optimization algorithm would carry device-specific logic, and adding a vendor would mean touching every algorithm.

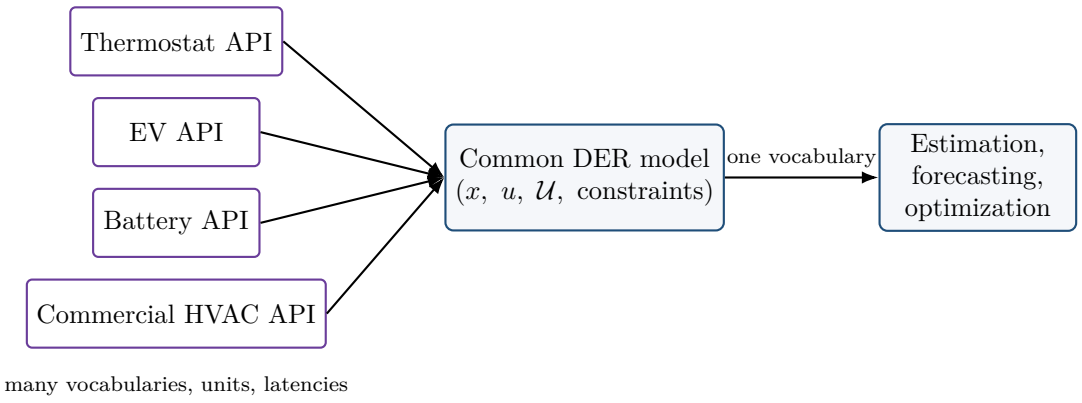


Figure 3.2: The integration layer. Each vendor API is translated once into the common DER model; every later stage sees only that model.

3.2.2 Observe

Once connected, the DERMS collects telemetry. For a thermostat the telemetry at step t might be

$$y_t = \{T_{\text{in}},\ T_{\text{set}},\ \text{HVAC state}\};$$

for a battery, $y_t = \{\text{SOC},\ P,\ \text{availability}\}$; for an EV, $y_t = \{\text{SOC},\ \text{plugged in},\ P_{\text{EV}}\}$.

The symbol is y_t, not x_t, and the distinction is deliberate. Telemetry is what is *observed*. The full physical state of Chapter 1 may be only partially observed, reported with delay, or reported with error. The relationship between the two is a measurement map,

$$y_t = h(x_t) + v_t, \tag{3.2}$$

where h selects and transforms the parts of the state the device exposes and v_t is measurement noise. The gap between x and y is what the next stage has to close.

3.2.3 Estimate the state

Suppose an EV API reports "plugged in: yes, charging power: 0 kW". It does not report the driver's planned departure time, yet t_{dep} is exactly what determines how long charging can be deferred. Part of the state must therefore be *estimated*: from the driver's history, from the day of the week, from the current state of charge.

The same holds for a building. The indoor temperature may be observed, but the thermal mass C, the envelope conductance $1/R$ of equation (1.13) in Chapter 1, the occupancy, and the short-term cycling behavior of the compressor are not. They have to be inferred from the temperature's response to past control.

Conceptually, then, the DERMS maintains a probability distribution over the state given everything observed so far,

$$p(x_t \mid y_{1:t}), \tag{3.3}$$

rather than assuming x_t is known. Figure 3.3 shows the structure: the state evolves by f, the observations are produced by h, and the estimator runs the two maps backward from observations to a belief about the state.

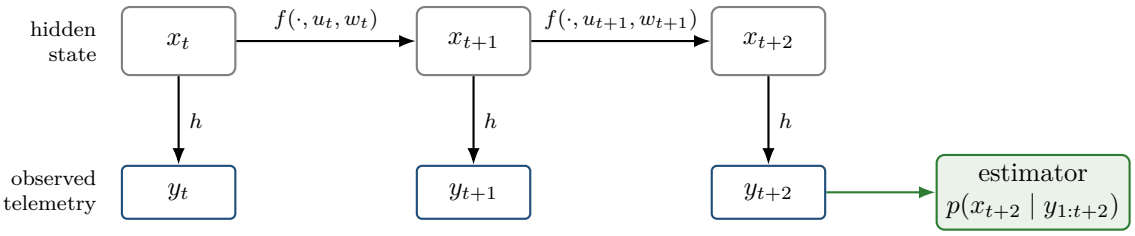


Figure 3.3: Hidden state and observed telemetry. The state advances by the transition map f; each observation is a noisy function h of the state. The estimator inverts the chain to produce a belief about the current state, which is what forecasting and optimization consume.

Many production systems implement this less formally than a full Bayesian filter, but the conceptual problem is the same: *what is the current condition of this resource?* An answer that ignores the unobserved part of the state will be confidently wrong about flexibility.

Principle 3.1. *A DERMS acts on an estimate of the state, not on the state. Telemetry y_t and state x_t are different objects, and the quality of everything downstream is bounded by the quality of $p(x_t \mid y_{1:t})$.*

3.2.4 Forecast

The next question is what the resource will be capable of later:

$$\hat{P}^{\text{avail}}(t+\tau), \tag{3.4}$$

the forecast of available power at a lead time τ. The inputs differ by device type, as Table 3.2 lists. This stage is the bridge to the forecasting track of the lesson sequence: every row is a prediction problem with its own data and its own uncertainty.

Table 3.2: Inputs to the availability forecast, by device type.

Device	Forecast depends on
Thermostat fleet	outdoor temperature, indoor temperatures, occupancy, historical cycling, recent event history (fatigue, overrides)
EVs	plug-in probability by hour, state of charge on arrival, departure behavior
Batteries	state of charge, expected customer usage, reserve constraints

3.3 Aggregation

Two stages turn many devices into resources a utility can reason about.

3.3.1 Aggregate

Utilities do not want to reason about 100 000 devices. They want answers such as "how much flexible capacity is there in Zone A from 17:00 to 19:00?". The DERMS therefore turns many individual devices into aggregate resources.

If device i has flexibility $F_i(t)$, the naive aggregate is the sum,

$$F_{\text{fleet}}(t) = \sum_i F_i(t). \tag{3.5}$$

The sum is fine for the expected value. It is wrong for the uncertainty, because device responses are not independent: every thermostat in a region experiences the same hot afternoon, so their available flexibilities are correlated. The variance of the sum is

$$\operatorname{Var}\Big(\sum_i F_i\Big) = \sum_i \operatorname{Var}(F_i) + 2\sum_{i<j} \operatorname{Cov}(F_i, F_j), \tag{3.6}$$

and the covariance term is not small. For N devices with a common per-device standard deviation σ and a common pairwise correlation ρ, the standard deviation of the fleet total, divided by the sum of the individual standard deviations, is

$$\frac{\sigma_{\text{fleet}}}{N\sigma} = \sqrt{\frac{1}{N} + \Big(1 - \frac{1}{N}\Big)\rho}. \tag{3.7}$$

Figure 3.4 plots this ratio against ρ for $N = 1\,000$. At $\rho = 0$ the fleet uncertainty is $1/\sqrt{N}$ of the naive sum: aggregation averages the noise away, which is the whole appeal of a fleet. At $\rho = 0.1$ the ratio has already risen tenfold, and at $\rho = 1$ the fleet is no more predictable than a single device scaled up.

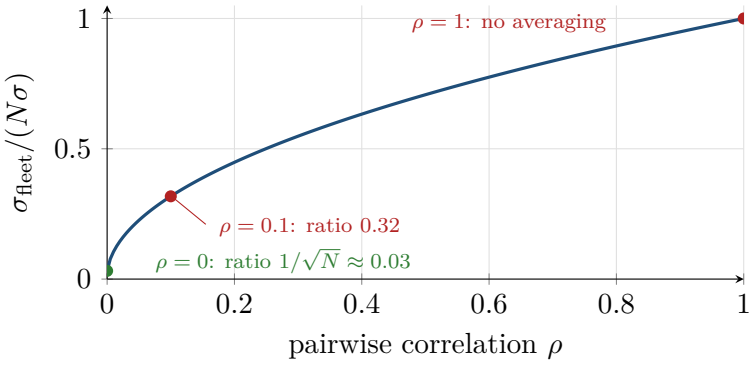


Figure 3.4: Fleet uncertainty relative to the naive sum of device uncertainties, equation (3.7), for $N = 1\,000$ devices. A small common correlation, such as shared weather, removes most of the benefit of aggregation.

This matters when the DERMS has to *guarantee* fleet capacity, as in the chance constraint of Chapter 2: the quantile that can be promised depends on σ_{fleet}, and σ_{fleet} depends on ρ far more than on N.

Principle 3.2. *Aggregation reduces uncertainty only to the extent that device responses are independent. Shared drivers such as weather put a floor under fleet uncertainty that no number of devices removes.*

3.3.2 Grouping

A DERMS also partitions the fleet into groups, by region, utility program, feeder, substation, device type, customer class, contractual availability, or expected flexibility. Figure 3.5 shows a typical tree. The reason is that grid value depends on location and on resource characteristics: a 5 MW reduction on one feeder may relieve a constraint while the same 5 MW elsewhere relieves nothing.

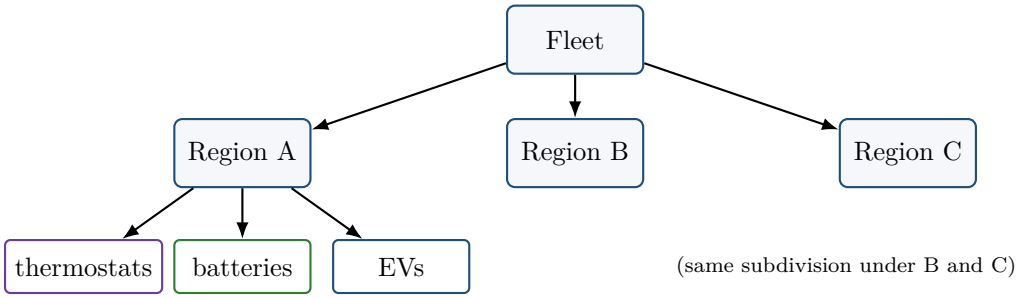


Figure 3.5: Grouping the fleet by region and then by device type. Each leaf is a resource with its own $F(r, t, d)$.

The fleet is therefore not a single number such as “100 MW” but a function of three arguments,

$$F(r, t, d), \tag{3.8}$$

with r the region or location, t the time, and d the DER type. Later chapters on aggregation and dispatch work with this object.

3.4 Decision and action

The next three stages choose actions, send them, and check what happened.

3.4.1 Optimize

Suppose the grid asks for 50 MW from 17:00 to 20:00, and the DERMS has thermostat flexibility, EV charging flexibility and battery discharge capacity. The question is

Which devices should be used, when, and by how much?

Mathematically,

$$\min_{u_{1:T}} J \quad \text{subject to} \quad x_{t+1} = f(x_t, u_t, w_t), \qquad u_t \in \mathcal{U}(x_t), \tag{3.9}$$

which is the EV problem of Chapter 2 generalized to every device at once. The objective is a sum of terms,

$$J = J_{\text{tracking}} + J_{\text{customer}} + J_{\text{cost}} + J_{\text{risk}}, \tag{3.10}$$

with, for example, $J_{\text{tracking}} = \sum_t (P_{\text{fleet}}(t) - P^{\text{tgt}}(t))^2$, a customer term penalizing discomfort or missed departure margins, a cost term for energy and battery degradation, and a risk term for the probability of shortfall. This stage is the center of the dispatch problem; the earlier stages exist to give it correct inputs.

3.4.2 Dispatch

After choosing the controls the DERMS sends commands. Table 3.3 lists what a command looks like for each device type. The sign of the battery command follows the convention of Chapter 1: negative is discharge.

Table 3.3: Example commands sent at dispatch.

Device	Command
Thermostat	$T_{\text{set}} : 22°\text{C} \to 23°\text{C}$
EV charger	$P_{\text{EV}} : 7.2\,\text{kW} \to 2\,\text{kW}$
Battery	$P_{\text{bat}} : 0 \to -5\,\text{kW}$ (discharge)

The commanded action and the realized action are different quantities,

$$u_t^{\text{commanded}} \neq u_t^{\text{realized}} \quad \text{in general,} \tag{3.11}$$

because a device may be offline, a message may be lost, a customer may override, the device's own local controller may refuse or modify the request, or the telemetry on which the command was based may have been stale. Dispatch therefore has to be monitored, which is the next stage.

3.4.3 Verify and re-optimize

Suppose the DERMS commanded 50 MW of reduction and the metered result is $P^{\text{del}} = 43\,\text{MW}$. The tracking error is

$$e_t = P^{\text{tgt}} - P^{\text{del}} = 7\,\text{MW}. \tag{3.12}$$

A system that stops at "forecast, optimize, dispatch, finished" records the error and does nothing. A closed-loop system responds: "7 MW short; dispatch another group". Concretely, at 17:05 the DERMS observes $P^{\text{tgt}} = 50\,\text{MW}$ and $P^{\text{del}} = 42\,\text{MW}$, and adds 4 MW of EV curtailment and

4 MW of battery discharge while the service continues. The resource mix changes during the event.

Figure 3.6 draws the loop. The DERMS is not executing a precomputed schedule; it is adapting to the realized behavior of the fleet, and "near real time" means that the measure-compare-re-optimize cycle is short compared with the duration of the service.

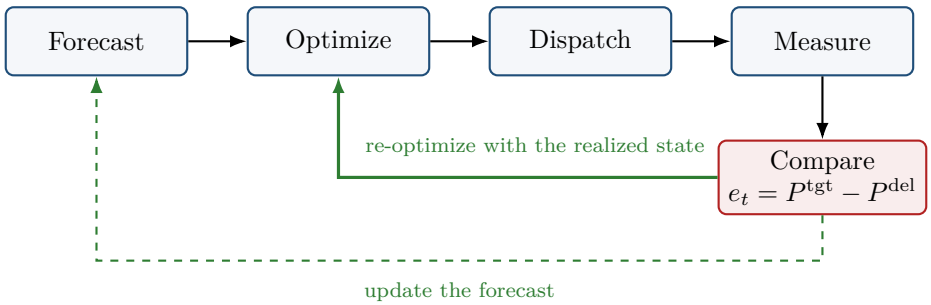


Figure 3.6: The verification loop. The comparison of target and delivered power feeds a re-optimization; the realized behavior also updates the forecast for the next cycle.

3.5 The DERMS as a control system

Seen as a whole, the chain has the structure of a known control scheme, and that structure says where forecasting belongs.

3.5.1 A control representation: model predictive control

The whole system can now be written compactly. There is a state x_t, an observation y_t, an external input w_t and a control u_t. The dynamics and the measurements are

$$x_{t+1} = f(x_t, u_t, w_t), \tag{3.13}$$

$$y_t = h(x_t) + v_t, \tag{3.14}$$

and the objective over a horizon of H steps is

$$\min_{u_{t:t+H}} \sum_{\tau=t}^{t+H} c(x_\tau, u_\tau) \quad \text{subject to the constraints.} \tag{3.15}$$

This is the structure of *model predictive control* (MPC). An MPC controller repeats five steps: estimate the current state; predict the future over the horizon; optimize the actions over that horizon; apply only the first action; observe the response; and repeat from a new estimate. Figure 3.7 draws the receding horizon. Even when a production DERMS is not literally one monolithic MPC solver, its architecture is this one, and the vocabulary of MPC (horizon, receding window, first-action application) describes it accurately.

3.5.2 Where forecasting lives

Forecasting is not a separate, decorative module whose output goes on a dashboard. It feeds the optimizer directly. If the optimizer believes the thermostat fleet can provide 40 MW and the fleet can actually provide 25 MW, the plan built on 40 MW fails during execution. The chain is

$$\text{forecast error} \rightarrow \text{dispatch error} \rightarrow \text{grid-service risk}, \tag{3.16}$$

which is why forecast quality has direct operational value: it determines control decisions and, through them, whether a committed service is delivered.

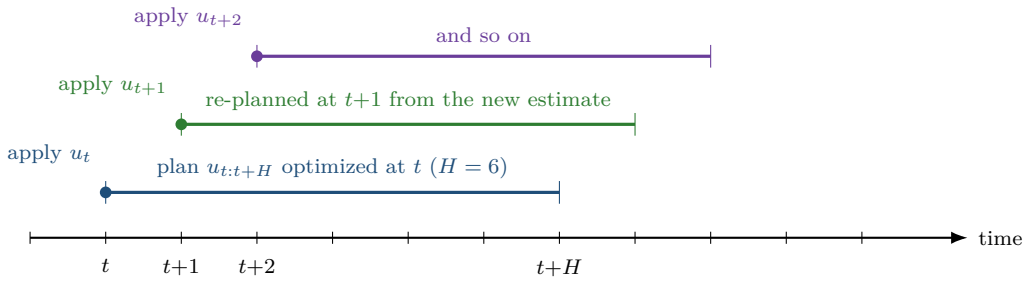


Figure 3.7: The receding horizon of model predictive control. At each step a plan over the next H steps is optimized from the current state estimate, only its first action is applied, and the plan is discarded and recomputed one step later.

3.6 Architecture and neighbors

This section places the DERMS among its neighbors in a utility's software and inside its own cyber-physical loop.

3.6.1 DERMS versus VPP, EMS and ADMS

Three neighboring terms are routinely confused with DERMS. Table 3.4 fixes the boundaries and Figure 3.8 places the systems in the utility's hierarchy.

Table 3.4: DERMS and its neighbors.

System	Objects it operates on	Typical functions
EMS	generators, transmission lines, substations	SCADA, state estimation, contingency analysis, economic dispatch, automatic generation control
ADMS	feeders, switches, voltage regulators, outages	feeder topology, switching, outage management, conservation voltage reduction, volt/VAR optimization
DERMS	thermostats, EVs, batteries, solar, distributed loads	connect, observe, estimate, forecast, aggregate, optimize, dispatch, verify
VPP	(not a system) the aggregated flexible resource the grid sees	provides grid services; operated *by* the DERMS

The DERMS/VPP distinction is the most important:

$$\text{DERMS} = \text{the system used to operate the VPP}, \qquad \text{VPP} = \text{the aggregated flexible resource}. \tag{3.17}$$

DERs are coordinated by the DERMS; the DERMS produces VPP behavior; the VPP delivers a grid service. The VPP is the resource seen by the grid; the DERMS is the machinery that makes that resource operable.

EMS and ADMS are the utility's own systems. An EMS runs the bulk transmission grid from a control center. An ADMS runs the distribution network itself: its topology, voltage and outages. A DERMS orchestrates the resources connected to that network. ADMS and DERMS overlap and, in some architectures, integrate tightly, because the resources the DERMS dispatches change the voltages and flows the ADMS manages.

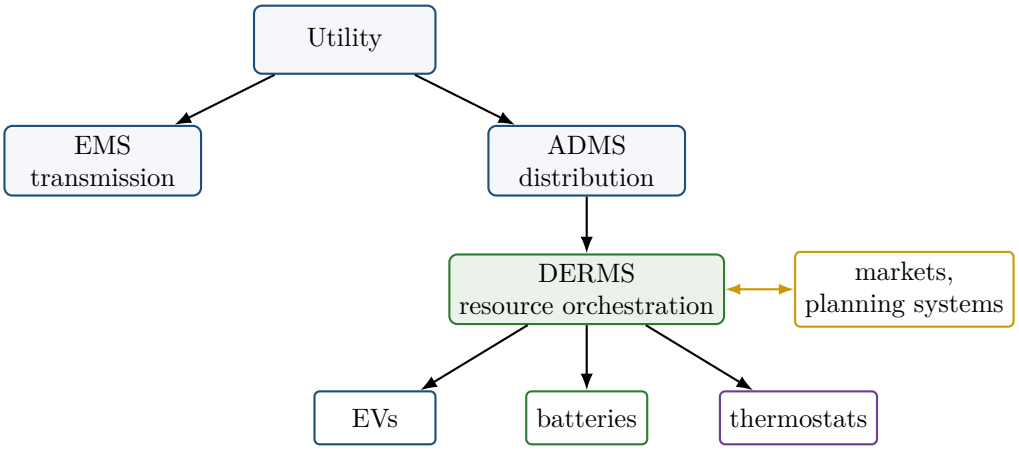


Figure 3.8: A simplified hierarchy. The EMS and ADMS are the utility's grid-operating systems; the DERMS sits beneath the ADMS and above the devices, and also exchanges information with market and planning systems.

3.6.2 Centralized and hierarchical control

With a million DERs, does a central optimizer compute every control signal individually? It can, in principle: $u_1, u_2, \ldots, u_{1\,000\,000}$ computed centrally. In practice a hierarchical structure scales better:

system objective → regional fleet targets → device-group targets → local or device control.

The central optimizer then works with group targets $P^{\text{tgt}}_{\text{group}}$ rather than every compressor state, and local logic decides which devices in the group respond. Figure 3.9 contrasts the two architectures. The hierarchical one is why the grouping of Section 3.3.2 exists, and the chapter on fleet aggregation returns to it.

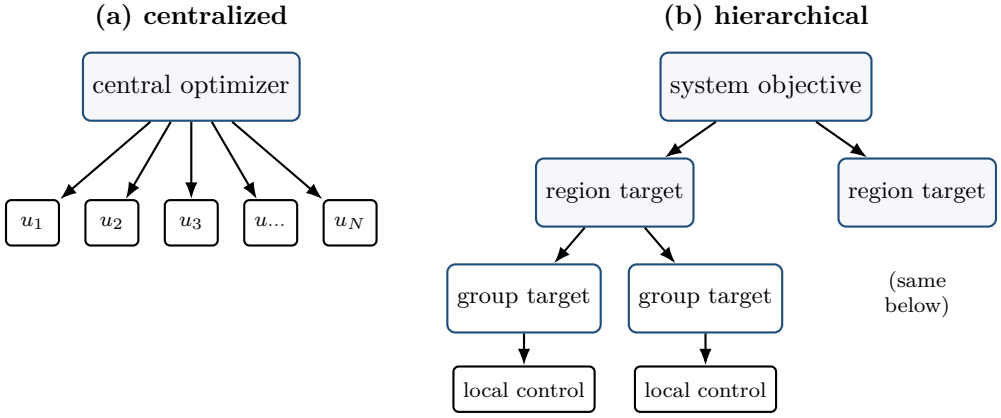


Figure 3.9: (a) A central optimizer computing every device's action. (b) A hierarchy of targets, with device-level decisions made locally. The hierarchy scales because each level optimizes over a few targets rather than a million actions.

3.6.3 A cyber-physical system

A DERMS lives in two worlds. The cyber world is APIs, telemetry, communication, databases, forecasts and algorithms. The physical world is temperatures, batteries, power flows, charging and customer behavior. The DERMS sends a digital command u_t, but its value comes from the change it produces in a physical system, $x_{t+1} = f(x_t, u_t, w_t)$. The compact statement is

DERMS = cyber control of distributed physical systems.

A reader who thinks naturally in physical states and models will find that every stage of Section 3.1.2 is either a model of the physical world (estimate, forecast, aggregate) or a decision about it (optimize, dispatch), with the cyber stages (connect, observe, verify) as the interface.

3.6.4 The core loop in one figure

Figure 3.10 compresses the chapter into a single loop. A grid need enters the optimizer; the optimizer emits controls u to the DERs; the DERs respond physically; telemetry y returns; state estimation and forecasting convert telemetry into a belief about the fleet; and that belief re-enters the optimizer. This is the picture to retain.

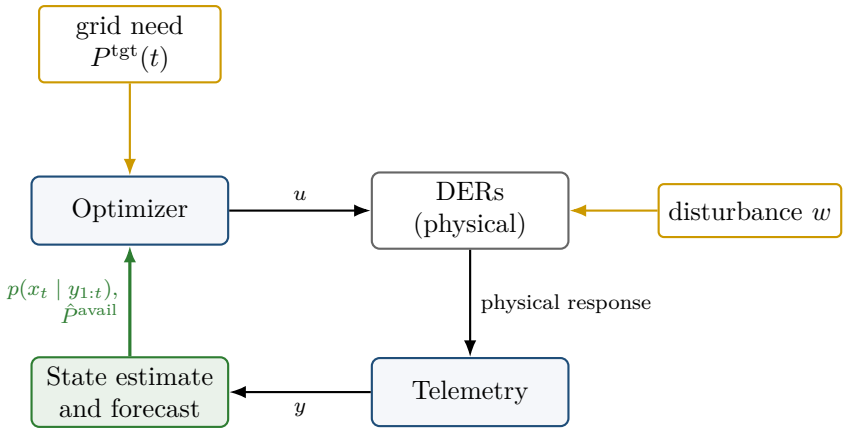


Figure 3.10: The DERMS core loop. Commands go out along the top, physical responses and telemetry come back along the bottom, and the estimate of the fleet's state closes the loop into the optimizer.

3.7 The working definition

Definition 3.1 (DERMS)**.** A DERMS is the operational software layer that connects distributed energy resources, estimates and forecasts their flexibility, aggregates them into useful grid resources, optimizes how they should be dispatched, sends control signals, and verifies the realized response.

The concept to retain is that a DERMS turns heterogeneous devices into controllable grid capability, through

$$\text{observation} + \text{forecasting} + \text{aggregation} + \text{optimization} + \text{control}.$$

The VPP is what the grid ultimately sees; the DERMS is what makes that VPP operable. Table 3.5 lists, for each stage of the chain, the question it answers and the object it hands to the next stage.

3.8 Exercises

Exercise 3.1. A battery vendor reports `energy_remaining` = 8.1 kWh for a pack with 13.5 kWh usable. Convert this to the common model's SOC. If the owner's reserve is 20 %, what is E^{avail} for discharge?

Table 3.5: The eight stages, the question each answers, and its output.

Stage	Question	Output
Connect	How do I talk to this device?	common DER model
Observe	What does it report?	telemetry y_t
Estimate	What is its actual condition?	$p(x_t \mid y_{1:t})$
Forecast	What can it do later?	$\hat{P}^{\text{avail}}(t+\tau)$
Aggregate	How much is there, where, of what kind?	$F(r,t,d)$ with uncertainty
Optimize	Which devices, when, how much?	plan $u_{t:t+H}$
Dispatch	Send it.	commands
Verify	Did it happen?	e_t, fed back

Exercise 3.2. Using equation (3.7), how many devices N are needed to bring $\sigma_{\text{fleet}}/(N\sigma)$ below 0.1 when $\rho = 0$? Show that no N achieves this when $\rho = 0.02$.

Exercise 3.3. At 17:05 the target is 50 MW and the delivered reduction is 42 MW. Available uncommitted resources are 6 MW of EV curtailment and 5 MW of battery discharge. Write the re-dispatch, and state what happens to the battery's capability for the rest of the event.

Exercise 3.4. For each function below, say whether it belongs to the EMS, the ADMS or the DERMS, and why: automatic generation control; volt/VAR optimization on a feeder; forecasting EV plug-in probability; switching to isolate a faulted feeder section; grouping thermostats by substation.

3.9 Connection to the TERRA framework

The eight-stage chain of this chapter is, almost stage for stage, the workflow of a graph-based physical model with inference on top.

- **Connect is building the graph.** The common DER model of Section 3.2.1 is the declaration of components, their states and their constraints in one vocabulary; in a conservation-graph framework that is the model definition, with each vendor's API reduced to the component's state variables and the edges it attaches to.
- **Observe and estimate are measurement and inference.** Equation (3.2), $y_t = h(x_t) + v_t$, is the measurement model; equation (3.3), $p(x_t \mid y_{1:t})$, is the posterior over the state. Inferring an EV's departure time or a building's thermal mass from partial telemetry is exactly the problem of estimating unmeasured states and parameters on a graph from the measurements that are available, with a posterior rather than a point estimate.
- **Forecast is forward integration from the posterior.** The availability forecast $\hat{P}^{\text{avail}}(t+\tau)$ is the predicted edge flow after integrating the transition map forward from the current belief, with the disturbance w drawn from its own forecast. The uncertainty of the forecast is the propagated uncertainty of the posterior.
- **Aggregate is the graph at larger scale, with the covariance kept.** Equation (3.6) is the statement that the posterior over a sum is not the sum of marginal posteriors. A model that carries the joint posterior over all devices, including the shared weather that correlates them, reports σ_{fleet} correctly; one that sums marginals underestimates it by the factor of Figure 3.4.

- **Optimize and dispatch are decisions on the model.** The MPC structure of Section 3.5.1 uses the graph as its prediction model and the posterior as its initial condition. The hierarchy of Section 3.6.2 is a decomposition of the graph into regions with aggregate targets at the interfaces.
- **Verify is conditioning on new data.** The tracking error e_t is a residual between prediction and measurement; feeding it back is updating the posterior with the new observation, after which the next forecast and optimization start from a corrected belief.

In summary, a DERMS is a loop of model-building, inference, prediction, decision and conditioning on a population of physical devices. A framework that treats a physical system as a conservation graph with probabilistic inference supplies the middle four stages of that loop directly, and the operational software supplies the connections at either end.

Chapter 4

Virtual power plants: when does a collection of DERs become a grid resource?

4.1 Many devices and one resource

Chapter 3 fixed the boundary: the virtual power plant (VPP) is the aggregated resource created by coordinating DERs, and the DERMS is the system used to operate it. This chapter is about the resource itself, and its central statement is:

Key idea

A VPP is not many DERs. It is many DERs coordinated so that the grid can treat them as *one* resource.

The word "one" carries the whole chapter. A single resource has a capacity, a duration, a location and a reliability, and can be committed to a utility or a market on those terms. A pile of devices has none of those properties until something computes them. The sections below develop, in order, what capacity means for a fleet, why aggregation helps and where it stops helping, what a service commitment contains, why heterogeneous portfolios are stronger than homogeneous ones, how the fleet's internal complexity is hidden from the grid, and how the commitment decision is made under uncertainty.

4.2 Capacity

The first task is to say what capacity means for a fleet, which turns out to be three different numbers, and why aggregation narrows the gap between them only up to a point.

4.2.1 Nameplate, available and deliverable capacity

Consider one home battery that can provide 5 kW for two hours. Useful locally; negligible at utility scale. Now consider 10 000 such batteries. A naive calculation gives

$$10\,000 \times 5\,\text{kW} = 50\,\text{MW},$$

which is utility scale. But that is not automatically a 50 MW VPP, because at any given moment some batteries are offline, some have low state of charge, some are reserved for backup, some owners have opted out, some communications fail, and some can respond for only twenty minutes. Installed capacity and dispatchable capacity are different quantities.

Let battery i have nameplate power $P_i^{\max}$. The fleet's nameplate capacity is

$$P^{\text{name}} = \sum_i P_i^{\max}, \tag{4.1}$$

while its available capacity is

$$P^{\text{avail}} = \sum_i a_i \, P_i^{\max}, \qquad a_i \in [0, 1], \tag{4.2}$$

where a_i is an *availability factor*: 0 if the device is unavailable, 1 if fully available, and in between if partially constrained (for example by a state of charge near its reserve, as in Figure 1.2 of Chapter 1). Two inequalities follow, and the second is the more important:

$$P^{\text{name}} \neq P^{\text{avail}}, \tag{4.3}$$

$$P^{\text{avail}} \neq P^{\text{reliably deliverable}}. \tag{4.4}$$

The first says that what is installed is not what is available now. The second says that what is available now, on average, is not what can be *promised*. Figure 4.1 draws the three numbers as nested bars; the rest of the chapter is about the gap between the second and the third.

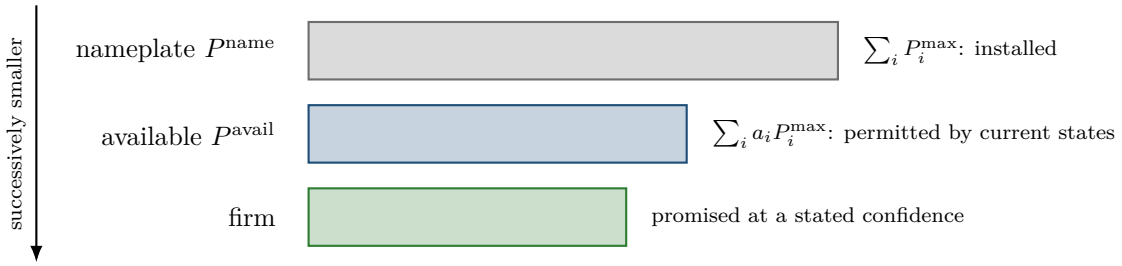


Figure 4.1: Three capacities of the same fleet. Nameplate counts what is installed; available applies the current availability factors; firm is the amount that can be delivered with a stated probability. Bar lengths are illustrative.

4.2.2 Reliability and derating

Suppose the forecast says $P^{\text{avail}} = 50\,\text{MW}$ and the utility asks "can you guarantee 45 MW for two hours?". That is a different question. Available fleet power is a random variable with a distribution $p(P^{\text{avail}})$, and the utility's question is a chance constraint of the kind met in Chapter 2,

$$\Pr(P^{\text{avail}} \geq 45\,\text{MW}) \geq 0.95. \tag{4.5}$$

If that holds, the useful VPP capacity is 45 MW even though the mean is 50 MW. This is the fleet analog of *derating* a conventional generator: the capacity credited to a plant is below its nameplate to account for forced outages, and the capacity credited to a VPP is below its mean availability to account for everything that can keep devices from responding.

4.2.3 Diversity from aggregation

Aggregation is powerful because device behavior is heterogeneous. In a population of 1 000 thermostats, not all compressors cycle at the same time: one home's air conditioner switches off as another's switches on. Each device's power is an on/off trace; the fleet's power is the sum,

$$P_{\text{fleet}}(t) = \sum_i P_i(t), \tag{4.6}$$

and the sum is far smoother than any term. Figure 4.2 shows the effect with three devices: each panel on the left is a binary trace, and the trace on the right is their sum. With three devices the sum still steps; with thousands it is nearly continuous. The individual signals are noisy; their aggregate is predictable. That predictability is one of the main sources of value in a fleet.

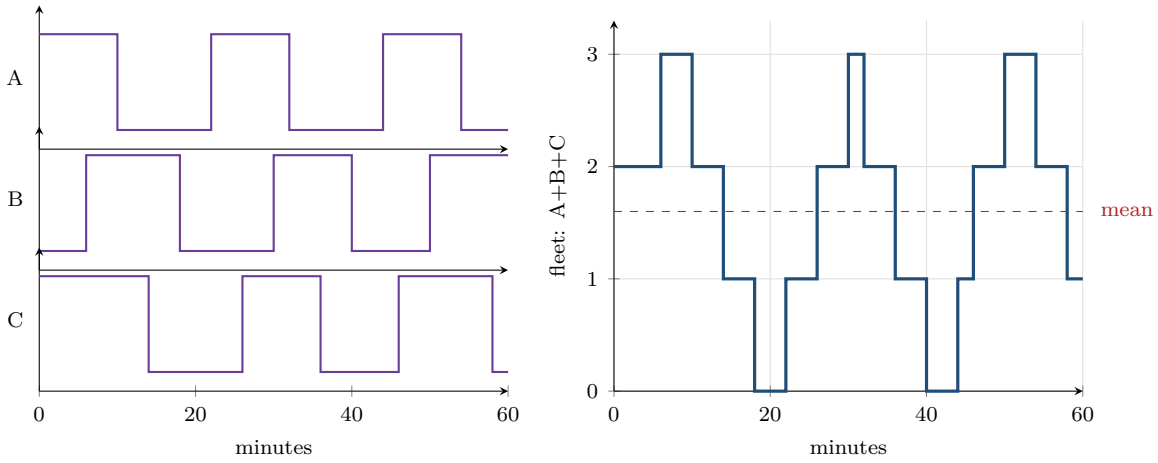


Figure 4.2: Diversity. Left: three compressors cycling out of phase. Right: their sum, which varies around a mean rather than between zero and full. With thousands of devices the sum becomes a smooth curve. Traces are illustrative.

4.2.4 Correlation limits diversity

Aggregation does not eliminate uncertainty, because device responses are not independent. If all 100 000 thermostats are in the same heat wave, the same T_{out} moves all of them at once and their availabilities are correlated. Chapter 3 gave the variance of the fleet total, equation (3.6), and showed in Figure 3.4 that a pairwise correlation as small as 0.1 removes most of the averaging benefit of a thousand devices. The same figure applies here without change: the diversity of Section 4.2.3 is the $\rho = 0$ end of that curve, and weather pushes a thermostat fleet toward the other end. This is why weather forecasting is central to a thermostat-based VPP: the weather is the common cause, and forecasting it is the only way to know where on the curve tomorrow's fleet will be.

4.3 The service product

A resource is something that can be committed. This section defines the commitment, the duration dependence of capacity, and the portfolio that makes a commitment holdable.

4.3.1 The service commitment

A collection of devices becomes operationally meaningful when the operator can say

From 17:00 to 19:00 we can provide 80 MW of load reduction.

That sentence contains a magnitude, a time window, a direction, and (implicitly or explicitly) a confidence. A VPP resource is therefore not a number but a tuple,

$$R = (P,\ t_{\text{start}},\ t_{\text{end}},\ \text{location},\ \text{reliability}), \tag{4.7}$$

rather than $R = 80\,\text{MW}$. Table 4.1 lists the fields and what each one constrains.

Table 4.1: The fields of a VPP service commitment.

Field	What it fixes
Magnitude P	the power change, in MW
Window $[t_{\text{start}}, t_{\text{end}}]$	when it starts and how long it must last; sets the duration τ
Direction	reduction, increase, or both (Chapter 2)
Location	which feeder, substation or zone; sets which devices count
Reliability	the probability with which P is delivered over the window

4.3.2 Capacity as a function of duration

Fleet A can provide 100 MW for ten minutes. Fleet B can provide 60 MW for four hours. Which is bigger? The question is underspecified; they provide different products. VPP capacity is always tied to a service specification, and the cleanest way to express that is the deliverable capacity as a function of required duration,

$$C(\tau) = \text{the largest constant power that can be held for } \tau. \tag{4.8}$$

For fleet A, $C(15\,\text{min}) = 100\,\text{MW}$ while $C(4\,\text{h})$ is far smaller; for fleet B, $C(\tau) = 60\,\text{MW}$ out to four hours. Figure 4.3 draws both. The fleet-level $C(\tau)$ is the aggregate of the device-level $F(t, \tau)$ of Chapter 1, and it carries the same message: a resource is a curve, not a number.

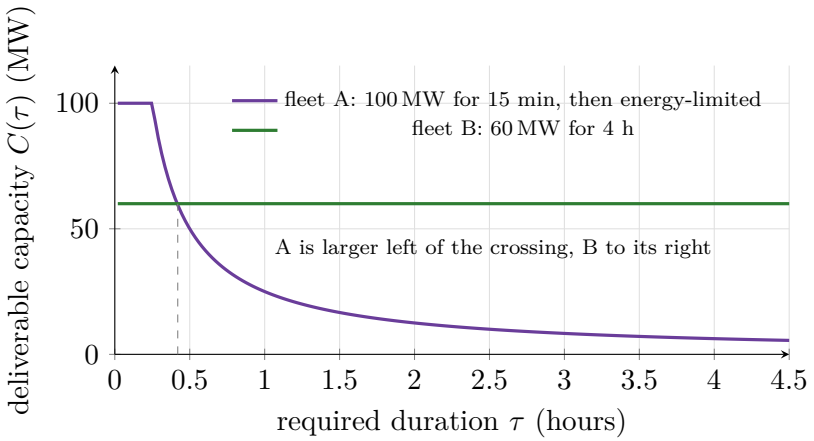


Figure 4.3: Deliverable capacity versus required duration for two fleets. Fleet A (purple) holds 100 MW only briefly and then falls as its energy is spent; fleet B (green) holds 60 MW for four hours. Neither is "bigger"; which one a service needs depends on τ. Curves are illustrative.

4.3.3 Heterogeneous portfolios

A VPP portfolio typically mixes thermostats, EVs and batteries, and each has different dynamics. Table 4.2 summarizes what each class does well and what limits it, in the vocabulary of the earlier chapters. The limits are complementary: the thermostat's limit is duration, the EV's is energy, the battery's is state of charge, and they do not bind at the same time.

Table 4.2: Strengths and limits of the three main DER classes.

Class	Good at	Limited by	Chapter
Thermostats	temporary load reduction	comfort band, hence duration	5
EVs	load shifting	departure requirement, hence energy	later
Batteries	fast, bidirectional response	state of charge, hence duration	later

4.3.4 Portfolio complementarity

Suppose the utility requests 50 MW for four hours. Thermostats alone cannot sustain it; batteries alone run out of energy; EVs alone may not have enough plugged-in load. Together they can, by handing the service from one class to another as each one's limit approaches, as in Table 4.3 and Figure 4.4. The fleet holds 50 MW for four hours even though no single class could, which is the strongest argument for building heterogeneous portfolios.

Table 4.3: A hand-off schedule for 50 MW over four hours (MW).

Hour	Thermostats	EVs	Batteries	Total
1	30	0	20	50
2	20	30	0	50
3	0	30	20	50
4	0	25	25	50

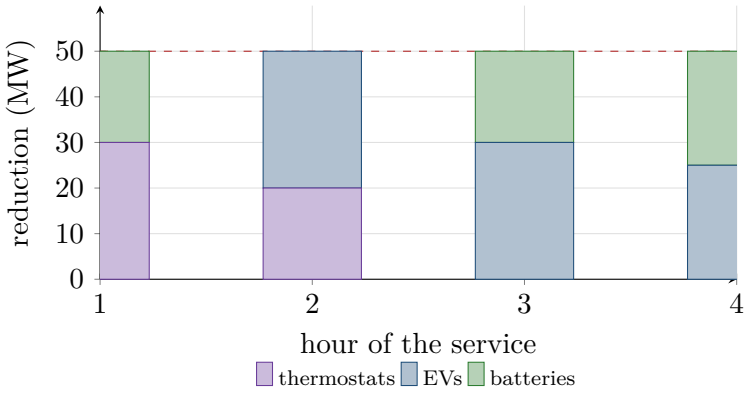


Figure 4.4: The hand-off schedule of Table 4.3. The thermostat contribution is front-loaded and fades; EVs and batteries take over. The dashed line is the 50 MW request.

4.3.5 Portfolio optimization

The hand-off is the output of an optimization. Let $u_{i,t}$ be the dispatch of resource i at step t, so the fleet response is

$$P_{\text{fleet},t} = \sum_i u_{i,t}, \tag{4.9}$$

and let $P_{\text{target},t}$ be the request. The coordination problem is

$$\min_u \sum_t \left(P_{\text{fleet},t} - P_{\text{target},t}\right)^2 \quad \text{subject to} \quad u_{i,t} \in \mathcal{U}_i(x_{i,t}), \tag{4.10}$$

where $\mathcal{U}_i(x_{i,t})$ is the state-dependent feasible set of resource i from Chapter 1. Equation (4.10) is the mathematical center of VPP coordination. The constraint is what makes it interesting:

the feasible sets shrink as resources are used, so the optimizer must plan the hand-off rather than spend everything in hour one.

4.4 The VPP as a resource abstraction

The VPP differs from a generator in kind, not only in size; this section states the difference and the abstraction that hides it.

4.4.1 Why a VPP is not a generator

A conventional generator can be modeled, roughly, by a fixed range,

$$P^{\min} \le P_g \le P^{\max}, \tag{4.11}$$

plus ramp-rate and other operating constraints. A VPP's feasible power range is not fixed:

$$P_{\mathrm{VPP}}(t) \in \mathcal{F}(x_t, w_t). \tag{4.12}$$

The region $\mathcal{F}$ changes with weather, device states, customer behavior, state of charge, indoor temperatures, connectivity and previous dispatch. Figure 4.5 contrasts the two: the generator's range is a fixed band, while the VPP's range is a band whose edges move with time and, crucially, move in response to the VPP's own actions.

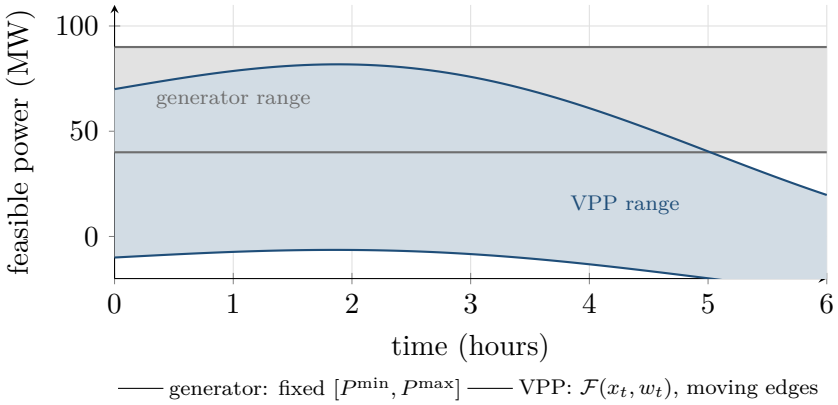


Figure 4.5: Feasible power of a generator (gray band, fixed) and of a VPP (blue band, time-varying). The VPP's lower edge dips below zero where the fleet can add load. The drift of the blue band with time stands for the dependence on prior dispatch. Curves are illustrative.

Principle 4.1. *VPP capacity is* endogenous and time-varying*: the feasible region depends on the state, and the state depends on what the VPP has already done.*

4.4.2 The hidden internal state and the resource abstraction

The grid sees one number, $P_{\mathrm{VPP}}(t)$. Internally the VPP contains the states of every enrolled device,

$$x_t = \begin{bmatrix} x_{1,t} \\ x_{2,t} \\ \vdots \\ x_{N,t} \end{bmatrix}, \tag{4.13}$$

with N in the hundreds of thousands. The utility does not want to manage $x_{1,t}, \ldots, x_{N,t}$. It wants a simplified resource abstraction such as $0 \le P_{\mathrm{VPP}} \le 80\,\mathrm{MW}$ over a window. Hiding the internal complexity behind that abstraction is one of the DERMS's most important functions.

4.4.3 Model reduction

There is a useful mathematical name for this. The physical fleet is a state in a very high-dimensional space, $x \in \mathbb{R}^N$ (or larger). The utility needs only a low-dimensional representation

$$z = \phi(x), \qquad z = \{P^{\text{avail}},\ E^{\text{avail}},\ \text{ramp capability},\ \text{duration},\ \text{location}\}, \tag{4.14}$$

so the DERMS implements a map from millions of device states to a small grid-resource description, Figure 4.6. This is aggregation viewed as *model reduction*: z keeps exactly the quantities a dispatch decision needs and discards the rest. The map ϕ is not a projection in the linear-algebra sense; computing E^{avail} or "duration" requires simulating the devices forward, which is why the reduction is a job for the DERMS and not for the utility.

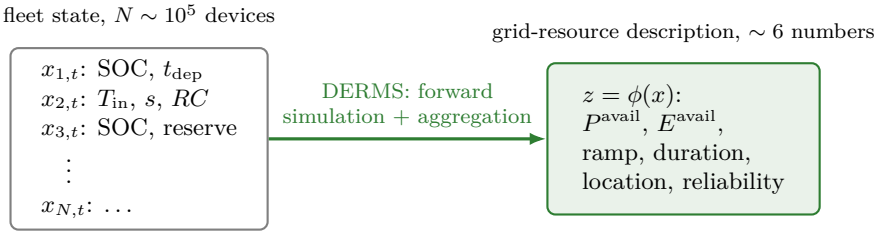


Figure 4.6: Model reduction. The DERMS maps the full fleet state to the handful of quantities a grid operator needs to make a dispatch decision.

4.5 Commitment under uncertainty

How much to enroll, how much to commit, and how to account for customers and tail risk are decisions under uncertainty, treated here.

4.5.1 Overbooking and the coverage ratio

Suppose a utility needs 100 MW. The aggregator may enroll devices whose nominal flexibility totals 150 MW, because it expects some customers to opt out, some devices to be offline and some forecasts to be wrong. This is the same logic as airline overbooking: when each enrolled unit has uncertain availability, deliberately hold more potential capacity than the commitment. Thus

$$P^{\text{enroll}} > P^{\text{commit}}, \qquad \text{coverage ratio} = \frac{P^{\text{enroll}}}{P^{\text{commit}}} = \frac{150}{100} = 1.5. \tag{4.15}$$

The coverage ratio is a risk-management decision, and its right value depends on the fleet's uncertainty: a ratio of 1.5 is generous for a battery fleet with a_i near one and thin for a thermostat fleet in a heat wave.

4.5.2 Overbooking is not free

Enrolling too little leads to $P^{\text{del}} < P^{\text{commit}}$, which is a failed commitment. But dispatching the enrolled fleet too aggressively has its own costs: customers become uncomfortable, opt-outs increase, batteries cycle unnecessarily, and future availability falls. The optimizer must therefore balance

$$\text{reliability} \quad \text{versus} \quad \text{customer impact},$$

which is one of the recurring trade-offs of VPP operation. It appears in the objective of Chapter 3 as the customer term J_{customer} beside the tracking term.

4.5.3 Customer behavior as part of the resource model

Suppose a thermostat customer overrides an event with probability $p^{\text{ovr}} = 0.15$. Under the simplest model, the expected response of a nominal 3 kW thermostat is

$$\mathbb{E}[P] = 3\,(1 - 0.15) = 2.55\,\text{kW}, \tag{4.16}$$

and at fleet scale

$$\mathbb{E}[P_{\text{fleet}}] = \sum_i P_i\,(1 - p_i^{\text{ovr}}). \tag{4.17}$$

The override probability is not a constant. It depends on temperature, time of day, event duration and the number of prior events a customer has experienced. So even "customer behavior" is a forecastable dynamic quantity, and it belongs inside the resource model rather than outside it as a fudge factor.

4.5.4 Performance guarantees and tail risk

Suppose the VPP commits 100 MW into a market or utility program and delivers 85 MW. The 15 MW shortfall creates penalties, reliability concerns, lost revenue and reduced trust. The operator therefore cares not only about $\mathbb{E}[P]$ but about $\Pr(P < 100)$, and often about the *size* of the shortfall when it occurs, measured by a tail statistic such as the conditional value at risk (CVaR): the expected delivery in the worst few percent of outcomes. All of these are properties of the distribution of P, not of its mean, which is why probabilistic forecasting is a requirement rather than a refinement.

Jargon: CVaR

The conditional value at risk at level α is the average outcome in the worst $(1 - \alpha)$ fraction of cases. For a VPP with $\alpha = 0.95$, it is the mean delivered power on the worst 5 % of days. Unlike a shortfall probability it says how bad the bad days are.

4.5.5 A probabilistic commitment model

Take available capacity to be

$$P^{\text{avail}} \sim \mathcal{N}(120, 10^2)\ \text{MW}. \tag{4.18}$$

Committing 100 MW leaves a wide margin: the shortfall probability is $\Phi((100 - 120)/10) = \Phi(-2) \approx 0.023$. Committing 118 MW raises revenue but the shortfall probability rises to $\Phi(-0.2) \approx 0.42$. The commitment decision is therefore an optimization,

$$\max_{P^{\text{commit}}} \left[\text{revenue}(P^{\text{commit}}) - \text{expected penalty}(P^{\text{commit}})\right]. \tag{4.19}$$

With revenue proportional to the commitment, $r\,P^{\text{commit}}$, and a penalty proportional to the expected shortfall, $\pi\,\mathbb{E}[(P^{\text{commit}} - P^{\text{avail}})^+]$, the expected shortfall of a normal variable has the closed form $\sigma\,[\varphi(z) + z\,\Phi(z)]$ with $z = (P^{\text{commit}} - \mu)/\sigma$, where φ and Φ are the standard normal density and distribution function. Figure 4.7 plots the objective for $r = 1$ and $\pi = 3$ (relative units): it rises linearly while the commitment is safely below the mean, bends over as the shortfall risk grows, and peaks at 116 MW, which is where the marginal penalty $\pi\,\Phi(z)$ equals the marginal revenue r, that is, at $\Phi(z) = 1/3$. Forecasting and optimization are inseparable here: the optimum depends on σ as much as on μ, and a narrower forecast moves the optimum toward the mean.

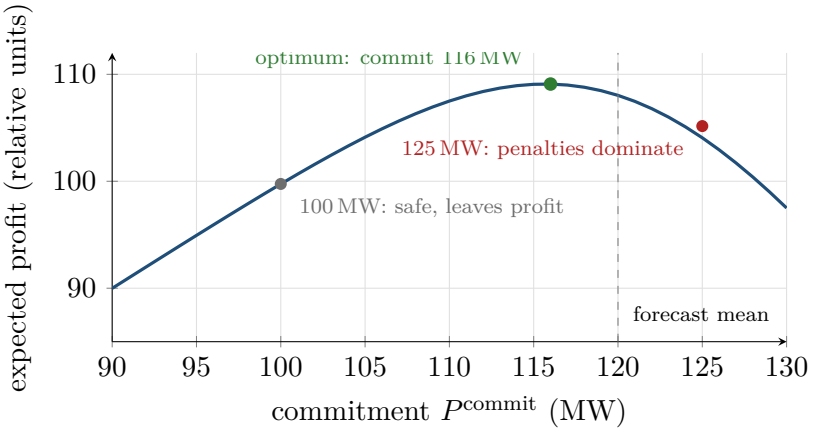


Figure 4.7: Expected profit versus committed capacity for $P^{\text{avail}} \sim \mathcal{N}(120, 10^2)$ MW, revenue $r = 1$ per MW committed and penalty $\pi = 3$ per MW short (relative units). Data from `scripts/ch45_data.py`. The optimum lies below the forecast mean by $0.43\,\sigma$.

4.6 Operation and forecasting

The chapter closes its body with the two views of one resource, the forecasts that feed the commitment, and a worked example of expected versus firm capacity.

4.6.1 The abstraction seen by the grid, and what the DERMS must answer

From the grid's side the VPP should look like a resource with an available megawatt figure and a set of constraints, which is dispatched and produces a response. The operator should not need to know whether 50 MW came from 5 000 batteries, 50 000 thermostats or 10 000 EVs. From the DERMS's side the same 50 MW is the answer to five questions: which devices, how much, for how long, at what customer cost, with what reliability. Figure 4.8 draws the two views of one resource. The VPP is a simple external resource built from extremely complex internal coordination, and the DERMS is the layer at which the complexity stops.

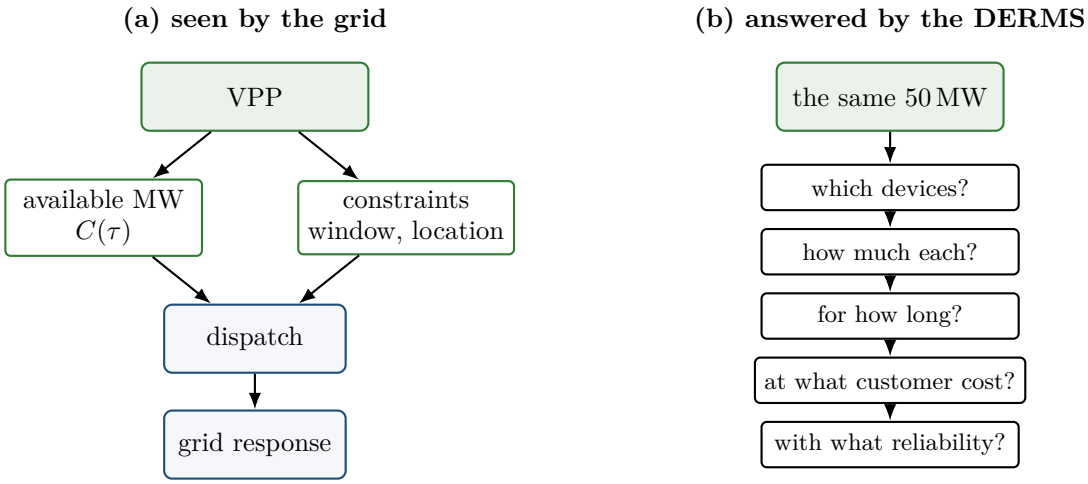


Figure 4.8: One resource, two views. (a) The grid sees capacity, constraints and a response. (b) The DERMS must answer five questions to produce that response.

4.6.2 Forecasts as the input to commitment

If the utility asks for 75 MW at 18:00 tomorrow, the platform needs, before any optimization, forecasts of the available power of each class, P_{T} for thermostats, P_{E} for EVs and P_{B} for batteries, with $P^{\mathrm{total}} = P_{\mathrm{T}} + P_{\mathrm{E}} + P_{\mathrm{B}}$. These are distributions, not numbers, and they are not independent, so the actual input to the commitment decision is a joint distribution,

$$p(P_{\mathrm{T}}, P_{\mathrm{E}}, P_{\mathrm{B}} \,|\, \text{information}) \,, \tag{4.20}$$

from which the optimizer decides what commitment is prudent. This is where the forecasting track and the operations track of the lesson sequence converge.

4.6.3 Worked example: expected and firm capacity

Table 4.4 gives tomorrow's 18:00 forecast for the three classes, as an expected value and as a 90 %-reliable value (the 10th percentile of each class's distribution).

Table 4.4: Forecast flexibility by class for 18:00 tomorrow.

Resource	Expected flexibility	90 %-reliable flexibility
Thermostats	40 MW	30 MW
EVs	30 MW	22 MW
Batteries	25 MW	24 MW
Total	95 MW	76 MW

The expected total is $40 + 30 + 25 = 95$ MW, and a naive operator would call that the available capacity. If reliability matters the prudent number is nearer the sum of the reliable values, $30 + 22 + 24 = 76$ MW. The gap of 19 MW is the price of a guarantee.

One refinement is worth stating, because it shows that the sum of the three reliable values is itself a modeling choice. The 10th percentile of a sum is not the sum of the 10th percentiles. If the three classes are perfectly correlated (a hot day hurts all three together), the percentile of the sum *is* the sum of the percentiles, 76 MW. If they are independent, their errors partly cancel and the 10th percentile of the total is higher: under a normal model with the standard deviations implied by the table, the total has mean 95 MW and standard deviation 10.0 MW, so its 10th percentile is about 82 MW. The firm capacity therefore lies between 76 and 82 MW depending on how correlated the classes are, which is the same lesson as Section 4.2.4 at the portfolio level. Either way,

$$\text{expected capacity} \neq \text{firm capacity}. \tag{4.21}$$

4.7 The working definition

Definition 4.1 (Virtual power plant)**.** A VPP is an aggregated, forecastable, dispatchable and verifiable grid resource built from many DERs. The DERMS hides the heterogeneous physical devices behind a simpler operational abstraction, and that abstraction must answer: *how much can be delivered, for how long, where, and with what confidence?*

Each adjective in the definition names a chapter: aggregated (Sections 4.2.3 to 4.4.3), forecastable (Sections 4.6.2 and 4.6.3), dispatchable (Sections 4.3.4 and 4.3.5), verifiable (Chapter 3). Table 4.5 collects the chapter's quantities.

Table 4.5: The quantities that describe a VPP, from installed to committed.

Quantity	Definition	Section
P^{name}	$\sum_i P_i^{\max}$	4.2.1
P^{avail}	$\sum_i a_i P_i^{\max}$	4.2.1
firm capacity	$\Pr(P^{\text{avail}} \geq P) \geq 1 - \epsilon$	4.2.2
$C(\tau)$	largest constant power for duration τ	4.3.2
R	$(P, t_{\text{start}}, t_{\text{end}}, \text{location}, \text{reliability})$	4.3.1
coverage ratio	$P^{\text{enroll}}/P^{\text{commit}}$	4.5.1
P^{commit}	$\arg\max[\text{revenue} - \text{expected penalty}]$	4.5.5

4.8 Exercises

Exercise 4.1. A fleet of 10 000 batteries of 5 kW each has availability factors distributed as follows: 70 % of units at $a_i = 1$, 20 % at $a_i = 0.5$, 10 % at $a_i = 0$. Compute P^{name} and P^{avail}.

Exercise 4.2. Using Figure 4.3, at what duration do fleets A and B have equal deliverable capacity? If a service requires 80 MW for 20 minutes, which fleet can provide it? For 60 MW for 3 hours?

Exercise 4.3. Repeat the commitment optimization of Section 4.5.5 with $\sigma = 5$ MW instead of 10 MW. Where is the optimum, and by how much does the expected profit at the optimum change? Explain the result in one sentence.

Exercise 4.4. For Table 4.4, compute the 10th percentile of the total under a normal model with a common pairwise correlation of $\rho = 0.5$ between the three classes. Where does it fall relative to 76 and 82 MW?

4.9 Connection to the TERRA framework

A VPP is the conservation graph of Chapter 1 at its largest scale, together with one new operation: summarizing it.

- **Availability factors are state functions.** Each a_i in equation (4.2) is a function of device i's state, the same $P^{\text{avail}}(x)$ of Chapter 1. On a graph, P^{avail} of the fleet is a derived quantity evaluated on the current state, and its uncertainty is the posterior uncertainty of those states propagated through the sum.

- **Firm capacity is a quantile of a predicted flow.** The chance constraint (4.5) and the firm numbers of Table 4.4 are quantiles of the distribution of the aggregate grid-edge flow. The 76-versus-82 MW discussion of Section 4.6.3 is the statement that this distribution must be computed jointly, with the correlations the shared weather induces, not assembled from marginals.

- $C(\tau)$ **is a family of forward simulations.** The deliverable capacity for duration τ is found by integrating the fleet forward under a constant request and recording the largest request that stays feasible; it is the fleet-level version of the $F(t, \tau)$ query of Chapter 1.

- **Model reduction is a query interface.** The map $z = \phi(x)$ of equation (4.14) is a set of queries on the full model that return the few numbers a decision needs. The full state stays inside the model; the grid sees only the answers. This is the natural division of labor between a physical model and the operator who uses it.

- **The endogenous feasible region is the model's own dynamics.** Figure 4.5 is what a forward simulation of a state-dependent system produces: the feasible set at $t+1$ is computed from the state at $t+1$, which the control at t produced. No separate "capacity model" is needed; the capacity is read off the physics.

In summary, the VPP is what a conservation-graph model of a fleet looks like from outside: a small set of predicted, uncertain quantities offered as a resource. The next chapter goes inside the most important component of that graph for an aggregator, the thermostatically controlled building.

Chapter 5

Thermostatically controlled loads

5.1 The question

The earlier chapters established that DER flexibility is state-dependent, time-dependent, customer-constrained and uncertain. Thermostats, with the heating and cooling plant behind them, are the cleanest example of all four at once, and they are the largest resource class for a residential aggregator. The question of this chapter is

How does a thermostat become a controllable grid resource?

The answer runs from the physics of a building, through the logic of the thermostat, to the behavior of a fleet of a hundred thousand of them.

Jargon: thermostatically controlled load (TCL), HVAC

A *thermostatically controlled load* is any load switched on and off by a thermostat to hold a temperature: air conditioners, heat pumps, electric furnaces, water heaters, refrigerators. *HVAC* stands for heating, ventilation and air conditioning, the building's climate plant. This chapter uses an air conditioner in cooling mode throughout; heating is the same with the signs reversed.

5.2 The building and the thermostat

The resource is the building, the plant and the controller together. This section models each in turn.

5.2.1 The physical system

A house has indoor air, walls and furnishings, outdoor weather, and an HVAC system. The indoor temperature changes because heat enters or leaves the building. The simplest model that captures this is a single thermal mass behind a single thermal resistance,

$$C\,\frac{\mathrm{d}T_{\text{in}}}{\mathrm{d}t} = \frac{T_{\text{out}} - T_{\text{in}}}{R} + Q_{\text{int}} - Q_{\text{HVAC}}, \tag{5.1}$$

where T_{in} and T_{out} are the indoor and outdoor temperatures, R is the thermal resistance of the envelope (°C per kW of heat flow), C is the effective thermal capacitance of the building (kWh per °C), Q_{int} is the internal heat gain from people, appliances and sunlight, and Q_{HVAC} is the heat the air conditioner removes. This is equation (1.13) of Chapter 1 with the internal gain added. Figure 5.1 draws it as the electrical circuit it is analogous to: a capacitor to ground, a resistor to the outdoor "voltage", and two current sources. The analogy is exact, and it is why R and C carry those names.

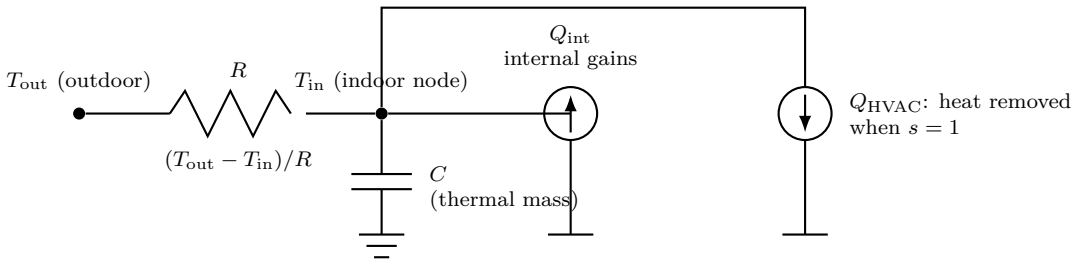


Figure 5.1: The first-order thermal model of a building as a circuit. Temperature plays the role of voltage and heat flow the role of current: heat leaks in through R from the outdoor temperature, internal gains add heat, the air conditioner removes it when running, and the thermal mass C integrates the net flow.

5.2.2 The time constant

The product RC is the building's thermal time constant. A building with good insulation (large R) and a lot of thermal mass (large C) changes temperature slowly; one with poor insulation and little mass changes quickly. Two houses at the same $T_{\text{in}} = 22°\text{C}$ can therefore have very different flexibility. Figure 5.2 shows the indoor temperature after the compressor is switched off in two houses with the same starting temperature and the same outdoor conditions but time constants of one hour and four hours: the fast house reaches the comfort limit in 13 minutes, the slow house in 54.

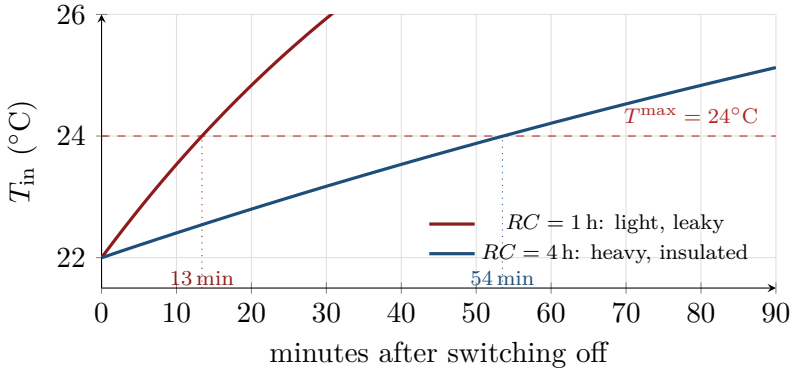


Figure 5.2: Indoor temperature after the compressor stops, from 22°C with $T_{\text{out}} = 32°\text{C}$, for two buildings that differ only in RC. Same starting state, four times the flexibility.

Principle 5.1. *The current indoor temperature alone does not determine HVAC flexibility. The building's dynamics, summarized by RC, determine how long any temperature headroom lasts.*

5.2.3 The thermostat is a switching controller

A basic thermostat does not command an arbitrary cooling power. It switches the compressor on or off,

$$s_t \in \{0, 1\}, \tag{5.2}$$

with $s_t = 1$ meaning the air conditioner is on. If the compressor draws electrical power P_{AC} when running, the electrical demand is

$$P_t = s_t \, P_{\text{AC}}, \tag{5.3}$$

so at the level of one device the load is discrete: either P_{AC} or zero.

5.2.4 Deadband logic

With the setpoint at $T_{\text{set}} = 22°\text{C}$, the thermostat does not switch exactly at 22°C every second. It uses a *deadband* (also called hysteresis), for example $21.5°\text{C} \le T_{\text{in}} \le 22.5°\text{C}$, and a simplified cooling controller behaves as

$$s_t = \begin{cases} 1, & T_{\text{in}} > 22.5°\text{C}, \\ 0, & T_{\text{in}} < 21.5°\text{C}, \\ s_{t-1}, & \text{otherwise.} \end{cases} \tag{5.4}$$

Inside the band the compressor keeps whatever state it had. This prevents rapid switching, which would wear the compressor, and it produces the familiar cycling pattern. Figure 5.3 shows the rule as a loop in the (T_{in}, s) plane together with the resulting temperature and power traces: the temperature saws between the two thresholds and the power is a train of pulses.

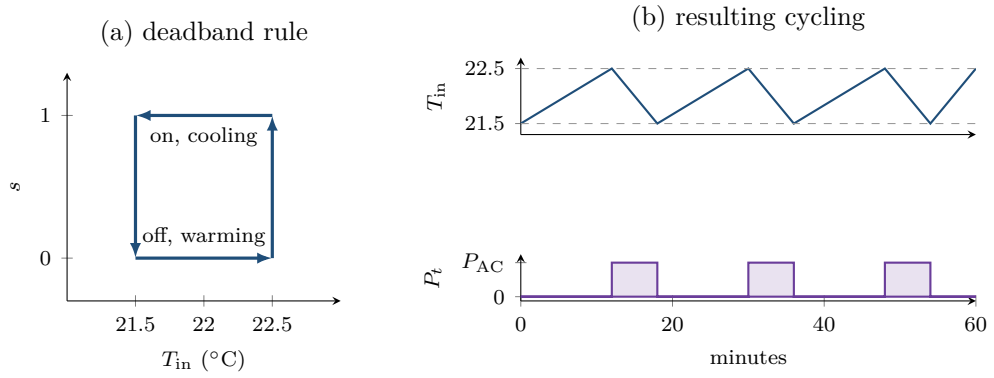


Figure 5.3: (a) The deadband rule of equation (5.4) as a loop: the compressor switches on at the upper threshold and off at the lower one, and holds its state in between. (b) The temperature sawtooth and power pulse train the rule produces. Traces are illustrative.

5.2.5 Intermittent load and duty cycle

Because of the cycling, the statement "this air conditioner is a 4 kW resource" does not mean it draws 4 kW continuously. Its average demand is

$$\bar{P} = D\,P_{\text{AC}}, \qquad D = \frac{\text{time the compressor is on}}{\text{total time}}, \tag{5.5}$$

where D is the *duty cycle*. With $D = 0.4$ and $P_{\text{AC}} = 4\,\text{kW}$ the average is 1.6 kW.

The duty cycle depends strongly on weather. In the steady state of equation (5.1), with internal gains neglected, the heat removed must equal the heat leaking in, $D\,Q_{\text{HVAC}} = (T_{\text{out}} - T_{\text{in}})/R$, so

$$D = \frac{T_{\text{out}} - T_{\text{in}}}{R\,Q_{\text{HVAC}}}, \tag{5.6}$$

which rises linearly with the outdoor temperature until it saturates at 1, when the air conditioner can no longer keep up. Figure 5.4 plots it for a plant that just reaches full duty at $T_{\text{out}} = 42°\text{C}$. A hot afternoon gives $D \approx 0.8$; a mild day gives 0.2.

If the compressor is already off, raising the setpoint gives $\Delta P = 0$: there is nothing to shed at that moment. This is the first major lesson of the chapter.

Principle 5.2. *Thermostat flexibility depends on whether the underlying HVAC load is actually active, which the duty cycle measures and the weather sets.*

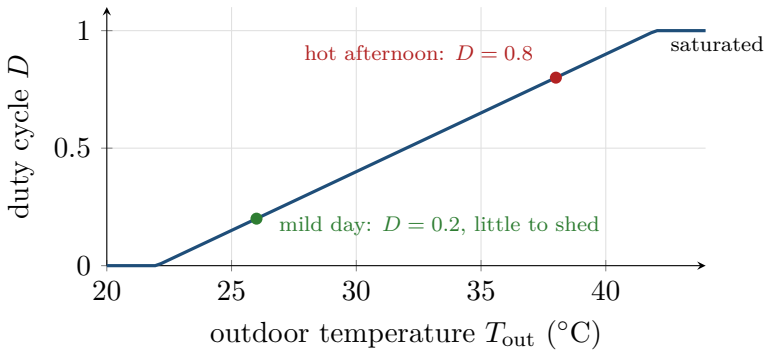


Figure 5.4: Steady-state duty cycle versus outdoor temperature, equation (5.6), with $T_{\text{in}} = 22°\text{C}$ and $RQ_{\text{HVAC}} = 20°\text{C}$. The curtailable load of a thermostat is proportional to D, so it is a weather quantity.

5.3 What control does

With the model in hand, the effect of a setpoint change can be followed from the instant of the command to the restart of the compressor and beyond.

5.3.1 What control does

With the normal setpoint at 22°C, the DERMS raises it during an event to 24°C. The control is the setpoint offset,

$$u_t = \Delta T_{\text{set},t} = +2°\text{C}. \tag{5.7}$$

This widens the allowed temperature range before the air conditioner must restart. If the compressor was on it turns off, the power changes from 4 kW to zero, and the instantaneous response is $\Delta P = 4\,\text{kW}$. But only temporarily.

5.3.2 Why the response decays

Once cooling stops, $Q_{\text{HVAC}} \to 0$, and under hot conditions $\mathrm{d}T_{\text{in}}/\mathrm{d}t > 0$: the house warms. Eventually T_{in} reaches the new upper threshold and the compressor restarts, and the 4 kW reduction is gone. Figure 5.5 shows the full history of the reduction for one house: a step to the full value, a plateau until the first restart, and then cycling at the new, higher setpoint, where the average reduction is the difference between the old and new duty cycles times P_{AC}. The average over any window therefore decays from the full value toward that small sustained value, which is the per-device origin of the fleet curve in Figure 1.3 of Chapter 1.

5.3.3 Stored thermal energy

The physical interpretation is that a cool building holds stored cooling. If T_{in} is well below the upper comfort limit, the building can be allowed to warm for some time without running the air conditioner. The building acts as a thermal battery: for cooling,

cooler building $\Longrightarrow$ more upward temperature headroom $\Longrightarrow$ more future load-shed capability.

This is the view that makes pre-cooling, the thermal state of charge, and the battery analogy of the following sections natural.

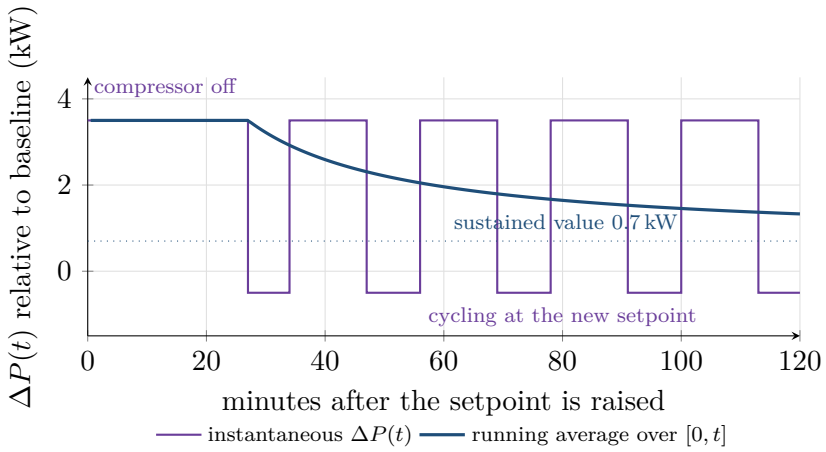


Figure 5.5: Load reduction of one house after a +2°C setpoint change, against a 3.5 kW baseline average. The instantaneous reduction (purple) is a full 3.5 kW until the first restart at 27 minutes, then alternates as the compressor cycles at the higher setpoint; the running average (blue) decays toward the 0.7 kW that the higher setpoint sustains. Parameters as in the worked example of Chapter 1.

5.3.4 Pre-cooling

If the grid expects a peak at 18:00, the controller need not wait. At 16:00 it can cool the house harder, T_{in} : 22.5°C → 21.5°C, raising consumption before the event so that from 18:00 to 20:00 the HVAC can stay off longer. Figure 5.6 shows the load profile: a bump before the peak period and a deeper, longer reduction during it. Load was deliberately increased earlier to reduce it later, which is load shifting, and it is the single-house version of the demand flexibility of Chapter 2. The temperature trajectories behind this profile were shown in Figure 1.8 of Chapter 1.

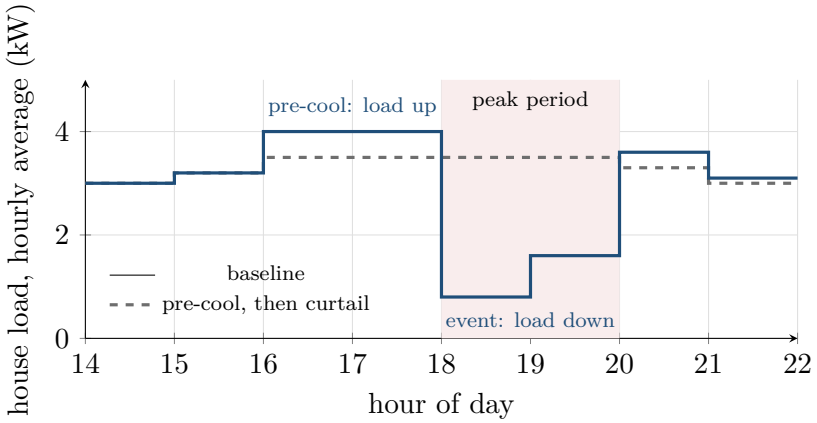


Figure 5.6: Hourly-average load of one house with pre-cooling (blue) against its baseline (dashed). Consumption rises for two hours before the peak so that it can fall during it. Values are illustrative.

Simple demand response is “peak arrives, turn the air conditioners down”. Pre-cooling is “predict the peak, pre-condition the buildings, reduce load during the peak”, and it needs forecasts of future weather, future grid load, the thermal response and the customer constraints before the event, which is why it is a problem of optimal control rather than of switching.

5.4 The state-space model

This section writes the model in the standard state-space form of the earlier chapters, with the thermostat's rule and the customer's band in their proper places.

5.4.1 State-space form

Discretizing equation (5.1) with a step Δt gives

$$T_{t+1} = a\,T_t + b\,T_{\text{out},t} + c\,Q_{\text{int},t} - d\,s_t, \tag{5.8}$$

with electrical power $P_t = s_t\,P_{\text{AC}}$. For the exact discretization of the first-order model the coefficients are

$$a = e^{-\Delta t/RC}, \qquad b = 1 - a, \qquad c = R\,(1-a), \qquad d = R\,Q_{\text{HVAC}}\,(1-a), \tag{5.9}$$

so that with the compressor off ($s_t = 0$) and no internal gain the temperature relaxes toward T_{out} with the factor a per step, and with it on the temperature is pulled down by d per step. In the notation of Chapter 1 the state is $x_t = T_t$ and the control is either s_t or $T_{\text{set},t}$, depending on how the control is represented, which is the subject of the next section.

5.4.2 Control is indirect and hierarchical

A DERMS usually does not command $s_t = 0$ for every compressor. It commands a setpoint, $u_t = T_{\text{set},t} = 24°\text{C}$, and the local thermostat determines s_t with its own logic:

$$s_t = g(T_t, u_t), \tag{5.10}$$

$$T_{t+1} = f(T_t, s_t, T_{\text{out},t}). \tag{5.11}$$

This is a two-level, hierarchical control system, Figure 5.7: the DERMS chooses the setpoint, the thermostat chooses the switching, and the physics responds. Two consequences follow. The DERMS's model of the house must include the thermostat's rule, not just the thermal dynamics, because the rule decides when the compressor restarts. And the realized response to the same command differs across houses, because each thermostat applies the rule to its own temperature.

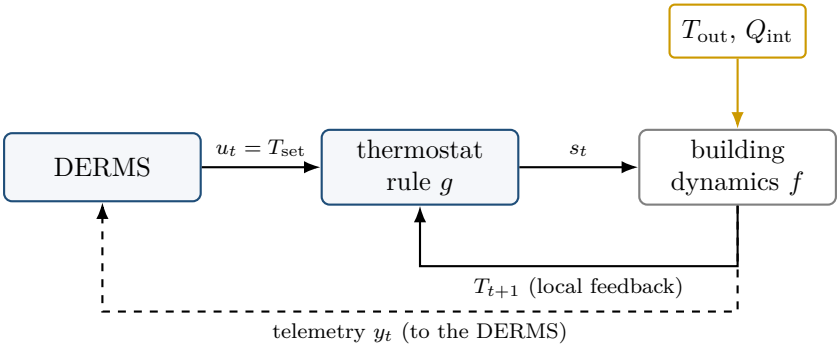


Figure 5.7: Hierarchical control of a thermostatic load. The DERMS sets the setpoint; the thermostat applies its deadband rule to the local temperature; the building responds. The inner loop is fast and local, the outer loop is slow and remote.

5.4.3 The comfort constraint and the thermal state of charge

If customers allow $21°\text{C} \le T_{\text{in}} \le 25°\text{C}$, the DERMS cannot keep the air conditioner off indefinitely; $T_{\text{in},t} \le T^{\max}$ must hold, and that makes the flexibility finite. A home at 21°C has far more headroom than one at 24.8°C. The position of the temperature within the comfort band is, in effect, a state of charge. For cooling, define

$$\text{SOC}^{\text{th}} = \frac{T^{\max} - T_{\text{in}}}{T^{\max} - T^{\min}}, \tag{5.12}$$

which is 1 at $T_{\text{in}} = T^{\min} = 21°\text{C}$ (fully "charged" with cooling) and 0 at $T_{\text{in}} = T^{\max} = 25°\text{C}$ (nothing left to give). Figure 5.8 draws it. Commercial systems do not necessarily represent buildings this way, but conceptually it is exactly right: it makes a thermostat look like a battery.

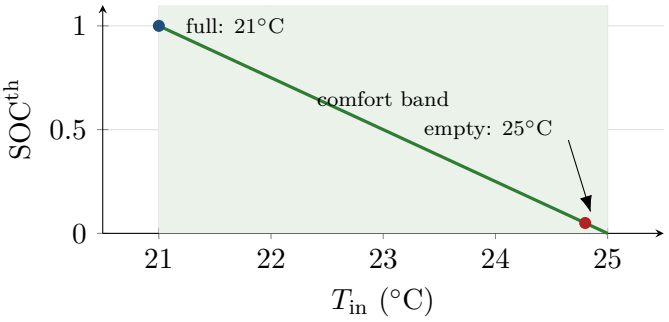


Figure 5.8: Thermal state of charge, equation (5.12), for a comfort band of 21 to 25°C. A home at 21°C (blue) is fully charged; one at 24.8°C (red) is almost empty.

5.4.4 The battery analogy

Table 5.1 sets the two devices side by side. The battery's state advances by the energy it moves; the building's state advances by the heat it exchanges. In both, current control affects future capacity: discharging a battery now leaves less energy later, and curtailing cooling now leaves a warmer building and less curtailment later. This is the intertemporal structure of Chapter 1 in its two most common forms.

Table 5.1: Battery and building, term by term.

	Battery	Building (cooling)
State	SOC_t	T_t, or SOC^{th}_t
Dynamics	$\text{SOC}_{t+1} = \text{SOC}_t - P_t\,\Delta t/E^{\max}$	$T_{t+1} = f(T_t, s_t, T_{\text{out},t})$
Capacity	$E^{\max}$	$C\,(T^{\max} - T^{\min})$
Bounds	$\text{SOC}^{\text{res}} \le \text{SOC} \le 1$	$T^{\min} \le T_{\text{in}} \le T^{\max}$
"Self-discharge"	small, slow	heat leak $(T_{\text{out}} - T_{\text{in}})/R$, fast when hot
Control	P_t, continuous	s_t via T_{set}, discrete
Using it now	less energy later	warmer building, less curtailment later

5.5 Fleet behavior

Two phenomena appear only at fleet scale: the rebound of many buildings released at once, and the loss of diversity that a common command causes.

5.5.1 Rebound and staggered recovery

Suppose thousands of homes are curtailed from 17:00 to 19:00. At 19:00 their indoor temperatures are elevated. If the event ends abruptly, many compressors restart at once and the fleet load spikes above baseline: the rebound of Chapter 2, now explained by the thermostat rule. Poor control can simply move the peak.

A DERMS need not release every thermostat at once. It can stagger the recovery: group A at 19:00, group B at 19:10, group C at 19:20, group D at 19:30. The aggregate recovery is then smoother. This introduces another optimization problem, *how to restore comfort without creating a new peak*, and it is a clean example of why fleet-level orchestration matters. Section 5.5.2 shows the effect quantitatively.

5.5.2 Aggregation smooths, synchronization does not

At the single-house level $P_i(t)$ is binary. For $N = 100\,000$ houses the fleet load $P_{\text{fleet}}(t) = \sum_i P_i(t)$ is nearly continuous, because the compressors cycle at different times: instead of 0 or 4 kW one sees 157.2, 156.8, 157.4 MW. The fleet can be treated almost as a continuous resource. Almost.

The danger is synchronization. If the DERMS sends the same command to every thermostat at the same moment, all of them shift together; the random phases that produced the smooth sum are destroyed, and the population switches on and off in unison, producing large oscillations in load. Figure 5.9 shows this with a simulated fleet of 2 000 houses, each following equation (5.8) with the deadband rule and its own randomly drawn RC and cooling rate, under $T_{\text{out}} = 32°\text{C}$. Before the command the fleet draws a steady 4.6 MW of its 8 MW nameplate (a duty cycle near 0.58). At minute 60 every setpoint is raised by 2°C: all compressors stop, the load drops to zero, and when the houses reach the new threshold they restart in a bunch, so the load returns as a wave rather than as a steady value. At minute 180 the event ends. If every setpoint is restored at once (gray), every compressor is above its threshold and the fleet goes to 100 % on, 8 MW, almost twice its baseline, and then oscillates. If the fleet is released in four groups fifteen minutes apart (blue), the peak is 6.7 MW and the oscillation is damped.

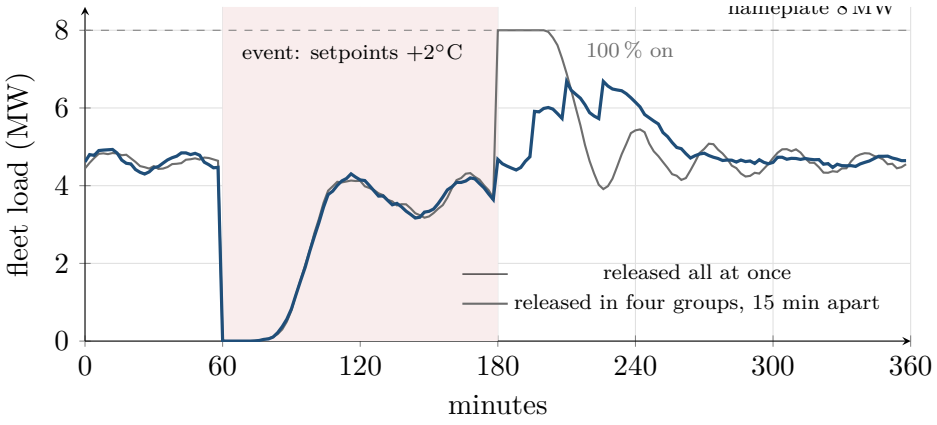


Figure 5.9: Simulated load of 2 000 thermostatic houses (equation (5.8) with the deadband rule, RC drawn uniformly from 90 to 240 minutes, $T_{\text{out}} = 32°\text{C}$, 4 kW per house). A common +2°C command at minute 60 synchronizes the fleet; the restart arrives as a wave. Releasing everyone at minute 180 (gray) sends the fleet to 100 % on; releasing four groups fifteen minutes apart (blue) caps the rebound and damps the oscillation. Data from `scripts/ch45_data.py`.

Principle 5.3. *Diversity is a resource that commands can destroy. Good fleet control deliberately preserves or recreates it, by staggering commands and releases, so that the fleet remains a smooth resource after being used.*

5.6 Forecasting and dispatch

The last section of the body treats the thermostat fleet as a resource to be forecast and dispatched.

5.6.1 Forecasting thermostat flexibility

If the utility asks how many megawatts the thermostat fleet can deliver tomorrow at 18:00, a crude model multiplies the fleet size, the compressor power and the duty cycle,

$$P^{\text{flex}} = N \times P_{\text{AC}} \times D = 100\,000 \times 3.5\,\text{kW} \times 0.6 = 210\,\text{MW}. \tag{5.13}$$

This is an upper-ish instantaneous estimate. It ignores customer opt-outs, duration, the temperature state of each house, communication failures, comfort constraints and control history. The quantity that matters is really a function of all of these,

$$P_t^{\text{flex}} = f\big(T_{\text{out},t},\ T_{\text{in},t},\ T_{\text{set},t},\ D_t,\ \text{building parameters},\ \text{customer behavior},\ \text{prior dispatch}\big), \tag{5.14}$$

and at fleet level not every house is known. The practical forecast therefore relies heavily on learned relationships,

$$\hat{P}_t^{\text{flex}} = g(\text{weather},\ \text{time},\ \text{historical events},\ \text{fleet composition},\ \text{recent telemetry}), \tag{5.15}$$

which is why a forecasting engineer is central in this domain. Table 5.2 lists the variables such a forecast draws on. Not every implementation exposes every one, but conceptually these are the variables that govern the physical process.

Table 5.2: Variables that govern thermostat-fleet flexibility.

Group	Variables
Weather	outdoor temperature, humidity, solar irradiance
Calendar	time of day, day of week, season
Fleet history	historical fleet load, prior event behavior, override rates
Device telemetry	setpoints, indoor temperatures, HVAC runtime, availability
Building	inferred RC and cooling capacity per house

5.6.2 Weather: more power, less duration

On a 25°C day many air conditioners are off and the potential flexibility is small. At 38°C most are running and the instantaneous flexibility is enormous. But hot weather also makes buildings warm faster when cooling is curtailed, so the time to the first restart shrinks. Figure 5.10 plots both effects against outdoor temperature: the shed-able power rises with T_{out} through the duty cycle, while the sustainable duration falls with T_{out} through the time constant. The two curves move in opposite directions, which is the subtle point of this section.

Principle 5.4. *Hotter weather increases instantaneous thermostat flexibility and decreases its sustainable duration. A forecast of megawatts alone cannot capture this; the forecast must be of $F(t, \tau)$.*

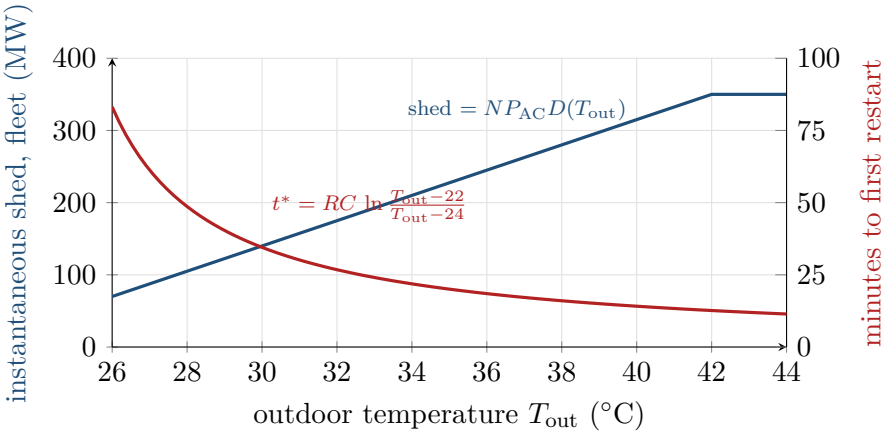


Figure 5.10: Two effects of outdoor temperature on a 100 000-thermostat fleet with $P_{\mathrm{AC}} = 3.5\,\mathrm{kW}$ and $RC = 2\,\mathrm{h}$. The shed-able power (blue, left axis) rises with the duty cycle; the minutes until the first compressor restart after a $22 \to 24$°C setpoint change (red, right axis) fall. Hot days give more power for less time.

5.6.3 Power flexibility and energy flexibility

For thermostats two quantities must be distinguished: the instantaneous power reduction $\Delta P(t)$, and the total energy shifted,

$$E^{\mathrm{shift}} = \int \Delta P(t)\,\mathrm{d}t. \tag{5.16}$$

A fleet might reduce load by 100 MW initially but sustain that only briefly, so forecasting megawatts is insufficient for a service that needs megawatts over time. In Figure 5.5 the instantaneous reduction is 3.5 kW and the energy shifted in the first hour is the area under the purple curve, about 2.0 kWh; a battery promising 3.5 kW for an hour would deliver 3.5 kWh.

5.6.4 Dynamic load shaping by rotation

Instead of curtailing every house from 17:00 to 19:00, the DERMS can divide the fleet into groups and curtail them in turn: group A from 17:00 to 17:30, group B from 17:30 to 18:00, and so on. Each group's response decays as in Figure 5.5, but by rotating the groups the aggregate reduction can be held nearly constant, Figure 5.11. This is the idea a later chapter develops as dynamic load shaping; thermostats make the motivation obvious.

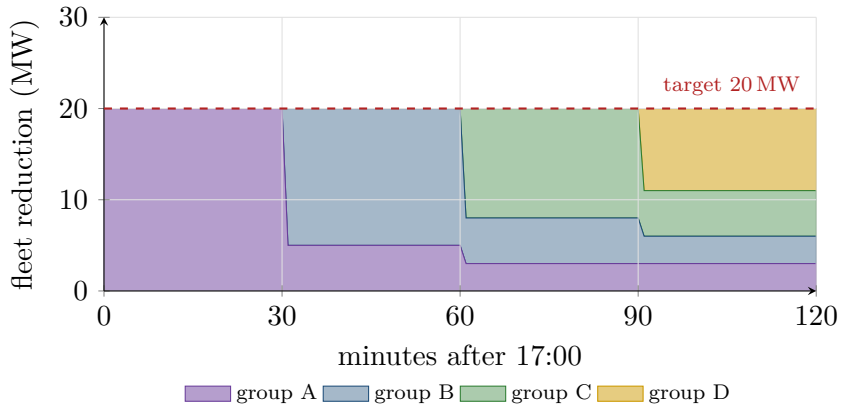


Figure 5.11: Rotating four groups through half-hour curtailments. Each group's contribution decays after its turn, but the stacked total stays near the 20 MW target. Values are illustrative.

5.6.5 A fleet optimization problem

Let $u_{i,t} \in \{0, 1\}$ indicate whether thermostat i is curtailed at step t. The DERMS wants $\sum_i P_i\, u_{i,t} \approx P_{\text{target},t}$, and might minimize

$$J = \sum_t \Big(\sum_i P_i\, u_{i,t} - P_{\text{target},t} \Big)^2 + \lambda \sum_{i,t} \text{discomfort}_{i,t}, \tag{5.17}$$

subject to $T_{i,t} \le T_{i,\max}$ for every house and step. The first term tracks the grid target; the second, weighted by λ, charges for customer discomfort (for example the excursion of $T_{i,t}$ above the normal setpoint); the constraint makes the comfort band hard. The rotation of Figure 5.11 and the staggered release of Figure 5.9 are both solutions of problems of this form.

5.6.6 Forecast and dispatch are coupled

If the optimizer believes group A can deliver 20 MW for an hour and the true capability is 12 MW, the optimizer solves the wrong problem. Conversely, if the forecast is perfect but the optimizer ignores rebound, the dispatch still fails, because the peak reappears at the release. Forecast quality and dispatch quality are therefore tightly coupled, as Chapter 3 stated in general and this chapter shows in the particular.

5.7 What to retain

Definition 5.1 (Thermostatic flexibility)**.** A thermostat controls electrical demand by exploiting the thermal storage of a building. The building's thermal state determines how long that flexibility lasts.

Thermostat-based flexibility is governed by

$$\text{thermal dynamics} + \text{thermostat logic} + \text{weather} + \text{comfort constraints} + \text{customer behavior},$$

and at fleet scale, aggregation turns thousands of intermittent binary loads into a smooth, dispatchable resource, but only if forecasting and control are done well. Table 5.3 maps each governing factor to the section and the object that captures it.

Table 5.3: The five governing factors of thermostatic flexibility.

Factor	Object	Section
Thermal dynamics	R, C, equation (5.1)	5.2.1, 5.2.2
Thermostat logic	deadband rule (5.4), $s_t = g(T_t, u_t)$	5.2.4, 5.4.2
Weather	$D(T_{\text{out}})$, $t^*(T_{\text{out}})$	5.2.5, 5.6.2
Comfort constraints	$T^{\min} \le T_{\text{in}} \le T^{\max}$, SOC^{th}	5.4.3
Customer behavior	p^{ovr}, opt-outs, override history	5.6.1

5.8 Exercises

Exercise 5.1. A house has $R = 2.5$°C/kW and $C = 1.2$ kWh/°C. Compute RC in hours and the coefficients a and b of equation (5.9) for $\Delta t = 5$ minutes.

Exercise 5.2. Using equation (5.6) with $R\,Q_{\text{HVAC}} = 20$°C and $T_{\text{in}} = 22$°C, find the outdoor temperature at which a fleet of 100 000 units of 3.5 kW has 150 MW of curtailable load. At that temperature, how long until first restart after a +2°C setpoint change if $RC = 2$ h?

Exercise 5.3. Define the thermal state of charge for a *heating* thermostat with comfort band $[T^{\min}, T^{\max}]$, so that it equals 1 when the building is fully "charged" with heat.

Exercise 5.4. In Figure 5.9, estimate from the plot the energy (MWh) that the fleet consumed above baseline in the hour after the synchronized release, and compare it with the energy not consumed during the event. Explain the sign of the difference using equation (5.1).

5.9 Connection to the TERRA framework

The thermostatic building is the first component in this book whose model is two coupled conservation laws, and that is exactly the case a multi-conservation graph framework exists for.

- **Two conservation nodes, one component.** The house has an electrical node, where $P_t = s_t\, P_{\mathrm{AC}}$ is drawn from the grid, and a thermal node, where equation (5.1) balances the heat flows of Figure 5.1. The compressor is the edge that couples them: it consumes electrical power on one node and removes heat on the other, with a coefficient of performance relating the two. A conservation graph carries both balances in one model, so the rebound, the pre-cooling bump and the duty cycle all follow from the same equations rather than being added as rules.
- **R and C are parameters to be inferred.** Most houses do not report their thermal mass or envelope conductance. They do report temperature trajectories, and the response of T_{in} to a known switching history identifies RC and the cooling rate; this is the parameter-inference problem on a graph with a posterior over R and C per house. Figure 5.2 shows why the inference matters: the same state with a different RC is a four-fold different resource.
- **The thermostat rule is part of the component.** The hierarchical structure of Figure 5.7 is a local controller inside a component: the graph's time integration evaluates g at each step to decide s_t before advancing T_t. A framework with first-class controls makes the rule a declared part of the model, which is what allows the restart time, the rebound and the synchronization of Figure 5.9 to be predicted rather than observed after the fact.
- **The thermal state of charge is a derived state.** $\mathrm{SOC}^{\mathrm{th}}$ of equation (5.12) is a function of the node temperature and the comfort bounds, and the battery analogy of Table 5.1 is the observation that an accumulator component (energy stored) and a thermal-mass component (heat stored) are the same kind of object on a graph with different units.
- **The fleet figure is a population forward simulation.** Figure 5.9 was produced by integrating 2 000 copies of the component with parameters drawn from a distribution and summing an edge flow; this is a graph with 2 000 houses attached to one aggregate node. The same run from a posterior over the parameters rather than a chosen distribution is the forecast $\hat{P}_t^{\mathrm{flex}}$ of equation (5.15), with physics instead of a learned g.

In summary, the thermostatic load is a two-domain component with a local switching rule and two unmeasured parameters, and every phenomenon of this chapter (decay, pre-cooling, rebound, synchronization, the weather trade-off) is a forward simulation of that component. The next chapter turns to the battery, where the same structure appears with one conservation law instead of two.

Chapter 6

Batteries as DERs: state of charge, power, energy and dispatch

6.1 The question

Batteries follow thermostats in the lesson sequence because they make the same central idea much easier to see mathematically: flexibility depends on an internal state and on future constraints. For a thermostat the state was a temperature and the constraint a comfort band; for a battery the state is a state of charge and the constraints are explicit numbers on a label. The question of the chapter is

What can a battery provide now without preventing it from doing what it must do later?

Jargon: BESS, inverter, C-rate

A *battery energy storage system* (BESS) is the battery cells, the power-electronic *inverter* that converts between the battery's direct current and the grid's alternating current, and the controller. The inverter rating sets the power limit; the cells set the energy limit. The ratio of power to energy, in units of 1/h, is sometimes called the *C-rate*: a 10 kW, 40 kWh battery is a 0.25C system, meaning it empties in four hours at full power.

6.2 State and dynamics

Four quantities describe a battery: its power rating, its energy capacity, its state of charge, and the equation that links them.

6.2.1 Power and energy

A battery might be rated $P^{\max} = 10\,\text{kW}$ and $E^{\max} = 40\,\text{kWh}$. These are different quantities: $P^{\max}$ is how fast the battery can charge or discharge, $E^{\max}$ is how much energy it can store. If the battery discharges at 10 kW, a full 40 kWh battery sustains that for

$$\frac{E^{\max}}{P^{\max}} = \frac{40\,\text{kWh}}{10\,\text{kW}} = 4\ \text{h}$$

in the idealized case. The lesson of Chapter 1 is immediate: a megawatt capability without a duration is incomplete information.

6.2.2 State of charge

Define SOC_t as the fraction of usable stored energy currently available, so that $\mathrm{SOC}_t = 0.75$ means roughly 75 % charged and the stored energy is

$$E_t = \mathrm{SOC}_t \, E^{\max} = 0.75 \times 40 = 30\,\mathrm{kWh}. \tag{6.1}$$

This is the analog of the thermal state of Chapter 5, and for the battery it is the obvious first state variable: $x_t = \mathrm{SOC}_t$.

6.2.3 Battery dynamics

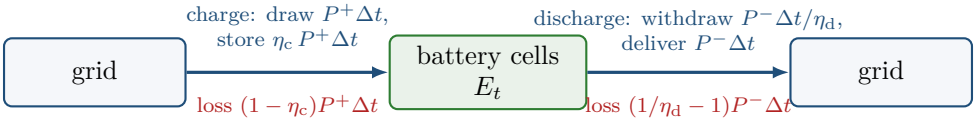
Let $P_t > 0$ denote charging and $P_t < 0$ discharging, and split the power into its charging part $P_t^+ = \max(P_t, 0)$ and its discharge magnitude $P_t^- = \max(-P_t, 0)$. A simplified energy balance over a step Δt is

$$E_{t+1} = E_t + \eta_{\mathrm{c}} \, P_t^+ \, \Delta t - \frac{1}{\eta_{\mathrm{d}}} \, P_t^- \, \Delta t, \tag{6.2}$$

where η_{c} is the charging efficiency (the fraction of drawn energy that is stored) and η_{d} the discharging efficiency (the fraction of withdrawn energy that reaches the grid). Dividing by $E^{\max}$ gives the same statement in state-of-charge form,

$$\mathrm{SOC}_{t+1} = \mathrm{SOC}_t + \frac{\eta_{\mathrm{c}} \, P_t^+ \, \Delta t - P_t^- \, \Delta t / \eta_{\mathrm{d}}}{E^{\max}}. \tag{6.3}$$

Figure 6.1 shows why the two efficiencies enter on opposite sides: losses are taken from the grid-side energy on the way in and from the battery-side energy on the way out, so delivering 1 kWh to the grid costs $1/\eta_{\mathrm{d}}$ kWh of stored energy. Equation (6.3) is the single most important equation of the chapter, because it says that *dispatch now changes flexibility later.*



Example with $\eta_{\mathrm{c}} = \eta_{\mathrm{d}} = 0.95$: 10 kWh drawn stores 9.5 kWh;
delivering those 9.5 kWh back to the grid yields $9.5 \times 0.95 \approx 9.0$ kWh.

Figure 6.1: Energy flow through a battery over one charge and one discharge. Losses are charged on the grid side going in and on the battery side going out, which is why η_{c} multiplies and η_{d} divides in equation (6.2).

6.2.4 One hour of dispatch

Take $E^{\max} = 40$ kWh, $\mathrm{SOC}_t = 0.75$, so $E_t = 30$ kWh, and assume ideal efficiency for the moment. Discharging at $P = 10$ kW for one hour gives $E_{t+1} = 30 - 10 = 20$ kWh and

$$\mathrm{SOC}_{t+1} = \frac{20}{40} = 0.50.$$

One hour of dispatch moved the battery from 75 % to 50 %. The battery's future capability cannot be separated from its dispatch history.

6.3 Constraints

The constraints are explicit numbers, which is what makes the battery the cleanest case; this section takes them one by one.

6.3.1 The first constraint: power

The battery cannot charge or discharge arbitrarily fast. The inverter imposes

$$-P^{\text{dis}} \le P_t \le P^{\text{ch}}, \tag{6.4}$$

and for a symmetric battery $-10 \le P_t \le 10\,\text{kW}$. This is the instantaneous power capability. It says nothing about whether that power can be maintained.

6.3.2 The second constraint: state of charge

The state of charge is kept inside operating limits,

$$\text{SOC}^{\min} \le \text{SOC}_t \le \text{SOC}^{\max}, \tag{6.5}$$

for example $0.10 \le \text{SOC}_t \le 0.95$. The ends are avoided for battery protection, degradation, reserve requirements or device-specific operating rules. The energy available for discharge is therefore not $\text{SOC}_t\, E^{\max}$ but

$$E^{\text{dis,avail}} = (\text{SOC}_t - \text{SOC}^{\min})\, E^{\max}, \tag{6.6}$$

and symmetrically the room for charging is $(\text{SOC}^{\max} - \text{SOC}_t)\, E^{\max}$.

6.3.3 Example with a reserve limit

With $\text{SOC}_t = 0.60$, $\text{SOC}^{\min} = 0.20$ and $E^{\max} = 40\,\text{kWh}$ the usable discharge energy is

$$(0.60 - 0.20) \times 40 = 16\,\text{kWh},$$

and at the maximum discharge power of 10 kW the battery sustains full power for about $16/10 = 1.6\,\text{h}$. So $P^{\text{avail}} = 10\,\text{kW}$, but only for about 1.6 hours. Figure 6.2 draws the resulting power–duration capability, the same shape as Figure 1.4 of Chapter 1 but now with the reserve included: the energy limit is set by the usable energy, not by the nameplate. This is why a capacity statement needs both MW and MWh.

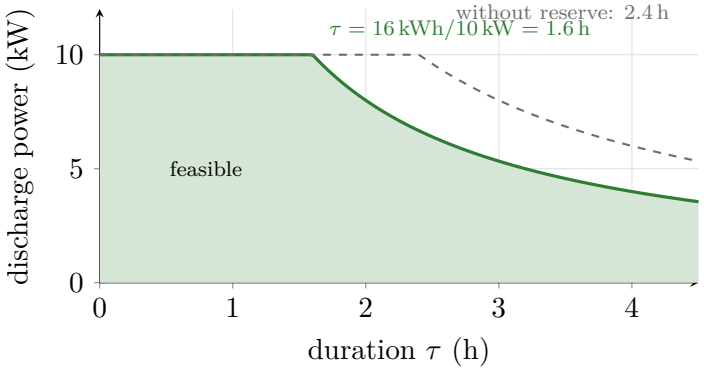


Figure 6.2: Power–duration capability of the 10 kW, 40 kWh battery at SOC = 0.60. With the owner's 20 % reserve (green) the usable energy is 16 kWh; ignoring the reserve (dashed) would credit 24 kWh and overstate the duration by half.

6.3.4 Flexibility in both directions

A battery is especially useful because it is bidirectional. If it is idle, it can discharge ($P_t < 0$), which reduces net grid load, or charge ($P_t > 0$), which increases it. With $P_{\text{grid}} = P_{\text{load}} + P_{\text{bat}}$, charging raises net demand and discharging lowers it. That makes batteries useful for peak shaving, renewable absorption, price arbitrage, contingency support and balancing.

Two quantities describe the two directions. *Upward grid support* $F_t^{\uparrow}$ is the ability to reduce net load or inject more power; *downward grid support* $F_t^{\downarrow}$ is the ability to increase net load or reduce injection. For a battery both depend strongly on the state of charge:

$$F_t^{\uparrow} = \min\Big(P^{\text{dis}},\ \frac{(\text{SOC}_t - \text{SOC}^{\min})\,E^{\max}\,\eta_{\text{d}}}{\tau}\Big), \qquad F_t^{\downarrow} = \min\Big(P^{\text{ch}},\ \frac{(\text{SOC}^{\max} - \text{SOC}_t)\,E^{\max}}{\eta_{\text{c}}\,\tau}\Big), \tag{6.7}$$

for a required duration τ. Near $\text{SOC}^{\min}$ the discharge flexibility is small, $F_t^{\uparrow} \approx 0$, while the charging flexibility is large; near $\text{SOC}^{\max}$ the opposite holds. Figure 6.3 plots both against the state of charge for a one-hour requirement.

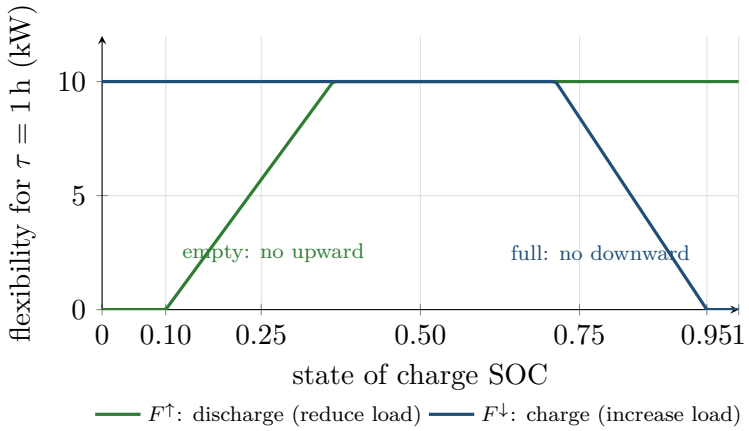


Figure 6.3: Upward and downward flexibility of the 10 kW, 40 kWh battery for a one-hour requirement, equation (6.7), with limits $\text{SOC}^{\min} = 0.10$, $\text{SOC}^{\max} = 0.95$ and $\eta_{\text{c}} = \eta_{\text{d}} = 0.95$. The state of charge decides which direction is available.

> **Principle 6.1.** *The state of charge determines direction-dependent flexibility: a battery near empty can only absorb, a battery near full can only deliver, and only a battery in the interior can do both.*

6.3.5 The thermostat analogy

For a thermostat, a cool building means more curtailment headroom; for a battery, a high state of charge means more discharge headroom. The two states play the same role, and Table 5.1 of Chapter 5 set them side by side. The battery is simply much cleaner mathematically: its state is directly measured, its limits are numbers, and its dynamics are equation (6.3) with no weather term.

6.3.6 Customer reserve constraints

For a residential battery the owner may want 30 % reserved for outage backup, so that $\text{SOC}_t \geq 0.30$ at all times. Even if the battery could physically go lower, the VPP should not use that energy. The useful grid capacity is constrained by the customer contract, which is the "customer-constrained" property of Chapter 1 in its simplest form: the reserve is a number the customer chose, and it replaces $\text{SOC}^{\min}$ in equations (6.6) and (6.7).

6.4 Dispatch

Given the model and its constraints, this section asks how a battery should be used: when, for what objective, and at what cost in losses and wear.

6.4.1 A future obligation and the opportunity cost

Suppose the utility expects a major peak at 18:00. At 15:00 the battery has SOC = 0.70 and prices are already somewhat high. Should it discharge now? Perhaps not: if too much energy is used now, SOC at 18:00 may be insufficient for the more valuable later event. The real decision is

Is the current dispatch worth the opportunity cost of reduced future flexibility?

That is the intertemporal problem of Chapter 2 with the state equation (6.3) making the cost explicit: every kWh discharged at 15:00 is a kWh of $(\mathrm{SOC}_{18:00} - \mathrm{SOC}^{\min})\,E^{\max}$ that is not there at 18:00.

6.4.2 Price arbitrage

If the price at 14:00 is \$30/MWh and at 18:00 is \$150/MWh, storage should charge in the cheap period and discharge in the expensive one. The storage scheduling problem is

$$\min_{P_1,\dots,P_T} \sum_t c_t\,P_t\,\Delta t \quad \text{subject to equations (6.3), (6.4) and (6.5),} \tag{6.8}$$

where a negative P_t (discharge) earns revenue. Figure 6.4 shows a day's solution for the 10 kW, 40 kWh battery with $\eta_c = \eta_d = 0.95$ against an illustrative price profile: the battery fills in the cheapest overnight hours and empties through the evening price peak. Over the day it draws 35.8 kWh and returns 32.3 kWh, the difference being the round-trip loss of Section 6.4.5, for a net gain of about \$3.3. A residential battery earns little from arbitrage alone; the value of the same cycle is far larger when the discharge also serves a committed grid service.

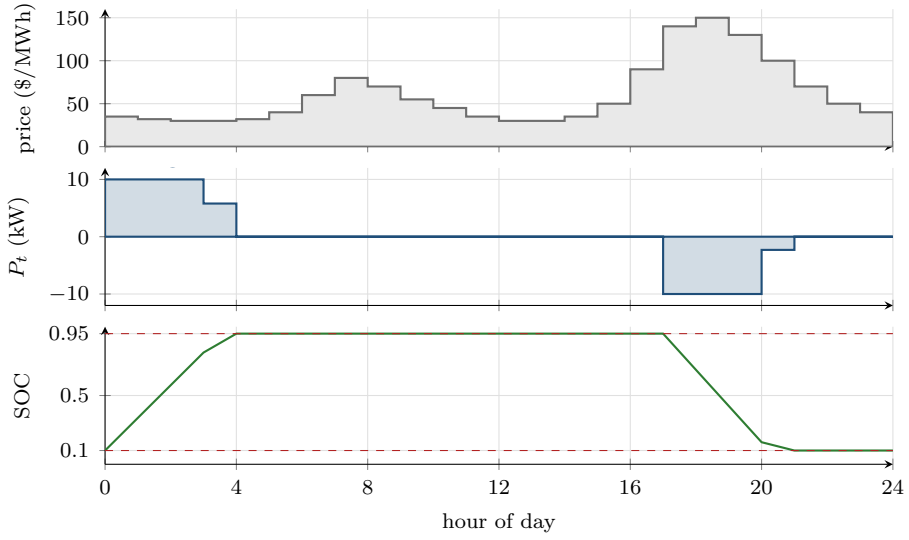


Figure 6.4: A day of price arbitrage for the 10 kW, 40 kWh battery. Top: hourly price. Middle: dispatch, charging at full power in the cheapest hours and discharging through the evening peak. Bottom: state of charge between its limits. Data from `scripts/ch67_data.py`.

6.4.3 A multi-objective dispatch

Inside a VPP the objective is rarely price alone. It is a sum,

$$J = J_{\text{grid tracking}} + J_{\text{degradation}} + J_{\text{customer}} + J_{\text{market cost}}, \tag{6.9}$$

and the battery has to balance meeting a utility dispatch target, preserving the backup reserve, limiting degradation, and capturing market value. Optimal battery dispatch is multi-objective, which is the objective of Chapter 3 specialized to one device class.

6.4.4 Degradation

Cycling is not free; usage reduces battery life. A simple way to include this is a cost proportional to the energy throughput,

$$C^{\text{deg}} = \lambda \sum_t |P_t|\,\Delta t, \tag{6.10}$$

and the optimizer solves min $J_{\text{grid}} + C^{\text{deg}}$. The model is crude (real degradation depends on depth of discharge, temperature and rate as well as throughput) but it does the essential job: it stops the algorithm from cycling the battery aggressively for tiny benefits. In the arbitrage of Figure 6.4, a λ of \$50/MWh would wipe out the \$3.3 gain and leave the battery idle, which is the correct decision for a cell that costs more to wear than the price spread is worth.

6.4.5 Round-trip efficiency

With $\eta_{\text{c}} = 0.95$ and $\eta_{\text{d}} = 0.95$ the round-trip efficiency is

$$\eta_{\text{RT}} = \eta_{\text{c}}\,\eta_{\text{d}} = 0.9025, \tag{6.11}$$

so about 90 % of the energy put in comes back out: charge 10 kWh and recover about 9 kWh, as Figure 6.1 showed. An arbitrage therefore has to overcome the loss: with the prices of Section 6.4.2, discharging at \$150 recovers $0.9025 \times 150 \approx$ \$135 per MWh charged, so the spread must exceed about 10 % of the discharge price before the cycle pays.

6.4.6 Response speed

Batteries change power very quickly compared with thermal loads and with generators that have long ramp times: $P_t \to P_{t+1}$ can move across the full range in under a second, subject to inverter limits and control-system settings. This makes them the natural resource for fast services. The point to retain for this chapter is

battery power is highly controllable, but battery energy is finite.

6.4.7 The feasible region

At a given state the feasible power range is $P_t \in \mathcal{U}(\text{SOC}_t)$, the state-dependent feasible set of Chapter 1. Table 6.1 gives it at three states for the 10 kW battery with limits $[0.10, 0.95]$ and a one-hour step, and Figure 6.5 shows the whole band: the interval of allowed powers as a function of the state of charge, which is the one-hour section of Figure 6.3 drawn as a region.

6.5 Fleet and forecasting

At fleet scale the battery acquires the two properties it seemed to lack: a distribution of states, and a dependence on forecasts.

Table 6.1: Feasible power at three states of charge (one-hour step).

State	Feasible P_t	Meaning
$\mathrm{SOC} = 0.50$ (interior)	$-10 \le P_t \le 10\,\mathrm{kW}$	both directions
$\mathrm{SOC} = 0.10$ (at $\mathrm{SOC}^{\min}$)	$0 \le P_t \le 10\,\mathrm{kW}$	charging only
$\mathrm{SOC} = 0.95$ (at $\mathrm{SOC}^{\max}$)	$-10 \le P_t \le 0$	discharging only

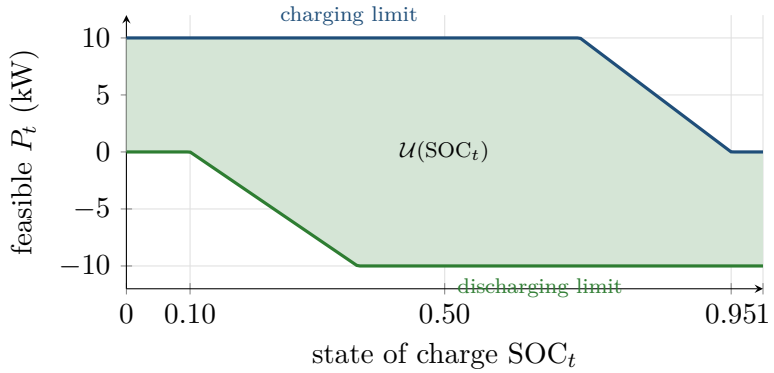


Figure 6.5: The feasible control set $\mathcal{U}(\mathrm{SOC}_t)$ of the 10 kW, 40 kWh battery over one hour. The band narrows to charging-only at the lower limit and discharging-only at the upper limit.

6.5.1 Fleet aggregation

With $N = 10\,000$ batteries, each with its own SOC_i, $P_{i,\max}$, $E_{i,\max}$ and customer constraints, the fleet power and stored energy are

$$P_{\mathrm{fleet},t} = \sum_i P_{i,t}, \qquad E_{\mathrm{fleet},t} = \sum_i \mathrm{SOC}_{i,t}\, E_{i,\max}, \tag{6.12}$$

but not all of that energy is usable; the usable part is

$$E^{\mathrm{usable}} = \sum_i (\mathrm{SOC}_i - \mathrm{SOC}_{i,\min})\, E_{i,\max}. \tag{6.13}$$

These sums begin to define a VPP-level storage resource.

6.5.2 An average state of charge can mislead

Two four-battery fleets both have an average state of charge of 50 %. Fleet A is $\{50, 50, 50, 50\}$ %; fleet B is $\{95, 95, 5, 5\}$ %. With limits $[0.10, 0.95]$ and 10 kW units, fleet A can deliver $4 \times 10 = 40$ kW in either direction. Fleet B can discharge only from its two full batteries (20 kW) and charge only into its two empty ones (20 kW); its two full units are at $\mathrm{SOC}^{\max}$ and cannot absorb, its two empty units are below $\mathrm{SOC}^{\min}$ and cannot deliver. Figure 6.6 draws the two distributions and the resulting capability.

Principle 6.2. *The average state of charge does not describe fleet flexibility. The distribution of states across the fleet does, because the limits bind unit by unit.*

6.5.3 Forecasting still matters

A battery looks deterministic: SOC_t, $P^{\max}$ and $E^{\max}$ are known. For customer-owned batteries a good deal remains to forecast: whether the device will be online, the household's demand,

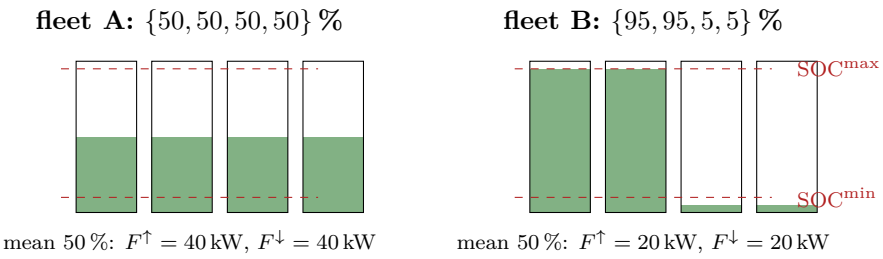


Figure 6.6: Two fleets with the same average state of charge and half the capability. Each box is a 10 kW unit, filled to its state of charge; the dashed lines are the operating limits. The distribution, not the mean, sets what the fleet can do.

rooftop PV production, the customer's backup needs, opt-out behavior, the future state of charge, and the dispatch requests that will arrive. Residential solar charges the battery through the afternoon, so $\mathrm{SOC}_{18:00}$ depends on $P_{\mathrm{PV}}(t)$, and weather forecasting re-enters the problem through the battery's own charging source.

6.5.4 Solar and battery

Suppose $\mathrm{SOC}_{12:00} = 0.40$, the expected afternoon solar production is 20 kWh and the expected household consumption is 8 kWh. The net 12 kWh of surplus charges the 40 kWh battery by 0.30, so $\mathrm{SOC}_{18:00} \approx 0.70$. If clouds cut solar production to 10 kWh the surplus is 2 kWh and $\mathrm{SOC}_{18:00} \approx 0.45$. Figure 6.7 draws the two trajectories. The difference between them is 10 kWh of usable energy, an hour of full-power discharge, and it is decided by the weather rather than by anything the battery reports.

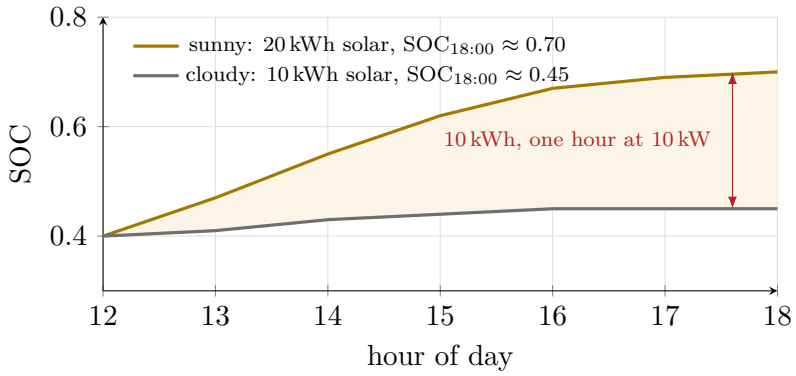


Figure 6.7: State of charge of a solar-coupled home battery through the afternoon under a sunny and a cloudy forecast. The evening capacity offered to the VPP depends on which one occurs. Trajectories are illustrative.

> **Principle 6.3.** *Future battery flexibility must be forecast, not merely measured now, whenever the battery's charging source or the customer's use is uncertain.*

6.5.5 Representing the uncertainty

Instead of a number $\mathrm{SOC}_{18:00} = 0.70$, the platform holds a distribution $\mathrm{SOC}_{18:00} \sim p(\mathrm{SOC})$, which induces a distribution of available discharge power $P^{\mathrm{avail}}_{18:00}$, and the operator cares about $\Pr(P^{\mathrm{avail}} \geq P^{\mathrm{commit}})$. The forecasting track and the operations track meet here exactly as they did for thermostats in Chapter 5; only the source of the uncertainty differs.

6.5.6 Batteries as the balancing resource

Inside a heterogeneous VPP the battery's speed and bidirectionality give it a particular job: filling the gaps left by the other resources. If the fleet must follow $P^{\text{tgt}}(t)$, the thermostat response is strong and then decays, the EV response is variable, and the battery supplies the difference,

$$P_{\text{bat}}(t) = P^{\text{tgt}}(t) - P_{\text{thermostat}}(t) - P_{\text{EV}}(t). \tag{6.14}$$

Figure 6.8 shows the three contributions stacked against a flat target. Because the battery is the residual, its energy budget must cover the integrated shortfall of the others over the window, which is the practical meaning of "MW and MWh" for a portfolio.

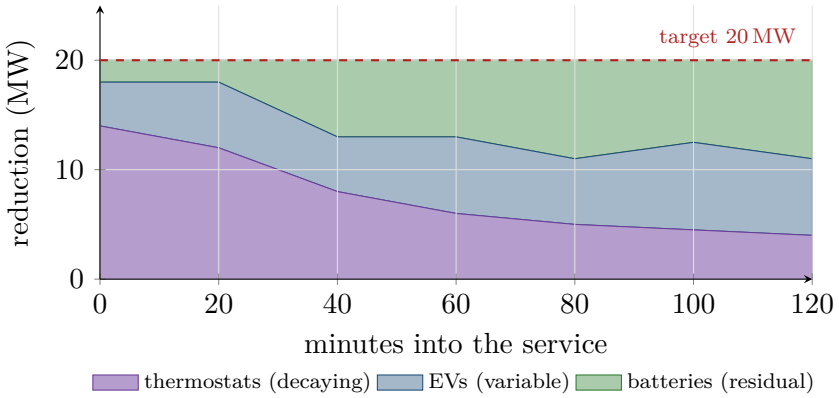


Figure 6.8: Batteries as the balancing resource, equation (6.14): the thermostat contribution decays, the EV contribution varies, and the battery contribution is whatever is needed to hold the 20 MW target. Values are illustrative.

6.5.7 Worked fleet example

A VPP has 1 000 batteries of 10 kW and 40 kWh, with an average SOC $= 0.70$ and $\text{SOC}^{\text{min}} = 0.20$. Nominal power capacity is $1\,000 \times 10 = 10$ MW; usable energy per battery is $(0.70 - 0.20) \times 40 = 20$ kWh, so fleet usable energy is $1\,000 \times 20 = 20$ MWh. Nominally, then,

10 MW for roughly 2 hours

is available, which is a far more useful statement than "a 10 MW battery fleet". Two cautions from earlier sections apply: the 20 MWh assumes the average describes every unit (Section 6.5.2), and delivering it to the grid returns $0.95 \times 20 = 19$ MWh after discharge losses (Section 6.4.5).

6.6 Comparison and control formulation

The chapter closes its body by setting the battery beside the thermostat and stating the control problem in full.

6.6.1 Thermostat and battery compared

Table 6.2 compares the two resource classes feature by feature. From a control perspective they are surprisingly similar: both fit $x_{t+1} = f(x_t, u_t, w_t)$ with a bounded state and a customer constraint. The differences are in the measurability of the state, the explicitness of the dynamics, and the battery's bidirectionality.

Table 6.2: Thermostat and battery as flexible resources.

Feature	Thermostat	Battery
Main state	indoor temperature	SOC
Power response	HVAC cycling	charge or discharge
Stored quantity	thermal energy	electrochemical energy
Customer constraint	comfort band	reserve, SOC limits
Duration limit	thermal drift	stored energy
Rebound	HVAC recovery	recharge requirement
Bidirectional	limited, indirect	yes
Dynamics	building-dependent, inferred	explicit, equation (6.3)

6.6.2 The control formulation

For battery i, $x_{i,t} = \mathrm{SOC}_{i,t}$ and $u_{i,t} = P_{i,t}$, with $\mathrm{SOC}_{i,t+1} = f(\mathrm{SOC}_{i,t}, P_{i,t})$ from equation (6.3). At fleet scale,

$$\min_{\{P_{i,t}\}} J \quad \text{subject to} \quad \mathrm{SOC}_{i,\min} \le \mathrm{SOC}_{i,t} \le \mathrm{SOC}_{i,\max}, \qquad -P_i^{\mathrm{dis}} \le P_{i,t} \le P_i^{\mathrm{ch}}, \tag{6.15}$$

for every i and t. This is a clean optimal-control formulation, and because the dynamics are linear and the constraints are boxes, it is a linear or quadratic program for most objectives, which is why battery fleets are the easiest VPP resource to dispatch well.

6.7 What to retain

Definition 6.1 (Battery flexibility)**.** Battery flexibility is jointly constrained by power and energy. Power says how much right now; state of charge and energy say for how long and what remains afterward.

The DERMS must therefore think intertemporally,

$$\text{dispatch now} \rightarrow \text{SOC later} \rightarrow \text{future flexibility},$$

and at VPP scale the problem is

$$\text{forecast state} + \text{allocate limited energy} + \text{meet grid objectives}.$$

Table 6.3 lists the chapter's constraints and the quantity each one limits.

Table 6.3: The constraints on a battery and what each limits.

Constraint	Equation	Limits
Inverter rating	(6.4)	power now
State-of-charge limits	(6.5)	energy available in each direction
Customer reserve	$\mathrm{SOC} \ge \mathrm{SOC}^{\mathrm{res}}$	energy the VPP may use
Dynamics	(6.3)	how use now changes limits later
Efficiency	(6.11)	energy returned per energy stored
Degradation	(6.10)	how often cycling is worth it

6.8 Exercises

Exercise 6.1. A battery has $E^{\max} = 13.5\,\text{kWh}$, $P^{\max} = 5\,\text{kW}$, $\eta_c = \eta_d = 0.92$, limits $[0.10, 0.95]$ and $\text{SOC}_t = 0.55$. Compute the usable discharge energy, the energy that reaches the grid, and the duration at full power. Then compute the same for charging.

Exercise 6.2. Using the price profile of Figure 6.4, show that the greedy rule (charge at full power when the price is at most \$35/MWh, discharge at full power when it is at least \$100/MWh) is optimal for a battery that starts and ends at $\text{SOC}^{\min}$, and compute the daily gain exactly.

Exercise 6.3. Fleet B of Section 6.5.2 is rebalanced by charging the two empty units from the two full ones through the grid. How much energy is lost in the transfer at $\eta_{\text{RT}} = 0.9025$, and what is the fleet's capability afterward?

Exercise 6.4. In Section 6.5.4, suppose the solar forecast is $\mathcal{N}(20, 5^2)\,\text{kWh}$. What is the distribution of $\text{SOC}_{18:00}$, and what is the probability that at least 15 kWh of usable energy (reserve 0.20) is available at 18:00?

6.9 Connection to the TERRA framework

The battery is the simplest component in this book, and it is the one on which a conservation-graph model is most transparent.

- **The battery is an accumulator on one node.** Equation (6.2) is a conservation statement for stored energy: inflow minus outflow minus losses equals the change in storage. On a graph it is a single edge from the grid node into a storage component, with efficiencies as loss terms on the edge. The thermostat of Chapter 5 needed two conservation laws; the battery needs one.
- **Limits are bounds on the stored quantity.** Equations (6.4) and (6.5) are bounds on an edge flow and on a stored state. The feasible set $\mathcal{U}(\text{SOC}_t)$ of Figure 6.5 is the set of edge flows that keep the state inside its bounds over the next step, which the model evaluates by one forward step.
- **The fleet distribution is the joint state.** Figure 6.6 is the observation that fleet capability is a function of the vector of states, not of their mean. A model that carries every unit's state answers the question by evaluating the bounds unit by unit; one that carries only the fleet total cannot distinguish fleet A from fleet B.
- **Solar coupling is a second edge into the same node.** The sunny and cloudy trajectories of Figure 6.7 are forward simulations with the PV edge driven by two different forecasts of the disturbance. The posterior over $\text{SOC}_{18:00}$ of Section 6.5.5 is the same simulation run from the distribution of the forecast rather than from two of its values.
- **The dispatch is a linear program on the graph.** Equation (6.15) with the linear dynamics is the problem the model hands to an optimizer: edge flows as decisions, conservation and bounds as constraints, the posterior as the initial state.

In summary, a battery is a bounded accumulator with lossy edges, and every result of this chapter (duration, direction, arbitrage, the misleading mean, the solar dependence) is read off that component by forward simulation. The next chapter adds the one thing a stationary battery never does: it leaves.

Chapter 7

EV managed charging: flexibility as a scheduling problem

7.1 The question

Electric vehicles combine the battery dynamics of Chapter 6 with a new kind of uncertainty: the resource can physically leave. A stationary battery is always there; an EV is connected only while it is parked and plugged in, and the times at which that happens are decided by the driver. The question of the chapter is

How can EV charging be shifted without preventing the driver from having enough energy when the car is needed?

Jargon: managed charging, V2G, Level 2

Managed charging (also smart charging) means the aggregator controls *when* and how fast a plugged-in EV charges; energy flows only into the car. *Vehicle-to-grid* (V2G) adds discharge from the car into the grid. A *Level 2* home charger delivers roughly 7 kW on a 240 V circuit; the 7.2 kW used throughout this chapter is a typical rating.

7.2 One vehicle's night

One commuter's evening and night contain the whole idea of managed charging; this section follows it from plug-in to departure.

7.2.1 Uncontrolled charging

A driver arrives home at 18:00 with SOC = 0.30 and plugs in. The charger draws 7.2 kW from the moment of connection until the battery is full, so the load profile is a rectangle starting at 18:00. When thousands of drivers do this at the same time, EV charging adds a block to the evening peak that is already the hardest hour of the day. Managed charging asks whether the charging actually needs to happen immediately. Often it does not.

7.2.2 The deadline

Suppose the driver leaves at 07:00 the next morning and requires SOC ≥ 0.80 at departure. The real constraint is

$$\mathrm{SOC}(07{:}00) \geq 0.80, \tag{7.1}$$

not "start charging at 18:00". The gap between the two statements is the flexibility.

7.2.3 The problem is temporal

With $E^{\max} = 60\,\text{kWh}$, $\text{SOC}_0 = 0.30$ and $\text{SOC}^{\text{req}} = 0.80$, the energy that must be stored is

$$E^{\text{need}} = (0.80 - 0.30) \times 60 = 30\,\text{kWh}, \tag{7.2}$$

and at 7.2 kW the ideal charging time is $30/7.2 \approx 4.17\,\text{h}$ (or about 4.5 h at $\eta = 0.92$). The car is plugged in from 18:00 to 07:00, a window of 13 h. Only about 4.2 h of charging is needed inside a 13 h window. Figure 7.1 shows the window and the block of charging that can slide anywhere inside it.

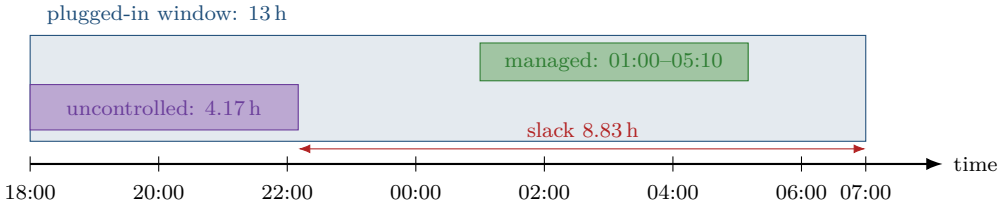


Figure 7.1: One EV's night. The charging block (4.17 h at 7.2 kW) can be placed anywhere inside the 13 h plugged-in window; the difference, 8.83 h, is the slack available to the aggregator.

7.2.4 Charging slack

The conceptual quantity is

$$\text{slack} = \text{time available} - \text{time required to charge} = 13 - 4.17 = 8.83\ \text{h}. \tag{7.3}$$

The DERMS has almost nine hours of scheduling freedom for this one car. That is a large resource, and it was created by nothing more than the difference between when the car could charge and when it must.

7.2.5 Managed charging

Instead of 18:00 to 22:10, the DERMS may choose 01:00 to 05:10. The driver still gets $\text{SOC} \geq 0.80$ at departure, but the charging has moved away from the evening peak. EV flexibility usually comes from delaying or reshaping charging, not from eliminating energy use: the same 30 kWh is delivered, at a different time.

7.2.6 Comparison with thermostats

A thermostat shifts cooling in time subject to $T^{\min} \leq T \leq T^{\max}$; an EV shifts charging in time subject to $\text{SOC}(T^{\text{dep}}) \geq \text{SOC}^{\text{req}}$. Both are flexible loads, but the customer constraint differs in kind: the thermostat's is a *comfort band* that must hold at every instant, the EV's is an *energy deadline* that must hold at one instant. A deadline is easier to satisfy, because anything is allowed before it, and harder to miss gracefully, because missing it by an hour means the driver cannot leave.

7.3 The model

This section writes the EV as a battery with one extra variable, the plug-in indicator, and shows how much that variable changes.

7.3.1 State dynamics

While plugged in, the EV's state of charge follows the battery equation of Chapter 6 with charging only,

$$\mathrm{SOC}_{t+1} = \mathrm{SOC}_t + \frac{\eta\, P_t\, \Delta t}{E^{\max}}, \tag{7.4}$$

with charging power P_t as the control, $u_t = P_t$. The familiar structure $x_{t+1} = f(x_t, u_t)$ appears again.

7.3.2 The availability variable

Now add $a_t \in \{0, 1\}$, equal to 1 when the vehicle is plugged in, and let it gate the power:

$$0 \le P_t \le a_t\, P^{\max}. \tag{7.5}$$

If $a_t = 0$ then $P_t = 0$: no command can charge a car that is not there. Figure 7.2 shows the gate over a day. This one variable changes everything. A stationary battery is usually still there; the EV may not be, and the plug-in pattern is set by the driver's life, not by the grid.

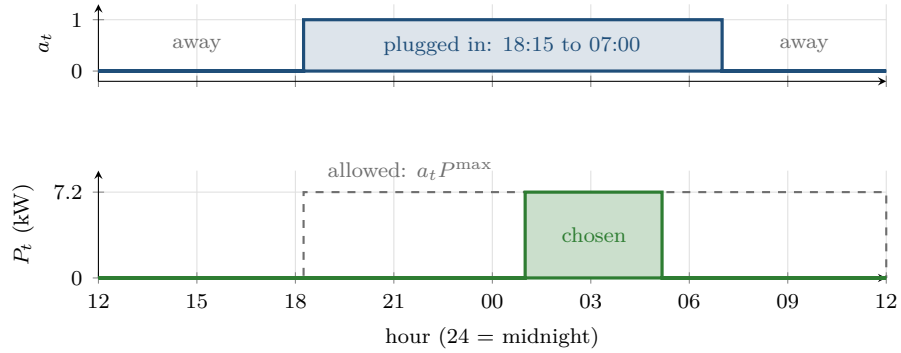


Figure 7.2: The availability gate. Top: the plug-in indicator a_t over a day. Bottom: the allowed power $a_t\, P^{\max}$ (dashed) and one managed schedule inside it (green). Outside the window no schedule is possible.

7.3.3 Arrival and departure are uncertain

The DERMS may not know when the vehicle will arrive. The arrival time T^{arr} is a random variable with a forecast distribution $p(T^{\mathrm{arr}})$, for example centered on 18:15 with an hour of spread; likewise the departure time T^{dep}. Figure 7.3 shows typical commuter distributions. With both ends of the window random, EV scheduling is a stochastic problem before any price or grid objective is added.

7.3.4 The required state of charge is uncertain too

A commuter usually needs $\mathrm{SOC}^{\mathrm{req}} = 0.70$, but tomorrow may bring a longer trip. The system may have to estimate $p(\mathrm{SOC}^{\mathrm{req}})$ as well. The customer constraint itself is uncertain, which has no counterpart in a stationary utility battery and only a weak one in a thermostat (where the comfort band is at least fixed).

7.4 Scheduling

Choosing when each car charges is a scheduling problem; this section states it, adds the local constraints, and introduces the urgency rule that solves it well in practice.

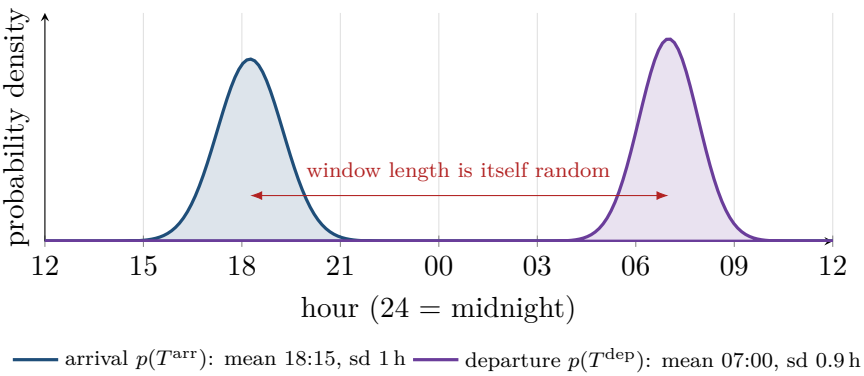


Figure 7.3: Forecast distributions of arrival and departure for a commuter. The plugged-in window is the gap between two random times, so the slack of equation (7.3) is random too.

7.4.1 The basic optimization

With an electricity price c_t, the managed-charging problem for one EV is

$$\begin{aligned} &\min_{P_t} \sum_t c_t\, P_t\, \Delta t \\ &\text{subject to} \quad \mathrm{SOC}_{t+1} = \mathrm{SOC}_t + \frac{\eta P_t \Delta t}{E^{\max}}, \qquad 0 \le P_t \le a_t P^{\max}, \qquad \mathrm{SOC}_{T^{\text{dep}}} \ge \mathrm{SOC}^{\text{req}}. \end{aligned} \tag{7.6}$$

This is the core managed-charging problem: the EV problem of Chapter 2 with the availability gate added.

7.4.2 Other objectives

If price is irrelevant the objective can be peak reduction, $\min \max_t P_{\text{grid}}(t)$, which shifts charging away from system peaks; or a local limit, $P_{\text{tr}}(t) \le P_{\text{tr}}^{\max}$ on a distribution transformer, which constrains charging by the local infrastructure. Same EVs, different objective, as in Table 2.1 of Chapter 2.

7.4.3 Local constraints

Ten EVs behind the same transformer, each able to draw 7.2 kW, would draw 72 kW charging simultaneously. If the transformer has only 40 kW of remaining capacity the constraint is

$$\sum_i P_{i,t} \le 40\,\text{kW}, \tag{7.7}$$

and the DERMS must coordinate the ten chargers. The problem has become an *allocation*: which cars get the 40 kW now, and which wait. The urgency metric of Section 7.4.6 is the natural way to decide.

7.4.4 Fleet-level managed charging

For $N = 100\,000$ EVs the DERMS holds, for each vehicle i, the state $\mathrm{SOC}_{i,t}$, the departure T_i^{dep}, the requirement $\mathrm{SOC}_i^{\text{req}}$ and the rating $P_{i,\max}$, and chooses every $P_{i,t}$. The fleet charging load is $P_{\text{fleet},t} = \sum_i P_{i,t}$, and the challenge is to shape that aggregate while satisfying every customer's departure requirement.

7.4.5 Earliest-deadline intuition

A simple scheduling idea: charge the vehicles with the most urgent deadlines first. If EV A leaves at 06:00, EV B at 08:00 and EV C at 10:00, then all else equal A is the most urgent. This is the earliest-deadline-first rule of classical scheduling theory. "All else equal" hides the energy side, which the next section restores.

7.4.6 An urgency metric

Let EV i still need E_i^{rem} and have T_i^{rem} before departure. A rough urgency ratio is

$$U_i = \frac{E_i^{\text{rem}} / P_{i,\max}}{T_i^{\text{rem}}}, \tag{7.8}$$

the charging time still needed divided by the time left. Its interpretation:

- $U_i \ll 1$: a lot of flexibility;
- $U_i \approx 1$: very little; the car must charge almost continuously from now on;
- $U_i > 1$: the target cannot be met even at the maximum rate.

Figure 7.4 places vehicles in the plane of time left versus charging time needed; lines of constant urgency are rays from the origin, and the line $U = 1$ separates feasible from infeasible.

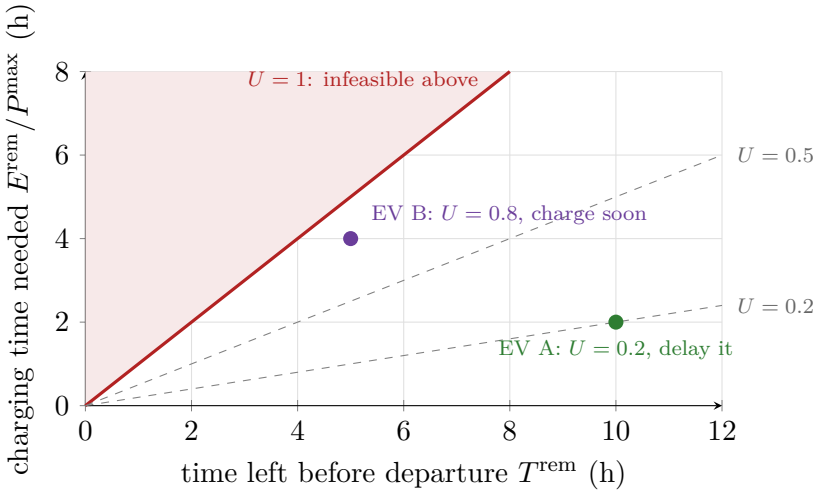


Figure 7.4: Urgency, equation (7.8). Each EV is a point; urgency is the slope of the ray from the origin. EV A (14 kWh needed at 7 kW, 10 h left) and EV B (28 kWh needed, 5 h left) from Section 7.4.7.

7.4.7 Example

EV A needs $E^{\text{rem}} = 14\,\text{kWh}$ at $P^{\max} = 7\,\text{kW}$, so 2 h of charging, with departure in 10 h: $U_A = 0.2$, very flexible. EV B needs 28 kWh at 7 kW, so 4 h, with departure in 5 h: $U_B = 0.8$, much less flexible. If load must be curtailed, EV A is the one to delay.

7.5 Direction, forecasting and risk

Three further aspects of the resource are treated here: the direction of its flexibility, the abstraction that organizes it, and the forecasts and chance constraints that govern it.

7.5.1 Customer flexibility is asymmetric

If an EV is already charging, its power can usually be reduced from 7.2 kW to zero: downward-load flexibility. But if it is already charging at its maximum, consumption cannot be increased further. At other times an idle, plugged-in EV offers upward-load flexibility, from zero to 7.2 kW, which can absorb excess renewable generation. As for the battery of Chapter 6, the direction available depends on the current operating point, with the extra condition that the car must be plugged in at all.

7.5.2 Renewable absorption

When solar output is very high at 13:00 the DERMS can advance EV charging rather than curtail the solar: $P_{\mathrm{EV}}(13{:}00)$ rises and $P_{\mathrm{EV}}(18{:}00)$ falls. The fleet moves energy consumption from the evening to midday, which is the demand flexibility of Chapter 2 with the EV as the instrument. This works only for cars plugged in at midday (workplace chargers, home chargers of people who are home), which is why the plug-in pattern of Figure 7.3 matters at both ends of the day.

7.5.3 An energy queue

A useful abstraction: each car arrives with an amount of work E_i that must be completed before a deadline T_i^{dep}, and the DERMS decides when to process it. EV charging is *deadline-constrained energy scheduling*, Figure 7.5. This is probably the cleanest conceptual description of the resource, and it connects managed charging to a large body of scheduling theory, from which rules such as earliest-deadline-first and least-slack-first are borrowed.

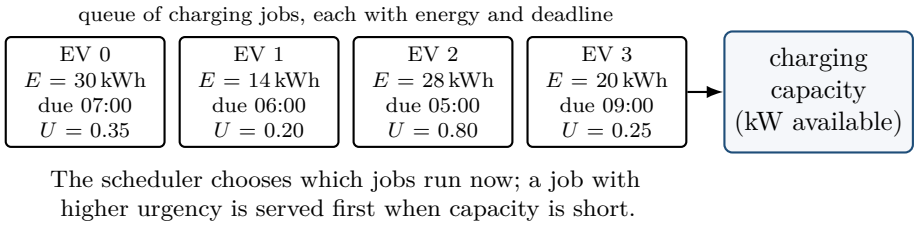


Figure 7.5: EV charging as an energy queue. Jobs arrive with an amount of work and a deadline; the scheduler decides their order and timing within the available capacity.

7.5.4 Three forecasts

Three quantities have to be forecast for tomorrow's fleet, and Table 7.1 lists them with the consequence of getting each wrong.

The departure time is the most distinctive: it acts almost like a hard horizon in the optimization, and unlike plug-in count or energy need it can change by nine hours on a single phone call.

7.5.5 Probabilistic constraints

With departure uncertain, the deterministic requirement $\mathrm{SOC}(T^{\mathrm{dep}}) \geq \mathrm{SOC}^{\mathrm{req}}$ at one known time becomes

$$\Pr\big(\mathrm{SOC}(T^{\mathrm{dep}}) \geq \mathrm{SOC}^{\mathrm{req}}\big) \geq 0.99, \tag{7.9}$$

a chance constraint. Under it the car must be ready not at the mean departure but at an early quantile of the departure distribution: with the distribution of Figure 7.3, the 1st percentile of

Table 7.1: The forecasts behind EV flexibility and their failure modes.

Quantity	Object	If overestimated
Plug-in count at 18:00	$p(N^{\text{plugged}})$	the VPP overcommits: cars assumed present are not
Energy need on arrival	$p(\text{SOC}^{\text{arr}})$	too little slack per car; a car at 20 % has far less flexibility than one at 80 %
Departure time	$p(T^{\text{dep}})$	the horizon ends early; a departure that moves from 07:00 to 22:00 removes most of the slack

departure is about 04:55, so charging must finish two hours earlier than a deterministic 07:00 plan would require, and the slack of equation (7.3) shrinks from 8.8 to 6.7 h. This is one place where uncertainty directly changes the control, not just the forecast.

7.6 Vehicle-to-grid and the fleet

The last body section extends the model to bidirectional charging and then to the fleet, ending with a simulation of three charging policies.

7.6.1 Vehicle-to-grid

So far $P_t \geq 0$: charging only. With V2G the power can be negative,

$$-P^{\text{dis}} \leq P_t \leq a_t\, P^{\text{ch}}, \tag{7.10}$$

and the EV behaves much more like the battery of Chapter 6, with the extra constraint that *it must still be ready to drive*. Consider an EV arriving at 17:00 with SOC = 0.90 whose driver needs only 0.60 at 07:00, on a day with a grid peak at 18:00. The DERMS can discharge from 0.90 to 0.75 during the peak and recharge overnight to 0.60 or higher; Figure 7.6 draws the trajectory. The EV fleet becomes a true bidirectional resource.

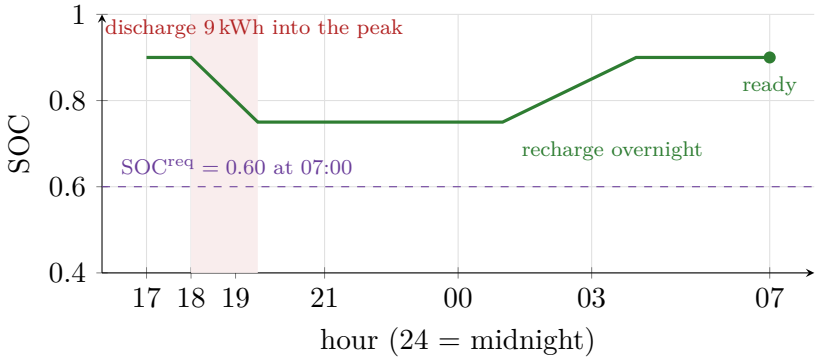


Figure 7.6: A vehicle-to-grid day for a 60 kWh EV. The car arrives nearly full, delivers 15 percentage points (9 kWh) during the 18:00 peak, and is recharged overnight well above the driver's requirement.

7.6.2 Managed charging is easier than V2G

Managed charging asks *when should energy that was going to be consumed anyway be consumed?* V2G asks *should the customer's battery energy also be discharged to the grid?* The second brings additional concerns: battery degradation, customer acceptance, interconnection rules, compensation, and inverter capability. Managed charging alone has substantial grid value, and most programs start there.

7.6.3 Diversity helps, correlation hurts

Not all drivers arrive at the same time, leave at the same time, or need the same energy. That diversity provides scheduling flexibility: with 100 000 cars of identical deadline and energy need, the optimizer could only move the whole block; with diverse cars it can reshuffle charging among them, which is the analog of cycling diversity in a thermostat fleet. But commuting creates strong patterns: a plug-in wave from 17:00 to 19:00 and a departure wave from 06:00 to 08:00, visible in Figure 7.3. Weather, weekdays, holidays and commuting schedules correlate behavior across the fleet, and the correlation lesson of Chapters 3 and 4 applies.

7.6.4 A fleet optimization, and three policies compared

For vehicles $i = 1, \dots, N$ the managed-charging problem against a fleet target is

$$\min_{P_{i,t}} \sum_t \Big(P_{\text{base},t} + \sum_i P_{i,t} - P_{\text{target},t} \Big)^2 \tag{7.11}$$

subject to, for every i,

$$\text{SOC}_{i,t+1} = \text{SOC}_{i,t} + \frac{\eta_i P_{i,t} \Delta t}{E_i}, \qquad 0 \le P_{i,t} \le a_{i,t} P_{i,\max}, \qquad \text{SOC}_{i,T_i^{\text{dep}}} \ge \text{SOC}_i^{\text{req}}.$$

This is a fairly complete representation of managed EV charging. Figure 7.7 shows what three policies do to the aggregate load of a simulated fleet of 1 000 commuter EVs (7.2 kW chargers, 60 kWh packs, arrivals near 18:15, departures near 07:00, needs from Table 7.1-style distributions):

- **uncontrolled**: every car charges on arrival; the fleet peaks at 5.9 MW at 19:45, on top of the evening system peak;
- **latest start**: every car charges as late as its own deadline allows; the peak is the same size, 6.0 MW, but moved to 05:15, which merely relocates the problem to the morning;
- **fleet cap with urgency ordering**: the fleet is held to a cap and, when demand exceeds it, the most urgent cars are served first. At 2.5 MW every car meets its deadline and the load is flat through the night; at 2.0 MW the load is flatter still but 1.6 % of cars fall short, so the cap is a reliability choice.

All three policies deliver the same 25.5 MWh. Managed charging changes only the shape.

7.6.5 Why EVs are attractive VPP resources

EVs have large batteries, substantial charging power, long plugged-in periods in many cases, and flexible charging deadlines. Even without V2G they provide significant controllable load through scheduling. The VPP does not need to create energy; it needs to decide when that energy is consumed.

7.6.6 The three DER classes compared

Table 7.2 compares the three classes developed in Chapters 5 to 7. The row that EVs add is "resource can leave": for thermostats and batteries the answer is no, for EVs it is yes, and that is why EVs introduce the concept that *resource availability itself is stochastic.*

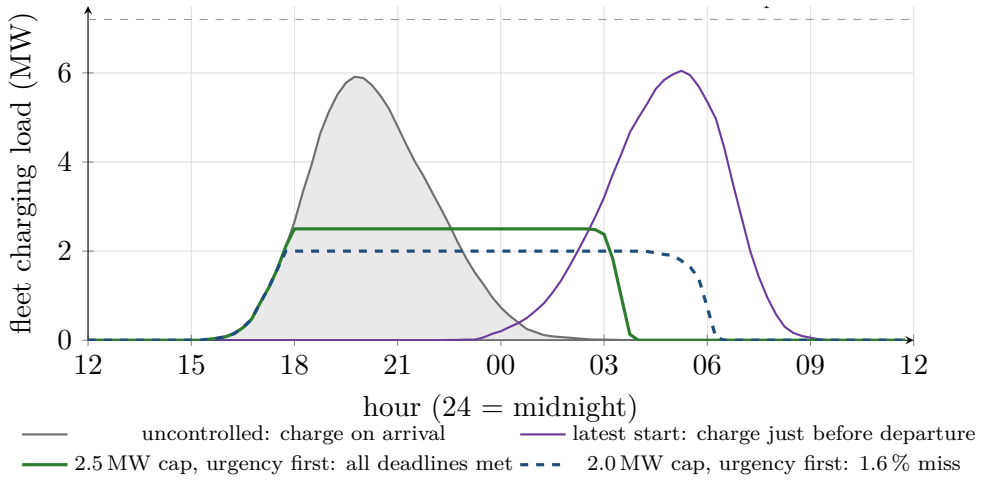


Figure 7.7: Aggregate charging load of 1 000 simulated commuter EVs under three policies. Uncontrolled and latest-start charging produce the same peak at opposite ends of the night; a fleet cap with urgency-first allocation flattens the load, and the cap level trades flatness against the fraction of deadlines met. Data from `scripts/ch67_data.py`.

Table 7.2: Thermostat, battery and EV as flexible resources.

Property	Thermostat	Battery	EV
Main state	temperature	SOC	SOC
Main control	setpoint, HVAC	charge/discharge	charging power
Main constraint	comfort band	SOC limits, reserve	departure SOC
Resource can leave?	no	no	yes
Main flexibility	thermal shifting	energy shifting	charging scheduling
Key forecast	weather, load	SOC, future use	arrival, departure, energy need

7.7 What to retain

Definition 7.1 (EV flexibility)**.** EV flexibility is the difference between when charging *could* occur and when it *must* occur. The resource is created by an energy requirement, a charging-rate limit and a deadline; the greater the slack between required charging time and plugged-in time, the greater the flexibility.

Unlike a stationary battery, the DERMS must forecast whether the resource will even be connected when it is needed. Table 7.3 lists the quantities of the chapter and the section that defines each.

Table 7.3: The quantities that describe an EV as a resource.

Quantity	Definition	Section
E^{need}	$(\text{SOC}^{\text{req}} - \text{SOC}_0)\, E^{\max}$	7.2.3
slack	window length minus $E^{\text{need}}/(\eta P^{\max})$	7.2.4
a_t	plug-in indicator; $0 \le P_t \le a_t P^{\max}$	7.3.2
U_i	$(E_i^{\text{rem}}/P_{i,\max})/T_i^{\text{rem}}$	7.4.6
chance constraint	$\Pr(\text{SOC}(T^{\text{dep}}) \ge \text{SOC}^{\text{req}}) \ge 0.99$	7.5.5
V2G bounds	$-P^{\text{dis}} \le P_t \le a_t P^{\text{ch}}$	7.6.1

7.8 Exercises

Exercise 7.1. An EV arrives at 19:30 with SOC $= 0.25$ on a 75 kWh pack, charges at 11 kW with $\eta = 0.92$, and must leave at 06:30 with SOC ≥ 0.90. Compute E^{need}, the charging time, the slack and the urgency at arrival. At what time does the urgency reach 1 if the car has not started charging?

Exercise 7.2. Ten EVs share a 40 kW transformer margin, each with a 7.2 kW charger. Five have $U = 0.9$ and five have $U = 0.2$. Which cars charge now, and at what power, under urgency-first allocation? What is the fleet load, and how does it change once the urgent cars finish?

Exercise 7.3. With departure $\sim \mathcal{N}(07{:}00, (0.9\,\text{h})^2)$ and the car of Section 7.2.3, by what time must charging finish to satisfy equation (7.9) at 0.99? At 0.90? State the slack in each case.

Exercise 7.4. In Figure 7.7, estimate the energy charged between 18:00 and 21:00 under the uncontrolled policy and under the 2.5 MW cap. Where did the difference go?

7.9 Connection to the TERRA framework

The EV is the battery component of Chapter 6 with a time-varying connection, and that single change exercises three parts of a graph framework.

- **The edge exists only sometimes.** The gate $0 \le P_t \le a_t P^{\max}$ of equation (7.5) is an edge whose capacity is switched by a schedule. A framework with schedule-driven, time-callable components represents the plug-in window directly: the charging edge carries flow only while $a_t = 1$, and the car's accumulator is simply disconnected from the grid node otherwise.

- **The deadline is a terminal constraint on a state.** $\mathrm{SOC}(T^{\mathrm{dep}}) \geq \mathrm{SOC}^{\mathrm{req}}$ is a bound on a stored quantity at one future time. Checking a schedule is a forward simulation to T^{dep}; finding one is an optimization whose only coupling across time is the accumulator, which is what makes the problem a linear program and the urgency rule of equation (7.8) a good heuristic for it.
- **Arrival, departure and energy need are uncertain inputs, not parameters.** The distributions of Figure 7.3 are disturbances w on the schedule itself. Propagating them through the model gives the distribution of a_t, and hence of the fleet's feasible set, at every hour; the chance constraint of equation (7.9) is a quantile of the terminal state under that propagation.
- **The fleet figure is a population of such components.** Figure 7.7 is 1 000 copies of the EV component with parameters drawn from distributions, summed at one grid node and run under three control policies. The urgency-first cap is a per-step allocation of a shared edge capacity among components, which is the same structure as the transformer constraint of equation (7.7).
- **V2G is the same component with the edge made bidirectional.** Equation (7.10) relaxes a bound; the accumulator, the deadline and the schedule are unchanged, which is why V2G adds operational rather than modeling complexity.

In summary, the EV is a bounded accumulator behind a switched edge with a terminal constraint and an uncertain schedule. The next chapter steps back from devices to ask how much any of them actually delivered, which is the problem of baselines and measurement and verification.

Chapter 8

Baselines and measurement and verification: how much did the fleet actually deliver?

8.1 The counterfactual problem

Chapters 5 to 7 built three resource classes and Chapter 4 committed them to a service. Every one of those chapters assumed, implicitly, that the delivered response is known. It is not. After a dispatch the platform holds a metered load trace and a record of what it commanded, and from those it must answer

How much load reduction did the fleet actually cause?

The obstacle is that the quantity of interest is a difference between what happened and what would have happened, and only the first term is measured. The second term has to be constructed. This chapter is about how it is constructed, how the construction fails, and what the failures do to the rest of the system.

Jargon: M&V, baseline, counterfactual, average treatment effect

Measurement and verification (M&V) is the procedure that establishes what a dispatch delivered: it compares the metered load with an estimate of the load that would have occurred without the dispatch, and also checks whether the service requested was actually provided. The *baseline* is that estimate. The *counterfactual* is the thing the baseline estimates: the load trajectory in the world where no command was sent, which by construction never occurred and was never measured. The *average treatment effect* (ATE) is the causal quantity an experiment estimates: the mean difference in load between devices that were dispatched and comparable devices that were not.

Key idea

DER response is not measured. It is inferred, as the difference between a measurement and an estimated counterfactual, and the estimate can dominate the answer.

This section states what quantity M&V is after, shows that two defensible baselines can halve or double the reported performance, and explains why the problem is easier for a fleet than for one device.

8.1.1 The delivered response is a difference, not a measurement

During an event the meter records $P^{\text{act}}(t)$. That is not the quantity of interest. The quantity of interest is the load reduction caused by the dispatch,

$$\Delta P(t) = P^{\text{base}}(t) - P^{\text{act}}(t), \tag{8.1}$$

which is equation (2.2) of Chapter 2 written again because it is now the object of study rather than a definition in passing. In equation (8.1) the first term is $P^{\text{base}}(t)$, the load the premises or the fleet *would have* drawn had no command been sent. No instrument reports it. It is not a measurement with an error bar; it is an estimate produced by a model, and everything in this chapter follows from that one fact.

8.1.2 Two baselines, two verdicts

Consider a feeder whose load peaks in the late afternoon. At 17:00, just before an event, the metered load is 1 000 kW. The event runs from 17:00 to 19:00, and at 18:00 the meter reads $P^{\text{act}} = 850$ kW. Two baseline models are available for that hour:

Baseline model	$P^{\text{base}}(18{:}00)$	Estimated $\Delta P(18{:}00)$
Model A: hold the pre-event level	1 000 kW	150 kW
Model B: continue the natural decline	920 kW	70 kW

The measured load is identical in the two rows. The reported performance differs by a factor of more than two. Figure 2.2 of Chapter 2 made this point with two flat baselines at a single instant; Figure 8.1 makes it with *trajectories*, because the shape of the counterfactual, not just its level, is what the two models disagree about.

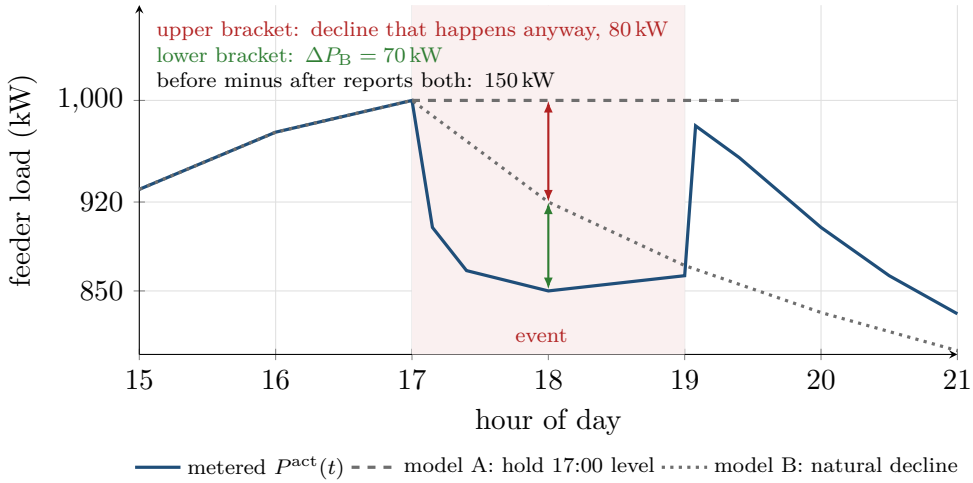


Figure 8.1: The counterfactual gap as a pair of trajectories. The blue curve is the metered load through a 17:00–19:00 event and its rebound; the dashed gray line is model A, which holds the pre-event level; the dotted gray curve is model B, which continues the feeder's natural evening decline. At 18:00 the whole 150 kW bracket is what before-minus-after reports, of which 80 kW is decline that model B says would have happened without any dispatch. Load values are the ones used in Section 8.1.2; the curve shapes between them are illustrative.

Principle 8.1. *VPP performance is partly a property of the counterfactual model. Two analysts with the same meter data and different baselines will report different performance, and neither has made an arithmetic error.*

8.1.3 Why before-minus-after fails

The shortest available baseline is the load just before the event: report $P^{\text{act}}(\text{before}) - P^{\text{act}}(\text{during})$. In Figure 8.1 that rule gives $1\,000 - 850 = 150\,\text{kW}$, which is exactly model A. The two are the same assumption stated twice: that load would have stayed flat.

Load does not stay flat. It follows a daily shape, so part of any observed drop is the shape and not the dispatch. On this feeder the natural decline between 17:00 and 18:00 is 80 kW, so attributing the full 150 kW to the fleet credits the dispatch with the evening. The error has the opposite sign just as often: during a morning ramp or a building heat wave the counterfactual load is *rising*, and a dispatch that holds load constant has delivered a real reduction that before-minus-after reports as zero. What M&V needs is the comparison of two trajectories over the same interval, not the comparison of two adjacent measurements.

8.1.4 The baseline is a forecast

Write the baseline with the hat notation of the rest of the book,

$$P^{\text{base}}(t) = \hat{P}^{\text{nc}}(t) = f\big(\text{weather}, \text{time of day}, \text{historical load}, \text{calendar}, \text{device population}\big), \tag{8.2}$$

where the superscript stands for "no control". Equation (8.2) is a forecasting model: the same regressors, the same fitting procedure and the same error metrics that the load-forecasting track of the lesson sequence develops. Nothing about it is specific to verification except *when* it is evaluated. Forecasting therefore enters the operational chain twice, once before the dispatch and once after it, and the quality of the second entry decides what the platform believes about its own performance.

8.1.5 Two forecasting problems, not one

The two entries are different problems and should not share a name. Before the event the platform forecasts *flexibility*: how much load change will be available at 18:00 tomorrow, which is a prediction about a quantity that depends on device states, weather and customer behavior, and which is used to decide what to commit and what to dispatch. After the event the platform estimates a *counterfactual*: what the load would have been at 18:00 today, which is a prediction about a world that did not occur and is used to decide what was delivered.

The two problems share regressors and differ in three ways. The flexibility forecast is conditioned on information available before the event; the baseline may condition on everything measured up to and including the event, which is strictly more information. The flexibility forecast predicts a *change* in load; the baseline predicts a *level*. And the flexibility forecast can be validated against a later observation, while the baseline can never be validated on the day it is used, because the observation it predicts is the one that was prevented. The last difference is the reason control groups and randomization appear in Section 8.5.3.

8.1.6 Device-level baselines are nearly uninformative

Take one thermostat during an event. Its compressor is off and its power is $P^{\text{act}} = 0$. Was that caused by the command? Possibly, and possibly not: the compressor cycles, and a duty cycle of $D = 0.4$ means it is off 60 % of the time without any intervention (Chapter 5). The two hypotheses are

$$P^{\text{base}} = 0: \qquad \Delta P = 0 - 0 = 0, \tag{8.3}$$

$$P^{\text{base}} = 4\,\text{kW}: \qquad \Delta P = 4 - 0 = 4\,\text{kW}. \tag{8.4}$$

Equations (8.3) and (8.4) are consistent with the same measurement. Observing $P^{\text{act}} = 0$ at one device at one instant therefore says almost nothing about the causal response of that device,

because the counterfactual of a binary cycling load is itself binary and unobserved. The same problem appears for an EV that was not going to charge in that hour anyway, and for a battery that was already idle.

8.1.7 Aggregation makes the baseline tractable

The fleet total is the sum over devices,

$$P_{\text{fleet}}(t) = \sum_{i=1}^{N} P_i(t), \tag{8.5}$$

and the cycling of individual devices, being out of phase, averages out in the sum, exactly as in Figure 4.2 of Chapter 4. The practical consequence for M&V is a change of question. Instead of

Would thermostat 73,412 have been running at 18:00?

which is a question about one Bernoulli draw, the platform asks

What would the aggregate power of these 50,000 thermostats have been at 18:00?

which is a question about a mean. The second is a far more tractable forecasting problem: the aggregate has a smooth, weather-driven, strongly predictable shape, and its relative forecast error falls with fleet size until correlation stops the improvement, which is the ρ curve of Figure 3.4 in Chapter 3 applied to the baseline instead of to the forecast. One aggregate baseline is both easier and more accurate than 50 000 device baselines summed.

8.2 Baseline methods

Four constructions are in common use, each conditioning on more information than the last and each with its own failure mode; this section develops them in order and then explains why all of them degrade under continuous control.

8.2.1 Historical-day averaging

The simplest baseline averages the same clock hour over N similar prior days,

$$P^{\text{base}}(t) = \frac{1}{N} \sum_{d=1}^{N} P_d(t), \tag{8.6}$$

with d indexing, for example, the previous ten weekdays with no event. Variants drop the highest and lowest days, or average the highest three of the last ten. Equation (8.6) needs no model and no weather data, which is why program rules often specify it.

Its weakness is that it conditions on the calendar and not on the weather. Suppose the event day reaches 37°C while the ten prior weekdays averaged 29°C. On a cooling-dominated feeder the HVAC load at 18:00 is much larger at 37°C, so the historical average underestimates the counterfactual. With the numbers of Section 8.2.2 the historical-day baseline for this feeder is 820 kW against a metered 850 kW, which reports $\Delta P = -30$ kW: the fleet appears to have *increased* load during an event in which it reduced 150 kW. The sign of the reported performance is wrong.

8.2.2 Weather-adjusted regression

The repair is to put the weather in the model. Fit, on non-event days,

$$P_t = \beta_0 + \beta_1 T_t + \beta_2 T_t^2 + \beta_3 \,\text{hour}_t + \beta_4 \,\text{weekday}_t + \varepsilon_t, \tag{8.7}$$

where T_t is the outdoor temperature T_{out}, the quadratic term captures the accelerating response of cooling load to heat, and the hour and weekday terms carry the daily and weekly shape. Then evaluate the fitted model at the event day's *actual* weather, with no control term, to obtain $\hat{P}^{\text{nc}}(t)$.

For the 18:00 hour of the feeder above, suppose the fit gives $\beta_0 = 704\,\text{kW}$ (the intercept with the 18:00 and weekday terms folded in), $\beta_1 = -10.5\,\text{kW/°C}$ and $\beta_2 = 0.5\,\text{kW/°C}^2$, so that the load minimum sits near 10.5°C and load rises with temperature above it. The model then predicts

$$T = 29°\text{C}: \;\; 704 - 304.5 + 420.5 = 820\,\text{kW}, \qquad T = 37°\text{C}: \;\; 704 - 388.5 + 684.5 = 1\,000\,\text{kW}.$$

The first value is what the ten cool prior days averaged, which is why equation (8.6) returned 820 kW. The second is the counterfactual for the event day, and against the metered 850 kW it gives $\Delta P = 150\,\text{kW}$. Figure 8.2 draws the two baselines and the metered load together.

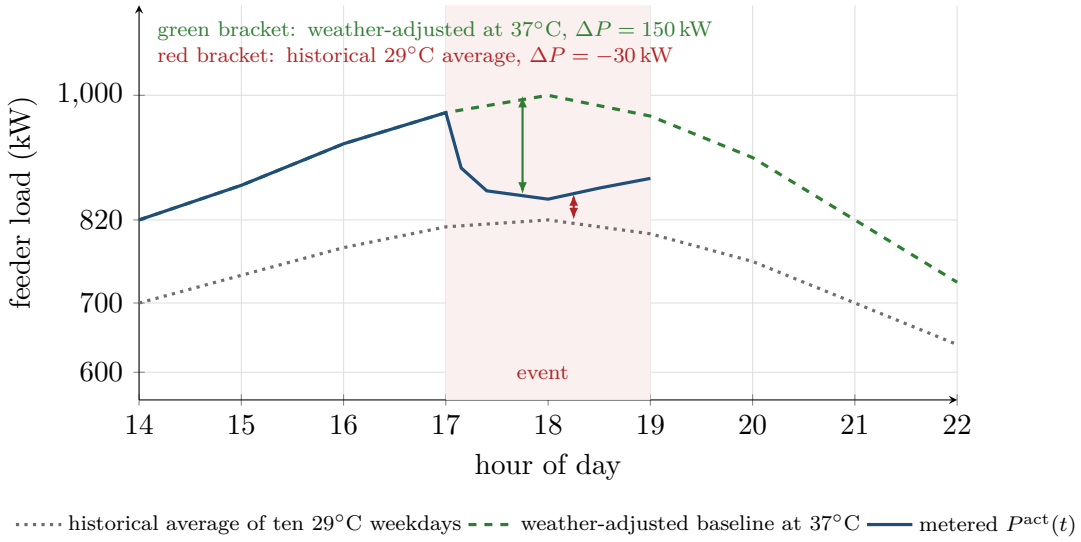


Figure 8.2: The same event evaluated against two baselines on a 37°C day. The dotted gray curve is the ten-weekday historical average, equation (8.6); the dashed green curve is equation (8.7) evaluated at the event day's temperature with $\beta_0 = 704\,\text{kW}$, $\beta_1 = -10.5\,\text{kW/°C}$ and $\beta_2 = 0.5\,\text{kW/°C}^2$; the blue curve is the meter. The weather-adjusted baseline reports $\Delta P = 150\,\text{kW}$ at 18:00 and the historical average reports $-30\,\text{kW}$. Daily shapes are illustrative; the 18:00 values are computed from the stated coefficients.

The weather-adjusted baseline also settles the ambiguity of Section 8.1.2 for this day, and in a direction worth noting. Model B extrapolated an evening decline that is observed on ordinary days; at 37°C the cooling load holds the feeder up through 18:00 and the decline does not occur. The higher baseline is the defensible one here. The lesson is not that higher baselines are right but that the choice between trajectories is a modeling question that weather data can answer and the meter cannot.

8.2.3 Pre-event adjustment

A fitted model carries a residual on any given day: occupancy, production schedules and unmodeled weather leave today's level above or below the fit. Pre-event adjustment removes

that residual using the last measurements before the event. Suppose the model predicts $P^{\text{base}}(16{:}00) = 900\,\text{kW}$ and the meter reads 960 kW. The model is under-predicting today's level by 60 kW, so shift the whole baseline,

$$P^{\text{adj}}(t) = P^{\text{mod}}(t) + \delta, \qquad \delta = \frac{1}{|W|}\sum_{t\in W}\left[P^{\text{act}}(t) - P^{\text{mod}}(t)\right] = 60\,\text{kW}, \tag{8.8}$$

where W is a window of pre-event hours. Multiplicative versions scale rather than shift. The idea is sound and the improvement is usually real: the adjustment conditions the counterfactual on today's observed level instead of on the average day.

8.2.4 Pre-event adjustment under pre-cooling

The adjustment assumes the pre-event window is untreated. It often is not. Pre-cooling, introduced in Chapter 1 and drawn in Figure 1.8, deliberately drives buildings below setpoint before an event so that more flexibility is available during it; a battery fleet charges before an event for the same reason (Chapter 6). Both raise pre-event load *because of the dispatch*.

Figure 8.3 shows what that does. With a 60 kW pre-cooling bump in the window W, equation (8.8) returns $\delta = 60\,\text{kW}$ and the adjusted baseline at 18:00 becomes 940 kW instead of 880 kW. Against a metered 800 kW the reported response rises from 80 kW to 140 kW, an overstatement of 75 %, and the extra 60 kW is the pre-cooling counted twice: once as extra baseline and once as the reduction it enabled. The rule that follows is that a baseline must be constructed from periods in which the control input was zero, and the platform knows when that was because it issued the commands.

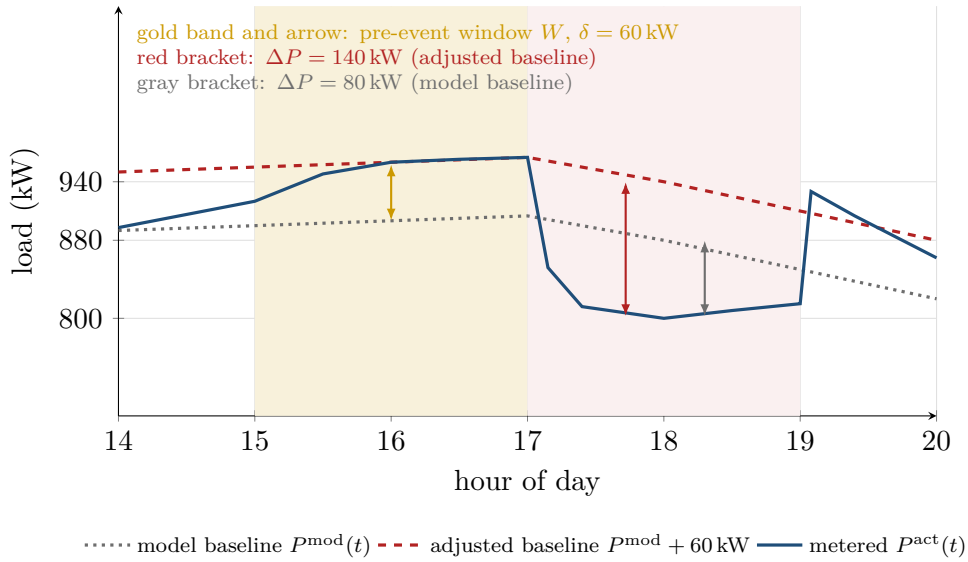


Figure 8.3: Pre-event adjustment contaminated by pre-cooling. The gold band is the pre-event window W, in which the fleet was deliberately pre-cooled; the meter (blue) therefore runs 60 kW above the model baseline (dotted) there, and equation (8.8) shifts the baseline to the dashed red curve. The reported reduction at 18:00 rises from 80 to 140 kW although the physical response is unchanged. Values are the ones used in Section 8.2.4; curve shapes are illustrative.

8.2.5 Continuous control removes the untreated period

All four constructions so far rest on the same structure: a period with no control, then a period with control, then a period with no control. Traditional demand response has that structure by

design, and Figure 2.1 of Chapter 2 is a picture of it. Continuous demand flexibility does not. If the DERMS is shaping EV charging every five minutes, holding thermostats inside a band all afternoon and cycling batteries against prices all day, then there is no untreated interval from which to build $\hat{P}^{\mathrm{nc}}$, and the question "what would have happened without control?" refers to a counterfactual several hours deep rather than two.

Three consequences follow. Historical days are themselves treated, so equation (8.6) averages controlled load. Pre-event windows do not exist. And the quantity being estimated is no longer a two-hour deviation but a whole-day trajectory under a different policy, which can only come from a physical model of the fleet simulated with the control input set to zero. This is why the M&V vocabulary of traditional demand response becomes strained for continuous VPP operation, and why verification moves from statistics on meter data toward inference in a model of the devices. Table 8.1 summarizes the four methods and the model based alternative with the failure mode of each.

Table 8.1: Baseline constructions and their failure modes.

Method	What it conditions on	Main failure mode
Historical-day average, equation (8.6)	the same hour on N similar prior days	ignores event-day weather; wrong sign on a hot day (Section 8.2.1)
Weather-adjusted regression, equation (8.7)	event-day temperature plus calendar terms	effects not among the regressors (occupancy, production) stay in the residual
Pre-event adjustment, equation (8.8)	metered load in the hours just before the event	credits pre-event control actions as baseline (Section 8.2.4)
Control group, Section 8.5.3	untreated devices measured during the event itself	bias if the treated and untreated groups are not comparable
Model-based counterfactual, Section 8.2.5	a physical fleet model simulated with zero control input	depends on the model; the only option under continuous control

8.3 Verification metrics

Verification is wider than the baseline: a fleet asked for a service can fail in several independent ways, and this section defines the metric for each and shows that the choice among them is fixed by the service, not by the analyst.

8.3.1 What verification covers

Computing $\Delta P(t)$ answers one question. M&V answers a second: *did the resource provide the service that was requested?* A VPP asked for 50 MW of reduction for two hours can fail by delivering 35 MW for two hours, 50 MW for forty minutes, 50 MW starting twenty minutes late, 50 MW reached through a ramp so slow that the first half hour is lost, or 50 MW on four days out of five. These are different failures with different operational consequences, and a single energy number conceals all of them. Verification therefore checks magnitude, timing, duration, ramp behavior, total energy and reliability, in whatever combination the service definition requires.

8.3.2 Tracking error

When the service is a trajectory, the natural metric is the deviation from it. With a requested response $P^{\mathrm{tgt}}(t)$ and an estimated delivered response $P^{\mathrm{del}}(t)$, the pointwise error is

$$e_t = P^{\mathrm{tgt}}(t) - P^{\mathrm{del}}(t), \tag{8.9}$$

the same quantity as equation (3.12) of Chapter 3, there used inside the control loop and here used after the fact. Verification summarizes it over the event by the mean absolute error, $\frac{1}{N}\sum_t |e_t|$, or by the root mean square error,

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N} e_t^2}\,, \tag{8.10}$$

which penalizes large excursions more heavily and is therefore the usual choice when the grid cares about the worst minute rather than the average minute. Note that $P^{\mathrm{del}}(t)$ in equations (8.9) and (8.10) is itself the baseline-dependent estimate $\Delta P(t)$ of equation (8.1). A tracking metric inherits the baseline's error at every step.

8.3.3 Energy delivery

When the service is an amount of energy, the metric integrates the response. A request of 20 MW for two hours is a request for

$$E^{\mathrm{tgt}} = 20\,\mathrm{MW} \times 2\,\mathrm{h} = 40\ \mathrm{MWh}, \tag{8.11}$$

and if the delivered response averages 17 MW then

$$E^{\mathrm{del}} = \int P^{\mathrm{del}}(t)\,\mathrm{d}t = 34\ \mathrm{MWh}, \qquad E^{\mathrm{short}} = E^{\mathrm{tgt}} - E^{\mathrm{del}} = 6\ \mathrm{MWh}. \tag{8.12}$$

Equation (8.12) is the quantity most settlement rules are written against, because energy is what a market can price. It is also the quantity that hides the next two failure modes completely.

8.3.4 Timing

Suppose the grid needs 20 MW from 18:00 to 19:00, and the fleet delivers 10 MW during that hour and 30 MW from 19:00 to 20:00. The shifted energy looks respectable: $10 + 30 = 40$ MWh of deviation against a 20 MWh request. Operationally the dispatch failed. The constraint that mattered held between 18:00 and 19:00, where the fleet supplied half of what was needed, and the 30 MW that arrived afterward was not merely useless but arrived in an hour when the system did not want it. The right amount of energy at the wrong time is not equivalent performance, and a verification procedure that reports only E^{del} cannot see the difference. This is the reason trajectory tracking, equations (8.9) and (8.10), is specified for services with a binding window.

8.3.5 Rebound

The post-event overshoot of Figure 2.3 in Chapter 2 is part of the fleet's effect on the grid and belongs in the accounting. If a thermostat event sheds $E^{\mathrm{shed}} = 20$ MWh and the recovery afterward consumes $E^{\mathrm{reb}} = 12$ MWh above baseline, then

$$E^{\mathrm{net}} = E^{\mathrm{shed}} - E^{\mathrm{reb}} = 20 - 12 = 8\ \mathrm{MWh}. \tag{8.13}$$

Whether 20 or 8 is the right number to report depends entirely on the service. If the utility bought energy reduction over the evening, the answer is 8 MWh and the rebound is a cost. If

the utility bought a reduction of the 18:00–19:00 peak, the answer is that the peak was reduced by the event's average power and the energy moved to 21:00 is an acceptable, possibly desirable, consequence, because 21:00 is not when the system was constrained. The same physical behavior is good performance under one contract and poor performance under another.

8.3.6 The service determines the metric

The preceding five subsections are five metrics, and nothing in the physics selects among them. The selection comes from the service, as Table 8.2 sets out. The practical implication is that a VPP cannot be summarized by a single performance percentage; the percentage is only defined once the product is named. A fleet that is excellent at peak shaving and poor at load following is not a fleet with one performance number, and an aggregator that reports one is reporting the metric that happens to flatter it.

Table 8.2: The service determines the verification metric.

Service	Metric	What counts as failure
Peak shaving	$\max_t P^{\mathrm{act}}(t)$ over the peak window	the peak is unchanged, however much energy was shifted
Energy reduction	$\int \Delta P(t)\,\mathrm{d}t$ over the settlement period, net of rebound, equation (8.13)	the rebound cancels the shed energy
Load following	$\mathrm{RMSE}(P^{\mathrm{tgt}}, P^{\mathrm{del}})$, equation (8.10)	the trajectory is tracked loosely even when the mean is right
Capacity	$\min_t P^{\mathrm{del}}(t)$ over the required interval	one dip below the committed level inside the window
Market product	whatever the product's tariff or market rule defines	any term of that definition, including telemetry granularity

8.4 Attributing underperformance

A shortfall is an observation, not a diagnosis. This section separates the three mechanisms that produce one, follows a single event through the four numbers it generates, and writes the error decomposition that makes the separation precise.

8.4.1 Three explanations for the same shortfall

Suppose the forecast said 50 MW of thermostat flexibility was available, the fleet was dispatched for it, and M&V reports 35 MW. Three mechanisms explain the 15 MW gap, and they call for opposite responses:

- **Forecast error.** Only 35 MW was physically available. The devices did what they could; the flexibility model was optimistic. The fix is in the forecast.
- **Control or communications error.** 50 MW was available, but some commands were never delivered, some devices rejected them, some customers overrode, and the realized dispatch reached only part of the fleet. The fix is in the dispatch path.
- **M&V error.** The fleet delivered close to 50 MW and the baseline underestimated the counterfactual, so the delivery was measured short. The fix is in the baseline.

Observed underperformance is therefore a sum,

$$\text{shortfall} = \text{forecast error} + \text{control and communications error} + \text{M\&V error}, \tag{8.14}$$

and the three terms are separately identifiable only with evidence beyond the aggregate meter. Device telemetry (Section 8.5.5) identifies the second term; a control group (Section 8.5.3) identifies the third.

8.4.2 The four numbers of one event

A single event produces four distinct power figures, and the habit of collapsing them into one is the most common reporting error in VPP operation. Consider a forecast capability of $F = 100\,\text{MW}$. The operator chooses to dispatch 80 MW, holding 20 MW in reserve. Device telemetry indicates that the commanded devices should have produced about 76 MW. Baseline-based M&V estimates 68 MW. The chain is

$$100 \to 80 \to 76 \to 68,$$

and Figure 8.4 labels each step with what it means. Only one of the three gaps is an error at all.

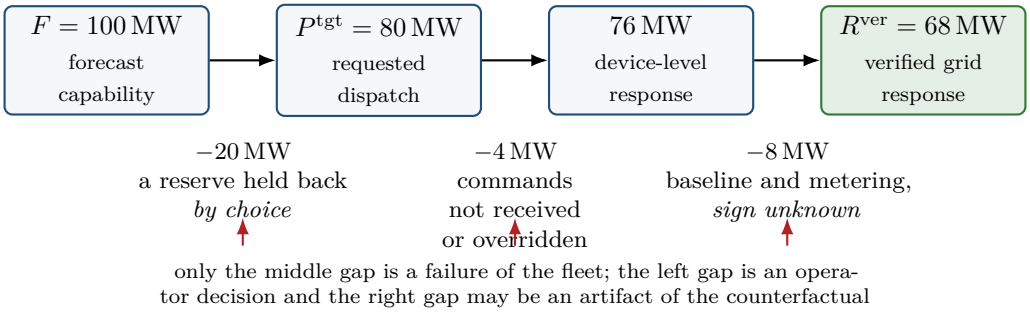


Figure 8.4: The four numbers of one event, $100 \to 80 \to 76 \to 68$ in megawatts, with the three gaps between them. Reporting a single "delivered" figure discards the information needed to decide what, if anything, to change.

The 20 MW from 100 to 80 is not a shortfall; it is the operator's own reserve decision, the coverage logic of Section 4.5.1 in Chapter 4. The 4 MW from 80 to 76 is a genuine control-path loss, and telemetry says where it went. The 8 MW from 76 to 68 is the discrepancy between what the devices report doing and what the baseline says the grid saw; it may be a real loss (device telemetry can overstate, for instance by assuming nominal compressor power), or it may be baseline bias. Which of the two it is cannot be decided from these four numbers alone, which is the point of Section 8.4.3.

8.4.3 The measured response decomposes

Write the verified response as the truth plus two errors,

$$R^{\text{meas}} = R^{\text{true}} + \varepsilon_{\text{base}} + \varepsilon_{\text{meas}}, \tag{8.15}$$

where R^{true} is the actual causal response, $\varepsilon_{\text{base}}$ is the error in the baseline and $\varepsilon_{\text{meas}}$ is metering and telemetry error. Equation (8.15) is the statement that the verified response is a random variable even when the physical response is not. The two error terms behave differently and must be treated separately:

- $\varepsilon_{\text{meas}}$ is usually small, roughly zero-mean, and averages down over devices and over time, because revenue-grade meters are accurate and their errors are nearly independent;

- $\varepsilon_{\text{base}}$ is often large, and above all it is *biased*: a baseline that is systematically low on hot days is low on every hot event, so the error does not average away over events and does not shrink with fleet size.

A bias that survives averaging is the dangerous kind, and Section 8.6.1 shows what it does to the rest of the platform. The 8 MW gap of Figure 8.4 is exactly $\varepsilon_{\text{base}} + \varepsilon_{\text{meas}}$ plus any real loss between the device and the point of common coupling, and separating them requires the experimental designs of Section 8.5.3.

8.4.4 Three numbers to keep distinct

The practice that follows from this section is to carry three numbers rather than one, listed in Table 8.3. They answer three different questions, they are known at three different times, and each feeds a different decision: the forecast capability F_t sets what can be committed, the requested response P_t^{tgt} is the operator's choice, and the verified response R_t^{ver} is what settlement pays for and what the models learn from. Reporting only the third makes a forecasting problem indistinguishable from a communications problem; reporting only the first describes a fleet that was never dispatched.

Table 8.3: The three numbers an event produces, and what each one is for.

Number	Symbol	Question it answers	Known when
Forecast capability	F_t	how much flexibility is believed to be available	before dispatch
Requested response	P_t^{tgt}	how much the operator chose to dispatch	at dispatch
Verified response	R_t^{ver}	how much is inferred to have been delivered	after the event, once a baseline exists

8.5 Probabilistic and experimental verification

If the baseline is inferred then the response is uncertain, and the uncertainty can be reported rather than hidden; if some devices are left undispatched then the counterfactual can be observed rather than modeled. This section develops both ideas and then the telemetry that makes the causal chain itself checkable.

8.5.1 Verification as inference

Collect the structure of the problem. The platform observes P^{act} and wants ΔP, which depends on the unobserved P^{base}. In the vocabulary of Chapter 3, P^{base} is a *latent* variable: it appears in the model, it is never measured, and it must be inferred from whatever is measured. M&V is therefore an instance of counterfactual inference, not an accounting step.

That framing changes the deliverable. A statistical baseline returns a point estimate and the response inherits it as a single number. A probabilistic baseline returns a distribution $p\Big(P^{\text{base}} \,\Big|\, \text{data}\Big)$, and equation (8.1) propagates it into a distribution over the response,

$$p(\Delta P \,|\, \text{data}) = p\Big(P^{\text{base}} = P^{\text{act}} + \Delta P \,\Big|\, \text{data}\Big)\,, \tag{8.16}$$

since P^{act} is known. The reported performance becomes an object with a width, and the width is informative: a wide posterior on a day with unusual weather is itself the statement that the event cannot be verified tightly.

8.5.2 Probabilistic M&V in one example

Take the feeder of Section 8.1.2 and suppose the baseline model reports

$$P^{\text{base}} \sim \mathcal{N}(1\,000,\ 40^2)\ \text{kW} \tag{8.17}$$

for the 18:00 hour, with the standard deviation coming from the regression's own predictive variance. The meter reads 850 kW. Then by equation (8.16),

$$\Delta P \sim \mathcal{N}(150,\ 40^2)\ \text{kW}, \tag{8.18}$$

so the expected reduction is 150 kW with about 40 kW of uncertainty contributed by the baseline alone. A contract term such as "deliver at least 100 kW" then becomes a probability rather than a verdict,

$$\Pr(\Delta P \geq 100\,\text{kW}) = 1 - \Phi\left(\frac{100 - 150}{40}\right) = \Phi(1.25) \approx 0.894, \tag{8.19}$$

where Φ is the standard normal distribution function. Figure 8.5 draws the density and the shaded region whose area is that number. The statement "we delivered 150 kW" and the statement "we met the 100 kW threshold with probability 0.89" rest on the same arithmetic; the second is the one an operator can act on, because it says how often a fleet verified this way will be found short.

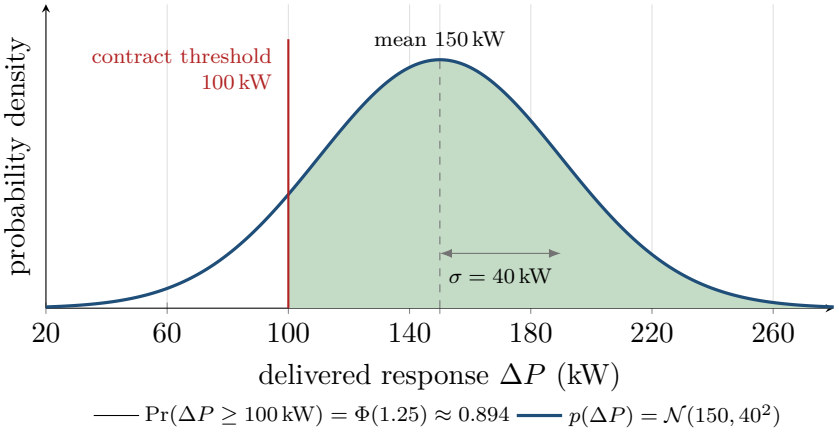


Figure 8.5: Probabilistic M&V for the event of Section 8.5.2. The curve is the density $p(\Delta P) = \exp[-(\Delta P - 150)^2/(2 \cdot 40^2)]/(40\sqrt{2\pi})$ implied by $P^{\text{base}} \sim \mathcal{N}(1\,000, 40^2)$ kW and a metered 850 kW. The shaded area right of the 100 kW threshold is $\Phi(1.25) \approx 0.894$: about an eleven percent chance that this event is verified as short even though the expected delivery exceeds the threshold by 50 kW.

8.5.3 Control groups

The posterior of Section 8.5.2 still rests on a model of the counterfactual. A control group replaces the model with a measurement. Enroll 100 000 thermostats, dispatch 90 000 of them and leave 10 000 undispatched during the event. The undispatched group experiences the same weather, the same hour and the same customer behavior as the dispatched group, but receives no command, so its metered load *is* a sample from the counterfactual. Scaling its mean per-device power up to the treated population gives a baseline that no historical day and no regression can match, because it is contemporaneous.

The method stands or falls on comparability. If the control group sits in a cooler part of the service territory, or contains older houses, or was assembled from the customers who had previously opted out most often, then its mean power differs from the treated group's counterfactual for reasons unrelated to the dispatch, and the difference enters $\varepsilon_{\text{base}}$ of equation (8.15) as a bias rather than as noise. A control group chosen by convenience can be worse than a regression, because its bias is invisible: the group looks like direct evidence.

8.5.4 Randomization and the average treatment effect

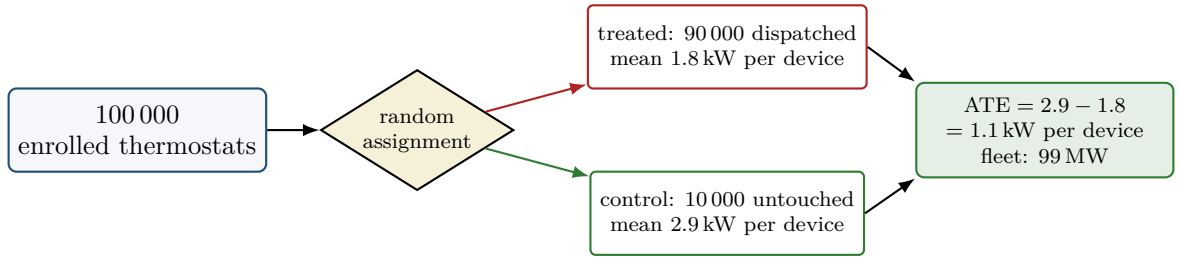
Random assignment removes the comparability problem by construction. If devices are allocated to the two groups at random then the groups are statistically identical in every respect, observed or not, so

$$\mathbb{E}\big[P^{\text{treat}}_{\text{no control}}\big] \approx \mathbb{E}\big[P^{\text{ctl}}\big], \tag{8.20}$$

and the causal effect per device is the difference of two observed means,

$$\text{ATE} = \mathbb{E}\big[P^{\text{ctl}}\big] - \mathbb{E}\big[P^{\text{treat}}\big]. \tag{8.21}$$

Figure 8.6 shows the design and the arithmetic with the mean powers of a worked case: 2.9 kW per control device, 1.8 kW per treated device, an ATE of 1.1 kW and a fleet response of $90\,000 \times 1.1\,\text{kW} = 99\,\text{MW}$. Equation (8.21) is a stronger causal estimate than any historical baseline, and its uncertainty is the ordinary standard error of a difference of means, which shrinks with the size of the control group.



Randomization makes the control group's metered load a sample of the treated group's counterfactual, equation (8.20); assignment by geography or by customer type does not.

Figure 8.6: A randomized control design for verification. The enrolled population is split at random; the control group is metered but never commanded, and the difference of per-device mean powers is the average treatment effect of equation (8.21). Mean powers are the ones used in Section 8.5.4.

Operational constraints limit how far this can be taken. Every device held in the control group is capacity not delivered, which is a direct cost when the commitment is tight; customers may object to being enrolled in a program that does not dispatch them; and a local constraint on one feeder may require dispatching every device behind it, which destroys randomization exactly where the verification is most needed. In practice aggregators randomize a small fraction, rotate which devices are held back across events in the manner of Figure 5.11 in Chapter 5, and use the result to calibrate the regression baselines that carry the rest of the year.

8.5.5 Device telemetry

A thermostat that reports compressor state, indoor temperature, setpoint and command receipt turns one aggregate number into a chain of observations. The platform can then distinguish three events that a meter cannot: a command that was never sent, a command that was sent

but never received, and a command that was received but produced no change in the HVAC state. Each points at a different repair, in the dispatch path, in the communications path or in the device and its building. Telemetry is thus the instrument that identifies the second term of equation (8.14), and it also sharpens the first: the same measurements are the inputs to the flexibility forecast of Chapter 5. Richer telemetry improves forecasting and verification together, which is why the two are usually procured as one capability.

8.5.6 The full causal chain

What verification is ultimately trying to establish is that a chain of five links held:

command issued → command received → device state changed → device power changed → aggregate grid effect.

Figure 8.7 draws the chain with the observation available at each link and the failure that breaks it. Checking the chain is more informative than checking the last link alone, because the aggregate meter confounds all four upstream failures into one number, while the links are separately observable with the telemetry of Section 8.5.5. A fleet whose commands all arrive and whose compressors all stop, but whose metered reduction is small, is telling the operator that the baseline is wrong; a fleet whose compressors never stopped is telling the operator something else entirely.

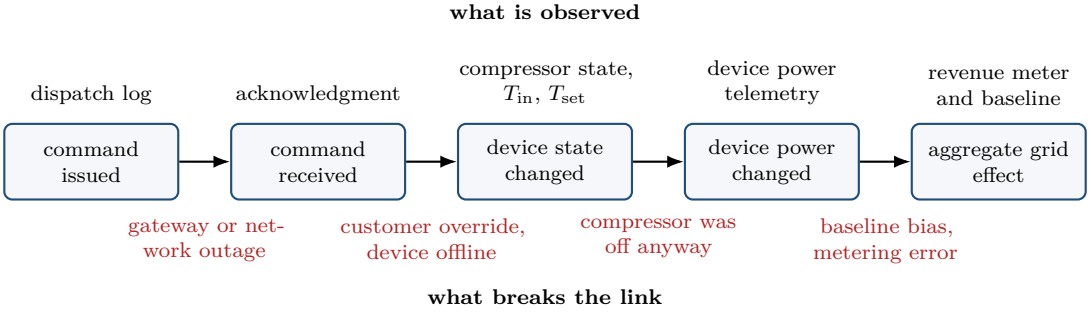


Figure 8.7: The causal chain verification must establish. Above each stage is the observation that confirms it; below each arrow is the failure that breaks it. The aggregate meter sees only the last stage, so the four upstream failures are confounded unless device telemetry is available.

8.6 Verification inside the learning loop

Verification is not the last step of an event; it is an input to the next one. This section shows how a biased baseline propagates into the forecast, places M&V in the platform's operating loop, and closes with a worked shortfall and the questions it raises.

8.6.1 Baseline bias contaminates the forecast

The DERMS updates its flexibility model from measured performance: after each event it compares what it predicted with what it verified and adjusts, so that tomorrow's forecast $\hat{P}_{t+1}$ is fitted in part to today's R_t^{ver}. Now suppose the baseline is biased low by 10 MW. The true response was 50 MW; M&V reports 40 MW; the model is told that a fleet in that state and that weather delivers 40 MW. The next forecast is low by the same amount, the optimizer commits less, the fleet is dispatched less aggressively, and the gap is never discovered, because the same biased baseline verifies the next event too.

The failure is self-confirming, which is what makes it serious. A zero-mean baseline error adds noise to the model fit and averages away over events; a bias does not, and it propagates in a specific direction. Underestimating capability leaves revenue and grid value unrealized and makes the fleet look worse than it is; overestimating it causes overcommitment and the penalties of Section 4.5.4 in Chapter 4. Either way the forecasting model has learned a property of the baseline rather than a property of the fleet.

Principle 8.2. *A biased baseline is not only a reporting error. It is a systematic error in the training data of the flexibility forecast, so bad verification degrades future forecasting and dispatch.*

8.6.2 The loop

Figure 8.8 places verification in the cycle it belongs to: forecast, optimize, dispatch, measure, infer the counterfactual, verify, update the models, forecast again. Figure 3.6 of Chapter 3 drew the fast, within-event version of this loop, in which a tracking error triggers a re-optimization minutes later. Figure 8.8 is the slow version, which closes over events rather than over minutes and carries model parameters rather than control actions. The two run at the same time and share the same measurement. The contamination path of Section 8.6.1 is the red edge: it is the only edge in the diagram along which an error in a counterfactual estimate reaches a future dispatch decision.

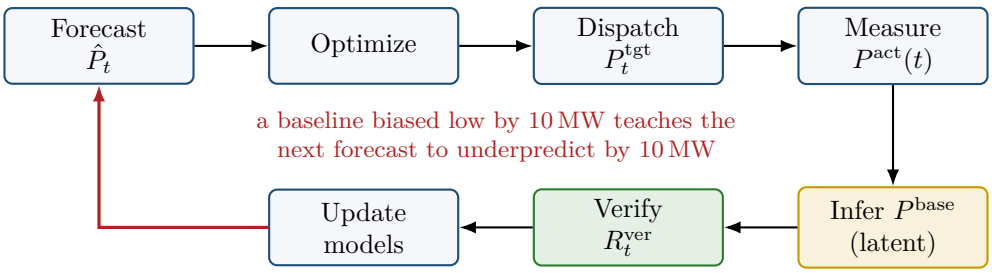


Figure 8.8: Verification inside the learning loop. Dispatch decisions run left to right along the top; the measurement, the inferred counterfactual and the verified response run right to left along the bottom; the red edge carries the updated models back into the next forecast, and carries any baseline bias with them.

8.6.3 Where the loop sits in a platform

In an operating aggregator the loop of Figure 8.8 is the product. The platform forecasts system-wide load and fleet capacity for the coming hours, optimizes dispatch schedules against the grid need and the device constraints, dispatches the DERs through the device integrations of Chapter 3, and then uses near-real-time performance information to adjust both the remainder of the current event and the models behind the next one. M&V is the function that lets the system answer two questions it cannot answer from the dispatch log alone: *did the portfolio behave as expected?* and therefore *should the next forecast or the next dispatch decision change?* Without it the platform is an open loop that issues commands and never learns whether they worked, which is the open-loop pipeline of Figure 2.6(a) in Chapter 2.

8.6.4 Worked example: a ten-megawatt shortfall

An event runs as follows. The flexibility forecast is $F = 120\,\mathrm{MW}$. The utility requests 100 MW. The DERMS dispatches enough devices for an expected response of 105 MW, taking a small margin against override and communications loss. During the event the metered fleet load is $P^{\mathrm{act}} = 900\,\mathrm{MW}$ and the baseline model estimates $P^{\mathrm{base}} = 990\,\mathrm{MW}$, so

$$R^{\mathrm{ver}} = P^{\mathrm{base}} - P^{\mathrm{act}} = 990 - 900 = 90\,\mathrm{MW}, \qquad \text{shortfall} = 100 - 90 = 10\,\mathrm{MW}. \tag{8.22}$$

The tempting conclusion is that the fleet underperformed by 10 MW and that the flexibility forecast should be reduced. Four questions must be answered first, and they are the four mechanisms of this chapter:

1. **Was the baseline biased?** A 990 MW counterfactual against a 900 MW meter is a one percent claim about a large number. If the baseline is low by 1 %, that is 10 MW and the entire shortfall is $\varepsilon_{\mathrm{base}}$ in equation (8.15). The randomized control group of Section 8.5.4 is what answers this; the regression's own predictive standard deviation, as in equation (8.17), says whether 10 MW is even distinguishable from zero.
2. **Did the dispatched devices respond?** Telemetry gives the fraction of commands acknowledged and the fraction of compressors or chargers that actually changed state, which is the second and third link of Figure 8.7.
3. **Were communications successful?** A gateway outage in one region produces a device-level response far below the dispatched amount while every surviving device performs normally, and it is visible in the acknowledgment rate rather than in the meter.
4. **Was the flexibility forecast too optimistic?** Only if the first three are clean does the 120 MW forecast become the suspect, and then the correction belongs in the flexibility model and not in the dispatch logic.

The four questions partition the shortfall, and the order matters: the cheapest and most common explanation is the baseline, and it is also the one that, left uninvestigated, corrupts the forecast through the red edge of Figure 8.8.

8.7 What to retain

Definition 8.1 (Measurement and verification)**.** DER response is not directly observable, because the no-control counterfactual is missing. Measurement and verification is therefore partly a forecasting and inference problem: the operational chain is forecast, dispatch, measure, infer the counterfactual, verify, learn, and the inferred step can dominate the reported result.

The chain is considerably richer than "did the meter go down?". It has a latent variable in the middle of it, three distinguishable error sources, a metric that depends on the service rather than on the physics, and a feedback edge along which a bias in the latent variable reaches tomorrow's dispatch. Table 8.4 lists the chapter's quantities with the subsection that defines each.

8.8 Exercises

Exercise 8.1. The feeder of Sections 8.1.2 and 8.2.2 meters $P^{\mathrm{act}}(18{:}00) = 850\,\mathrm{kW}$. Compute $\Delta P(18{:}00)$ under each of the three baselines in the chapter: the historical average of ten 29°C weekdays (820 kW), the natural decline of model B (920 kW), and the weather-adjusted value at 37°C (1 000 kW). Then suppose the pre-event meter at 16:00 runs 40 kW above the weather-adjusted model and apply equation (8.8). Which number would you report, and what evidence would change your answer?

Table 8.4: The quantities of baselines and verification.

Quantity	Definition	Section
$\Delta P(t)$	$P^{\text{base}}(t) - P^{\text{act}}(t)$	8.1.1
$P^{\text{base}}(t)$	$\hat{P}^{\text{nc}}(t)$, a forecast of the no-control load	8.1.4
historical-day baseline	$\frac{1}{N}\sum_d P_d(t)$	8.2.1
weather-adjusted baseline	$\beta_0 + \beta_1 T + \beta_2 T^2 + \dots$ at event-day weather	8.2.2
pre-event adjustment	$P^{\text{mod}}(t) + \delta$, δ from a pre-event window	8.2.3
e_t, RMSE	$P^{\text{tgt}} - P^{\text{del}}$ and its root mean square	8.3.2
E^{del}, E^{short}	$\int P^{\text{del}}\,\mathrm{d}t$ and $E^{\text{tgt}} - E^{\text{del}}$	8.3.3
E^{net}	$E^{\text{shed}} - E^{\text{reb}}$	8.3.5
R^{meas}	$R^{\text{true}} + \varepsilon_{\text{base}} + \varepsilon_{\text{meas}}$	8.4.3
F_t, P_t^{tgt}, R_t^{ver}	forecast, requested and verified response	8.4.4
$p(\Delta P \mid \text{data})$	posterior over the delivered response	8.5.1
ATE	$\mathbb{E}[P^{\text{ctl}}] - \mathbb{E}[P^{\text{treat}}]$	8.5.4

Exercise 8.2. Take $P^{\text{base}} \sim \mathcal{N}(1\,000, 40^2)$ kW as in equation (8.17), and now add independent metering error of standard deviation 10 kW on $P^{\text{act}} = 850$ kW. Write the distribution of ΔP and recompute $\Pr(\Delta P \geq 100\,\text{kW})$. What baseline standard deviation would be needed to raise that probability to 0.99, holding the metering error fixed? Comment on which of the two error terms of equation (8.15) is worth reducing.

Exercise 8.3. For the chain of Figure 8.4, state which of the three gaps belongs to each term of equation (8.14). A second event on the same fleet produces $100 \to 90 \to 88 \to 95$ megawatts. The verified response now exceeds the device-level response. Give two explanations, and say which measurement would distinguish them.

Exercise 8.4. In the design of Figure 8.6, the control group's mean power is 2.9 kW and the treated group's is 1.8 kW. Compute the ATE and the fleet response. Now suppose the control group's average outdoor temperature during the event was 1°C higher than the treated group's, and that per-device load rises by 0.09 kW per °C. By how much is the ATE biased, and by how much is the fleet response overstated? Compare that bias with the 8 MW gap of Figure 8.4.

8.9 Connection to the TERRA framework

Verification is the part of the lesson sequence that is most naturally expressed as inference on a graph, because its central quantity is a flow that was never measured.

- **The baseline is a latent edge flow.** $P^{\text{base}}(t)$ in equation (8.1) is the flow on the grid edge at the point of common coupling in the model run with the control input set to zero. It is not a new kind of object: it is the same edge flow the forward simulation produces, evaluated under a different input. Computing it is a simulation; computing its uncertainty is propagation of the posterior over device states and weather through that simulation.

- **The delivered response is a derived quantity with a posterior.** ΔP is a difference between a measured flow and a simulated one, so equation (8.16) is a query on the model rather than a reading from an instrument, and equation (8.19) is a tail probability of that query. This is the same machinery as the chance constraints of Chapters 2 and 4, applied after the fact instead of before it.

- **Pre-event adjustment is conditioning on a contaminated node.** Equation (8.8) conditions the counterfactual on measurements taken in a window where the control input was nonzero, which in graph terms means conditioning a zero-input simulation on data generated with input. A model that carries the control input explicitly rejects the contaminated window automatically, which is why the pre-cooling bias of Figure 8.3 is a modeling error rather than a statistical one.

- **The error decomposition is a measurement model.** Equation (8.15) is $y = h(x) + v$ with the baseline term written out: $\varepsilon_{\text{meas}}$ is the noise v of Chapter 3, and $\varepsilon_{\text{base}}$ is error in h itself. Separating them is exactly the distinction between noise and model misspecification that posterior inference makes, and a framework that holds both can report which one dominates.

- **Telemetry and the causal chain are extra measurement channels.** Each stage of Figure 8.7 is a component state or a component flow, and each observation above it is a measurement of that state. Adding those channels makes the upstream links identifiable from data, which the single aggregate flow at the boundary cannot do. The design question "what telemetry is worth buying?" is the identifiability question "which states does the current measurement set leave unresolved?".

- **The learning loop is parameter inference.** The red edge of Figure 8.8 updates the parameters of the flexibility model from verified responses. A biased $\varepsilon_{\text{base}}$ is a biased likelihood, so the parameter posterior converges to the wrong value with increasing confidence as events accumulate. Correct verification is therefore a precondition for the model to learn anything true about the fleet.

In summary, measurement and verification is posterior inference of an unmeasured boundary flow, together with a decision rule that compares that inference against a service definition. Treated as accounting it is a source of silent bias; treated as inference it supplies both the delivered response and the confidence with which it is known. The next chapter returns to the fleet itself and asks how many devices, and of what kinds, must be aggregated before these quantities become stable enough to commit.

Chapter 9

Fleet aggregation: how thousands of DERs become one manageable resource

9.1 The aggregation problem

Chapters 5 to 7 built one device at a time: a thermostat with a comfort band, a battery with a state of charge, an EV with a deadline. Chapter 4 asserted that a fleet of such devices can be offered to the grid as a single resource. This chapter is the bridge between the two statements. Its question is

How can 100 000 or more DERs be represented and controlled without treating each one as a separate optimization problem?

Jargon: aggregation, model reduction, histogram state, virtual battery, Minkowski sum

Aggregation is the construction of a small fleet-level description from many device-level states. *Model reduction* is the general name for replacing a high-dimensional state by a few quantities that suffice for a decision; aggregation is model reduction applied to a fleet. A *histogram state* replaces the list of individual states by the counts of devices in a set of bins. A *virtual battery* is an aggregate model that represents a whole fleet of flexible loads by power limits, an energy content and energy limits, as if it were one storage device. The *Minkowski sum* of two sets is the set of all sums of one element from each; the fleet's feasible set is the Minkowski sum of the devices' feasible sets.

Key idea

Fleet aggregation is model reduction for control: from the device states $x_1, \dots, x_N$ construct a compressed state $z = \phi(x)$ that keeps what determines dispatch and discards the rest.

9.1.1 The brute-force view

Take $N = 100\,000$ enrolled devices. Device i has a state x_i and a control u_i, with the meanings fixed in Chapter 1. The fleet state and fleet control are the stacked vectors

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}, \qquad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}, \tag{9.1}$$

which is equation (4.13) of Chapter 4 with the control side added. The dimensions are large but not absurd: a thermostat carries an indoor temperature and a compressor state, a battery a state of charge, an EV a state of charge and a plug-in indicator, so x has of order $2N = 200\,000$ components.

In principle a single optimizer could choose all 200 000 numbers every few minutes. In practice this is awkward in three separate ways. It is computationally heavy, because the optimization is re-solved on a control cycle of minutes. It is operationally heavy, because the optimizer's inputs would have to include fresh telemetry from every device, which Section 9.5.1 shows does not exist. And it is unnecessary, because the grid does not ask for 100 000 separate quantities; it asks for a few megawatt figures with a window and a location, as Chapter 4 established. The purpose of aggregation is to replace the device-level detail with a smaller fleet-level representation that is still sufficient for the decision being made.

9.1.2 The naive aggregate

The obvious candidate for that representation is the fleet's present power,

$$P_{\text{fleet}}(t) = \sum_{i=1}^{N} P_i(t). \tag{9.2}$$

It is useful: it is what the grid meters, and it is the quantity any tracking objective is written against. It is also insufficient as a fleet *state*, because two fleets with the same $P_{\text{fleet}}(t)$ can have completely different futures.

Consider two fleets of 10 000 residential batteries, each unit rated 5 kW with $E^{\max} = 13.5$ kWh and $\text{SOC}^{\min} = 0.10$. In fleet A most units sit at SOC ≈ 0.90; in fleet B most sit at SOC ≈ 0.20. Both fleets are idle, so

$$P^{\text{A}}_{\text{fleet}}(t) = P^{\text{B}}_{\text{fleet}}(t) = 0, \tag{9.3}$$

and a representation consisting of P_{fleet} alone cannot tell them apart. Their discharge capability is not comparable. The energy available above the reserve is $(\text{SOC} - \text{SOC}^{\min})\, E^{\max}$ per unit, so

$$E^{\text{avail}\,\text{A}} = 10\,000 \times (0.90 - 0.10) \times 13.5\,\text{kWh} = 108\ \text{MWh}, \qquad E^{\text{avail}\,\text{B}} = 13.5\ \text{MWh}.$$

At a 50 MW discharge both fleets can start, but fleet A sustains it for about 2.2 hours and fleet B for about 16 minutes. The present aggregate power is identical; the deliverable product, in the sense of $C(\tau)$ of Chapter 4 and Figure 4.3, is not. A sufficient fleet state must therefore carry information about the stored quantities, not only about the present flows.

9.1.3 What a fleet representation must preserve

The criterion is operational rather than aesthetic: a fleet representation is good enough when it predicts what the fleet can do next. The quantities that determine that are

$$z_t = \{P^{\text{avail}},\ E^{\text{avail}},\ \text{SOC and thermal-state distributions},\ \text{availability},\ \text{location}\}, \tag{9.4}$$

and the aggregation layer is the map

$$z_t = \phi(x_t), \tag{9.5}$$

from the detailed fleet state x_t to the compressed representation z_t. This is the map introduced as equation (4.14) in Chapter 4 and drawn there as Figure 4.6; the present chapter is about how ϕ is actually built, what it is allowed to throw away, and what happens when it throws away too much. Two features of ϕ are worth noting in advance. It is not a linear projection: computing E^{avail} or a sustainable duration requires simulating devices forward against their own limits. And it is not unique: different decisions tolerate different ϕ, which is the subject of Section 9.5.6.

9.2 Representations of a fleet

This section works through the candidate forms of ϕ in order of increasing fidelity, from a single average to a set of admissible power trajectories, and states what each one keeps and what it discards.

9.2.1 Averages

The cheapest compression replaces every device state by the fleet mean. For a battery fleet that means keeping $\overline{\mathrm{SOC}}$ alone, and it fails as soon as the limits bind unevenly. Chapter 6 gave the example: the four-unit fleets $\{50, 50, 50, 50\}\,\%$ and $\{95, 95, 5, 5\}\,\%$ have the same average state of charge and half the capability, because with limits $[0.10, 0.95]$ the two full units of the second fleet cannot absorb and the two nearly empty ones cannot deliver. Figure 6.6 of Chapter 6 draws the two distributions and the resulting $F^{\uparrow}$ and $F^{\downarrow}$; it applies here without change and is not repeated.

The reason the mean fails is worth stating precisely, because it also says when the mean is adequate. Aggregate capability is a sum of quantities of the form

$$F_i = \min\{P_i^{\max},\ \text{a function of } x_i\}, \tag{9.6}$$

that is, a sum of *clipped* functions of the device states. The mean of a clipped function is not the clipped function of the mean, and the discrepancy grows with the spread of the distribution and with the fraction of devices at a limit. A fleet whose states are tightly packed in the interior of their limits is described adequately by its mean; a fleet with mass at the limits is not. Since a fleet under active dispatch is continually pushed toward its limits, the second case is the normal one.

Principle 9.1. *Aggregation must usually preserve a distribution, not a mean, because device limits bind device by device and the limits are what make flexibility finite.*

9.2.2 Histograms as fleet state

The standard way to keep a distribution cheaply is to bin it. Instead of SOC_i for every unit, the aggregation layer keeps the count n_k of devices in state bin k. Figure 9.1 shows such a histogram for a fleet of 20 000 residential batteries: the horizontal axis is state of charge in percent, divided into six bins, and the height of each bar is the number of units in that bin. The first two bins are ten percentage points wide and the remaining four are twenty, so the bars are counts per bin and not densities, and the dashed line marks $\mathrm{SOC}^{\min} = 10\,\%$.

The compression is severe and almost free. Six numbers replace 20 000, a reduction by a factor of about 3 300, and the quantities a dispatcher needs can be read off them. Discharge power is available from every unit above the reserve,

$$P^{\mathrm{avail}} = (20\,000 - 500) \times 5\,\mathrm{kW} = 97.5\,\mathrm{MW},$$

and summing $(\mathrm{SOC}_k - \mathrm{SOC}^{\min})\,E^{\max}\,n_k$ over the bins at their midpoints gives

$$E^{\mathrm{avail}} = 0.81 + 10.8 + 37.8 + 40.5 + 24.84 = 114.75\ \mathrm{MWh}.$$

The mean-based representation of Section 9.2.1 would report $\overline{\mathrm{SOC}} = 52.4\,\%$ and hence 100 MW and 114.4 MWh. The energy figures nearly agree, because only 500 of 20 000 units are clipped; the power figures differ by 2.5 MW, because clipping is a step and every clipped unit loses its whole contribution. That contrast is the general rule of equation (9.6) in miniature: the mean

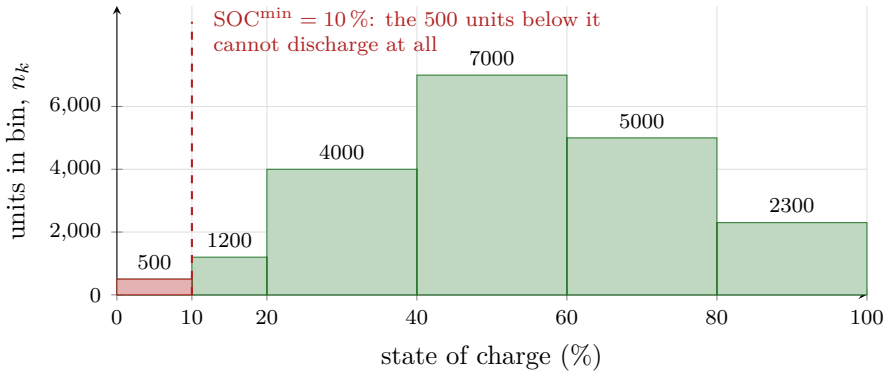


Figure 9.1: Histogram state of a fleet of 20 000 residential batteries. The six counts replace 20 000 individual states. Bars are counts per bin; the first two bins are 10 percentage points wide and the last four are 20. The 500 units in the lowest bin are below the reserve limit and contribute no discharge capability. Counts are illustrative.

survives for quantities that are nearly linear over the occupied range of states and fails for quantities that saturate.

The same construction applies to the other device classes. Thermostats are binned by $T_{\text{in}} - T_{\text{set}}$, the position inside the comfort band, which is the thermal state of charge SOC^{th} of Chapter 5; a bin near the upper limit holds units with almost no remaining headroom. EVs need a joint binning, because their flexibility depends on three quantities at once: departure deadline, energy still required, and charging power. A fleet of EVs is therefore represented by counts in a three-dimensional grid of bins, or equivalently by one two-dimensional histogram per deadline cohort, which is the object the urgency rule of Chapter 7 sorts.

9.2.3 Clustering

Binning is one way to group devices; clustering is the general one. Devices are grouped by similarity in whatever attributes drive their response. For thermostats those attributes are climate zone, building type, HVAC size, current temperature state and historically observed response. Instead of 50 000 thermostats the fleet is represented by perhaps $K = 50$ clusters, where cluster k holds N_k devices and is summarized by a representative state $\bar{x}_k$. The fleet response to a command is then approximated by

$$P_{\text{fleet}} \approx \sum_{k=1}^{K} N_k \, P_k(\bar{x}_k, u_k), \tag{9.7}$$

one simulated device per cluster, scaled by the cluster population.

The dimensionality gain. The optimizer now chooses $u_1^{\text{cl}}, \ldots, u_K^{\text{cl}}$ rather than $u_1, \ldots, u_N$. With $K = 50$ and $N = 100\,000$ the decision vector shrinks by a factor of 2 000, and the simulation underlying each evaluation of the objective shrinks by the same factor. A dispatch problem that was intractable on a five-minute cycle becomes a small optimization. This is why clustering, not device-level optimization, is what production platforms actually solve.

9.2.4 What clustering loses

Replacing a group by a representative discards the spread inside the group, and the spread is sometimes exactly what determines feasibility. Take two thermostats assigned to the same cluster, one at $T_{\text{in}} = 21°\text{C}$ and one at $T_{\text{in}} = 24.5°\text{C}$, both with an upper comfort limit $T^{\max} = 25°\text{C}$ and

a building thermal capacitance $C = 2\,\mathrm{kWh/°C}$. Their remaining headroom differs by a factor of eight. Replacing both by the cluster mean $\bar{T} = 22.75°\mathrm{C}$ overestimates the flexibility of the second and underestimates that of the first. Table 9.1 gives the three columns and the fleet total: the headroom in degrees, the thermal energy that headroom represents, and the duration for which each unit can sustain a displaced cooling load of 2 kW.

Table 9.1: Two thermostats in one cluster, with $T^{\max} = 25°\mathrm{C}$ and $C = 2\,\mathrm{kWh/°C}$. The cluster mean reproduces the total energy exactly and the sustainable duration not at all.

Quantity	Device 1	Device 2	Cluster model (two copies of the mean)
T_{in} (°C)	21.0	24.5	22.75 each
Headroom $T^{\max} - T_{\mathrm{in}}$ (°C)	4.0	0.5	2.25 each
Thermal energy $C(T^{\max} - T_{\mathrm{in}})$ (kWh)	8.0	1.0	4.5 each, total 9.0
Duration at 2 kW (h)	4.0	0.5	2.25 each

The totals agree: $8.0 + 1.0 = 9.0\,\mathrm{kWh}$ is reproduced exactly by $2 \times 4.5\,\mathrm{kWh}$. The durations do not. Asked for a two-hour event, the cluster model answers that both devices can serve it, whereas device 2 reaches its comfort limit after thirty minutes and drops out. The error is not in the energy accounting but in the *binding constraint*, and a dispatch built on the cluster model will discover it as a mid-event shortfall of the kind Chapter 5 described as flexibility decay (Figure 1.3 of Chapter 1). Aggregation always trades computational simplicity against state fidelity; the trade is acceptable only when the discarded spread does not change which constraint binds first.

9.2.5 Aggregate feasible sets and trajectory sets

A different family of representations describes not the fleet's state but the set of things it can do. The simplest version is an instantaneous feasible set,

$$P_{\mathrm{fleet}}(t) \in \mathcal{F}_t, \qquad \text{for example} \quad -50\,\mathrm{MW} \le P_{\mathrm{fleet}}(t) \le 80\,\mathrm{MW}, \tag{9.8}$$

with the negative end representing added load and the positive end reduction, as in the sign convention of Chapter 2. This is compact and it is what a market bid often looks like, but it is only a statement about one instant. It permits the fleet to sit at 80 MW forever, which no energy-limited fleet can do.

The object that matches the decision is a set of *trajectories*,

$$\big(P_t, P_{t+1}, \ldots, P_{t+H}\big) \in \mathcal{F}, \tag{9.9}$$

the power paths the fleet can actually sustain over the horizon H. Figure 9.2 contrasts the two. The horizontal axis is elapsed time since the start of the service and the vertical axis is fleet power in MW; the gray band is the instantaneous set of (9.8), the blue curve is the largest constant power the fleet can hold for a duration τ, and the two colored paths are candidate trajectories. The flat red path lies inside the gray band at every instant and is nevertheless infeasible: it exhausts the fleet's 120 MWh after 1.5 hours. The green step path respects the same energy budget and is admissible. The instantaneous set and the trajectory set differ by exactly the information that the fleet has a finite store.

9.2.6 The virtual battery

The trajectory set of (9.9) is a set in $\mathbb{R}^{H+1}$ and is not something to hand to a market system. The usual compact parameterization of it is the *virtual battery*: the whole heterogeneous fleet is

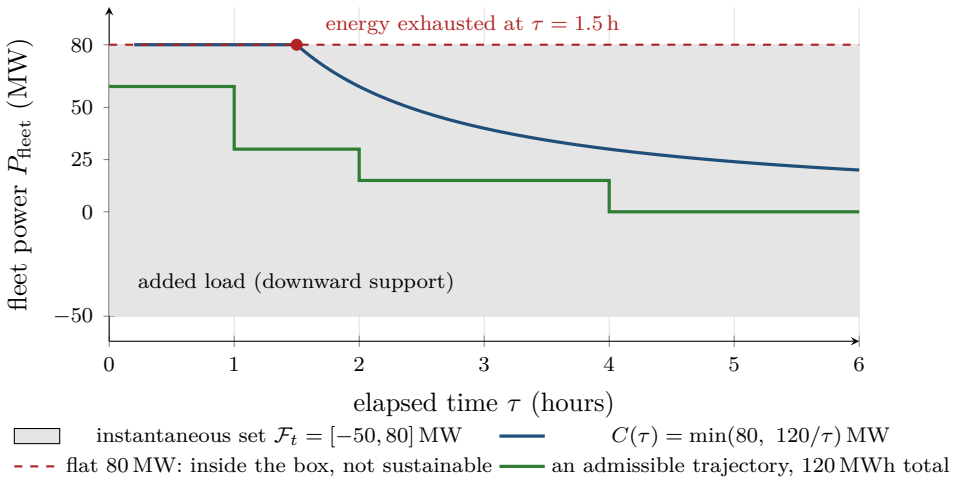


Figure 9.2: The instantaneous feasible set and the trajectory set for a fleet with 80 MW of reduction capability and 120 MWh of usable energy. The blue curve is the largest constant power sustainable for duration τ, $C(\tau) = \min(80, 120/\tau)$ MW, the fleet-level form of Figure 4.3 of Chapter 4. Values are illustrative.

modeled as one storage device,

$$P^{\min} \le P_t \le P^{\max}, \qquad E_{t+1} = E_t + P_t\,\Delta t, \qquad E^{\min} \le E_t \le E^{\max}. \tag{9.10}$$

Four numbers and an initial energy content describe the fleet, and the resulting feasible set is exactly the trajectory set drawn in Figure 9.2. The interpretation of E_t depends on the class. For a thermostat fleet it is aggregate thermal flexibility, the summed distance of indoor temperatures from their comfort limits weighted by building capacitance, as in Table 9.1. For an EV fleet it is aggregate deferred charging: the energy owed to drivers that has not yet been delivered. For a battery fleet it is the literal stored energy. The abstraction is genuinely useful, because every optimizer already knows how to handle one storage device.

Why the parameterization can fail. The virtual battery asserts that the energy in E_t is fungible: any part of it can be delivered at any time within the power limits. Real fleets violate this. Suppose the aggregation layer reports $E^{\text{avail}} = 100$ MWh at 17:00, and that half of it is the deferred charging of EVs whose drivers depart within thirty minutes. Those 50 MWh must be delivered almost immediately or not at all; they are not available for a 20:00 service, and committing them to one produces either a shortfall or a violated customer constraint. The aggregate model, read literally, implies that all 100 MWh are freely schedulable, and that is a statement the underlying devices do not support.

The repair is not to abandon the virtual battery but to use several of them. Partitioning the fleet by the constraint that actually binds, one virtual battery per deadline cohort or per comfort-headroom band, keeps the convenient form while restoring the constraints that matter. This is the operative question for every choice of ϕ:

what information can be discarded without changing the operational answer?

Table 9.2 collects the representations of this section with their dimension for the 20 000-unit fleet of Figure 9.1, what each preserves and what each discards.

9.3 Diversity and common-mode uncertainty

Aggregation is attractive partly because devices differ from one another. This section separates the part of fleet uncertainty that heterogeneity removes from the part it cannot touch.

Table 9.2: Representations of a fleet, in order of increasing fidelity, for the 20 000-unit battery fleet of Figure 9.1.

Representation	Dimension	Preserves	Discards
$P_{\text{fleet}}(t)$	1	the metered flow	everything about the future
Mean state	1–2	the center of the distribution	the clipping at device limits
Histogram	6–20	the shape of the state distribution	which device is where, and correlations between attributes
Clusters	K states	attribute structure and location	the spread inside each cluster
Instantaneous set	2	present power limits	the energy limit
Virtual battery	4–5	power and energy limits jointly	the fungibility assumption: deadlines inside E_t
Trajectory set	$\mathbb{R}^{H+1}$	everything the optimizer needs	nothing, but it is not compact

9.3.1 Diversity as a resource

In a fleet of 10 000 thermostats, at any instant some compressors are on and some are off, some units sit close to their comfort limit and some have hours of headroom. Chapter 4 showed in Section 4.2.3 that this makes the aggregate smooth even though every device is a two-state switch. Diversity does more than smooth: it is itself the control resource. A fleet of 100 000 identical devices in identical states offers one decision, since every unit must be commanded alike; a diverse fleet offers the choice of *which* units to use, and a well-chosen subset satisfies a request at far lower customer cost. The rotation scheme of Chapter 5, Figure 5.11, is exactly this: a sustained aggregate response assembled from short responses by different cohorts.

The same is true across classes and within the EV fleet. Drivers differ in arrival, departure, state of charge on arrival and energy needed, and the urgency ordering of Chapter 7 exploits those differences to hold a fleet cap while every deadline is met (Figure 7.7). Heterogeneity is what gives the optimizer room, and it is why an aggregator enrolls a portfolio rather than a clone army.

9.3.2 Averaging of independent errors

Diversity also reduces forecast uncertainty, up to a point. Let the error in the predicted flexibility of device i be ε_i with mean zero and variance σ^2, and let the errors be independent. The per-device error of the fleet mean then has variance

$$\operatorname{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i\right) = \frac{\sigma^2}{N}, \tag{9.11}$$

so its standard deviation falls as $\sigma/\sqrt{N}$. For $N = 10\,000$ that is a hundredfold reduction. Large fleets are, in this respect, far more predictable than their members: an individual thermostat's response to an event is nearly unforecastable, while the response of 10 000 of them is a smooth and repeatable curve. This is the statistical half of the case for aggregation.

9.3.3 Common-mode uncertainty

Equation (9.11) assumed independence, and the assumption is false for the dominant error source. Every thermostat in a region is driven by the same outdoor temperature, so an error in

the weather forecast moves all of their flexibilities in the same direction. The variance of a sum, repeated from equation (3.6) of Chapter 3, is

$$\mathrm{Var}\Big(\sum_i X_i\Big) = \sum_i \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j), \tag{9.12}$$

and the second term contains $N(N-1)$ covariances against only N variances. If the covariances are appreciable the second term dominates for any large N. Figure 3.4 of Chapter 3 quantifies it: for $N = 1\,000$ devices, a common pairwise correlation of $\rho = 0.1$ already removes most of the averaging benefit, and the ratio of fleet uncertainty to the naive sum rises from 0.03 to 0.32. Aggregation removes idiosyncratic uncertainty and leaves common-mode uncertainty untouched.

9.3.4 A weather forecast error

The concrete case is a biased weather forecast. Suppose tomorrow's outdoor temperature comes in 3°C above the forecast across the service territory, and that each thermostat's available flexibility changes by about 0.15 kW per degree of outdoor temperature. Every device in the fleet is then wrong in the same direction by 0.45 kW. For a fleet of 100 000 thermostats the aggregate forecast is off by

$$100\,000 \times 3^{\circ}\mathrm{C} \times 0.15\,\mathrm{kW}/^{\circ}\mathrm{C} = 45\,\mathrm{MW},$$

and the fraction by which it is off, 0.45 kW divided by the mean per-device flexibility, does not depend on N at all. Enrolling a million devices instead of a hundred thousand multiplies the error by ten along with the fleet.

Figure 9.3 contrasts the two regimes. Both panels show the forecast error of four individual devices over a day, in kW per device, as thin lines, with the fleet mean error as the thick line. In panel (a) the device errors are idiosyncratic: they take both signs at every hour and the fleet mean sits on zero. In panel (b) the same four traces are shifted by the 0.45 kW common bias; the individual traces look much the same, and the fleet mean is displaced to 0.45 kW and stays there. No amount of averaging moves the thick line in panel (b), which is why weather is the critical shared driver for a thermostat fleet and why the forecasting chapter that follows is about distributions rather than point predictions.

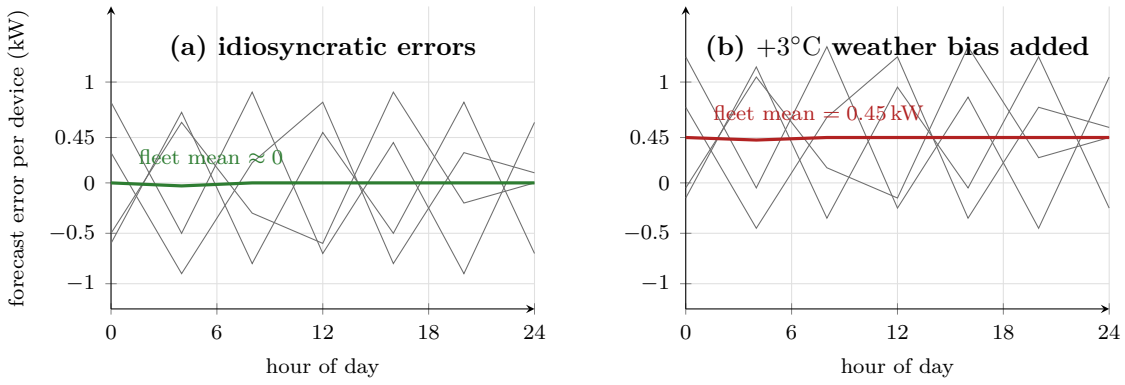


Figure 9.3: Idiosyncratic and common-mode forecast error. Thin lines are four devices, the thick line is the fleet mean. (a) Independent errors cancel in the mean, at a rate $\sigma/\sqrt{N}$, equation (9.11). (b) The same errors plus a common 0.45 kW bias from a 3°C outdoor-temperature error; the fleet mean is biased by the full amount for any N. Traces are illustrative.

Principle 9.2. *Fleet uncertainty has two parts. Aggregation divides the idiosyncratic part by N and leaves the common-mode part unchanged, so beyond a modest fleet size the only way to reduce fleet uncertainty is to forecast the shared drivers better.*

9.4 Location, service and hierarchy

A fleet aggregate is not a scalar. This section adds the two indices that a dispatchable aggregate must carry, electrical location and service capability, and then organizes the whole fleet into levels.

9.4.1 Geographic aggregation

A VPP with 100 MW of flexibility spread over a service territory cannot necessarily relieve a 20 MW constraint on one feeder. The devices that matter for that constraint are the ones electrically behind it, and they may total 5 MW. The aggregate must therefore be indexed by location,

$$P_{\text{fleet},r}(t) = \sum_{i \in r} P_i(t), \tag{9.13}$$

for regions r, and for distribution-level services by something finer still, the power at an individual feeder or transformer, $P_{\text{feeder}}(t)$. This is the $F(r, t, d)$ object of Chapter 3, equation (3.8), and it is the reason that chapter's grouping tree exists. Electrical location is part of the aggregation problem, not a label attached afterward: two devices with identical physics and different feeders belong in different aggregates.

9.4.2 Aggregation by service

The second useful index is not physical similarity but service capability. The question a grouping should answer is what the group can be sold to do, and a battery and an EV charger that both respond within seconds belong together for a fast-response product even though their physics differ. Table 9.3 gives the three groupings that cover most of what an aggregator offers, with the resources in each and the property that qualifies them.

Table 9.3: Aggregation by service. Membership follows response characteristics rather than device class.

Service group	Members	Qualifying property
Fast response	batteries; some EV chargers	response in seconds, bidirectional, no customer consultation
Long duration	managed EV charging; thermostats rotated across cohorts	hours of energy, assembled from short individual responses
Renewable absorption	EV charging; batteries	can increase consumption on command, absorbing surplus generation

The grouping cuts across the device classes of Chapters 5 to 7. Batteries appear in all three groups, thermostats only in the second and only by rotation, and the same EV appears in the long-duration group overnight and in the renewable-absorption group at midday. A device therefore belongs to several service aggregates at once, and the dispatch layer must avoid selling the same device's capability twice, a bookkeeping constraint that the hierarchy of the next subsection is designed to enforce.

9.4.3 Hierarchical aggregation

With location and service both in play, the aggregation is naturally built in levels. Figure 9.4 shows the structure for a million devices: 1 000 000 devices are summarized into 10 000 local groups, those into 500 feeder groups, those into 50 regional portfolios, and those into one VPP.

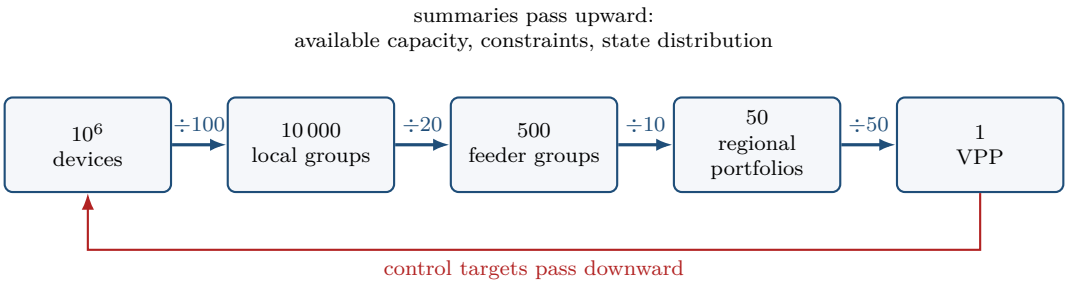


Figure 9.4: Hierarchical aggregation of a million devices into one VPP. Each level summarizes the level below it; the compression ratio between consecutive levels is small enough that every level's optimization is tractable. Counts are illustrative.

Each box is a level, the number in it is the number of objects at that level, and the labels on the arrows are the compression ratios between consecutive levels.

The point of the levels is that no single optimization is ever large. Each level solves a problem over the objects immediately below it: a hundred devices per local group, twenty local groups per feeder group, ten feeder groups per region, fifty regions in the VPP. The largest problem in the hierarchy has fifty decision variables, against the million of the brute-force formulation of Section 9.1.1. The levels also carry the natural boundaries: a feeder group is the unit at which the constraint of equation (9.13) applies, and a service group is the unit at which a market product is sold.

9.4.4 Hierarchical dispatch

The same structure runs in reverse for control. Suppose the top-level optimizer commits $P_{\text{VPP}} = 100\,\text{MW}$ of reduction and allocates it across three regions,

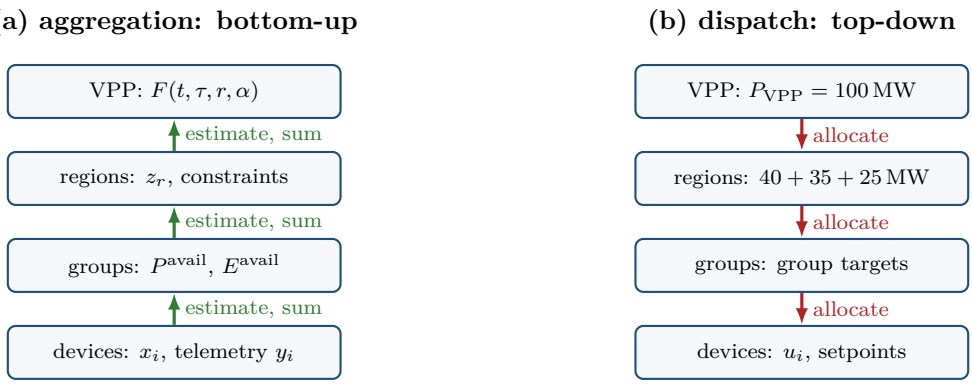
$$P_{\text{VPP}} = 100\,\text{MW} = \underbrace{40}_{\text{region A}} + \underbrace{35}_{\text{region B}} + \underbrace{25}_{\text{region C}}\ \text{MW}. \tag{9.14}$$

Each region then allocates its target among its own device groups, and each group decides which devices act, using local logic such as the urgency ordering of Chapter 7 or the cohort rotation of Chapter 5. A global objective becomes regional targets, regional targets become group targets, and group targets become device actions. Nothing in the chain requires one enormous centralized optimization, which is the architecture Chapter 3 contrasted with the centralized alternative in Figure 3.9.

The allocation in (9.14) is not arbitrary. It should follow the regions' available capacities, their marginal customer costs and any local network constraints, so the top level needs each region's summarized z_r, not its devices. The quality of the dispatch is therefore limited by the quality of the summaries, which is what makes ϕ an operational object rather than a reporting convenience.

9.4.5 Aggregation upward, dispatch downward

The two directions are different problems and it is worth naming them separately. Aggregation is bottom-up: it takes device states and telemetry and produces fleet capability. Dispatch is top-down: it takes a grid request and produces device actions. Figure 9.5 draws the two flows over the same four levels, with the content of the messages on each arrow: estimated states and capacities going up on the left, power targets coming down on the right, with the allocation of equation (9.14) shown at the regional level.



A DERMS runs both loops continuously: the left-hand flow is fleet state estimation, the right-hand flow is control, and the left-hand flow is what makes the right-hand allocation feasible.

Figure 9.5: The two directions of the same hierarchy. (a) Aggregation carries estimated states and capacities upward. (b) Dispatch carries targets downward, with the allocation of equation (9.14) at the regional level.

9.5 Probabilistic aggregation

Everything so far treated the device states as known. They are not, and the aggregate is consequently a random variable rather than a number. This section develops the probabilistic form of ϕ and the three ways a deterministic aggregate misleads.

9.5.1 Fleet state estimation

Not every device reports telemetry on every control cycle. Reports are lost, gateways go offline, vendor interfaces throttle, and some contracts carry no telemetry at all. The device state is therefore inferred rather than read, through the posterior

$$p(x_i \mid y_{i,1:t}) \tag{9.15}$$

of Chapter 3, and the fleet representation is the corresponding posterior over the compressed state,

$$p(z_t \mid y_{1:t}), \qquad z_t = \phi(x_t). \tag{9.16}$$

Aggregation is thus itself probabilistic: the output of ϕ is a distribution, and the quantities the grid is offered are summaries of it.

Missing devices. Take 100 000 enrolled devices of which 85 000 have fresh telemetry and 15 000 do not. The states of the silent 15 000 are not unknown in the sense of being arbitrary: they can be estimated from the last known state, the elapsed time, the weather since, and the observed behavior of similar devices in the same cluster. The fleet capability splits accordingly,

$$F_{\text{fleet}} = F_{\text{observed}} + F_{\text{estimated}}, \tag{9.17}$$

where the first term has small uncertainty and the second has large uncertainty, and the second is 15 % of the fleet. Treating F_{fleet} as a deterministic quantity throws that asymmetry away. A fleet with 85 % telemetry coverage and one with 100 % coverage may report the same expected capacity and should not be committed on the same terms.

9.5.2 Aggregate capacity as a distribution

The honest output of the aggregation layer is therefore not $F_{\text{fleet}} = 100\,\text{MW}$ but a distribution $p(F)$, from which the operator reads whichever summary the decision needs:

$$\mathbb{E}[F] = 100\,\text{MW}, \qquad \Pr(F > 85\,\text{MW}) = 0.95. \tag{9.18}$$

The first number is what the fleet delivers on average and the second is what it can be asked to guarantee. A commitment is written against the second, as the firm-capacity discussion of Chapter 4 established and as the commitment optimization of Figure 4.7 priced. The gap between the two, 15 MW here, is the cost of the uncertainty in ϕ, and it shrinks when telemetry coverage improves or when the weather forecast narrows.

9.5.3 Quantiles do not add

A deterministic aggregate can be summed; a reliable one cannot. Suppose resource group A offers 30 MW at 90 % reliability and group B offers 20 MW at 90 % reliability, meaning that each is exceeded with probability 0.90. The total at 90 % reliability is not 50 MW, because the quantile of a sum is a property of the joint distribution $p(F_{\text{A}}, F_{\text{B}})$ and not of the two marginals.

Put numbers on it with a normal model. Let $F_{\text{A}} \sim \mathcal{N}(40, 7.8^2)$ and $F_{\text{B}} \sim \mathcal{N}(25, 3.9^2)$ in MW, which reproduces the stated 90 %-reliable values, since $40 - 1.282 \times 7.8 = 30.0$ and $25 - 1.282 \times 3.9 = 20.0$. The total has mean 65 MW in every case, and its standard deviation and 10th percentile depend on the correlation ρ as in Table 9.4. Independent groups give 53.8 MW, perfectly correlated groups give exactly the naive sum of 50.0 MW, and intermediate correlations lie in between. For jointly normal margins, then, adding the reliable values is the perfectly-correlated case and is conservative by up to 3.8 MW.

Table 9.4: The 90 %-reliable total of two groups with $F_{\text{A}} \sim \mathcal{N}(40, 7.8^2)$ and $F_{\text{B}} \sim \mathcal{N}(25, 3.9^2)$ MW. The marginals, and hence the individual reliable values of 30 and 20 MW, are the same in every row.

Correlation ρ	σ_{total} (MW)	10th percentile (MW)	Interpretation
0 (independent)	8.7	53.8	errors partly cancel
0.5	10.3	51.8	partial common driver
1 (comonotone)	11.7	50.0	equals the sum of the two reliable values

The normality is doing work. The ordering above is not general. Suppose instead that group A delivers 40 MW with probability 0.92 and 5 MW otherwise, and group B delivers 25 MW with probability 0.92 and 3 MW otherwise. Each group's 90 %-reliable value is its good value, 40 and 25 MW, because the bad event has probability only 0.08. If the two bad events are the same regional weather event, the total is 65 MW with probability 0.92 and the 90 %-reliable total is 65 MW, the sum of the parts. If the bad events are independent, at least one occurs with probability 0.154, and the 90 %-reliable total falls to 43 MW. With discrete shared failure modes the direction of the error reverses: independence is now the worse case, because two chances of a partial failure are worse than one chance of a joint failure. The general statement is the one the lesson requires: fleet reliability is a property of the joint distribution, and any aggregation that reports only marginals cannot answer a reliability question.

9.5.4 Resource complementarity

Heterogeneity also acts through time. Thermostat flexibility is strongest on hot afternoons, when air conditioners are running and there is something to curtail. EV flexibility is strongest

overnight, when cars are plugged in with hours of slack. Battery flexibility is available almost whenever the units are charged. Figure 9.6 plots the three availability curves and their sum over a day: the horizontal axis is the hour, the vertical axis is available reduction in MW, the three colored curves are the individual classes, and the heavy black curve is their sum.

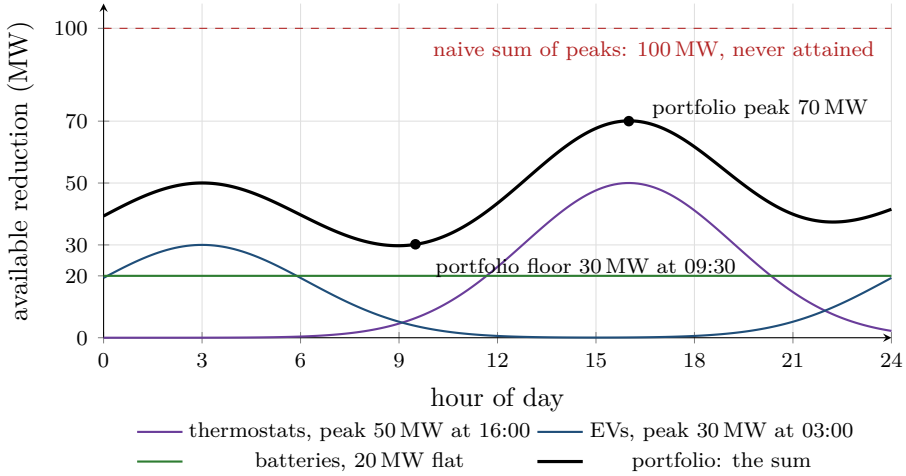


Figure 9.6: Temporal complementarity of three resource classes. Curves are $50\exp(-((t-16)/3.2)^2/2)$ for thermostats, $30[\exp(-((t-3)/3.2)^2/2)+\exp(-((t-27)/3.2)^2/2)]$ for EVs and a constant 20 for batteries, in MW. The sum of the three peaks is 100 MW, the peak of the sum is 70 MW, and the floor of the sum, 30 MW, exceeds every individual class's floor. Curves are illustrative.

Two readings of the figure matter. First, the naive sum of the three peak capacities, $50+30+20=100$ MW, is never available, because the peaks occur thirteen hours apart; the peak of the portfolio is 70 MW. Second, and more usefully, the portfolio's worst hour is far better than any single class's worst hour: at 09:30 the thermostats and the EVs are both near zero and the portfolio still offers 30 MW, against a floor of zero for a thermostat-only fleet. A heterogeneous portfolio is more valuable than any single class of the same nameplate, not because it is larger but because it is available for more of the day. This is the temporal counterpart of the hand-off schedule of Chapter 4, and it means the fleet capability is not the sum of the class capacities in either direction.

9.5.5 Hidden bottlenecks

The failure mode that makes coarse aggregation dangerous rather than merely imprecise is a location mismatch. Suppose a VPP reports 80 MW of total flexibility, of which 60 MW sits in region A and 20 MW in region B. A grid need of 40 MW arises in region B. The aggregate says the fleet is twice as large as the need; the fleet cannot deliver it, because only 20 MW of the capability is electrically in the right place. An aggregate that reports 80 MW against this need is not approximate, it is wrong, and the operator who acted on it has no fallback at the moment of use.

This is the characteristic danger of model reduction. A reduced model answers every question put to it, including the questions it has discarded the information to answer. The defense is not more accuracy but an explicit domain: the aggregation layer must report what constraint set its number is valid under, which in this case means reporting $P_{\text{fleet},r}$ per region rather than their sum.

Principle 9.3. *A reduced fleet model must carry the scope of its own validity. An aggregate that cannot say where its megawatts are, or for how long they last, will be used as though the answer were "anywhere" and "indefinitely".*

9.5.6 Matching the aggregation level to the decision

The conclusion of the preceding subsections is that there is no single correct aggregation. The right level of detail is the coarsest one that still determines the decision at hand. For a system-wide peak-shaving program, a broad aggregate is adequate: the decision is how many megawatts to reduce in a given hour, and location within the system is immaterial. For relieving one overloaded distribution transformer, the decision depends on a handful of devices behind that transformer, and no amount of fleet-level accuracy substitutes for knowing which ones they are. For a frequency-response product the relevant resolution is in time rather than space, since what matters is the response within seconds and not the energy over hours.

The practical consequence is that a platform maintains several reductions of the same fleet at once, each matched to a class of decision, and recomputes them from the same underlying device state. The map ϕ is therefore a family of queries on one model rather than a single preprocessing step, which is the view Chapter 4 took of equation (4.14).

9.6 The mathematical object

The chapter closes by naming what aggregation computes, why computing it exactly is hopeless, and what an engineering approximation to it must report.

9.6.1 The fleet feasible set as a Minkowski sum

Let device i in state x_i admit the controls $u_i \in \mathcal{U}_i(x_i)$, the state-dependent feasible set of Chapter 1. The fleet's feasible aggregate output is the set of all sums of admissible individual actions,

$$\mathcal{U}_{\text{fleet}} = \left\{ \sum_{i=1}^{N} u_i \;:\; u_i \in \mathcal{U}_i(x_i) \right\} = \mathcal{U}_1(x_1) \oplus \cdots \oplus \mathcal{U}_N(x_N), \tag{9.19}$$

which is the Minkowski sum of the device sets. Every statement in this chapter is an attempt to describe this one object cheaply: the instantaneous interval of (9.8) is a one-dimensional box containing it, the virtual battery of (9.10) is an inner approximation of it over a horizon, and the histogram of Figure 9.1 is a compressed description of the states that generate it.

Figure 9.7 shows the operation for two devices in a plane whose horizontal axis is a power change and whose vertical axis is the energy that change consumes over the horizon. Device 1's feasible set is a rectangle, device 2's is a triangle, and the sum is the hexagon on the right: each point of it is one point from the rectangle plus one point from the triangle. Two features generalize. The sum is larger and smoother than either part, which is diversity in geometric form. And the sum of convex sets is convex, which is the single fact that makes aggregate dispatch tractable when the device sets are convex.

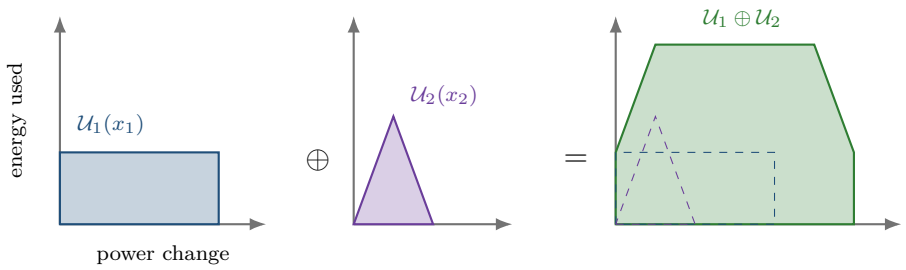


Figure 9.7: The Minkowski sum of two device feasible sets, equation (9.19). Each axis is a decision variable of one device: a power change and the energy it consumes over the horizon. The two summands are shown dashed inside the sum. The sum of convex sets is convex, and it is larger and smoother than either summand. Sets are illustrative.

9.6.2 Combinatorial hardness

Computing (9.19) exactly is out of reach as soon as the device sets are not convex. Suppose each device has only two admissible actions, $u_i \in \{0, 1\}$, which is a fair model of a compressor or a relay. The fleet has

$$2^N \text{ combinations}, \qquad 2^{100\,000} \approx 10^{30\,103}, \tag{9.20}$$

and enumeration is not merely slow, it is unrelated to anything physically possible. No increase in computing power addresses this; the only way through is structure. Four kinds of structure are used, and every one of them has appeared earlier in this chapter: device similarity, which makes most of the 2^N combinations equivalent and justifies clustering (Section 9.2.3); convex relaxation, which replaces the discrete set by the interval it spans and recovers a tractable sum (Section 9.2.5); statistical description, which replaces the exact combination by a distribution over states (Sections 9.2.2 and 9.5.1); and hierarchical decomposition, which never forms the fleet-wide problem at all (Section 9.4.3). Aggregation is, in this light, the set of techniques that make a 2^N problem into a sequence of small ones.

9.6.3 The equivalent-circuit analogy

Power engineering has solved this kind of problem before. A complicated subnetwork is replaced, for the purposes of everything outside it, by an equivalent that reproduces its behavior at the boundary: a Thévenin source and impedance in place of a whole network. The VPP is the same construction at a different boundary, as Figure 9.8 draws. On the left, inside the dashed boundary, are 100 000 devices with their own states, limits and customers. On the right is what the utility is given at the point of interconnection: a lower and upper power limit, an available energy, a ramp rate and a reliability level.

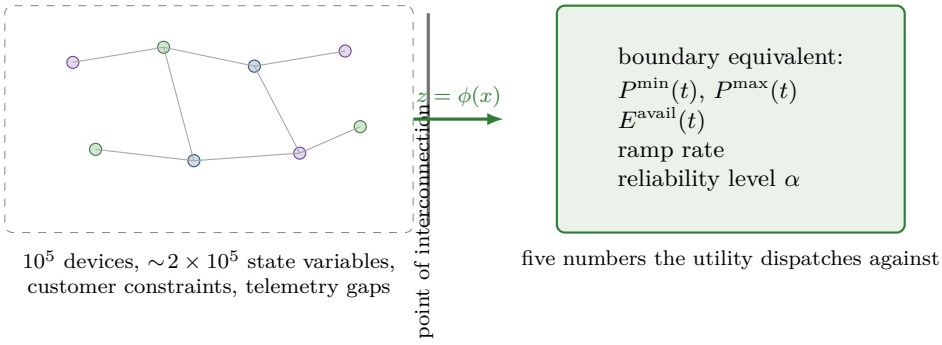


Figure 9.8: Aggregation as a boundary equivalent. Internally the fleet is 100 000 devices with their own physics and customers; externally the utility sees the power limits, energy, ramp rate and reliability that the aggregation map ϕ produces. The construction is the equivalent-circuit reduction of power engineering applied at the interconnection rather than at a network terminal.

The analogy is useful for two reasons beyond the picture. It sets the correctness criterion: an equivalent is valid when it reproduces the boundary behavior over the range of conditions in which it will be used, which is precisely the scope requirement of Section 9.5.5. And it places aggregation in the tradition of boundary-exchanging physical models, where a subsystem is characterized by what crosses its boundary rather than by its interior, which is also the view the TERRA framework takes.

9.6.4 What the optimizer asks of the aggregation layer

The requirements on ϕ can be stated as the six questions a dispatch optimizer puts to it:

- how much upward flexibility is available, meaning net load reduction or injection;
- how much downward flexibility, meaning added consumption;
- for how long each can be sustained;
- where it is, electrically;
- with what uncertainty; and
- what today's dispatch does to tomorrow's capability.

The first five are answered by the object

$$F(t, \tau, r, \alpha): \quad \text{start time } t,\ \text{duration } \tau,\ \text{region } r,\ \text{reliability } \alpha, \tag{9.21}$$

in the two directions $F^{\uparrow}$ and $F^{\downarrow}$. The sixth is different in kind: it asks for the state the fleet will be left in, which means ϕ must be attached to a model that can be simulated forward, not to a static summary. An aggregation layer that answers all six has done its job, and one that answers only the first has produced the number of Section 9.1.2.

9.6.5 Worked example

Take a VPP of 50 MW of thermostats, 30 MW of EVs and 20 MW of batteries, all figures nominal. The naive aggregate is 100 MW. Table 9.5 carries that number through the chapter's corrections to the representation that is actually dispatchable.

Table 9.5: From the naive aggregate to a dispatchable representation for a VPP of 50 MW of thermostats, 30 MW of EVs and 20 MW of batteries. Intermediate deductions are illustrative; the endpoints are the quantities the operator uses.

Stage	MW	What was applied
Naive sum of nameplates	100	equation (9.2)
Thermostat duration limit	−6	comfort headroom bounds the sustainable hours, Section 9.2.4
EV deadline conflicts	−5	some deferred energy is owed before the service window, Section 9.2.6
Battery state of charge	−3	units near their reserve, Section 9.2.1
Telemetry gaps and overrides	−1	the estimated share of equation (9.17)
Expected flexibility $\mathbb{E}[F]$	85	the mean of $p(F)$
95 %-reliable flexibility	70	$\Pr(F \geq 70\,\text{MW}) = 0.95$; implies $\sigma \approx 9\,\text{MW}$
Available in the constrained region	25	location index r, Section 9.4.1

Three numbers survive, and they answer different questions. The expected flexibility, 85 MW, is what the fleet delivers on an average day. The 95 %-reliable flexibility, 70 MW, is what can be committed; the 15 MW gap implies a standard deviation of about 9 MW under a normal model, since $1.645 \times 9 \approx 15$. And only 25 MW of the capability sits in the feeder region where the network constraint is, which is the bottleneck of Section 9.5.5 in numbers. A single scalar cannot represent all three, and the object that does is the indexed one of equation (9.21), $F(t, \tau, r, \alpha)$: 85 MW is F at $\alpha = 0.5$ summed over regions, 70 MW is F at $\alpha = 0.95$ summed over regions, and 25 MW is F at the constrained r.

9.7 What to retain

Definition 9.1 (Fleet aggregation)**.** Fleet aggregation is model reduction for control: from the device states $x_1, \ldots, x_N$ it constructs a compressed representation $z = \phi(x)$ that preserves what matters for dispatch. The design problem is to compress aggressively enough to make control scalable without discarding the constraints that determine the fleet's real flexibility.

The two failure directions are symmetric and both are expensive. Too little compression leaves an optimization that cannot be solved on a control cycle; too much leaves an aggregate that reports capability the fleet does not have, in the wrong place or for the wrong duration. Table 9.6 lists the chapter's objects with the section that defines each.

Table 9.6: The objects of fleet aggregation.

Object	Definition	Section
$P_{\text{fleet}}(t)$	$\sum_i P_i(t)$; insufficient as a state	9.1.2
$z = \phi(x)$	the compressed fleet representation	9.1.3
n_k	count of devices in state bin k	9.2.2
N_k, $\bar{x}_k$	cluster population and representative state	9.2.3
$\mathcal{F}_t$	instantaneous feasible set of P_{fleet}	9.2.5
$\mathcal{F}$	trajectory set over the horizon H	9.2.5
virtual battery	$P^{\min} \le P_t \le P^{\max}$, $E_{t+1} = E_t + P_t \Delta t$, $E^{\min} \le E_t \le E^{\max}$	9.2.6
$p(z_t \mid y_{1:t})$	probabilistic aggregation	9.5.1
$\mathcal{U}_{\text{fleet}}$	$\mathcal{U}_1(x_1) \oplus \cdots \oplus \mathcal{U}_N(x_N)$	9.6.1
$F(t, \tau, r, \alpha)$	the operational fleet representation	9.6.4

9.8 Exercises

Exercise 9.1. For the histogram of Figure 9.1, recompute P^{avail} and E^{avail} if the reserve limit is raised from $\text{SOC}^{\min} = 0.10$ to 0.25. By what percentage does each fall, and why do the two percentages differ? Then repeat the calculation from the mean state of charge alone and state the error in each quantity.

Exercise 9.2. A fleet is reported as a virtual battery with $P^{\max} = 40\,\text{MW}$ and $E^{\text{avail}} = 100\,\text{MWh}$ at 17:00. Half the energy belongs to EVs that depart by 17:30. Write the two-cohort virtual-battery model that represents this correctly, and give the largest constant power each model says can be held from 20:00 to 22:00.

Exercise 9.3. Using Table 9.4, compute the 90 %-reliable total for $\rho = 0.25$ and $\rho = 0.75$. Then suppose group B's standard deviation is 7.8 MW rather than 3.9 MW, with its mean adjusted so that its own 90 %-reliable value stays at 20 MW; how much does the spread between the $\rho = 0$ and $\rho = 1$ totals change, and why?

Exercise 9.4. In Figure 9.6, read off the hour at which the portfolio is weakest and the hour at which it is strongest. Which single class would you add 20 MW of, to raise the portfolio's floor the most, and what does your answer say about the difference between nameplate capacity and portfolio value?

9.9 Connection to the TERRA framework

Aggregation is the operation of answering a small number of questions about a large graph, and each of the chapter's objects corresponds to a specific part of such a model.

- **The fleet is one graph, not many models.** The stacked state of equation (9.1) is the state of a single conservation graph with N component subtrees joined at grid nodes. The aggregate P_{fleet} of equation (9.2) is the flow on the one edge that crosses the interconnection boundary, so the naive aggregate is a single edge flow and the fleet state is everything behind it.

- ϕ **is a set of queries, not a separate model.** Computing P^{avail} is an evaluation of component bounds at the current state; computing E^{avail} or $C(\tau)$ requires integrating the accumulators forward under a trial request, which is the forward simulation of Figure 9.2. Keeping several reductions of the same fleet, as Section 9.5.6 requires, is natural when they are queries on one model and awkward when they are separate aggregates maintained by hand.

- **Histograms and clusters are component populations.** A histogram bin or a cluster is a single component instance carrying a multiplicity N_k, with the sum of equation (9.7) as its contribution to the node balance. The fidelity loss of Table 9.1 is exactly the error made by evaluating a component's bounds at a mean state instead of integrating each instance, and it appears wherever a bound saturates.

- **The virtual battery is a fitted component.** Equation (9.10) is one bounded accumulator with a power bound, fitted so that its boundary behavior matches the fleet's. Its failure in Section 9.2.6 is a structural misfit, not a parameter error: the fleet has several accumulators with different deadlines and the fitted model has one, so partitioning the fleet into cohorts and fitting one accumulator per cohort is the modeling repair.

- **Probabilistic aggregation is posterior inference plus propagation.** Equation (9.16) is the posterior over unmeasured component states given partial telemetry, pushed through ϕ. The split of equation (9.17) into observed and estimated parts is the distinction between states a measurement pins down and states the model must infer, and the correlation structure of Table 9.4 comes from shared disturbances w, chiefly weather, acting on many components at once.

- **Location is topology, and hiding it is a modeling error.** The regional split of equation (9.13) and the bottleneck of Section 9.5.5 are statements about which conservation node a component's flow enters. A graph model carries that automatically; a scalar aggregate has to be told, which is why the boundary equivalent of Figure 9.8 must be defined per boundary rather than once for the fleet.

In summary, fleet aggregation is a query interface over a large conservation graph: the devices stay in the model, and what leaves it is a small set of predicted, uncertain, located and time-indexed quantities. Those predictions are only as good as the forecasts behind them, which is the subject of the next chapter.

Chapter 10

Forecasting flexibility: how much DER capacity will actually be available?

10.1 Load forecasting and flexibility forecasting

Chapter 4 established that a fleet becomes a grid resource only when somebody can state how much it will deliver, for how long, where, and with what confidence. Every one of those numbers is a forecast. This chapter is about producing them, and its question is

How much controllable capacity will the fleet actually have at a future time, and with what certainty?

The forecasting problem is not the one a utility load forecaster solves. A load forecast predicts the power a population of customers *will* draw. A flexibility forecast predicts how much that power *could be changed* on request. The two questions have different inputs, different targets, and different error consequences, and the second is the harder of the two because its answer depends on the internal state of every device, on customer behavior, and on what the operator itself has already done.

Jargon: load forecast, flexibility forecast, quantile, calibration, endogenous

A *load forecast* predicts future consumption, $\hat{P}_{\text{load}}(t+\tau)$: what the fleet will draw if nobody intervenes. A *flexibility forecast* predicts $\hat{F}(t+\tau)$: how much that draw could be changed by dispatching the fleet. A *quantile* of a forecast distribution is the value below which a stated fraction of outcomes fall; the 10th percentile $F_{0.10}$ is exceeded on 90 % of occasions. A forecast is *calibrated* when its stated probabilities match observed frequencies: outcomes fall inside a declared 90 % interval about 90 % of the time. A forecast target is *endogenous* when the act of using the forecast changes the quantity being forecast, which is the case here because dispatching the fleet today alters its state tomorrow.

Key idea

Load forecasting predicts what the fleet will do. Flexibility forecasting predicts what the fleet could be made to do. The second is a forecast of a *capability*, and a capability depends on state.

This section separates the two forecasting problems, defines flexibility as a two-sided range with a duration attached, and shows that forecasting it requires first forecasting the fleet's state.

10.1.1 Two different questions

The load forecast asks: how much power will this fleet consume at $t+\tau$? The flexibility forecast asks: how much could that consumption be changed at $t+\tau$ if the fleet is dispatched? The quantities are not interchangeable, and in each of the three device classes of Chapters 5 to 7 the relation between them is different.

Thermostats. High load usually does imply high flexibility, because load means compressors running and a running compressor can be switched off. The implication is not reliable: a fleet that has just been curtailed carries a high load from recovering buildings whose indoor temperatures are already at the top of the comfort band, so the load is large and the flexibility is small. This is the rebound of Chapter 2, Figure 2.3, seen from the forecaster's side.

Electric vehicles. A high charging load could mean many vehicles charging with hours of slack, or many vehicles charging because they must leave soon. The megawatt figure is the same in both cases and the flexibility differs by an order of magnitude.

Batteries. Current load says almost nothing. A battery drawing nothing may be full (large discharge capability, no charging capability) or empty (the reverse). Without the state of charge the load carries no information about the capability at all.

> **Principle 10.1.** *Load and flexibility are different functions of the same fleet. A load forecast is not a flexibility forecast, and no fixed ratio converts one into the other.*

10.1.2 Flexibility as a two-sided range

The definition used throughout this book is the difference between the baseline and the controlled power,

$$F_t = P_t^{\text{base}} - P_t^{\text{ctrl}}, \tag{10.1}$$

which is the quantity ΔP_t of Chapter 1 measured at fleet scale, with $F_t > 0$ meaning load reduction. Chapter 2 obtained P_t^{base} from a baseline model and Chapter 3 used ΔP_t for verification; here the same quantity is the forecast target.

A single signed number is not enough, because a fleet can usually move in both directions. The useful object is a feasible range,

$$F_t \in [F_t^{\downarrow},\ F_t^{\uparrow}], \tag{10.2}$$

for example $-20\,\text{MW} \le F_t \le 50\,\text{MW}$: the fleet can reduce load by up to 50 MW (upward grid support) or increase it by up to 20 MW (downward support, which absorbs surplus generation). The sign convention is the one fixed in the notation chapter: positive F reduces net load, so $F^{\uparrow}$ is the upper end of the range and $F^{\downarrow}$ is negative when the fleet can add load.

10.1.3 Flexibility has a horizon

The question "how much flexibility is available right now" asks for $F(t, 0)$ and is rarely the operational question. The operational question is "how much can be held tomorrow from 18:00 for two hours", which asks for

$$F(t_{\text{start}}, \tau) = \text{the largest change that can be } \textit{sustained} \text{ for a duration } \tau \text{ starting at } t_{\text{start}}. \tag{10.3}$$

A thermostat fleet may provide 100 MW for fifteen minutes but only 40 MW for three hours, because the buildings warm up and compressors restart; this is the device-level decay of

Chapter 1, Figure 1.3, and the fleet-level version is the $C(\tau)$ curve of Chapter 4, Figure 4.3. Figure 10.1 draws the two ideas of this section together: the vertical extent of the shaded region is the range of equation (10.2), and its narrowing to the right is the duration dependence of equation (10.3). The forecast target is therefore a trajectory or a curve, not one number.

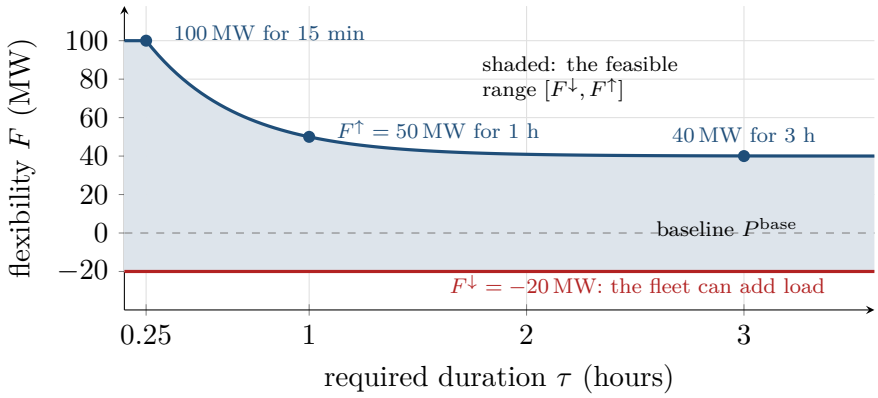


Figure 10.1: Flexibility as a range that depends on the required duration. The horizontal axis is the duration τ a response must be sustained; the vertical axis is the signed load change F, with zero at the baseline. The upper boundary is $F^{\uparrow}(\tau) = 40 + 60\exp[-(\tau - 0.25)/0.4186]$ MW, chosen so that it passes through 100 MW at 15 minutes, 50 MW at one hour and 40 MW at three hours; the lower boundary is held at -20 MW. Curves are illustrative. The fleet-level capacity curve $C(\tau)$ of Figure 4.3 is the upper boundary of this region.

10.1.4 State dependence

Flexibility is a function of the fleet's internal state, of exogenous conditions, and of the constraints the customers and the contract impose,

$$F_t = g(x_t,\ w_t,\ c_t), \tag{10.4}$$

with x_t the device or fleet state, w_t the disturbances and c_t the customer and program constraints. Table 10.1 instantiates the three arguments for the three device classes; every entry was derived in Chapters 5 to 7. The table is the reason flexibility forecasting is harder than load forecasting: a load forecast can be built from weather and calendar variables alone, whereas g cannot be evaluated without the states in the middle column.

Table 10.1: The arguments of the flexibility map g of equation (10.4), by device class.

Class	State x_t	Exogenous w_t	Customer constraint c_t
Battery	SOC_t, inverter state	site load, PV output	$\text{SOC}^{\min}$, $\text{SOC}^{\max}$, backup reserve SOC^{res}
EV	SOC_t, plug-in indicator a_t	arrival T^{arr}, departure T^{dep}, driving need	SOC^{req} at t_{dep}, charger rating $P^{\max}$
Thermostat	T_{in}, compressor state s_t, inferred RC	T_{out}, humidity, solar gain, occupancy	comfort band $[T^{\min}, T^{\max}]$, override right

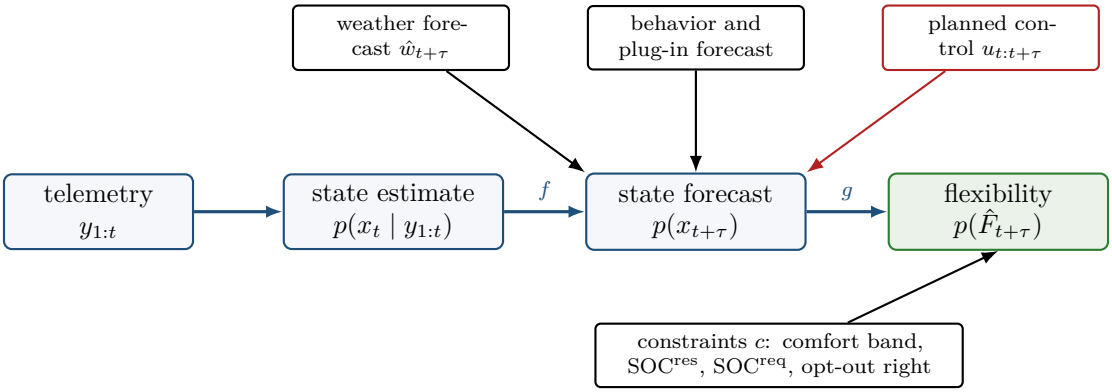
10.1.5 The future state must itself be forecast

Equation (10.4) evaluated at t gives the flexibility available now, which is a measurement problem. Evaluated at $t + \tau$ it requires the *future* state, which is a forecasting problem. Suppose

a battery reports SOC = 60 % at 12:00 and the operator needs the flexibility at 18:00. The 60 % figure cannot be used: the rooftop array may charge the battery during the afternoon, the household load may discharge it, and an earlier dispatch may spend it. The computation is therefore two-stage,

$$p(x_t \mid y_{1:t}) \;\longrightarrow\; p(x_{t+\tau}) \;\longrightarrow\; p(\hat{F}_{t+\tau}), \tag{10.5}$$

first a state forecast, then the map g applied to it. Figure 10.2 draws the pipeline: four boxes in a row for telemetry, the estimated present state, the forecast future state and the resulting flexibility distribution, with the three inputs that enter the state forecast drawn above it and the constraints that enter g drawn below. Two features of the diagram matter later in the chapter. The output is a distribution rather than a number, because the state forecast is uncertain; and one of the inputs to the state forecast is the operator's own planned control, which is the endogeneity of Section 10.5.



The red input is the operator's own plan: the forecast target depends on the control (Section 10.5).

Figure 10.2: The flexibility-forecasting pipeline of equation (10.5). Telemetry gives an estimate of the present state; the dynamics f and the forecast disturbances give a distribution over the future state; the map g of equation (10.4) turns that distribution into a distribution over flexibility.

10.2 Resource-specific forecasts

Because g differs by device class, the forecast is built class by class and then aggregated. This section works through the three classes and the aggregation.

10.2.1 A crude thermostat forecast

Take $N = 100\,000$ enrolled thermostats. A crude instantaneous estimate multiplies the fleet size by the average compressor power, the expected duty cycle, and the fraction of the running load that can actually be curtailed,

$$F_t = N\, P_{\mathrm{AC}}\, D_t\, r_t = 100\,000 \times 3.5\,\mathrm{kW} \times 0.60 \times 0.70 = 147\,\mathrm{MW}. \tag{10.6}$$

The three factors do different work. The duty cycle $D_t = 0.60$ says that at the moment in question about 60 % of the compressors are running, so the fleet's air-conditioning load is 210 MW rather than the 350 MW nameplate; this is the estimate of Chapter 5, equation (5.13). The curtailable fraction $r_t = 0.70$ then says that only 70 % of that running load can be removed, because some houses are already at the top of their comfort band, some customers have

overridden, and some devices are offline. Equation (10.6) is better than counting devices and still crude: it has no duration, no indoor temperatures, and no treatment of the dispatch history.

10.2.2 Drivers of thermostat flexibility

The predictors that matter are the outdoor temperature, humidity, the indoor temperature, time of day, the setpoint, recent HVAC runtime, the previous dispatch, and the required event duration. The relationship with weather is not monotonic in the forecast target, and this is the point at which a naive regression goes wrong. Very hot weather drives more compressors on, which raises the instantaneous response, but it also makes buildings warm faster once cooling is curtailed, which shortens the sustainable duration. Chapter 5 plotted the two effects against T_{out} and showed them moving in opposite directions. The consequence for the forecaster is that the sign of the weather coefficient depends on which target is being predicted: positive for $F(t, 15\,\text{min})$ and eventually negative for $F(t, 3\,\text{h})$. Defining the target precedes fitting the model, a point taken up in Section 10.6.

10.2.3 Electric-vehicle flexibility is a sum over vehicles

An EV fleet's flexibility is a sum of individual contributions,

$$F_{\text{E},t} = \sum_i f_i(a_{i,t},\ \text{SOC}_{i,t},\ T_i^{\text{dep}},\ \text{SOC}_i^{\text{req}},\ P_{i,\max}), \tag{10.7}$$

with each term governed by the five quantities of Chapter 7. The filtering this implies is severe. Of a fleet of 50 000 enrolled vehicles, at 20:00 perhaps 35 000 are plugged in, and of those perhaps 10 000 are actually drawing power. The remaining 25 000 plugged-in vehicles are either full or waiting under a managed-charging schedule; they contribute upward-load flexibility, not load reduction. Only the 10 000 charging vehicles can have their charging deferred, and how much of their load is deferrable depends on their slack.

10.2.4 The urgency split of a charging load

Suppose the aggregate charging load of those vehicles is 60 MW. The deferrable part is smaller, because the 60 MW is not homogeneous. Figure 10.3 is a stacked bar: the left bar divides the metered 60 MW into three groups by the urgency metric U_i of Chapter 7, and the right bar retains only the group whose charging can be deferred. Roughly 20 MW belongs to vehicles that will be unplugged within the hour, 10 MW to vehicles within a few percentage points of the state of charge their drivers require, and 30 MW to vehicles with hours of slack. Deferring the first two groups either fails outright or violates the departure requirement, so the usable flexibility is about 30 MW: half the metered load.

10.2.5 Battery flexibility through the state of charge

For a battery whose current state of charge is known, the instantaneous capability is a direct calculation from the bounds of Chapter 6: the discharge capability is limited by the inverter rating and by the energy above the reserve, and the charging capability by the rating and the headroom below $\text{SOC}^{\max}$. Forecasting is therefore trivial at $\tau = 0$ and nontrivial afterwards, because the future state of charge depends on the site load, the solar output, the backup reserve policy and any earlier dispatch. The forecast is built in the order of equation (10.5): obtain $p(\text{SOC}_{t+\tau})$ by propagating the accumulator forward under forecast site net load and a planned dispatch, then push that distribution through the capability bounds to obtain $p(F_{\text{B},t+\tau})$. Because the map from SOC to capability is piecewise linear and saturates at the inverter rating, the resulting distribution is not normal even when $p(\text{SOC}_{t+\tau})$ is: mass piles up at the rating whenever the state of charge is comfortably inside its bounds.

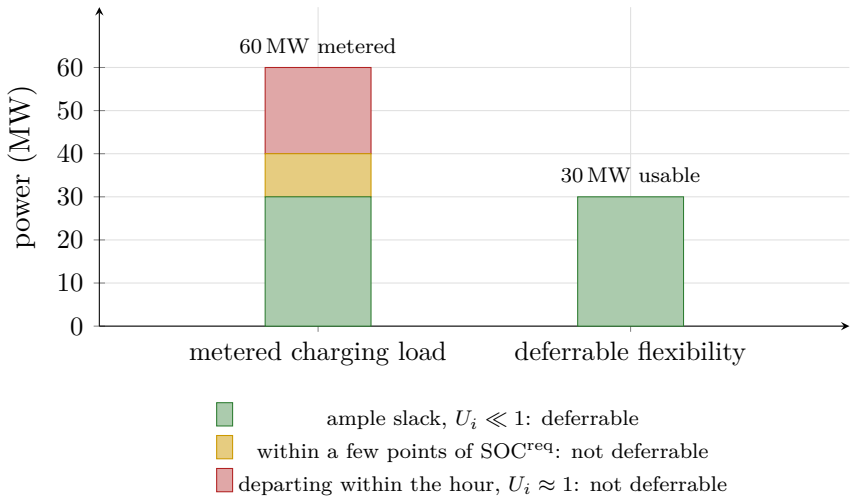


Figure 10.3: Why a 60 MW charging load is not 60 MW of flexibility. The left bar splits the metered load by urgency; the right bar keeps only the slack group. The split is illustrative, but the structure is general: the forecast needs information about urgency, not only about load.

10.2.6 Forecasting the aggregate

The portfolio forecast is the sum over classes,

$$F_{\text{fleet}} = F_{\text{T}} + F_{\text{E}} + F_{\text{B}}, \tag{10.8}$$

and the naive implementation adds the three central estimates, $40 + 30 + 25 = 95$ MW. That number is the one Chapter 4 examined and found insufficient: it hides how wide each term is and how the terms move together. The honest output is a distribution, $F_{\text{fleet}} \sim p(F)$, obtained from the joint distribution of the three components. The rest of the chapter is about what that distribution is used for, and the worked example of Section 10.6.4 computes one.

10.3 Probabilistic forecast targets

A deterministic flexibility forecast cannot support a commitment decision. This section shows why, defines the quantile and reliability language that replaces it, and states the two properties a probabilistic forecast must have to be operationally useful.

10.3.1 Why point forecasts are dangerous

Suppose the forecast reports $\mathbb{E}[F] = 100$ MW for tomorrow at 18:00 and the utility requests 95 MW. On the face of it the margin is comfortable. Now suppose the standard deviation is $\sigma_F = 25$ MW. Figure 10.4 plots the implied normal density with the request marked: the horizontal axis is delivered flexibility, the curve is the forecast density, and the shaded area is the part of the distribution that falls short of the request. The shaded area is

$$\Pr(F < 95\,\text{MW}) = \Phi\left(\frac{95 - 100}{25}\right) = \Phi(-0.2) = 0.42, \tag{10.9}$$

so the commitment fails on about two days in five. The expected size of the shortfall follows from the same normal model, $\sigma[\varphi(z) + z\Phi(z)]$ with $z = -0.2$, which is 7.7 MW. A mean of 100 MW is compatible with a near coin-flip of failure, and the mean alone does not reveal it.

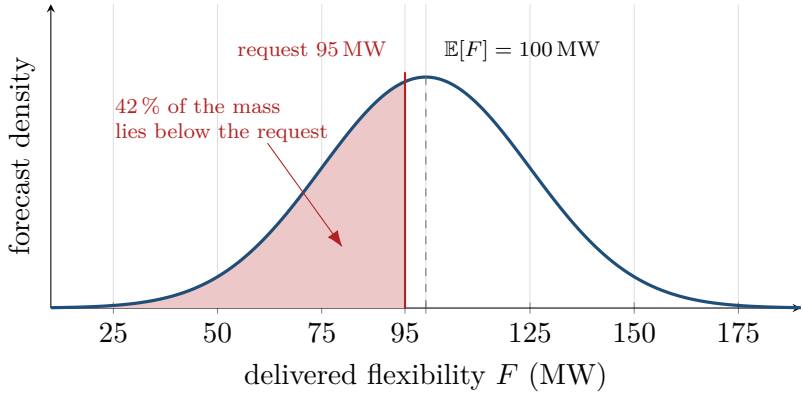


Figure 10.4: The danger of a point forecast. The density is $\mathcal{N}(100, 25^2)$ MW, the forecast for tomorrow at 18:00; the red line is a 95 MW request. The shaded area is equation (10.9): 42 % of the distribution is below the request, with an expected shortfall of 7.7 MW.

10.3.2 Quantile forecasts

The remedy is to forecast quantiles instead of a mean. A forecast might report

$$F_{0.10} = 70\,\text{MW}, \qquad F_{0.50} = 100\,\text{MW}, \qquad F_{0.90} = 130\,\text{MW}, \tag{10.10}$$

read as: a 10 % chance of less than 70 MW, a typical value of 100 MW, and a 10 % chance of more than 130 MW. If the distribution is normal these three numbers imply $\sigma_F = 30/1.2816 = 23.4$ MW, so they carry the same information as a mean and a spread; their advantage is that they remain meaningful when the distribution is skewed, which Section 10.2.5 showed it generally is. For a commitment decision the *lower* quantiles are the useful ones: the upper quantiles describe days on which the fleet has more capability than it promised, which creates no operational problem.

10.3.3 Reliable capacity

The commitment question of Chapter 4 has the form $\Pr(F \geq C) \geq 0.95$, and the largest such C is a single summary of the lower tail,

$$C_{95} = \max\{\, C : \Pr(F \geq C) \geq 0.95 \,\}, \tag{10.11}$$

the capacity exceeded with 95 % probability, which for a continuous distribution is the 5th percentile. A fleet with expected flexibility 100 MW might have $C_{95} = 75$ MW. Expected flexibility and reliable flexibility are different numbers, and the gap between them is set by the width of the distribution, not by its center. Figure 10.5 makes the point with two fleets of identical mean. Both panels show a forecast density against delivered flexibility; the upper panel is the wide fleet of equation (10.10), with its three quantiles marked, and the lower panel is a fleet half as uncertain, with its lower 5 % tail shaded. The wide fleet can promise only 61 MW at 95 % confidence; the narrow fleet, with the same 100 MW mean, can promise 75 MW. Reducing forecast uncertainty is therefore worth capacity directly, which is the commercial argument for better forecasting.

10.3.4 Calibration

A probabilistic forecast is only useful if its probabilities mean what they say. Suppose the model issues a 90 % interval of $[70, 120]$ MW. If the model is calibrated, realized flexibility falls inside that interval on about 90 % of occasions. If it falls inside on only 60 %, the model is overconfident,

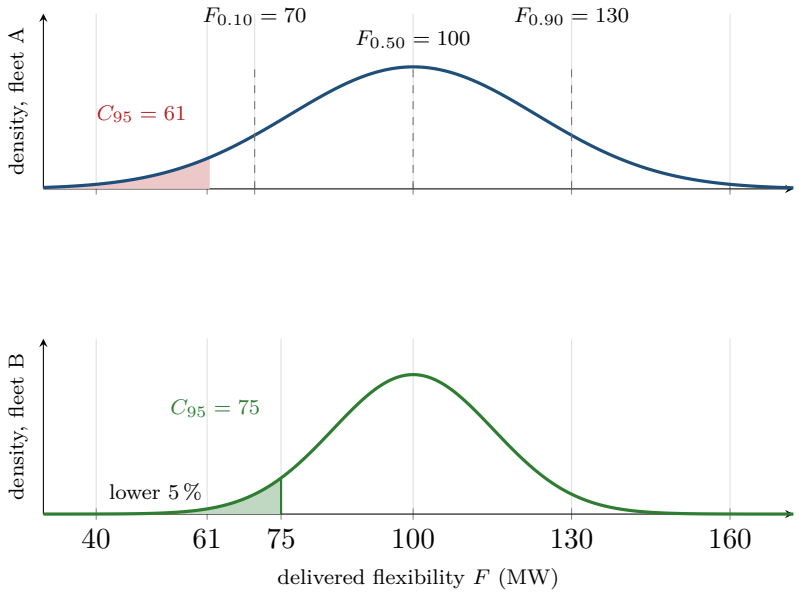


Figure 10.5: Two fleets with the same expected flexibility, 100 MW, and different uncertainty. Fleet A (top) has $\sigma = 23.4$ MW, the value implied by the quantiles of equation (10.10), and $C_{95} = 61$ MW. Fleet B (bottom) has $\sigma = 15.2$ MW and $C_{95} = 75$ MW. Both densities are normal; the shaded regions are the lower 5 % tails of equation (10.11).

and an optimizer that treats the stated interval as a 90 % interval will systematically take more risk than it believes it is taking. Under a normal model the size of the error is easy to state: a central 60 % interval spans $\pm 0.84\sigma$ and a central 90 % interval spans $\pm 1.64\sigma$, so an interval that achieves 60 % coverage while claiming 90 % understates the true standard deviation by a factor of about two.

Calibration is checked over many forecasts, by comparing the nominal coverage of each interval with the fraction of outcomes that actually fell inside it. Figure 10.6 plots that comparison: the horizontal axis is the stated coverage of an interval, the vertical axis is the coverage observed in the historical record, and the diagonal is perfect calibration. A curve lying below the diagonal is overconfident, because its intervals catch fewer outcomes than advertised; the marked point is the 90 %-interval case of the previous paragraph. Calibration and sharpness are separate goals: a forecast that always reports $[0, 500]$ MW is perfectly calibrated and useless, so the operational objective is the narrowest intervals that remain calibrated.

10.3.5 Forecast error is not symmetric

The second property is a matter of the loss function rather than of the model. Suppose the realized flexibility is 80 MW. A forecast of 60 MW causes the portfolio to be underused: an opportunity cost, and one the operator can usually absorb. A forecast of 100 MW leads to a commitment of 95 MW that cannot be met: penalties, a reliability event for the utility, and lost trust. The two errors have the same magnitude and very different costs, so the operational loss is asymmetric,

$$L(\hat{F}, F) = \begin{cases} \lambda_{\text{over}}\,(\hat{F} - F), & \hat{F} > F, \\ \lambda_{\text{under}}\,(F - \hat{F}), & \hat{F} < F, \end{cases} \qquad \lambda_{\text{over}} > \lambda_{\text{under}}. \tag{10.12}$$

Figure 10.7 draws it: the horizontal axis is the forecast error $\hat{F} - F$ and the vertical axis is the loss, so the function is two straight lines meeting at zero, with the over-forecast branch much

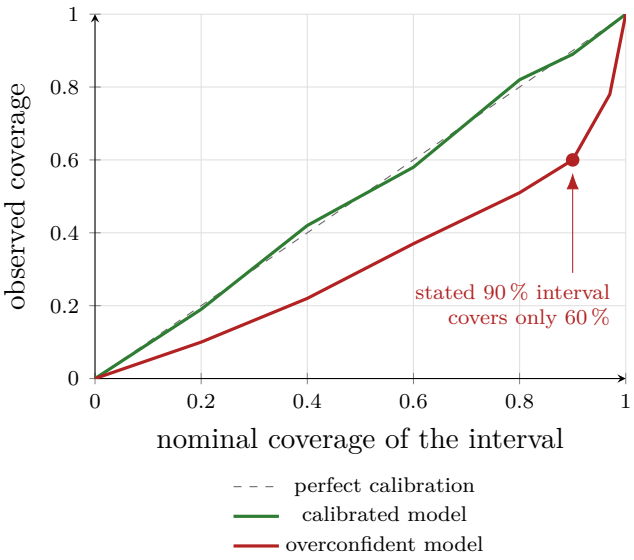


Figure 10.6: A calibration (reliability) diagram. Each point compares the stated coverage of a forecast interval with the fraction of historical outcomes that fell inside it. The dashed diagonal is perfect calibration; the red curve is overconfident, and its marked point is the $[70, 120]$ MW interval of Section 10.3.4. Curves are illustrative.

steeper. A symmetric criterion such as root mean square error treats the two branches alike and is therefore the wrong objective for this problem.

The asymmetry has a precise consequence. Minimizing the expected loss of equation (10.12) over $\hat{F}$ gives the first-order condition $\lambda_{\text{over}} \Pr(F < \hat{F}) = \lambda_{\text{under}} \Pr(F > \hat{F})$, so the optimal point forecast is the quantile at level

$$q^* = \frac{\lambda_{\text{under}}}{\lambda_{\text{under}} + \lambda_{\text{over}}}. \tag{10.13}$$

With $\lambda_{\text{over}} = 5$ and $\lambda_{\text{under}} = 1$, $q^* = 1/6$, and for the fleet of Figure 10.4 the best single number to report is $100 - 0.967 \times 25 = 76$ MW, not the mean. Read in reverse, equation (10.13) prices a reliability target: a 95 % reliability requirement, $q^* = 0.05$, is the behavior of an operator for whom over-forecasting is nineteen times as expensive as under-forecasting. This is the operational reason probabilistic forecasting pays: only a distribution lets the operator choose which quantile the cost structure actually calls for.

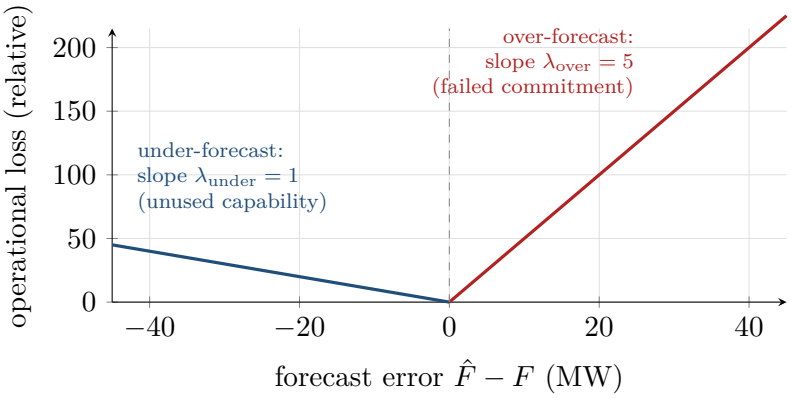


Figure 10.7: The asymmetric loss of equation (10.12) with $\lambda_{\text{under}} = 1$ and $\lambda_{\text{over}} = 5$ (relative units). The minimizing point forecast is the 1/6 quantile by equation (10.13), which for $\mathcal{N}(100, 25^2)$ MW is 76 MW.

10.4 Features and realizable flexibility

What goes into the model is as consequential as the model class. This section lists the features by resource class and then narrows the target from physical capability to what the platform can actually expect to collect.

10.4.1 Features are resource-specific, and history is a feature

Because the map g differs by class, so do the inputs. Table 10.2 lists them, grouped into state, exogenous and historical features, with the fleet-level features that apply to the aggregate in the last row. The practical architecture is one model per class, each with its own features, combined at fleet level, rather than a single model fitted to the aggregate: a single aggregate model cannot represent the fact that the same outdoor temperature raises thermostat flexibility and leaves battery flexibility untouched.

Table 10.2: Features of a flexibility forecast, by resource class.

Class	State features	Exogenous features	History features
Thermostats	T_{in}, T_{set}, recent HVAC runtime, duty cycle D	T_{out}, humidity, solar, hour, day type	prior events, override rate, verified response ratio
EVs	SOC_i, plug-in indicator, urgency U_i	hour, day type, workplace or home site	historical arrival and departure times, past charging profiles
Batteries	SOC, reserve SOC^{res}, inverter limits	site load forecast, PV forecast	cycling history, past dispatch, availability record
Fleet	counts available by class	weather field over the territory	record of delivered versus commanded response

The historical response ratio. The last column of Table 10.2 contains the single most useful correction available to a flexibility forecaster. If thermostat fleet A has historically delivered only 80 % of the reduction it was commanded, the ratio

$$r_{\text{hist}} = \frac{\text{verified response}}{\text{commanded response}} = 0.80 \tag{10.14}$$

belongs in the feature set, where the verified response is the measurement-and-verification quantity of Chapter 2. Including it changes what the model learns: the target becomes realizable flexibility rather than theoretical physical capability. The ratio is not a constant either, since it depends on event duration, weather and how recently the fleet was last called, so it is itself a small forecasting problem.

10.4.2 Physical and deliverable flexibility

Two quantities must be distinguished. Physical flexibility F^{phys} is what the devices could provide if every command arrived and every customer complied. Deliverable flexibility F^{del} is what the DERMS can expect once opt-outs, communication failures, control error and customer constraints are accounted for. Under a simple independent-factor model the second is the first

times a chain of derating factors,

$$F^{\mathrm{del}} = F^{\mathrm{phys}} \times \underbrace{0.95}_{\text{reachable}} \times \underbrace{(1 - 0.12)}_{\text{no opt-out}} \times \underbrace{0.80}_{r_{\mathrm{hist}}}, \tag{10.15}$$

which turns the 147 MW of equation (10.6) into about 98 MW, a derating of one third. The three factors are the fleet analogs of the availability factor a_i, the override probability p^{ovr} and the control error of Chapters 3 and 4. Always $F^{\mathrm{del}} \le F^{\mathrm{phys}}$, and the second is the number the utility is buying, so it is the quantity the model should be fitted to predict. Fitting to physical capability and then multiplying by a constant margin is worse, because the factors in equation (10.15) are themselves functions of weather, duration and dispatch history.

10.5 Endogeneity of the forecast target

The forecast target moves when the operator acts on the forecast. This section establishes the coupling, the scenario and multi-horizon constructions that cope with it, and the decomposition of the resulting error.

10.5.1 Previous dispatch changes the forecast

If a thermostat fleet was heavily curtailed from 15:00 to 16:00, then at 17:00 its thermal state is partly depleted: indoor temperatures sit near the top of the comfort band and the compressors are recovering, so a second curtailment yields less and lasts shorter. The same holds for every class. Batteries discharged earlier have a lower state of charge; EV charging deferred earlier has a higher urgency and therefore less remaining slack. In each case

$$F_{t+\tau} \text{ depends on } u_{t:t+\tau}, \tag{10.16}$$

the control history between now and then. The forecast is not independent of the optimization that consumes it.

10.5.2 The forecast-control coupling

This is a structural difference from weather forecasting. A weather forecast does not change tomorrow's weather; the forecaster is an observer of an exogenous process. A flexibility forecast is a forecast of a system the forecaster's own organization is driving, through the chain

$$u_t \longrightarrow x_{t+1} \longrightarrow F_{t+1}, \tag{10.17}$$

in which the control enters the state transition f and the state enters the flexibility map g. The forecast target is therefore partially endogenous, and the red input in Figure 10.2 is exactly this dependence. It also explains the drifting feasible band of Chapter 4, Figure 4.5: the edges of the band move with time because the fleet's own dispatch moves them.

10.5.3 There is no policy-free forecast

Concretely, suppose that with no dispatch at all the fleet would have $F(19{:}00) = 100$ MW. If 50 MW of battery flexibility is spent at 17:00, the batteries arrive at 19:00 with less stored energy and $F(19{:}00)$ might be 70 MW. Both numbers are correct forecasts of different futures. In general

$$F_{t+\tau} = F\big(x_{t+\tau}(u_{t:t+\tau})\big), \tag{10.18}$$

a function of a state that is itself a function of the control trajectory. A single number labeled "the flexibility forecast for 19:00" is not well defined without a stated dispatch assumption. At

this point the forecasting problem has acquired the structure of optimal control, which is the subject of the next chapter.

> **Principle 10.2.** *A flexibility forecast is conditional on a control trajectory. Reporting one unconditional number for a future hour is a category error; the forecast must be reported as a function of, or jointly with, the planned dispatch.*

10.5.4 Scenario forecasting

The practical construction is to forecast under several candidate control trajectories. Figure 10.8 shows the structure: one estimated present state on the left, three branches for three dispatch policies between now and 19:00, the resulting flexibility at 19:00 in the middle, and on the right whether a 70 MW request at 19:00 could then be met. With no dispatch before 19:00 the fleet retains 100 MW and the request is comfortable; with moderate use it retains 85 MW and the request is still met; with aggressive use it retains 60 MW and the request fails. The optimizer needs this table, not a single number, because its own decision determines which row it will live in.

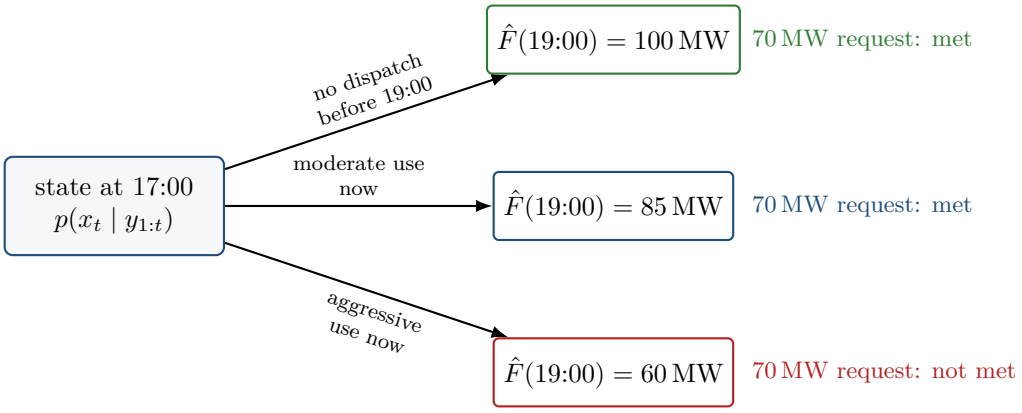


Figure 10.8: Scenario forecasting. The same estimated state at 17:00 yields three different flexibility forecasts for 19:00, one per candidate dispatch policy, and the feasibility of a later request depends on which policy is chosen. Values are illustrative and follow Section 10.5.4.

10.5.5 Multi-horizon forecasting

A virtual power plant needs the forecast at several horizons at once: $F_{t+5\,\mathrm{min}}$ for real-time correction, $F_{t+1\,\mathrm{h}}$ for intraday rebalancing, and $F_{t+24\,\mathrm{h}}$ for day-ahead commitment. These are different statistical problems even though they predict the same physical quantity, because the information that dominates changes with the horizon. Table 10.3 lists the three horizons with the information that dominates each, the decision each supports and the method that fits. The pattern across the rows is that current telemetry dominates at short horizons, where the state has had no time to move, while weather, calendar and customer behavior dominate at long horizons, where the initial state has largely been forgotten.

10.5.6 Updating forecasts in real time

Forecasts are revised as telemetry arrives. At 17:00 the forecast for 18:00 is 100 MW. By 17:30 three things have changed: fewer EVs are plugged in than assumed, the batteries are at a lower

Table 10.3: Three forecast horizons and what dominates each.

Horizon	Dominant information	Decision supported	Typical method
5 min	current telemetry: SOC, T_{in}, plug-in counts	real-time dispatch correction, regulation	persistence plus forward simulation of the state
1 h	telemetry plus near-term weather and the planned dispatch	intraday rebalancing, bid adjustment	regression or gradient boosting on telemetry features
24 h	weather forecast, calendar, customer behavior, fleet composition	day-ahead commitment, Chapter 4	probabilistic weather-driven model, scenarios

state of charge than the afternoon profile predicted, and thermostat load is slightly higher than forecast. The revision is the sum of the three effects,

$$\hat{F}(18{:}00) = 100 - 9 - 6 + 2 = 87\,\text{MW}, \tag{10.19}$$

where the EV and battery terms lower the estimate and the higher thermostat load raises it. The operating pattern is a loop, forecast, observe, update, which is the same loop as the model predictive control of Chapter 3; it is the forecasting half of that loop. Two properties make it work. The update is driven by the state, so the new telemetry acts through $p(x_t)$ rather than through a correction applied to the output; and the individual contributions are separable, as in equation (10.19), so the operator can see which class caused the revision.

10.5.7 Error decomposition

After the fact the forecast error is

$$e_F = F^{\text{act}} - \hat{F}, \tag{10.20}$$

and reporting its root mean square over a season is much less useful than attributing it. To first order the error separates into the five components of Table 10.4, each with a different remedy: a weather error is reduced by better weather inputs, a state error by better telemetry or a better dynamic model, an availability error by communication and enrollment work, a behavior error by program design, and a model error by a richer functional form. The practical value of the decomposition is that it converts a single aggregate metric into a list of actions, and the actions belong to different teams.

Table 10.4: Components of the flexibility-forecast error of equation (10.20).

Component	Source	How it is reduced
e_{weather}	the weather forecast was wrong, so $w_{t+\tau}$ was wrong	better numerical weather prediction, ensemble inputs
e_{state}	the predicted state was wrong: SOC, T_{in}, plug-in counts	more telemetry, better state estimation and dynamics
e_{avail}	devices could not be reached or did not respond	communication reliability, availability modeling
e_{behavior}	customers overrode, unplugged or behaved atypically	behavior models, program and incentive design
e_{model}	the map g is misspecified	richer model class, hybrid physics and machine learning

10.6 The forecast target and the interface to optimization

The last body section returns to the question that precedes every modeling choice, namely what exactly is being predicted, and then states the object the forecaster hands to the dispatch optimizer.

10.6.1 A simple thermostat forecasting model

A serviceable starting model for a thermostat fleet is a linear regression with a quadratic weather term,

$$F_t = \beta_0 + \beta_1 T_{\text{out},t} + \beta_2 T_{\text{out},t}^2 + \beta_3 P_t^{\text{base}} + \beta_4 H_t + \varepsilon_t, \tag{10.21}$$

where H_t encodes the hour of day, for example as a set of indicator variables or a pair of sine and cosine terms. The quadratic term is there for a physical reason: Section 10.2.2 showed that the weather effect is not monotonic once duration is taken into account, and $\beta_2 < 0$ lets the fitted curve turn over at high temperatures. The baseline term $\beta_3 P_t^{\text{base}}$ carries the duty cycle, since more running load means more to curtail. More expressive model classes are available, in rough order of flexibility: gradient boosting on the features of Table 10.2, recurrent networks for the sequence structure, transformers for long context, and hybrid models in which a physical building model supplies features to a learned correction. The choice of class is secondary. The first question is what the model is predicting.

10.6.2 Defining the operational target

The phrase "forecast flexibility" is ambiguous, and Table 10.5 lists five targets it can denote. They are genuinely different quantities: the first is a power that need last only minutes, the third is an energy, the fourth is a quantile rather than a central estimate, and the fifth is a vector indexed by location. A model fitted to one of them answers the others badly. The forecast target follows from the product the fleet is selling, which is the service specification of Chapter 4, so the ordering of work is: fix the service, define the target, then choose the model class.

Table 10.5: Five candidate forecast targets, all called "flexibility".

Target	Symbol	What it measures	Used for
Instantaneous shed	$F^{15\,\text{min}}$	largest reduction sustainable for 15 minutes	emergency response, fast reserve
Sustained response	$F^{1\,\text{h}}$	largest reduction sustainable for one hour	peak-hour shaving
Energy shift	E^{shift}	energy moved out of the window, in MWh	tariff arbitrage, load shaping
Reliable response	$F_{0.05} = C_{95}$	level exceeded with 95 % probability	capacity commitments
Locational response	$F(r, t, d)$	reduction available in zone r from class d	distribution constraints

10.6.3 The interface to optimization

What the forecaster hands to the dispatch optimizer is a joint distribution over the flexibility trajectory across the horizon,

$$p(F_{t:t+H} \,|\, \text{information at } t,\ u_{t:t+H})\,, \tag{10.22}$$

conditional both on present information and, by Section 10.5.3, on the planned dispatch. Three properties of the object matter to the optimizer. It is joint across time, because flexibility spent early is unavailable later. It is joint across classes, because the shared weather correlates them, which is the correlation result of Chapter 3, Figure 3.4, and the reason Chapter 4 insisted that firm capacity be computed from the joint distribution rather than assembled from marginals. And it is conditional on the control, so the optimizer and the forecaster solve one problem, not two. The optimizer's question, given that uncertain flexibility, what should be committed and dispatched, is the subject of the next chapter.

10.6.4 A worked portfolio example

Take tomorrow at 18:00 and the three-class forecast

$$F_{\mathrm{T}} \sim \mathcal{N}(50, 10^2), \qquad F_{\mathrm{E}} \sim \mathcal{N}(30, 8^2), \qquad F_{\mathrm{B}} \sim \mathcal{N}(25, 3^2) \tag{10.23}$$

in MW. The expected total is $50 + 30 + 25 = 105\,\mathrm{MW}$. Treating the three as independent, the variances add,

$$\mathrm{Var}(F) = 100 + 64 + 9 = 173\ \mathrm{MW}^2, \qquad \sigma_F = \sqrt{173} = 13.2\,\mathrm{MW}, \tag{10.24}$$

so $F \sim \mathcal{N}(105, 13.2^2)$. Figure 10.9 draws the three component densities and their sum against delivered flexibility, with the proposed 100 MW commitment marked. Two features of the picture are worth naming before interpreting it. The narrowest component, the battery term, has the tallest density, because a density integrates to one; and the sum is wider than any component while being narrower than the sum of the component widths, $10 + 8 + 3 = 21\,\mathrm{MW}$, which is what independence buys.

Now the decision: should the operator commit 100 MW? The shortfall probability is

$$\Pr(F < 100) = \Phi\!\left(\frac{100 - 105}{13.2}\right) = \Phi(-0.38) = 0.35, \tag{10.25}$$

with an expected shortfall of 3.1 MW, and the 95 %-reliable capacity is $C_{95} = 105 - 1.645 \times 13.2 = 83\,\mathrm{MW}$. A 100 MW commitment is therefore comfortable on the mean and fails about one day in three; if the contract requires 95 % reliability the commitment is 83 MW. If instead the three classes are perfectly correlated, the standard deviations add, $\sigma_F = 21\,\mathrm{MW}$, and C_{95} falls to 71 MW. The answer to "commit 100 MW?" is not contained in the expected value at all; it is contained in the lower tail, and the width of that tail depends on a correlation assumption that the forecast has to supply.

10.7 What to retain

Definition 10.1 (Flexibility forecast)**.** A flexibility forecast predicts what a fleet *could be made to do*, not what it will do. Because the answer depends on device states, customer behavior, weather, prior dispatch and future constraints, the useful output is the conditional distribution $p(F_{t:t+H} \mid \text{information}, u_{t:t+H})$ rather than a single number such as $\hat{F} = 100\,\mathrm{MW}$. At that point forecasting has become an input to control.

Table 10.6 collects the chapter's quantities with the section that defines each.

10.8 Exercises

Exercise 10.1. A fleet of 60 000 thermostats has $P_{\mathrm{AC}} = 4.0\,\mathrm{kW}$, an expected duty cycle of 0.55 and a curtailable fraction of 0.65. Compute the crude instantaneous flexibility of equation (10.6). Then apply the derating chain of equation (10.15) with a reachable fraction of 0.92, an opt-out rate of 0.15 and $r_{\mathrm{hist}} = 0.85$. What fraction of the physical figure survives?

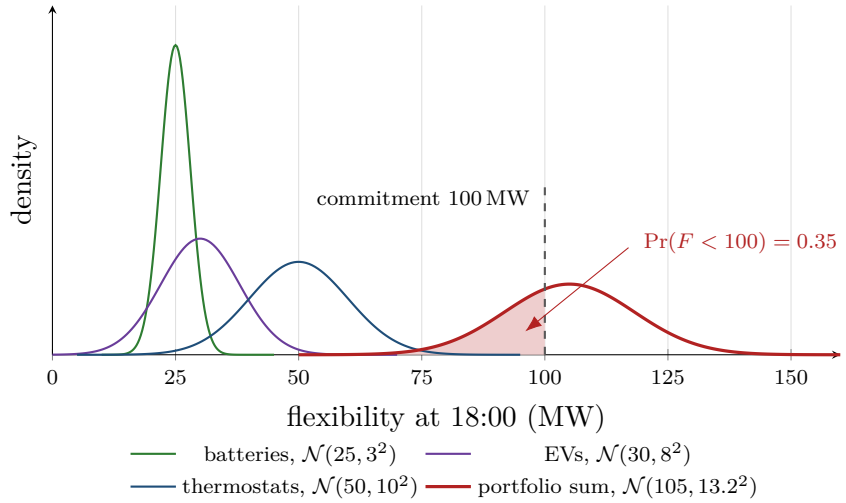


Figure 10.9: The portfolio of equation (10.23). The three component densities and their independent sum are plotted against delivered flexibility; the dashed line is the proposed 100 MW commitment and the shaded area is the shortfall probability of equation (10.25). All four densities are normal, with the parameters given in the legend.

Table 10.6: The objects of a flexibility forecast.

Quantity	Definition	Section
F_t	$P_t^{\text{base}} - P_t^{\text{ctrl}}$: load reduction, positive upward	10.1.2
$[F^{\downarrow}, F^{\uparrow}]$	feasible two-sided range of F_t	10.1.2
$F(t_{\text{start}}, \tau)$	largest change sustainable for duration τ	10.1.3
g	flexibility map, $F = g(x, w, c)$	10.1.4
F_q	the q quantile of the forecast distribution	10.3.2
C_{95}	largest C with $\Pr(F \geq C) \geq 0.95$	10.3.3
calibration	observed coverage equals nominal coverage	10.3.4
$L(\hat{F}, F)$	asymmetric loss; optimum at quantile $\lambda_{\text{under}}/(\lambda_{\text{under}} + \lambda_{\text{over}})$	10.3.5
r_{hist}	verified response divided by commanded response	10.4.1
F^{del}	F^{phys} after reachability, opt-out and response derating	10.4.2
e_F	$F^{\text{act}} - \hat{F}$, split into five components	10.5.7
$p(F_{t:t+H})$	the object handed to the dispatch optimizer	10.6.3

Exercise 10.2. For the portfolio of equation (10.23), find the commitment that satisfies $\Pr(F \geq P^{\text{commit}}) \geq 0.95$ under independence, and again under a common pairwise correlation of $\rho = 0.5$ between the three classes. State both answers and the revenue at stake if the commitment is paid at \$40 per MW.

Exercise 10.3. An operator faces $\lambda_{\text{over}} = 4$ and $\lambda_{\text{under}} = 1$, and the forecast distribution is $\mathcal{N}(100, 25^2)$ MW. Using equation (10.13), compute the loss-minimizing point forecast and compare it with the median and with C_{95}. What value of $\lambda_{\text{over}}/\lambda_{\text{under}}$ would make the loss-minimizing forecast equal to C_{95}?

Exercise 10.4. Using Figure 10.8, suppose the utility will request 80 MW at 19:00 and the aggregator can also earn revenue by dispatching before 19:00. Which of the three policies remain feasible? Then apply a real-time update in the style of equation (10.19) in which, at 18:30, EV availability costs 6 MW and the thermostat term adds 3 MW: which policies are still feasible?

10.9 Connection to the TERRA framework

Forecasting flexibility is, in a graph framework, not a separate model but a sequence of queries on the model that already exists.

- **Flexibility is a derived query, not a stored variable.** The map g of equation (10.4) is evaluated by pushing the component states through their bounds and summing at the grid node; the duration-dependent $F(t_{\text{start}}, \tau)$ of Figure 10.1 is a forward simulation under a constant request, recording the largest request that stays feasible. Nothing in the model stores "flexibility"; it is computed on demand.
- **The two-stage pipeline is inference followed by propagation.** Equation (10.5) is posterior inference of the present state from telemetry, then integration of the component dynamics forward under forecast disturbances, then the push-forward of that distribution through g. The non-normality noted in Section 10.2.5 is what a push-forward through saturating bounds produces, and it is a reason to carry the distribution rather than a mean and a variance.
- **Calibration is a test of the posterior predictive.** The diagram of Figure 10.6 checks whether the model's predictive intervals match observed frequencies. In a probabilistic framework this is the natural diagnostic, and failing it points at specific culprits: understated disturbance variance, neglected correlation between components, or a measurement model that assumes devices respond when they do not.
- **Realizability belongs in the measurement and actuation model.** The derating factors of equation (10.15) and the response ratio r_{hist} are not fudge factors but parameters of the map between the commanded edge flow and the delivered one. Estimating them from the verification record of Chapter 2 is parameter inference on that map, and it makes F^{del} the quantity the model predicts.
- **Endogeneity is just the dynamics.** Because the control enters the state transition, $F_{t+\tau}$ is conditional on $u_{t:t+\tau}$ by equation (10.18). The scenario tree of Figure 10.8 is three simulations of one model under three candidate control trajectories, which is also why no separate "capacity model" is needed: the capacity is read off the simulated physics.
- **Error attribution maps onto model components.** The five terms of Table 10.4 correspond to the model's parts: disturbance inputs, state estimates, the actuation model, the behavior model and the structure of g. A framework in which each part is explicit lets the error be attributed to a part instead of reported as one number.

In summary, a flexibility forecast is the predictive distribution of a derived quantity of a physical model, conditional on the operator's own planned actions. The next chapter takes that distribution as given and asks what to commit and how to dispatch against it.

Chapter 11

Dispatch and optimization: which DERs to use, when, and how much

11.1 The dispatch problem

The earlier chapters supplied the pieces of a decision without making it. Chapter 1 gave the device as a dynamical system with a state-dependent available power; Chapter 3 gave the platform that estimates fleet state and forecasts what the fleet can do; Chapter 4 gave the commitment the fleet has to honor; Chapters 5 to 7 gave three resource classes with different physics, different customer constraints and different time scales. What remains is the decision itself. The question of this chapter is

Given the fleet's state, its forecast capability and a target to follow, which resources should be used, at what time, and by how much?

Jargon: dispatch, stage cost, horizon, chance constraint, robust, scenario

Dispatch is the act of choosing and sending control actions to a set of resources for a period; the word comes from the bulk power system, where it names the choice of output level for each generator. The *stage cost* is the cost charged for one time step of a plan, written $c(x_t, u_t)$: the cost of being in state x_t and applying action u_t. The *horizon* H is how many steps ahead the optimizer plans. A *chance constraint* is a requirement that must hold with a stated probability rather than always, for example $\Pr(P^{\text{del}} \geq P^{\text{tgt}}) \geq 0.95$. A *scenario* is one complete realization of the uncertain inputs over the horizon, with a probability attached. A *robust* formulation requires feasibility for every value in a stated uncertainty set, with no probabilities at all.

The chapter establishes one idea in six steps: dispatch is not the problem of matching a number now, it is the problem of spending a finite, state-dependent resource over time. This opening section writes the naive version of the problem, shows what it leaves out, and introduces the stage cost that repairs it.

11.1.1 The simplest formulation

The grid asks the fleet for a trajectory $P^{\text{tgt}}(t)$. The aggregator holds three groups of controls: a thermostat command $u_{\text{T},t}$, an electric-vehicle charging command $u_{\text{E},t}$ and a battery command $u_{\text{B},t}$, each expressed as the megawatts of load change that group is asked to produce. The fleet's contribution is their sum,

$$P_{\text{VPP}}(t) = u_{\text{T},t} + u_{\text{E},t} + u_{\text{B},t}, \tag{11.1}$$

and the obvious objective is to make that sum follow the request,

$$\min_{u} \sum_{t} \left[P_{\mathrm{VPP}}(t) - P^{\mathrm{tgt}}(t) \right]^2 . \tag{11.2}$$

This is a least-squares tracking problem: make the fleet follow the requested trajectory. Each of the three aggregate commands is itself an allocation across thousands of devices, by the group rotation of Chapter 5, the urgency ordering of Chapter 7 or the power split of Chapter 6; for this chapter the three aggregates are the decision variables and the within-group allocation is taken as already solved. Equation (11.2) is a correct statement of part of the problem and an incomplete statement of the problem.

11.1.2 Tracking alone is not sufficient

Consider a 50 MW request from 18:00 to 20:00, and a fleet with 30 MW of thermostat flexibility, 25 MW of electric-vehicle flexibility and a battery fleet rated 20 MW with 100 MWh of usable energy, starting at SOC = 0.80 against a reserve of $\mathrm{SOC}^{\mathrm{res}} = 0.20$. Two dispatches meet the request exactly:

- **Option A**, battery-heavy: 20 MW from the batteries, 20 MW from the vehicles, 10 MW from the thermostats;
- **Option B**, battery-sparing: 5 MW from the batteries, 20 MW from the vehicles, 25 MW from the thermostats.

Both sum to 50 MW at every instant, so both give equation (11.2) the value zero. The objective cannot tell them apart. Figure 11.1 shows why they are not the same decision. Its top panel is the power delivered to the grid, identical under both options. Its middle panel is the battery fleet's state of charge over the same hours, with the reserve drawn as a horizontal line. Its bottom panel is the mean indoor temperature of the thermostat fleet, with the customers' upper comfort limit $T^{\max} = 25°\mathrm{C}$ drawn as a horizontal line.

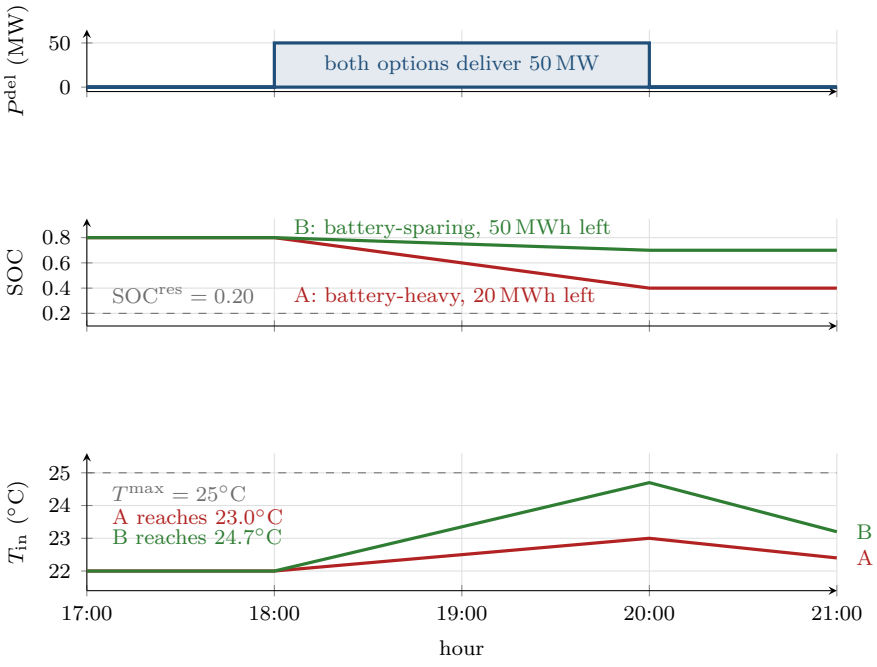


Figure 11.1: Two dispatches that deliver the same 50 MW. Option A takes 20 MW from a 20 MW, 100 MWh battery fleet and 10 MW from the thermostats; Option B takes 5 MW and 25 MW. The delivered power is identical; the battery energy and the comfort margin left at 20:00 are not. Trajectories are illustrative, drawn as straight ramps over the event.

The panels diverge. Option A ends the event at SOC = 0.40, only $0.20 \times 100 = 20$ MWh above the reserve, with the indoor temperature at 23.0 °C, two degrees below the limit. Option B ends with 50 MWh above the reserve and the temperature at 24.7 °C, three tenths of a degree below the limit. If a second request arrives at 20:00, Option A can answer it with thermostats and Option B can answer it with batteries, and if that second request is for fast, deep, short-duration support, only Option A's battery depletion hurts. Each option has spent a different part of the fleet's future capability. The same holds for the vehicles: curtailing charging raises the fleet's mean urgency U of Chapter 7 and removes the slack that would otherwise absorb a later request.

Key idea

Flexibility is a stock, not a flow. A dispatch that meets the target exactly can still be a bad decision, because it chose which part of the stock to spend.

11.1.3 The stage cost

To distinguish the two options the objective must charge for the use of flexibility, not only for the tracking error. The general form is the finite-horizon optimal-control problem

$$\begin{aligned} \min_{u_{0:H}} \quad & \sum_{t=0}^{H} c(x_t, u_t) \\ \text{subject to} \quad & x_{t+1} = f(x_t, u_t, w_t), \qquad u_t \in \mathcal{U}(x_t), \end{aligned} \tag{11.3}$$

with x the fleet state (every battery's state of charge, every building's indoor temperature, every vehicle's state of charge and deadline), u the control, w the disturbances (weather, base load, driver behavior) and $\mathcal{U}(x_t)$ the state-dependent set of admissible actions introduced in Chapter 1. The stage cost $c(x_t, u_t)$ is the price charged for one step: the cost of being in state x_t and applying u_t. For a DER fleet it decomposes into five terms,

$$c(x_t, u_t) = c_{\text{track}} + c_{\text{comfort}} + c_{\text{deg}} + c_{\text{EV}} + c_{\text{risk}}, \tag{11.4}$$

one for the grid's objective and four for the ways the fleet pays for meeting it: customer discomfort, battery wear, loss of driver margin, and exposure to future shortfall. The first term is what equation (11.2) already had; the rest are what it was missing. In Figure 11.1, Option A carries a large c_{deg} and a small c_{comfort}, Option B the reverse, and the optimizer chooses between them by comparing the two sums.

11.2 Costs and constraints

This section writes each term of the stage cost explicitly, separates the terms that are costs from the requirements that are constraints, and assembles the complete dispatch problem. Figure 11.2 shows the four penalty shapes side by side; each subsection below describes one of its panels.

11.2.1 Tracking cost

The tracking term is quadratic in the delivery error,

$$c_{\text{track}} = \left(P_{\text{VPP}}(t) - P^{\text{tgt}}(t)\right)^2 = e_t^2, \tag{11.5}$$

with e_t the tracking error of Chapter 3. Against a 50 MW target, delivering 48 MW costs $2^2 = 4$ and delivering 40 MW costs $10^2 = 100$: a five-times-larger error costs twenty-five times more. Panel (a) of Figure 11.2 is this parabola, with the two points marked. The square is

a modeling choice with an operational meaning. A grid service is usually procured with a tolerance band, and a small error inside the band is harmless while a large error outside it risks the whole commitment; squaring makes the optimizer spend aggressively to remove large errors and accept small ones. It also makes the term symmetric, so over-delivery is penalized as much as under-delivery, which is appropriate when the service is a trajectory to be followed and inappropriate when it is a floor to be exceeded.

11.2.2 Comfort cost

Thermostat curtailment is paid for in indoor temperature. The standard form is quadratic in the deviation from the customer's preferred setpoint,

$$c_{\text{comfort}} = \lambda_{\text{T}} \left(T_{\text{in}} - T_{\text{set}}\right)^2, \tag{11.6}$$

summed over the fleet or evaluated at the fleet mean. The weight λ_{T} sets how strongly discomfort is penalized relative to tracking error, and it is the term that encodes the statement that one megawatt of thermostat flexibility is not free. Panel (b) of Figure 11.2 draws the parabola for a preferred temperature of 22 °C and shades the region beyond $T^{\max} = 25°\text{C}$, which is not a cost at all but a constraint, as Section 11.2.5 explains. Inside the band the quadratic form says that the first half-degree of drift is nearly free and the last half-degree is expensive, which matches both the survey evidence on thermal comfort and the override behavior of Chapter 5: customers tolerate small deviations and intervene at large ones.

11.2.3 Degradation cost

Battery use is paid for in cell life. The simplest usable form charges for throughput,

$$c_{\text{deg}} = \lambda_{\text{B}} \left|P_{\text{B}}\right| \Delta t, \tag{11.7}$$

which is linear in the magnitude of the battery power and therefore proportional to the energy moved in the step, in either direction. Panel (c) of Figure 11.2 is the resulting V shape. A linear throughput charge is a crude model of a process that depends on depth of discharge, cycle count, cell temperature and the state-of-charge range in which the cycling happens, as Chapter 6 described; a more sophisticated c_{deg} makes those dependencies explicit, at the price of a non-convex term. The structural point survives every refinement: battery cycling has a cost, so a dispatch that reaches for the battery first because the battery has no comfort penalty is reaching for something that is not free either.

11.2.4 Electric-vehicle margin cost

The driver's requirement $\text{SOC}(T^{\text{dep}}) \geq \text{SOC}^{\text{req}}$ is a hard requirement, and Section 11.2.5 keeps it as one. But a dispatch that satisfies it with no margin is more fragile than one that satisfies it comfortably, because the departure time may move earlier or the charger may fault. A one-sided penalty expresses this,

$$c_{\text{EV}} = \lambda_{\text{E}} \left[\text{SOC}^{\text{req}} - \text{SOC}_t\right]_+^2, \tag{11.8}$$

where $[\,\cdot\,]_+$ is the positive part, so the term is zero once the vehicle is already above its requirement and grows quadratically below it. Panel (d) of Figure 11.2 shows the hinge at SOC^{req}. The effect on the dispatch is to push curtailment toward the vehicles that are far above their requirement and away from those that are close to it, which is the urgency ordering of Chapter 7 arrived at from the cost side rather than from the scheduling side.

The four panels of Figure 11.2 have four different shapes, and that is the substance of the dispatch problem: the optimizer is choosing among resources whose marginal cost curves differ in form, not only in level. A symmetric parabola, a parabola truncated by a wall, a V and a hinge do not substitute for one another at a fixed exchange rate.

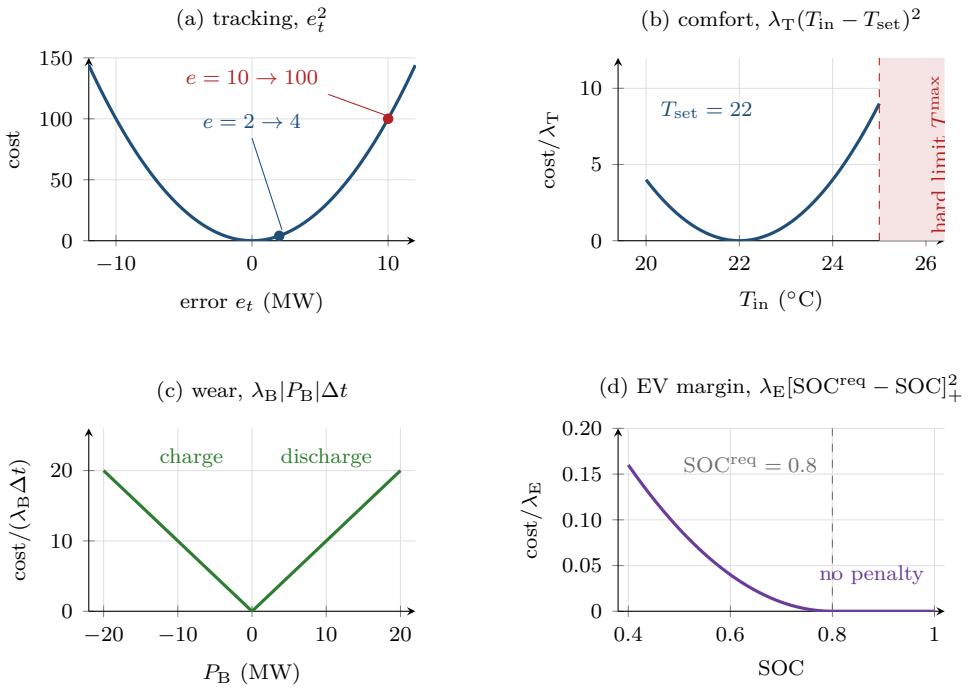


Figure 11.2: The four penalty shapes of equation (11.4), drawn from equations (11.5), (11.6), (11.7) and (11.8) with $T_{set} = 22°C$, $T^{max} = 25°C$ and $SOC^{req} = 0.8$. The weights are divided out so the shapes can be compared; the shaded region of panel (b) is a constraint, not a cost.

Weights are exchange rates. The four terms of equation (11.4) carry different units: MW^2 for tracking, $°C^2$ for comfort, MWh for throughput, and a dimensionless squared state-of-charge deficit for the vehicle margin. They can be added only after each is converted to one currency, and the weights λ_T, λ_B and λ_E are those conversion rates. Calibrating them is a business decision disguised as a modeling parameter: the correct λ_B is the dollar cost of a megawatt-hour of battery life, the correct λ_T is the dollar cost of a degree of customer dissatisfaction including its effect on enrollment, and both must be expressed in the same units as the penalty for missing the commitment. A sound calibration fixes the ratios so that, at the intended operating point, the marginal cost of the last megawatt taken from each resource is equal; otherwise the optimizer systematically overuses whichever resource was accidentally made cheap.

> **Principle 11.1.** *Every soft term in the dispatch objective is a price, and a dispatch is only as good as its prices. A weight set by convenience rather than by valuation does not remove a trade-off, it hides it.*

11.2.5 Costs and constraints are different objects

A cost says: avoid this if possible. A constraint says: you may not do this. The distinction is not stylistic, because the two enter the optimization differently and fail differently. Keeping the indoor temperature near 22 °C is a preference, represented by $(T_{in} - 22)^2$ in the objective and traded against tracking; keeping it at or below 25 °C is a limit, represented by $T_{in} \le T^{max}$ in the constraint set and never traded. A battery reserve $SOC \ge SOC^{res} = 0.20$ is likewise a constraint: the owner has withheld that energy, and no tracking benefit buys it. Table 11.1 puts the chapter's quantities on both sides of the line.

Two consequences follow. First, a problem with hard constraints can be infeasible, and an operational system must decide in advance what to do when it is; the usual answer is to relax

Table 11.1: Which requirements enter the dispatch as soft costs and which as hard constraints.

Quantity	Soft cost (avoid if possible)	Hard constraint (may not be violated)
Indoor temperature	$\lambda_{\mathrm{T}}(T_{\mathrm{in}} - T_{\mathrm{set}})^2$	$T^{\min} \le T_{\mathrm{in}} \le T^{\max}$, e.g. $T_{\mathrm{in}} \le 25\,^{\circ}\mathrm{C}$
Battery energy	$\lambda_{\mathrm{B}}\lvert P_{\mathrm{B}}\rvert \Delta t$ for wear	$\mathrm{SOC}^{\mathrm{res}} \le \mathrm{SOC} \le \mathrm{SOC}^{\max}$
Battery power	none	$-P^{\mathrm{dis}} \le P_{\mathrm{B}} \le P^{\mathrm{ch}}$
Vehicle charge	$\lambda_{\mathrm{E}}[\mathrm{SOC}^{\mathrm{req}} - \mathrm{SOC}]_+^2$ for lost margin	$\mathrm{SOC}(T^{\mathrm{dep}}) \ge \mathrm{SOC}^{\mathrm{req}}$ at departure
Tracking	$(P_{\mathrm{VPP}} - P^{\mathrm{tgt}})^2$	none, unless the contract sets a band
Transformer loading	none	$\sum_i P_i \le P_{\mathrm{tr}}^{\max}$

the tracking term, that is, to miss the target rather than violate a customer limit. Second, moving a requirement from the constraint side to the cost side changes the product being sold. A thermostat program that charges for comfort rather than bounding it will occasionally take a house to 27 °C when the tracking benefit is large enough, and the customer did not agree to that.

11.2.6 The basic virtual-power-plant dispatch problem

Collecting the pieces gives the problem the DERMS of Chapter 3 actually solves,

$$\begin{aligned} \min_{u_{t:t+H}} \quad & \sum_{k=t}^{t+H} \Big[\big(P_{\mathrm{VPP}}(k) - P^{\mathrm{tgt}}(k)\big)^2 + \lambda_{\mathrm{T}}\, c_{\mathrm{T},k} + \lambda_{\mathrm{B}}\, c_{\mathrm{B},k} + \lambda_{\mathrm{E}}\, c_{\mathrm{E},k} \Big] \\ \text{subject to} \quad & T_{\mathrm{in}}(k+1) = f_{\mathrm{T}}\big(T_{\mathrm{in}}(k),\ u_{\mathrm{T},k},\ T_{\mathrm{out}}(k)\big), \\ & \mathrm{SOC}_{k+1} = f_{\mathrm{B}}\big(\mathrm{SOC}_k,\ P_{\mathrm{B},k}\big), \\ & \mathrm{SOC}_{\mathrm{EV},k+1} = f_{\mathrm{E}}\big(\mathrm{SOC}_{\mathrm{EV},k},\ P_{\mathrm{EV},k}\big), \\ & \text{and every bound of Table 11.1.} \end{aligned} \tag{11.9}$$

The three state equations are the thermostat model of Chapter 5, the battery accumulator of Chapter 6 and the vehicle accumulator of Chapter 7. This is the virtual-power-plant control problem in compact form: one objective, three state equations, one bound set. Everything remaining in the chapter either chooses the horizon H, decides what to do about the unknown w, or interprets the solution.

11.3 Horizon and model predictive control

Problem (11.9) has a parameter that has not yet been discussed, the horizon H. This section shows that the horizon changes the answer, names the scheme that handles it, and explains why the resulting plan is deliberately discarded one step after it is made.

11.3.1 The optimization horizon

Suppose the fleet of Section 11.1.2 is asked for 50 MW from 18:00 to 22:00, four hours rather than two. Optimize only the next five minutes and the cheapest answer is to use the batteries: over five minutes a battery incurs a small throughput charge and no comfort cost at all, while thermostat curtailment starts warming houses immediately. Optimize over four hours and the

answer changes, because the four-hour problem can see that 20 MW of battery power exhausts the usable energy before the request ends. Figure 11.3 draws both plans. Its top panel is battery power against time; its bottom panel is the battery fleet's state of charge, with the reserve as a horizontal line.

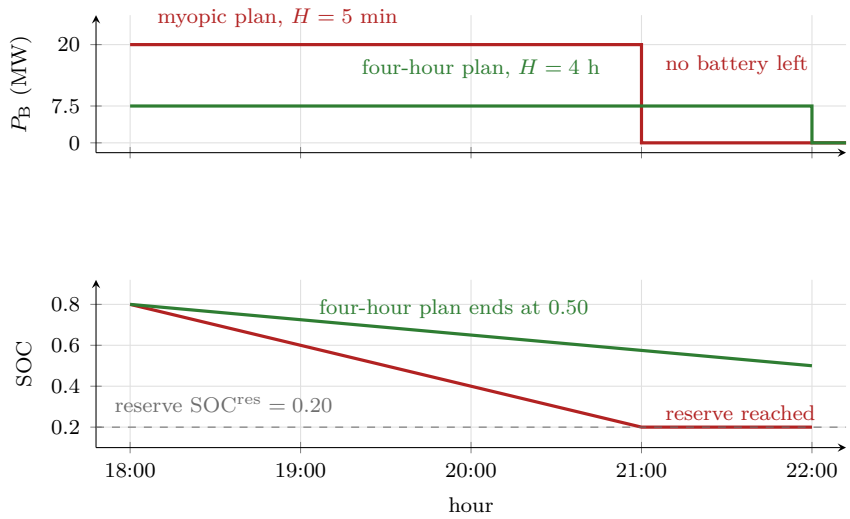


Figure 11.3: The horizon changes the plan. A 20 MW, 100 MWh battery fleet serves part of a 50 MW request from 18:00 to 22:00. The myopic plan discharges at the 20 MW rating and reaches the reserve after $0.60 \times 100/20 = 3$ h; the four-hour plan paces the battery at $30/4 = 7.5$ MW and ends at SOC $= 0.50$.

The myopic plan discharges at 20 MW and reaches the reserve at 21:00, after which the battery contributes nothing and the last hour of the request must be met by thermostats and vehicles that have meanwhile been left idle and are now the only resources available. The four-hour plan spreads 30 MWh of battery energy evenly, 7.5 MW for four hours, and ends at SOC $= 0.50$. Both plans are optimal for the problem they were given; they differ because the myopic problem cannot represent the consequence of its own action. The general statement is a rule for choosing H: the horizon must be at least as long as the energy time constant of the slowest resource the optimizer is allowed to spend. For the battery fleet that is $E^{\max}/P^{\max} = 100/20 = 5$ h; for a thermostat fleet it is the building time constant RC of Chapter 5, a few hours; for a vehicle fleet it is the plugged-in window, up to thirteen hours in the example of Chapter 7. A horizon shorter than that time constant makes the optimizer treat an energy-limited resource as though it were only power-limited, which is exactly the error in the myopic plan.

Principle 11.2. *The horizon must cover the time over which the fleet's state can be exhausted. Below that length the optimizer cannot see the cost of what it is spending, and it will spend it.*

11.3.2 Model predictive control

Problem (11.9) is solved repeatedly, in the loop Chapter 3 introduced as model predictive control. At time t the platform observes the telemetry and estimates the fleet state x_t; it forecasts the disturbances $w_{t:t+H}$ over the horizon; it solves the optimization; it applies only the first action u_t; it advances to $t+1$, collects new telemetry, updates the state estimate and the forecasts, and solves again. The cycle is *observe, forecast, optimize, act, repeat.* The receding

window it produces is drawn as Figure 3.7 of Chapter 3 and is not redrawn here: at each step a plan over H steps is computed, its first action is applied, and the rest is discarded.

Two properties of the scheme matter for dispatch. It is the mechanism by which a long horizon buys foresight without requiring the forecast to be correct far ahead: the hours at the end of the horizon only have to be good enough to shape the first action, and they will be recomputed many times before they arrive. And it closes the loop on state rather than on output, because the quantity fed back is the estimated fleet state, from which the admissible set $\mathcal{U}(x_t)$ is recomputed every step.

11.3.3 The plan is not executed to its end

At 16:00 the optimizer builds a plan through 22:00. By 17:00 four things have changed. The weather forecast has moved, so the thermostat fleet's available flexibility is no longer what was assumed. Vehicle availability has changed, as cars arrive earlier or later than their forecast arrival distributions implied. The measured battery state of charge differs from the plan's prediction, because the fleet did not track its schedule exactly. And some customers have overridden their thermostats, removing themselves from the controllable set. The plan built at 16:00 is now a plan for a fleet that no longer exists.

Re-optimizing continuously is the response: plan far ahead so that the current action is informed, act one step, then rebuild the plan from measurements. The planned trajectory is best understood not as a schedule to be executed but as a device for making the present action account for the future. Nothing beyond the first step is a commitment.

11.4 Dispatch under uncertainty

Problem (11.9) contains the disturbance w, which is not known when the problem is solved. This section works through the three ways of handling that, from the simplest to the most conservative, and ends with the choice among them.

11.4.1 Deterministic dispatch and how it fails

The simplest treatment replaces the unknown disturbance by a single forecast $\hat{w}_t$ and solves

$$x_{t+1} = f(x_t, u_t, \hat{w}_t) \tag{11.10}$$

as though the forecast were exact. If the flexibility forecast says $\hat{F}_{\mathrm{T}} = 40\,\mathrm{MW}$ from thermostats, $\hat{F}_{\mathrm{E}} = 25\,\mathrm{MW}$ from vehicles and $\hat{F}_{\mathrm{B}} = 20\,\mathrm{MW}$ from batteries, the optimizer works with 85 MW of capability and treats each number as a fact. The resulting problem is a single deterministic optimization, convex in the cases above and cheap to solve, which is why deterministic dispatch is the default in practice.

Its failure mode is direct. Against a 75 MW target the optimizer sees 85 MW of capability, commits 75 MW and plans a comfortable 10 MW of headroom. The event arrives and the thermostat fleet delivers 25 MW rather than 40 MW, because the afternoon was cooler than forecast and the compressors were not running as much as assumed. Realized capability is $25 + 25 + 20 = 70\,\mathrm{MW}$ against a 75 MW commitment, so the fleet is 5 MW short with nothing left to call on. The plan was optimal for a fleet that did not materialize, and no amount of re-optimization at 18:00 creates 5 MW that does not exist. The deterministic formulation produced a commitment with no stated reliability, and therefore one that was wrong roughly half the time.

11.4.2 Derating

The first repair is a safety margin applied to the forecast before the optimizer sees it,

$$F^{\text{usable}} = \alpha \hat{F}, \qquad \alpha < 1, \tag{11.11}$$

which is the derating of Chapter 4 applied inside the dispatch rather than at the commitment stage. With $\alpha = 0.85$ the 85 MW forecast becomes 72.25 MW of usable capability, and a commitment of 70 MW would have survived the realized 70 MW. Derating is one line of code and it works. It is also crude in a specific way: α is a single number applied to resources whose forecast errors have different distributions, so it over-derates the reliable resources and under-derates the unreliable ones; and it fixes no reliability level, so nobody can say what probability of shortfall $\alpha = 0.85$ buys. Section 11.4.4 computes that probability for one case and finds it is not a round number.

11.4.3 Chance constraints

The honest version of a safety margin is a probability requirement written into the problem,

$$\Pr(P^{\text{del}} \geq P^{\text{tgt}}) \geq 0.95, \tag{11.12}$$

read as: choose the dispatch so that the probability of meeting the target is at least 95 %. This is a chance constraint, the same object that bounded the vehicle deadline in Chapter 7 and firm capacity in Chapter 4, now applied to the dispatch decision itself. Its practical importance is that it is the point at which a probabilistic forecast enters the optimization directly: the constraint cannot be evaluated from a point forecast at all, because $\Pr(\cdot)$ requires a distribution. The reliability level becomes an explicit parameter, chosen by whoever owns the commercial risk, rather than an accident of a derating factor.

11.4.4 A Gaussian example

Let the fleet's total capability over the service window be forecast as $F \sim \mathcal{N}(100, 10^2)$ MW: an expected 100 MW with a standard deviation of 10 MW. The dispatch that satisfies equation (11.12) at 95 % commits the fifth percentile of that distribution,

$$P^{\text{commit}} = 100 - 1.645 \times 10 = 83.55 \approx 83.6\,\text{MW}, \tag{11.13}$$

where 1.645 is the standard normal 95 % quantile. Figure 11.4 draws the forecast density against capability in megawatts, shades the 5 % lower tail, and marks the resulting commitment, the ad hoc derate of Section 11.4.2 and the mean.

Two readings of the figure matter. First, the expected capability is 100 MW and the 95 %-reliable capability is about 84 MW; the 16 MW gap is the price of reliability, and it is set by the forecast's spread, not by the forecast's center. Second, the derate of Section 11.4.2 is drawn beside it: $\alpha = 0.85$ gives 85 MW, whose shortfall probability is $\Phi((85 - 100)/10) = \Phi(-1.5) = 0.067$, so the derate delivers 93.3 % reliability rather than 95 %. The derate is a chance constraint with an unstated and uncontrolled confidence level. Had the spread been $\sigma = 20$ MW, the same $\alpha = 0.85$ would deliver only $1 - \Phi(-0.75) = 0.77$, while the 95 % requirement would demand a commitment of $100 - 1.645 \times 20 = 67.1$ MW. The derating factor carries no information about which case it is in.

11.4.5 Scenario optimization

A density is not always the convenient description. When the uncertainty is driven by a few discrete futures, the natural representation is a set of scenarios, each a complete realization of the disturbances over the horizon with a probability attached. For the thermostat fleet of Section 11.4.1, three scenarios describe the afternoon:

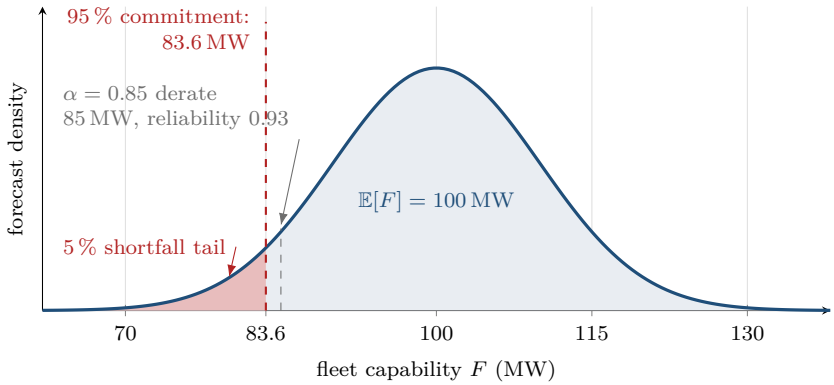


Figure 11.4: A probabilistic capability forecast $F \sim \mathcal{N}(100, 10^2)$, with density $\varphi((F - 100)/10)/10$. The 95 %-reliable commitment is the fifth percentile, 83.6 MW; the ad hoc 15 % derate commits 85 MW, whose reliability is $1 - \Phi(-1.5) = 0.933$.

- **Scenario 1**, hot weather, compressors running hard: $F_{\mathrm{T}} = 50$ MW, probability 0.25;
- **Scenario 2**, moderate weather: $F_{\mathrm{T}} = 35$ MW, probability 0.50;
- **Scenario 3**, high opt-out rate: $F_{\mathrm{T}} = 25$ MW, probability 0.25.

The optimizer is then asked for one dispatch that performs well across all three,

$$\min_{u} \; \mathbb{E}_{\omega}\big[J(u,\omega)\big] = \min_{u} \; \sum_{\omega} p_{\omega}\, J(u,\omega), \tag{11.14}$$

with ω indexing the scenario. This is stochastic optimization. The decision variable is a single u, not one per scenario, because the dispatch has to be chosen before the scenario is known. Figure 11.5 plots the three scenario levels of thermostat flexibility against time over the service window, with the forecast mean marked and the interval between the extreme scenarios shaded.

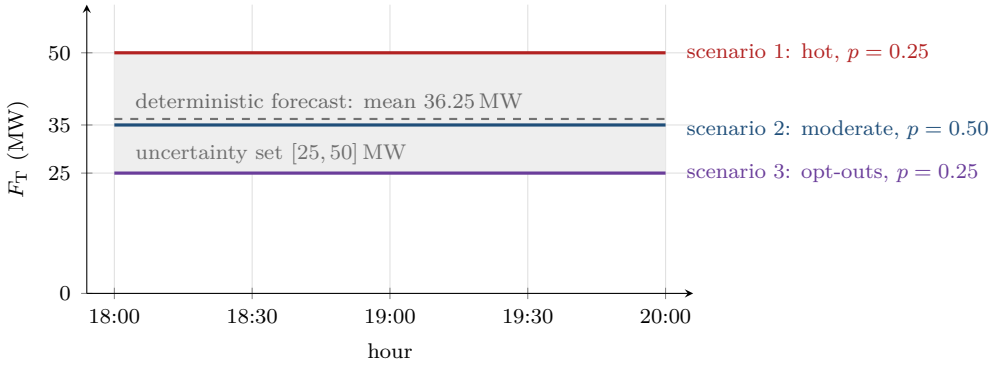


Figure 11.5: Three scenarios for thermostat flexibility over an 18:00 to 20:00 window, with probabilities 0.25, 0.50 and 0.25. The dashed line is their mean, 36.25 MW, which a deterministic formulation would use; the shaded band is the uncertainty set a robust formulation would use. Levels are held constant over the window for clarity.

The three scenario levels have mean 36.25 MW and standard deviation 8.9 MW, so the scenario set carries roughly the same information as a density with that spread. The difference is that scenarios can represent futures that are not a smooth cloud around a mean, for example a forecast that the weather either does or does not break.

Example 11.1. Take the vehicles and batteries as firm at $25 + 20 = 45$ MW, so total capability is $45 + F_{\mathrm{T}}$, that is 95, 80 or 70 MW in the three scenarios. Let the service pay \$40 per MW of

commitment and charge \$120 per MW of shortfall. The expected net value of a commitment P^{commit} is

$$40\,P^{\text{commit}} - 120\sum_{\omega} p_\omega \big[P^{\text{commit}} - \text{cap}_\omega\big]_+ .$$

At $P^{\text{commit}} = 95$ the expected shortfall is $0.5(15) + 0.25(25) = 13.75$ MW and the value is $3800 - 1650 = \$2150$. At $P^{\text{commit}} = 80$ the expected shortfall is $0.25(10) = 2.5$ MW and the value is $3200 - 300 = \$2900$. At $P^{\text{commit}} = 70$ no scenario falls short and the value is \$2800. The expected-value function is piecewise linear with kinks at the three capability levels, so its maximum sits at one of them, here at 80 MW. A deterministic formulation using the mean capability would have committed 81.25 MW, worth \$2838; a robust formulation would have committed 70 MW.

11.4.6 Robust optimization

The third philosophy removes probabilities entirely. The uncertain quantity is only required to lie in a set, for example $F_{\text{T}} \in [25, 50]$ MW, the shaded band of Figure 11.5, and the dispatch is required to work for every value in that set,

$$\min_{u} \; \max_{F \in \mathcal{F}} \; J(u, F), \tag{11.15}$$

which for a monotone problem means planning on the adverse endpoint $F_{\text{T}} = 25$ MW. The result is conservative: in the example above it leaves 10 MW of commitment unsold relative to the stochastic answer, and \$100 of expected value on the table. Robust formulations are the right choice when failure is expensive out of proportion to its probability, which is the usual situation for a reliability service whose non-delivery triggers a penalty, a disqualification from the program or a load-shedding event. They are the wrong choice when the uncertainty set is wide and the penalty is mild, because then the adverse endpoint is a future that will almost never happen and the dispatch is paying for it every day.

11.4.7 The three philosophies

Table 11.2 sets the three approaches side by side with the commitment each produces in the running example, and names the failure mode each accepts.

Table 11.2: Three ways to dispatch under uncertainty, with the commitment each chooses for the three-scenario example of Section 11.4.5.

Philosophy	What the optimizer is given	Commitment	Failure mode accepted
Deterministic	one forecast $\hat{F}$, treated as exact	81.25 MW	fails whenever the realization falls below the forecast, roughly half the time
Stochastic	scenarios with probabilities, or a density $p(F)$	80.00 MW	best on average; still fails in the adverse scenario, by design
Robust	an uncertainty set $\mathcal{F} = [25, 50]$ MW	70.00 MW	never fails inside the set; leaves capability unsold every day

The three are not rival theories but three settings of one dial, the weight placed on the adverse tail. A chance constraint at 95 % sits between the stochastic and robust extremes and is the usual operational compromise, because it names the reliability the commercial contract actually requires.

11.5 Coordinating heterogeneous resources

Everything so far has treated the three resource classes symmetrically. They are not symmetric. This section works through how a dispatch assigns roles among thermostats, vehicles and batteries, why the assignment changes with the fleet state, and what happens when the fleet underperforms its plan.

11.5.1 The dispatch of three resource classes

Take a 50 MW target from 18:00 to 20:00 with $F_{\mathrm{T}} = 30\,\mathrm{MW}$ of thermostat flexibility, $F_{\mathrm{E}} = 25\,\mathrm{MW}$ of vehicle flexibility and $F_{\mathrm{B}} = 20\,\mathrm{MW}$ of battery power available. The sum, 75 MW, comfortably exceeds the target, so the question is not whether the target can be met but which 50 of the 75 megawatts to use. Each class brings a different liability. Thermostat flexibility is cheap per megawatt and large, but it decays within the event as indoor temperatures drift toward the comfort limit, by the mechanism of Chapter 5. Vehicle flexibility is large and nearly free in the middle of a long plug-in window, but it is bounded by departure deadlines that approach during the event. Battery power is precisely controllable and immediately available, but it is limited by stored energy and it incurs the degradation charge of equation (11.7). There is no universal ordering among them; the right one depends on the current states and on what the fleet expects to be asked for next.

11.5.2 Merit order is state-dependent

Conventional dispatch orders generators by marginal cost into a merit order and fills the demand from the cheapest upward. The DER analog is tempting and partly correct. On a normal evening, with vehicles that have large slack and batteries near full, the conceptual order is: first the low-impact vehicle shifting, which costs the driver nothing; second a moderate thermostat curtailment, which costs a fraction of a degree; third the batteries, used to close whatever tracking error remains. Run the same fleet two hours later, with vehicle deadlines close and the weather mild, and the order reverses: the vehicles can no longer be delayed without risking a deadline, the thermostats have little band left, and the batteries become the first resource rather than the last. Table 11.3 states both orderings.

Table 11.3: The merit order of DER flexibility reverses with the fleet state. Both rows describe the same fleet at different times.

Fleet state	First	Second	Third
Early evening: vehicles have large slack, batteries near full, comfort bands wide	vehicle shifting, lowest customer impact	moderate thermostat curtailment	batteries, to close the tracking error
Late evening: vehicle deadlines near, comfort bands nearly spent	batteries, no customer impact	the few vehicles that still have slack	thermostats, only while the band allows

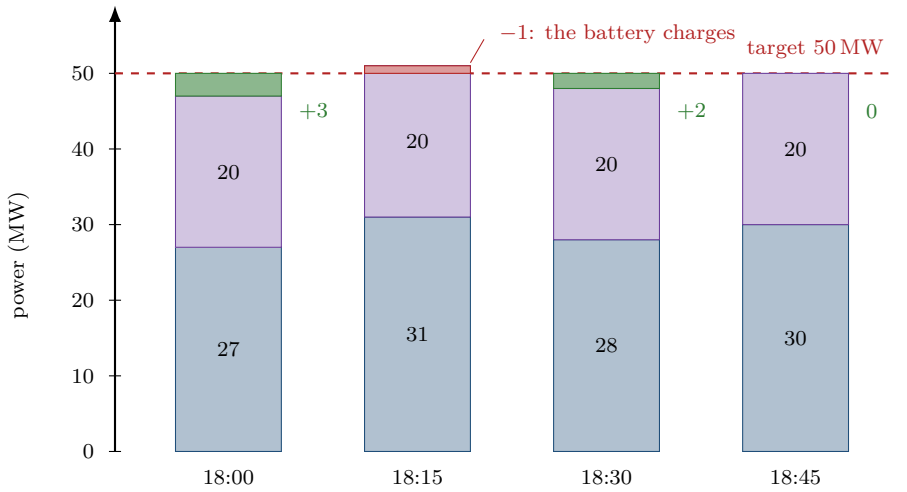
The lesson is that a static merit order is the wrong data structure for a DER fleet. The marginal cost of a megawatt from any class is a function of that class's current state, so the ordering has to be recomputed every dispatch cycle from the state estimate, which is what solving problem (11.9) with the current x_t does.

11.5.3 Batteries as fast balancing resources

Batteries are the only class in the fleet whose output can be commanded to a setpoint and reached within seconds, so the natural assignment is to let the slow and uncertain resources carry the bulk of the energy and let the batteries absorb the difference between plan and realization. Suppose the thermostat response was forecast at 30 MW and the realized response over four successive fifteen-minute intervals is 27, 31, 28 and 30 MW, while the vehicle contribution holds steady at 20 MW. The battery command is then whatever closes the gap,

$$P_{\mathrm{B},t} = P^{\mathrm{tgt}}(t) - P_{\mathrm{T},t} - P_{\mathrm{E},t}, \tag{11.16}$$

which for a 50 MW target gives 3, −1, 2 and 0 MW. Figure 11.6 shows the four intervals as stacked bars: each bar carries the thermostat contribution at the bottom, the vehicle contribution above it, and the battery contribution on top, against the 50 MW target drawn as a horizontal line.



Blue: thermostats, as realized. Purple: vehicles, steady at 20 MW. Green, or red when charging: the battery segment, whatever remains to reach the target.

Figure 11.6: Batteries closing the tracking error from equation (11.16). The thermostat fleet was forecast at 30 MW and realizes 27, 31, 28 and 30 MW over four fifteen-minute intervals, so the battery is commanded to +3, −1, +2 and 0 MW. In the second interval the fleet over-delivers and the battery absorbs 1 MW.

In the second interval the thermostats over-deliver, the stack exceeds the target, and the battery command is negative: the battery charges, absorbing the excess, which is the behavior that makes the aggregate trackable at all. The pattern generalizes into a division of labor: slow, uncertain and energy-rich resources provide bulk energy, and a small, fast, precisely controllable resource provides the tracking. The battery's energy requirement in this role is modest, because it moves only the forecast error and not the whole service; over the four intervals above the net battery energy is $(3 - 1 + 2 + 0) \times 0.25 = 1$ MWh.

11.5.4 Thermostats as energy-like resources

Thermostats are the opposite kind of resource: slow to respond, imprecise in aggregate, and able to shift a large amount of energy at low cost, provided the shift is not asked to last too long. The binding limit is the comfort band, which converts duration into discomfort and makes a long sustained curtailment expensive even when its first minutes were nearly free. The remedy developed in Chapter 5 is rotation: instead of asking one cohort for 30 MW for two hours, ask

three cohorts for 10 MW each in sequence, so that no house spends more than forty minutes off and the aggregate reduction stays nearly constant. Figure 5.11 of Chapter 5 shows the stacked result and is not redrawn here. From the dispatch optimizer's point of view, rotation is the solution the problem produces by itself: with the comfort cost of equation (11.6) and a horizon long enough to see the drift, spreading the curtailment across cohorts has a lower objective value than concentrating it, because the comfort penalty is convex in the deviation.

11.5.5 Vehicles as deadline-constrained resources

Vehicle flexibility is bounded neither by power nor by comfort but by a deadline, and the amount usable right now is a property of the fleet's state rather than of its size. A fleet drawing 40 MW of charging power may offer only 25 MW of curtailable load, because the remaining vehicles are close enough to their departure requirements that any delay risks a deadline. The dispatch rule that follows is to curtail preferentially the vehicles with the largest slack, in the sense of Chapter 7, and to release them back to charging as their deadlines approach; the one-sided cost of equation (11.8) produces exactly this ordering without the rule having to be stated separately. Because the usable fraction depends on state, the same 40 MW fleet offers a different amount at 20:00 than at 23:00, and the control policy, not only the control action, is state-dependent. Table 11.4 summarizes the three roles.

Table 11.4: Role of each resource class in a coordinated dispatch.

Class	Role	Limited by	Useful time scale
Thermostats	bulk energy shifting at low cost	comfort band, response decay, override	tens of minutes to a few hours
Vehicles	deadline-constrained energy shifting	departure requirement, plug-in window	hours
Batteries	fast balancing and error closure	stored energy, degradation charge	seconds to tens of minutes

11.5.6 Intertemporal coupling and reserve

Every state constraint in Table 11.1 couples the present action to future feasibility, through the three dynamics

$$\begin{aligned} \mathrm{SOC}_{t+1} &= \mathrm{SOC}_t - \frac{P_{\mathrm{B},t}\,\Delta t}{E^{\max}}, \\ \mathrm{SOC}_{\mathrm{EV},t+1} &= \mathrm{SOC}_{\mathrm{EV},t} + \frac{\eta\, P_{\mathrm{EV},t}\,\Delta t}{E^{\max}}, \\ T_{\mathrm{in}}(t+1) &= f_{\mathrm{T}}\big(T_{\mathrm{in}}(t),\ u_{\mathrm{T},t},\ T_{\mathrm{out}}(t)\big). \end{aligned} \tag{11.17}$$

Each says that today's control changes tomorrow's admissible set $\mathcal{U}(x_{t+1})$. Dispatch is therefore not a sequence of independent allocation problems but a genuine optimal-control problem over time, which is the formal content of the claim in Section 11.1.2 that flexibility is a stock.

Reserve. One practical consequence is that a fleet should rarely be dispatched to its limit. If 100 MW of flexibility is available, scheduling 80 MW and holding 20 MW unscheduled buys room to react to three things that regularly happen: part of the scheduled 80 MW does not materialize because the forecast was wrong; individual devices fail, drop offline or are overridden; and the target itself increases, because the grid condition that caused the request got worse. Without reserve the fleet has no response to any of them. Maximum utilization is not optimal utilization,

and the 20 MW held back is the operational counterpart of the derating of Section 11.4.2 and the chance constraint of Section 11.4.3, expressed as held capacity rather than as a reduced commitment.

11.5.7 Closed-loop re-optimization

The loop of Section 11.3.2 earns its cost when the fleet misses its plan, which it does routinely. A 50 MW target is planned as 20 MW of thermostat curtailment, 20 MW of vehicle curtailment and 10 MW of battery discharge. Five minutes into the event the telemetry returns 17, 16 and 10 MW, a total of 43 MW and an error of 7 MW. The optimizer is re-run from the measured state and the new forecast, and finds the 7 MW where it is cheapest: 5 MW more from the battery, taking it from 10 to 15 MW, and 2 MW more of vehicle curtailment, taking it from 16 to 18 MW, which restores 50 MW. Nothing about this requires the original plan to have been good; it requires only that the measurement arrive and the problem be re-solved. The correction is checked against the same constraints as the original plan: 15 MW is inside the battery's 20 MW rating, and the extra energy drawn has to fit above the reserve, which bounds how long the correction can be held.

A receding-horizon plan. The plans themselves change, not only the commands. Figure 11.7 shows two rows on a common 17:00 to 20:00 time axis. The upper row is the plan made at 17:00, divided into three hourly blocks with the resource mix each block intends to use. The lower row is the plan made one hour later, after the battery's measured state of charge came in below the 17:00 prediction.

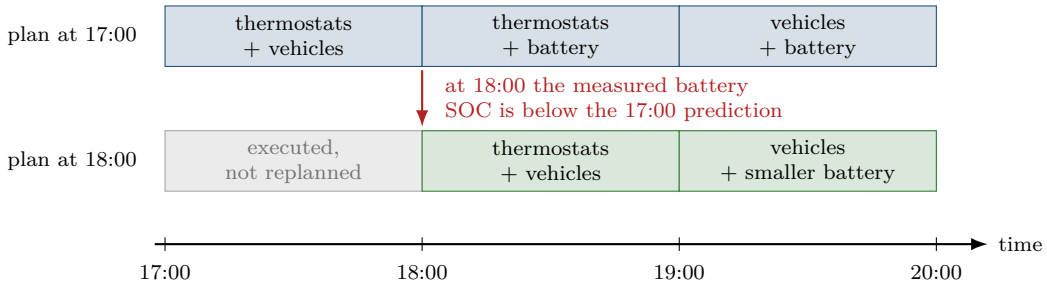


Figure 11.7: The plan evolves. The 17:00 plan intended to use the battery from 18:00 to 19:00; the 18:00 plan, built from a lower measured state of charge, moves that hour onto thermostats and vehicles and reduces the battery's share of the final hour. Resource mixes are illustrative.

The 17:00 plan intended thermostats and vehicles from 17:00 to 18:00, thermostats and battery from 18:00 to 19:00, and vehicles and battery from 19:00 to 20:00. The 18:00 plan keeps the past hour as executed history, replaces the 18:00 to 19:00 block with thermostats and vehicles, and reduces the battery's share of the final hour. No step of the original plan beyond the first was ever a commitment, so no step had to be broken: the plan is the output of the optimizer, and the optimizer runs again with better information.

11.6 Economics and the value of a forecast

The objective used so far was a tracking error with penalty weights. In a market the objective is money, and once it is, the quality of the flexibility forecast acquires a measurable value. This section writes the market objective, shows how to evaluate a forecast by the decisions it produces, and closes the loop the book has been assembling.

11.6.1 A market objective

If the fleet is paid π_t per megawatt delivered, the dispatch problem can be written as a profit maximization,

$$\max_{u} \sum_{t} \Big[\pi_t \, P^{\mathrm{del}}(t) - C_{\mathrm{deg},t} - C_{\mathrm{comfort},t} - C_{\mathrm{shortfall},t} \Big], \tag{11.18}$$

which is equation (11.9) with its sign flipped and its weights replaced by prices. The tracking term has become a shortfall charge $C_{\mathrm{shortfall}}$, which is what the contract actually penalizes, and the exchange rates λ of Section 11.2 have become dollar costs that someone can be asked to justify. The structure of the problem does not change; what changes is that the trade-offs are now stated in the units in which the business is run, so a dispatch can be compared against the alternative of not dispatching at all.

11.6.2 Forecasts and optimization are evaluated together

A flexibility forecast is an input to equation (11.18), and the only thing that matters about an input is the decision it produces. That observation contradicts the usual way forecasts are compared. If model A has a lower root-mean-square error than model B, model A is the better predictor of the mean; it is not automatically the better forecast for dispatch. Model B may have slightly worse mean error and much better calibrated uncertainty, and a calibrated uncertainty is what equation (11.12) needs. A fleet operated on model B can then commit at a known reliability, suffer fewer delivery failures and earn more, while remaining the worse model by root-mean-square error.

Example 11.2. Let the fleet's true capability be $F \sim \mathcal{N}(100, 10^2)$ MW, with capacity paid at $\pi = \$40$ per MW and shortfall charged at $L = \$120$ per MW. The expected net value of a commitment k is $\pi k - L\,\mathbb{E}\big[(k - F)_+\big]$, whose derivative in k is $\pi - L\,\Phi\big((k - 100)/10\big)$. Setting it to zero puts the optimum at the quantile where $\Phi\big((k - 100)/10\big) = \pi/L = 1/3$, that is $k = 95.7$ MW, worth \$3564. A point forecast of 100 MW cannot locate that quantile: the dispatcher has to guess a derate, and α of 0.95, 0.90 and 0.85 is worth \$3563, \$3500 and \$3365. The spread across plausible derates, about 6 %, is the cost of not knowing the distribution. Change the fleet so that $\sigma = 20$ MW and the optimal commitment moves to 91.4 MW while every derate is unchanged, because a point forecast carries no information about which fleet it is describing.

11.6.3 A point forecast and a distribution

The comparison can be made concrete with two forecasts of the same quantity. Forecast A says $F = 100$ MW. Forecast B gives a distribution $p(F)$ under which $\Pr(F > 80) = 0.98$ and $\Pr(F > 100) = 0.50$; these are the Gaussian forecast of Figure 11.4, since $\mathcal{N}(100, 10^2)$ puts probability $1 - \Phi(-2) = 0.977$ above 80 MW and 0.50 above its mean. Forecast A and Forecast B agree on the center and differ on everything else. Forecast B is far more useful, because it lets the decision layer choose its own risk tolerance: a market product with a severe non-delivery penalty is bid at 83.6 MW, a product with the mild penalty of Section 11.6.2 at 95.7 MW, and a product that is only indicative at 100 MW. All three decisions come from one forecast, and none of them can be derived from Forecast A. The chain is

$$\text{probabilistic forecasting} \;\rightarrow\; \text{risk-aware optimization}, \tag{11.19}$$

and it is the reason the forecasting layer of a DERMS is asked for distributions rather than numbers.

Principle 11.3. *A forecast is worth what the decisions it supports are worth. Report the distribution, and let the optimizer, which knows the prices and the penalties, choose the quantile.*

11.6.4 The decision loop

The pieces now close on themselves. Telemetry y_t arrives from the fleet; state estimation turns it into $\hat{x}_t$; the forecasting layer turns the state and the weather into a distribution $p(F_{t:t+H})$ of what the fleet can do over the horizon; optimization turns that distribution, the target and the prices into a plan $u^{\star}_{t:t+H}$; only the first action u_t is applied; the fleet and the grid respond; and new telemetry y_{t+1} arrives. Figure 11.8 draws the six stages as boxes with arrows in the order just given, and the arrow from the last box back to the first closes the loop.

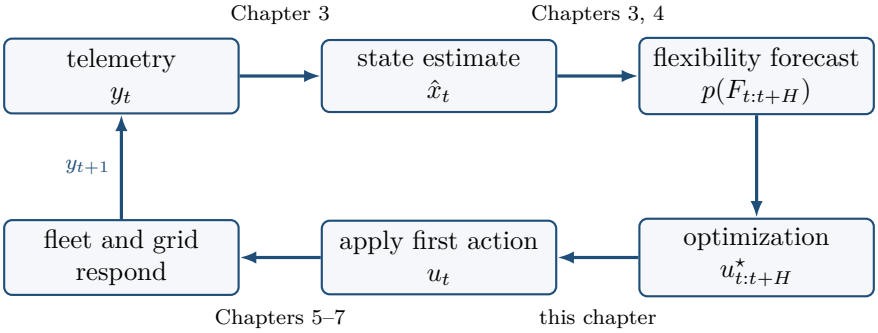


Figure 11.8: The DERMS decision loop, $y_t \to \hat{x}_t \to p(F_{t:t+H}) \to u^{\star}_{t:t+H} \to u_t \to y_{t+1}$. Each stage is a chapter of this book; the optimization stage is the subject of this one, and only the first action of its plan is ever applied.

Every stage of the loop has appeared before: the telemetry and state estimation of Chapter 3, the flexibility forecast of Chapters 3 and 4, the device responses of Chapters 5 to 7. The optimization stage is what this chapter supplied, and it is the stage at which the fleet's physics, the customers' limits and the grid's request are reconciled into a number that gets sent.

11.7 What to retain

Definition 11.1 (Dispatch)**.** Dispatch is the decision of how to spend flexibility over time. Flexibility is a finite, state-dependent stock, so the optimizer chooses which resources to use, by how much and when, balancing grid performance against customer constraints, future flexibility, uncertainty and cost. Dispatch is an intertemporal optimal-control problem, not an allocation of a quantity available now.

The chapter's content is equation (11.9) and the four decisions that surround it: what the stage cost charges for, how long the horizon is, how the unknown disturbance is represented, and which resource plays which role. Table 11.5 lists the objects and the section that defines each.

11.8 Exercises

Exercise 11.1. Take Options A and B of Section 11.1.2, with the trajectories of Figure 11.1. Using $\lambda_{\mathrm{T}} = 2$ cost units per $^{\circ}\mathrm{C}^2$ per hour and $\lambda_{\mathrm{B}} = 0.5$ cost units per MWh of battery throughput, compute the comfort cost and the degradation cost of each option over the two-hour event. Treat each temperature deviation as a linear ramp, so that its mean square over the

Table 11.5: The objects of a DER dispatch problem.

Object	Symbol or form	Meaning	Section
Stage cost	$c(x_t, u_t)$	cost of one step of the plan	11.1.3
Tracking cost	$(P_{\mathrm{VPP}} - P^{\mathrm{tgt}})^2$	delivery error, squared	11.2.1
Comfort cost	$\lambda_{\mathrm{T}}(T_{\mathrm{in}} - T_{\mathrm{set}})^2$	price of thermostat flexibility	11.2.2
Wear cost	$\lambda_{\mathrm{B}}\lvert P_{\mathrm{B}}\rvert \Delta t$	price of battery throughput	11.2.3
Margin cost	$\lambda_{\mathrm{E}}[\mathrm{SOC}^{\mathrm{req}} - \mathrm{SOC}]_+^2$	price of lost driver slack	11.2.4
Horizon	H	steps the optimizer plans ahead	11.3.1
Derating	$\alpha\hat{F}$, $\alpha < 1$	margin with unstated reliability	11.4.2
Chance constraint	$\Pr(P^{\mathrm{del}} \geq P^{\mathrm{tgt}}) \geq 0.95$	stated reliability of delivery	11.4.3
Reliable capability	$\mathbb{E}[F] - z_{1-\epsilon}\sigma$	the quantile that can be committed	11.4.4
Scenario objective	$\sum_\omega p_\omega J(u, \omega)$	expected cost over futures	11.4.5
Robust objective	$\max_{F \in \mathcal{F}} J(u, F)$	worst case in an uncertainty set	11.4.6
Reserve	schedule 80 of 100 MW	unscheduled capacity held back	11.5.6
Market objective	$\sum_t \pi_t P^{\mathrm{del}}(t) - \mathrm{costs}$	dispatch stated in money	11.6.1
Optimal quantile	$\Phi(z) = \pi / L$	where marginal revenue meets marginal penalty	11.6.2

event is one third of the square of its final value. Which option does the optimizer choose, and by how much? What value of λ_{B} makes the two options tie?

Exercise 11.2. A fleet's capability over a service window is forecast as $F \sim \mathcal{N}(90, 12^2)$ MW. Find the commitment that satisfies $\Pr(P^{\mathrm{del}} \geq P^{\mathrm{commit}}) \geq 0.95$, and the commitment at 0.99. Which single derating factor α reproduces the 95 % answer, and what reliability would that same α give if the standard deviation were 6 MW instead of 12 MW?

Exercise 11.3. Using the battery fleet of Figure 11.3, a 50 MW target is to be met from 18:00 to 22:00. The vehicles can provide 20 MW for the whole window; the thermostats can provide 25 MW but only for 90 minutes before the comfort band binds. Compute the energy the battery fleet must supply and compare it with the energy available above the reserve. By how much does the plan fall short, and name two ways to close the gap.

Exercise 11.4. Repeat the re-optimization of Section 11.5.7 when the battery fleet starts the event at SOC = 0.35 with $E^{\max} = 100$ MWh, a 20 MW rating and $\mathrm{SOC}^{\mathrm{res}} = 0.20$. Give the new battery command, and the longest time the corrected 50 MW dispatch can be held before the reserve binds. What should the optimizer do at that point?

11.9 Connection to the TERRA framework

Dispatch is the first chapter in which the graph model is not merely simulated but optimized, and the mapping is close enough that the two formulations share objects.

- **The tracking term is a conservation residual at the grid node.** $P_{\mathrm{VPP}} - P^{\mathrm{tgt}}$ in equation (11.5) is the mismatch between the flow the fleet puts on its grid edge and the

flow the grid asked for. Minimizing its square is penalizing a residual at one node, and the sum (11.1) is that node's balance over the three class edges.

- **The state equations are the components' accumulators.** The three dynamics of equation (11.17) are the bounded accumulators of Chapters 5 to 7: a thermal capacitance, a battery energy store and a vehicle energy store. Evaluating a candidate plan is a forward simulation of those components over the horizon; finding the plan is the same simulation wrapped in an optimizer.

- **Hard constraints are component bounds; soft costs are not.** Table 11.1 is the division a graph framework already makes: bounds on states and edge flows belong to the components and are enforced, while the λ-weighted terms are an objective layered on top. A framework that represents bounds explicitly makes the infeasibility of Section 11.2.5 a detectable condition rather than a silently violated preference.

- **The feasible set is read off the state, not stored.** $\mathcal{U}(x_t)$ in equation (11.3) and the state-dependent merit order of Table 11.3 are the same statement: what the fleet can do now is a function evaluated on the current state, so it is recomputed every cycle instead of being carried as a separate capacity model.

- **Chance constraints and scenarios are the posterior pushed through the model.** The quantile of equation (11.13) and the scenario expectation of equation (11.14) both require the distribution of a future edge flow. That distribution comes from propagating the posterior over component states and parameters, inferred from telemetry, through the forward model, which is exactly the pairing of inference with simulation that the framework provides.

- **The loop is estimate, predict, optimize, act.** Figure 11.8 is a graph model used in closed loop: the posterior state estimate is the model's inference step, the flexibility distribution is a query on the model, and the dispatch is an optimization over the model's admissible controls. Reserve, in Section 11.5.6, is the practice of leaving slack in the bounds so that the next solve has somewhere to go.

In summary, a dispatch optimizer is a conservation graph with an objective attached: a residual to be driven to zero at one node, prices on the use of each component's stored quantity, bounds that may not be crossed, and a horizon long enough that the stored quantities cannot be exhausted without the optimizer seeing it. The next chapter keeps the same machinery and changes the request, replacing a target trajectory to be followed with a load shape to be produced continuously.

Chapter 12

Dynamic load shaping: making aggregate load follow a trajectory

12.1 From events to trajectories

The earlier chapters treated a demand-response action as an event: a magnitude, a window, and a check afterward on what was delivered. That description is adequate when the request is a single number. It stops being adequate as soon as the grid asks for a *shape*. The question of this chapter is

How can the aggregate load of a fleet be made to follow a desired trajectory over time?

Jargon: load shaping, cohort, reference signal, pre-conditioning, ramp rate

Dynamic load shaping is the coordination of groups of DERs at different times, and across DER types, so that the fleet's aggregate power follows a prescribed time profile rather than producing a one-time reduction. A *cohort* is a subset of the fleet that is commanded together and treated as one unit by the schedule; the word is used here for a group of thermostats, but it applies to any partition of a fleet into separately commanded blocks. A *reference signal* is the desired trajectory as seen by a controller, written r_t in control notation and P_t^{tgt} here. *Pre-conditioning* is deliberately moving a device's state before a service begins so that more flexibility is available during it: pre-cooling a building, pre-charging a battery, advancing EV charging. A *ramp rate* is the time derivative of a power quantity, in MW per minute or MW per hour; a *ramp constraint* bounds its magnitude.

Key idea

Dynamic load shaping is trajectory control. The fleet does not have to produce a clean response device by device; it has to produce a clean response *in aggregate*, and the schedule is what makes imperfect parts add up to a specified whole.

This opening section establishes why a shape, rather than a number, is the right object: it contrasts the flat block that programs ask for with the decaying response that devices actually give, writes the tracking objective, shows that no single resource class can hold a four-hour service, and gives the portfolio schedule that can.

12.1.1 The flat block and the decaying response

A traditional demand-response request has the form *reduce load by 50 MW from 17:00 to 19:00.* Geometrically that is a rectangle: constant height, sharp edges. The physical devices that must produce it do not behave that way. A thermostat fleet switched off at 17:00 gives a strong initial

reduction and then decays, because indoor temperatures drift up through the comfort band and compressors restart one after another, as Figure 1.3 of Chapter 1 and the fleet simulation of Figure 5.9 in Chapter 5 both show. A battery fleet holds its full power until its energy runs out and then gives nothing.

Figure 12.1 puts the two shapes on one axis. The horizontal axis is hours after the start of the service and the vertical axis is load reduction in megawatts. The dashed horizontal line is the 50 MW request. The solid curve is the reduction a single thermostat fleet can sustain, falling from 50 MW at the start to 20 MW after two hours. The shaded region between them is the shortfall, and it is not a small correction: by the end of the second hour the fleet is delivering less than half of what was asked. Nothing is wrong with the devices. The request and the physics simply have different shapes.

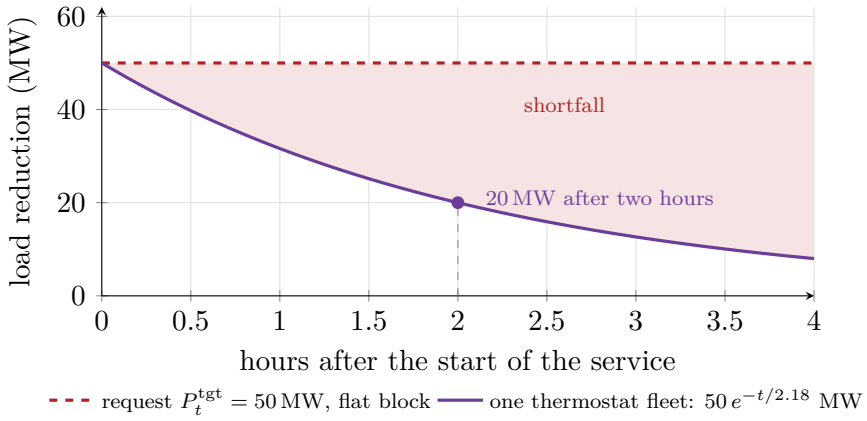


Figure 12.1: The shape of the request and the shape of the response. The dashed line is a 50 MW flat block; the purple curve is $50\,e^{-t/2.18}$ MW, an exponential decay whose time constant 2.18 h is chosen so that the fleet delivers 20 MW at two hours, the value quoted for the thermostat fleet in Section 12.1.3. The shaded area is the shortfall. Values are illustrative.

12.1.2 The target trajectory and the tracking objective

The way out is to stop treating the request as a scalar. Let P_t^{tgt} be the desired fleet response at step t, a whole time series rather than one number, and let P_t^{del} be the response the fleet actually delivers at that step. (In the language of Chapter 4, P_t^{del} is the grid-visible quantity $P_{\text{VPP}}(t)$ measured against the baseline.) The coordination problem is to make one time series follow the other:

$$\min_{u} \sum_{t=1}^{H} \left(P_t^{\text{del}} - P_t^{\text{tgt}}\right)^2, \tag{12.1}$$

where H is the horizon and u collects the commands sent to every cohort and every device class at every step. This is the portfolio optimization of equation (4.10) in Chapter 4, read as a statement about shapes: equation (12.1) is a least-squares fit of a trajectory, not the maximization of an instantaneous megawatt figure.

Two consequences follow immediately. First, a response that is too large is penalized as much as one that is too small, because the square is symmetric; overshooting a target is a tracking failure, which is not how a one-shot reduction program would have scored it. Second, the objective couples the steps only through the constraints, which is where all the difficulty lives: using a resource at step t changes what it can do at step $t + 1$.

12.1.3 The limits of a single resource class

Take a concrete service: 50 MW for four hours, 17:00 to 21:00. Examine each class against it in turn.

- **Thermostats.** The fleet can give 50 MW at the start, but only about 20 MW after two hours, for the reasons of Chapter 5: the comfort band is finite and compressors must restart. Capacity is large, duration is short.
- **Batteries.** The fleet can give 50 MW, and give it precisely, but only for about 1.5 h before state of charge reaches its floor. Capacity and precision are good, energy is the binding limit.
- **EVs.** About 30 MW of charging load is flexible, and the amount varies through the evening as cars plug in and reach their energy requirements. Duration is good, magnitude is limited and uncertain.

No class meets the service on its own: the first fails on duration, the second on energy, the third on magnitude. This is the complementarity argument of Section 4.3.4 in Chapter 4, and it is the reason a shaping problem exists at all. If one class could simply hold the block, there would be nothing to schedule.

12.1.4 Portfolio shaping over a four-hour service

Together the three classes can hold the service, provided each one's contribution is allowed to change from hour to hour. Table 12.1 gives a schedule that does it, and Figure 12.2 draws the same numbers as stacked bars. In the figure the horizontal axis is the hour of the service, the vertical axis is reduction in megawatts, and the three stacked segments are the thermostat, EV and battery contributions; the dashed line is the 50 MW target.

Table 12.1: A four-hour portfolio schedule holding 50 MW. Contributions are in MW; every row sums to the target.

Hour	Thermostats	EVs	Batteries	Total
17:00–18:00	30	10	10	50
18:00–19:00	25	15	10	50
19:00–20:00	15	20	15	50
20:00–21:00	10	25	15	50

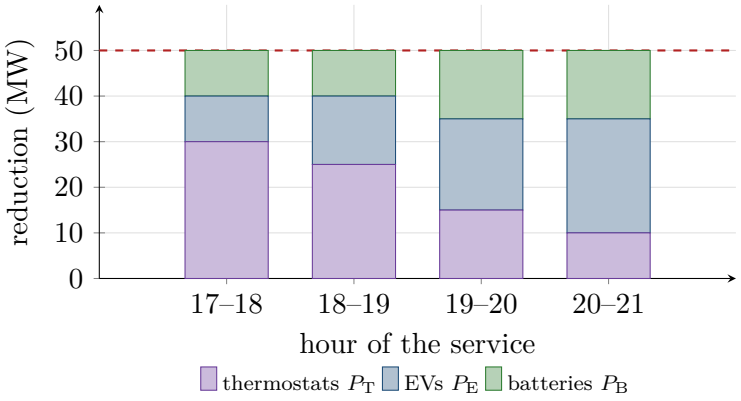


Figure 12.2: The portfolio schedule of Table 12.1. The thermostat contribution falls as the comfort band is consumed, the EV contribution rises as more cars are plugged in, and the battery contribution is held back early so that energy remains late. The dashed line is the 50 MW target. Values are illustrative.

The arithmetic of the table is unremarkable: $P_{\mathrm{T}} + P_{\mathrm{E}} + P_{\mathrm{B}} = 50\,\mathrm{MW}$ in every hour. What matters is that no column is constant. The thermostat column falls because its resource decays; the EV column rises because the plugged-in population grows through the evening; the battery column is deliberately small in the first two hours so that energy is left for the last two. A schedule in which each class's contribution changes over time, chosen so that the sum follows a target, is dynamic load shaping in its simplest form.

12.2 Cohort rotation, rebound and pre-conditioning

Shaping also operates *inside* one resource class, by splitting the fleet into cohorts and moving the service between them. This section develops rotation, the recovery load it creates, the rebound-aware accounting that corrects for it, and the complementary technique of changing device states before the service starts.

12.2.1 Cohort rotation

Consider 60 000 enrolled thermostats divided into four cohorts A, B, C and D of 15 000 each, and a two-hour service from 17:00 to 19:00. Instead of curtailing all 60 000 for two hours, the schedule curtails one cohort at a time for thirty minutes, as in Table 12.2. While one cohort is curtailed the other three are recovering: their compressors run, their indoor temperatures return toward setpoint, and their comfort budget is restored for a later turn.

Table 12.2: A rotation schedule for four thermostat cohorts of 15 000 devices each over a two-hour service.

Interval	Curtailed	Recovering
17:00–17:30	A	B, C, D
17:30–18:00	B	A, C, D
18:00–18:30	C	A, B, D
18:30–19:00	D	A, B, C

Rotation does two things at once. It spreads customer impact, because no household is asked to sit through the full two hours, which lowers override rates and makes the program easier to keep enrolled. And it keeps the fleet from exhausting itself: the full comfort budget is never spent at the same moment, so the fleet retains response capability throughout the window. The stacked result is already drawn as Figure 5.11 in Chapter 5, where four cohorts rotate through half-hour curtailments and the total stays near its target even though each cohort's own contribution decays after its turn.

12.2.2 Why rotation extends duration

The accounting is worth doing explicitly. One cohort of 15 000 thermostats delivers about 20 MW, and can hold it for about thirty minutes before temperatures reach the top of the comfort band. Acting alone, that cohort is a resource of

$$20\,\mathrm{MW} \times 0.5\ \mathrm{h} = 10\ \mathrm{MWh}, \tag{12.2}$$

available only in the half hour in which it is used. Rotating four such cohorts through consecutive half hours produces

$$20\,\mathrm{MW}\ \text{for}\ 2\ \mathrm{h} = 40\ \mathrm{MWh}, \tag{12.3}$$

at the same 20 MW magnitude but four times the duration. Nothing about any individual device changed. The transformation came from the schedule.

Principle 12.1. *Aggregation with rotation converts short-duration device flexibility into long-duration fleet flexibility: the fleet's $C(\tau)$ extends far beyond any single device's $F(t, \tau)$, at constant magnitude, purely through sequencing.*

This is the fleet-level mechanism behind the duration dependence of Figure 4.3 in Chapter 4. A fleet that can be partitioned into n cohorts, each good for τ_0, is a candidate resource for durations up to roughly $n\tau_0$ at the per-cohort magnitude, rather than a resource of n times the magnitude for τ_0.

12.2.3 Recovery load is part of the schedule

The accounting of Section 12.2.2 is optimistic, because a recovering cohort is not neutral. When cohort A is released at 17:30, its 15 000 compressors restart to pull indoor temperatures back down, and the cohort's load rises *above* its baseline. That is the rebound of Chapter 2, drawn there as Figure 2.3: during an event the load falls below baseline and then overshoots it while the devices recover the state they gave up.

The consequence for shaping is that the aggregate response is not the sum of the curtailed cohorts' reductions. Writing the fleet response as a sum over cohorts,

$$P_t^{\mathrm{del}} = \sum_k \Delta P_{k,t}, \qquad \Delta P_{k,t} = P_{k,t}^{\mathrm{base}} - P_{k,t}^{\mathrm{ctrl}}, \tag{12.4}$$

the curtailed cohort contributes a positive $\Delta P_{k,t}$ and every recovering cohort contributes a negative one. A naive reading of Table 12.2 would record $20 + 0 + 0 + 0 = 20\,\mathrm{MW}$ in the second interval; the correct reading records the curtailed cohort's reduction plus the released cohort's recovery, which is negative. Badly managed recovery of one cohort can cancel a large part of another cohort's reduction, and a rotation schedule that ignores recovery can deliver far less than its arithmetic promises. The schedule therefore coordinates curtailment and recovery together, not curtailment alone.

12.2.4 Rebound-aware shaping

Put numbers on it. Cohort A delivers $+20\,\mathrm{MW}$ during its half-hour turn. After release it rebounds to $-8\,\mathrm{MW}$ relative to its baseline: the recovery is an *increase* in load of $8\,\mathrm{MW}$. If the schedule must hold a net $20\,\mathrm{MW}$ during the second interval, then cohort B has to provide

$$P_{\mathrm{B}}^{\mathrm{gross}} = 20 + 8 = 28\,\mathrm{MW} \qquad \text{so that} \qquad 28 - 8 = 20\,\mathrm{MW\ net}. \tag{12.5}$$

Figure 12.3 draws the two intervals. The horizontal axis is the two half-hour intervals; the vertical axis is the contribution to the fleet response in megawatts, positive upward for a reduction and negative downward for recovery load above baseline. In the first interval a single green bar of $+20$ reaches the dashed net target. In the second interval the blue bar is cohort B's gross contribution of $+28$ and the red bar below the axis is cohort A's recovery of -8; their sum, marked on the target line, is again $20\,\mathrm{MW}$.

Two practical points follow. The gross requirement is larger than the net requirement whenever anything in the fleet is recovering, so a shaping plan sized on net megawatts will be short. And the gross requirement is only computable if the rebound is *predicted*, which makes a rebound model a prerequisite for shaping rather than an afterthought for measurement and verification. The size and timing of the recovery depend on the thermal time constant RC of the buildings, the outdoor temperature, and how deep into the comfort band the curtailment went, all of which the thermostat model of Chapter 5 already carries.

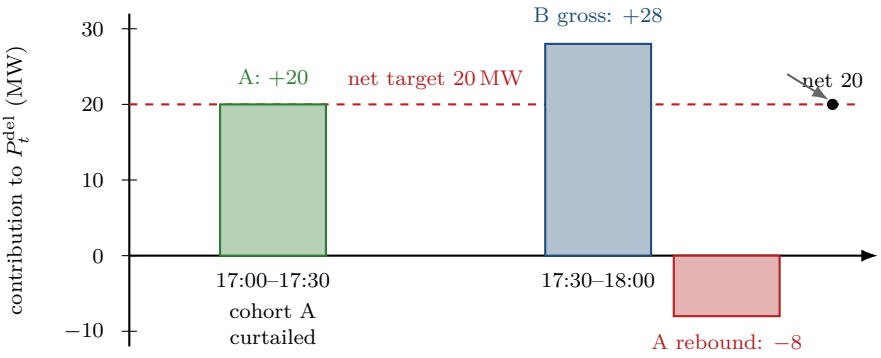


Figure 12.3: Rebound-aware accounting for two intervals of the rotation of Table 12.2. Bars above the axis are reductions, the bar below it is cohort A's recovery load after release. To hold 20 MW net while A rebounds by 8 MW, cohort B must deliver 28 MW gross. Values are illustrative.

12.2.5 Pre-conditioning

Rotation and rebound accounting work with the states the fleet happens to have. Pre-conditioning changes them. The idea is to apply control *before* the service begins, so that the state at the start of the service, x_{t_0}, is more favorable than it would otherwise have been:

- pre-cool buildings, so indoor temperatures start at the bottom of the comfort band rather than at setpoint, and the thermostat fleet has further to drift before compressors must restart;
- pre-charge batteries, so state of charge starts near $\text{SOC}^{\max}$ and the energy limit binds later;
- advance EV charging, so cars arrive at the service window with their energy requirements nearly met and their charging load is free to be reduced.

In each case the quantity being manipulated is an initial condition. Flexibility was defined in Chapter 1 as a function of the state, $F(t, \tau)$ evaluated at x_t; pre-conditioning is the recognition that x_t is itself partly a decision. Control becomes proactive rather than reactive: the fleet is prepared for the service, not merely dispatched into it.

12.2.6 Pre-cooling and the cost of preparation

A worked case makes the trade-off visible. A reduction is required beginning at 18:00. Without preparation, the thermostat fleet can sustain 30 MW for one hour before enough compressors have restarted to erode the response. If the fleet is pre-cooled from 16:00 to 17:00, lowering indoor temperatures across the enrolled buildings, the same 30 MW is sustained for two hours. The duration doubled, and nothing about the devices changed except where in the comfort band they started. Figure 1.8 of Chapter 1 shows the mechanism for one building: pre-cooling by half a degree moves the first compressor restart from 27 to 33 minutes, and a deeper pre-cool or a slower building moves it further.

Pre-cooling is not free. Between 16:00 and 17:00 the compressors run harder than they otherwise would, so fleet load is *higher* before the service than the uncontrolled baseline. The optimizer therefore faces a genuine trade: accept an increase in load in an earlier hour in exchange for a larger or longer reduction later. Whether that trade is worth making depends on the objective. It is worth making when the earlier hour is unconstrained and the later hour is the system peak, and it is a mistake when the pre-cooling hour is itself near a local limit, or when energy prices in the pre-cooling hour are high enough to outweigh the value of the longer service. Nothing in the trade can be evaluated without a model of both hours, which is

why pre-conditioning belongs inside the shaping optimization and not in a separate operating procedure.

12.2.7 Upward shaping

Dynamic load shaping does not always reduce load. The goal is a trajectory, and a trajectory may call for more consumption rather than less. On a day with high midday solar output, the useful action at noon is to *add* load: target +40 MW of additional EV charging, so that energy that would otherwise be curtailed at the generator is absorbed by the fleet. The same fleet is then asked for −40 MW relative to uncontrolled charging at 18:00, when the solar output has gone and the system peak has arrived.

Over the day the fleet has moved 40 MW of charging load from the evening to midday, which is the load shifting of Chapter 2 with the EV as the instrument and the shape, rather than the reduction, as the specification. In the notation of Chapter 6, this is downward grid support $F^{\downarrow}$ at noon and upward support $F^{\uparrow}$ in the evening, scheduled as one trajectory. The sign of P_t^{tgt} changes during the day, and a shaping optimizer handles that without modification, because equation (12.1) never assumed a sign.

12.3 Shaping as tracking control

The problem stated in Section 12.1.2 has a standard name in control engineering. This section makes the identification, then describes what is unusual about the actuator: it is a portfolio of devices with different delays, durations, uncertainties and costs, operated by a hierarchy of optimizers running at different rates.

12.3.1 The reference signal and the tracking error

In the language of feedback control, write

$$r_t = P_t^{\mathrm{tgt}}, \qquad y_t = P_t^{\mathrm{del}}, \qquad e_t = r_t - y_t, \tag{12.6}$$

where r_t is the reference signal, y_t is the measured output and e_t is the tracking error. The controller's job is to drive e_t toward zero. This is a textbook servo problem, and the least-squares objective of equation (12.1) is the usual quadratic penalty on e_t.

Framing shaping this way brings a large body of standard machinery with it: integral action to remove steady offsets, gain selection to trade response speed against overshoot, and the distinction between feedforward (acting on a forecast of the disturbance) and feedback (acting on the measured error). The one part that is not standard is the actuator. A conventional servo has a single input with a known transfer function. A VPP has a portfolio.

12.3.2 Heterogeneous actuators

A conventional actuator maps a command directly to an output, $u_t \mapsto y_t$, with a delay and a gain that can be written down. A single VPP command instead fans out to thermostat setpoints, EV charging rates and battery inverter dispatches, and each of those channels has its own delay, its own sustainable duration, its own uncertainty and its own cost. Table 12.3 collects the four properties for the three classes, in the terms the earlier chapters established.

The table is the reason the allocation of the tracking burden is itself an optimization. A controller that treated the three channels as interchangeable would use the battery for everything, because it is the fastest and most certain, and would exhaust it. A controller that ignored delay would assign a fast correction to thermostats and see it arrive minutes late. The optimizer has to place each part of the reference signal on the channel whose delay, duration, uncertainty and cost suit it.

Table 12.3: The three DER classes as tracking actuators. Values are typical orders of magnitude, not program specifications.

Class	Delay	Sustainable duration	Uncertainty	Cost of use
Thermostats	minutes: command propagation plus compressor response	0.5–2 h, set by the comfort band	high: weather, occupancy and customer overrides move the whole fleet together	customer comfort; opt-outs reduce future availability
EVs	seconds to minutes once connected	hours, bounded by the departure deadline	high in the plugged-in count, low per connected car	low: energy is shifted, not denied, if the deadline holds
Batteries	seconds	set by energy, typically 1–4 h	low: state of charge is metered and the inverter is precise	degradation per cycle; the most expensive megawatt in the portfolio

12.3.3 Fast and slow resources

That placement has a simple and robust default, which follows from Table 12.3:

- **batteries** are fast and precise, so they are used to correct tracking error, not to carry bulk response;
- **EVs** are flexible over long windows, so they are used for energy shifting across hours;
- **thermostats** offer large fleet capacity but state-dependent and uncertain response, so they are used for bulk temporary load shaping.

The resulting division of labor is that the slow, cheap, high-capacity resources provide the bulk of the trajectory while the fast, expensive, precise resource absorbs the residual. This is multi-timescale control, and it is the same structure a power system uses at the bulk level, where slow economic dispatch sets the schedule and fast automatic generation control removes the residual imbalance.

12.3.4 Multi-timescale optimization

The division of labor among resources is mirrored by a division of labor among optimizers. Shaping is not solved once; it is solved repeatedly, at several rates, by nested loops with different information and different authority. Table 12.4 lists the four levels, what each decides, and what it knows that the level above it did not.

Table 12.4: The levels of a shaping controller and what each one decides.

Level	Decision	New information	Resources
Day ahead	how much flexibility of each class to preserve for tomorrow's window	weather forecast, enrollment, market or utility request	all classes, as budgets
Hour ahead	which cohorts and which device classes to use, hour by hour	updated weather, realized plug-in counts, measured states	all classes, as schedules
Minutes ahead	revision of the schedule against telemetry	measured P_t^{del}, realized rebound, opt-outs	thermostat cohorts, EV rates
Seconds to minutes	correction of the residual tracking error e_t	the metered error itself	batteries

The levels are ordered by speed and inversely by scope: the day-ahead problem sees the whole horizon but the coarsest information, and the real-time loop sees the best information

but can only act on a small, fast slice of the portfolio. A single monolithic optimizer is not a practical substitute, because the fast loop must close in seconds while the day-ahead problem takes minutes to solve and needs forecasts that do not exist at that resolution.

12.3.5 A hierarchical example

Follow one service through the hierarchy. The utility requests 100 MW from 18:00 to 21:00 tomorrow evening.

Day ahead. The platform reserves 50 MW of thermostat flexibility, 30 MW of EV flexibility and 30 MW of battery flexibility: 110 MW of budget against a 100 MW request, the coverage margin of Section 4.5.1 in Chapter 4.

Hour ahead. The day came in cooler than forecast and fewer thermostats are cycling, so the thermostat fleet offers only 40 MW rather than 50; meanwhile more cars are plugged in than expected and the EV fleet offers 35 MW rather than 30. The hour-ahead optimizer re-solves with the revised budgets, shifting about 10 MW of the burden from thermostats to EVs and batteries.

Real time. Within the hour the battery controller closes the loop on the measured response,

$$P_{\mathrm{B}}(t) = P_t^{\mathrm{tgt}} - P_{\mathrm{T}}(t) - P_{\mathrm{E}}(t), \tag{12.7}$$

taking whatever the slower classes did not deliver, subject to its own power and energy limits. Figure 12.4 draws the three loops: each box is one optimizer with its rate and its decision, the solid arrows are the budgets and schedules passed downward, and the dashed arrow on the left is the telemetry that flows back up and causes the upper loops to re-solve.

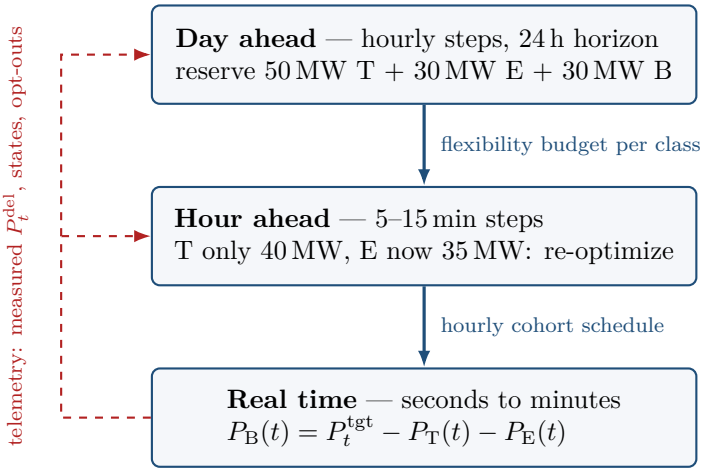


Figure 12.4: Hierarchical shaping for a 100 MW service from 18:00 to 21:00. Each box is one optimizer, labeled with its rate and its decision; solid arrows carry budgets and schedules downward, and the dashed arrow carries telemetry upward, which is what makes the upper loops re-solve with better information.

The example shows the purpose of the hierarchy. The day-ahead reservation was wrong in both directions, by 10 MW on thermostats and 5 MW on EVs, and the service was still delivered, because each level corrected the level above it with information that level did not have.

12.4 Shapes and services

Once the target is a trajectory rather than a number, the trajectory can be anything the grid needs. This section works through the shapes that matter in practice: flat blocks, ramps,

renewable-following profiles, ramp-rate limits, and shapes required at a specific location on the distribution network.

12.4.1 Arbitrary target shapes

Nothing in equation (12.1) requires P_t^{tgt} to be constant. A utility may want a flat block during the evening peak, or a ramp that grows through the afternoon, or a profile that mirrors renewable output hour by hour. Each is a different reference signal fed to the same optimizer.

Figure 12.5 draws three of them on one axis. The horizontal axis is the hour of the day and the vertical axis is the required fleet response P_t^{tgt} in megawatts, with positive values meaning a load reduction and negative values meaning an increase. The red step is a flat 50 MW block from 16:00 to 21:00. The blue line is an evening ramp rising from zero at 16:00 to 60 MW at 21:00. The gold curve is a renewable-following target: negative around midday, where the fleet should absorb surplus generation, and strongly positive in the evening, where generation falls away.

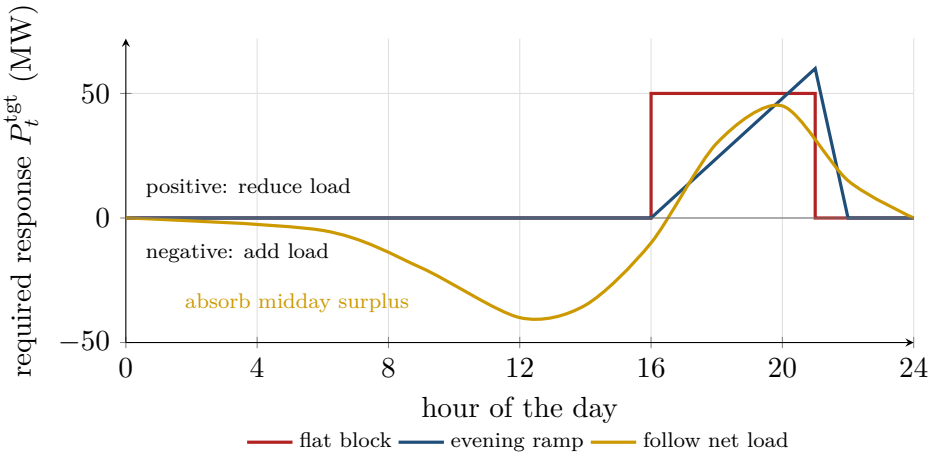


Figure 12.5: Three target trajectories for the same fleet, all handled by equation (12.1) without modification. Positive values are load reductions, negative values are deliberate increases. Shapes are illustrative.

Peak shaving is the special case in which P_t^{tgt} is a block placed on the system peak. Dynamic load shaping is the general case, and it is general in a useful way: supporting a new service means writing a new reference signal, not building a new control scheme.

12.4.2 Load following

The most important of the general shapes is the one that follows net load. Define net grid load as demand minus non-dispatchable renewable generation,

$$P_{\text{net}}(t) = P_{\text{demand}}(t) - P_{\text{renew}}(t), \tag{12.8}$$

which is the quantity the dispatchable fleet of a system actually has to serve. The shaping problem is then to choose $P^{\text{del}}(t)$ so that

$$P_{\text{net}}(t) - P^{\text{del}}(t) \tag{12.9}$$

is smoother than $P_{\text{net}}(t)$ itself, by whatever measure of smoothness the system cares about: a lower peak, a smaller peak-to- trough range, or a bounded rate of change. With that objective the VPP is no longer a load-reduction program. It is an active balancing resource, and the fleet's role is to take variability out of the residual that conventional generation must follow.

12.4.3 Ramp control

Smoothness often means the rate of change, not the level. On a system with substantial solar generation, output falls steeply in the hour or two before sunset while demand is still rising, so $\mathrm{d}P_{\mathrm{net}}/\mathrm{d}t$ becomes large and positive. Conventional generation can be ramped, but only within its own rate limits, and a steep net-load ramp is expensive to follow and can threaten reliability if the ramp capability is short.

A fleet can reduce the ramp seen by generation by reducing flexible load or discharging batteries progressively through the ramp hour. The requirement is a bound on the derivative of the residual,

$$\left| \frac{\mathrm{d}}{\mathrm{d}t} \left(P_{\mathrm{net}}(t) - P^{\mathrm{del}}(t) \right) \right| \leq R^{\max}, \tag{12.10}$$

which is a constraint on the shape, not on the magnitude, of the fleet response. Figure 12.6 shows the effect. The horizontal axis is the hour of the afternoon and evening; the vertical axis is net load in megawatts. The gray curve is the uncontrolled net load, whose steepest hourly change is 75 MW/h between 17:00 and 18:00. The green curve is the same net load with the fleet shaped to satisfy equation (12.10) with $R^{\max} = 50$ MW/h: the fleet adds load in the early afternoon and reduces it through the evening, so the steepest hourly change falls to 45 MW/h. The two profiles carry essentially the same energy over the window, 3555 and 3550 MWh; the fleet moved load in time rather than removing it.

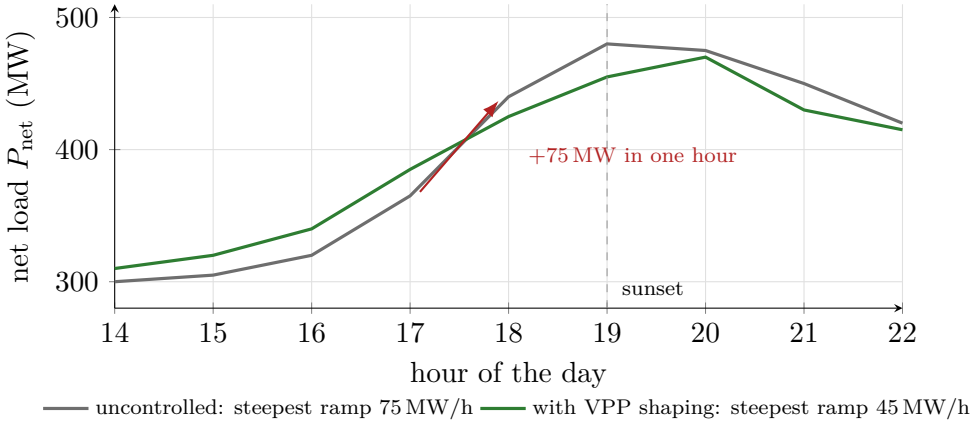


Figure 12.6: Ramp control. The gray profile is an uncontrolled evening net load with a 75 MW/h ramp into the sunset hour; the green profile is the same system with the fleet shaped to satisfy equation (12.10) at $R^{\max} = 50$ MW/h, reducing the steepest hourly change to 45 MW/h at essentially unchanged energy. Profiles are illustrative.

Ramp management is a different product from peak reduction, and a fleet can be good at one and poor at the other. A battery fleet with little energy is an excellent ramp resource and a poor capacity resource; a thermostat fleet is the reverse.

12.4.4 Distribution-level shaping

Shaping targets need not come from the bulk system. Consider a distribution transformer rated 5 MW and a forecast that the uncontrolled load behind it will reach 5.6 MW in the evening. The overload is 0.6 MW, and the DERMS needs exactly that much local flexibility during exactly those hours:

$$P^{\mathrm{local}}_{\mathrm{flex}}(t) \;\geq\; \max\left(0,\; \hat{P}_{\mathrm{local}}(t) - 5\,\mathrm{MW}\right), \tag{12.11}$$

where $\hat{P}_{\mathrm{local}}(t)$ is the forecast load behind the transformer. The target is small, local and time-specific, and it has a hard edge: 0.5 MW of reduction does not solve a 0.6 MW overload.

12.4.5 Location in the optimization

A system-wide capacity figure cannot answer a local question. Suppose the fleet has 100 MW of total flexibility, of which only 0.3 MW sits behind the overloaded transformer of Section 12.4.4. The relevant number for the constraint is 0.3 MW, and the other 99.7 MW is irrelevant to it; power reduced on a different feeder does not flow through this transformer. The overload is not solvable with the enrolled fleet, and the honest output of the optimization is that fact, not an aggregate reassurance.

Flexibility must therefore be indexed by location as well as by time and type. The notation of Chapter 3 already allows it, $F(r, t, d)$ for region r, time t and device type d, and shaping uses all three indices: the target is a trajectory *at a place*. A fleet that is well sited for bulk energy services may be useless for the distribution constraints that actually bind, and measuring enrollment only in total megawatts hides that completely. The constraints that follow from network location are the subject of the next chapter.

12.5 Feasibility over the horizon and feedback

A schedule that is correct at each instant can still be wrong over the horizon, because using a resource now removes it later. This section covers depletion, the failure of greedy dispatch, the reserve that uncertainty requires, the feedback loop that removes residual error, and the protection of the resource that closes that loop.

12.5.1 Resource depletion

A battery fleet asked for 20 MW delivers it comfortably in the first hour. State of charge falls as it does so, and by the second hour only about 10 MW remains available. The resource was not consumed by anything external; it was consumed by its own use. The same holds for thermostats, whose comfort band is a budget, and for EVs, whose energy requirement must still be met by departure.

The consequence for shaping is a constraint on the whole trajectory, not on individual steps. The target must be feasible across the entire horizon,

$$u_t \in \mathcal{U}(x_t), \qquad x_{t+1} = f(x_t, u_t, w_t), \qquad t = 1, \dots, H, \tag{12.12}$$

with the feasible set at each step depending on the state that earlier steps produced. This is the endogenous capacity of Figure 4.5 in Chapter 4, seen from the scheduling side: the optimizer cannot minimize the instantaneous error at each step independently, because doing so changes the sets $\mathcal{U}(x_t)$ that later steps will face.

12.5.2 Greedy and horizon-aware dispatch

Table 12.5 contrasts the two strategies on the four-hour, 50 MW service of Section 12.1.4. A greedy dispatch minimizes the error in the current hour with the cheapest and most certain resource, which is the battery fleet; it therefore spends 50 MW of battery in hour one. The battery fleet holds 50 MW for about 1.5 h, so about 75 MWh is available; hour one consumes 50 MWh of it and hour two the remaining 25 MWh, after which the battery is at its floor. From hour three the greedy schedule has only thermostats and EVs left, the thermostat fleet has already been drawn down, and the service falls short: 40 MW in hour three and 35 MW in hour four, a cumulative shortfall of 25 MWh against the 200 MWh requested.

The horizon-aware columns are the schedule of Table 12.1. It holds 50 MW in all four hours, and it does so while using only $10 + 10 + 15 + 15 = 50$ MWh of the battery's 75 MWh budget, leaving headroom. The difference between the two halves of the table is not better forecasting

Table 12.5: Greedy and horizon-aware dispatch of the same 50 MW, four-hour service. Contributions are hourly averages in MW; T, E and B are thermostats, EVs and batteries. Values are illustrative.

	Greedy				Horizon-aware			
Hour	T	E	B	Total	T	E	B	Total
1	0	0	50	50	30	10	10	50
2	25	0	25	50	25	15	10	50
3	30	10	0	40	15	20	15	50
4	15	20	0	35	10	25	15	50
Energy (MWh)				175				200

or better hardware. It is only that the right-hand schedule was chosen against the horizon and the left-hand one against the current step.

Principle 12.2. *In a system whose feasible set depends on its own history, greedy instantaneous tracking is not merely suboptimal; it can make a target that was feasible over the horizon infeasible after the first step.*

12.5.3 Uncertainty and reserve

Even a horizon-aware schedule can fail, because the response of each class is uncertain. The thermostat contribution is not a number but a draw from a distribution, $F_{\mathrm{T}}(t) \sim p(F_{\mathrm{T}})$, whose spread comes from weather, occupancy, overrides and communications. A deterministic schedule that exactly exhausts the expected capability of every class has no margin: any adverse draw produces a shortfall.

The standard remedy is to hold a reserve. If the battery fleet can provide 15 MW, the schedule may deploy only 5 MW of it and keep the remaining 10 MW unassigned, available for correction:

$$P_{\mathrm{B}}^{\mathrm{sched}} = 5\,\mathrm{MW}, \qquad R_t = P_{\mathrm{B}}^{\mathrm{avail}} - P_{\mathrm{B}}^{\mathrm{sched}} = 15 - 5 = 10\,\mathrm{MW}. \tag{12.13}$$

This is the fleet analog of operating reserve in bulk power systems, where a quantity of fast capacity is deliberately left undispatched so that it can respond to deviations. The size of R_t is a risk decision of the same kind as the coverage ratio of Chapter 4: too small and the fleet cannot correct, too large and capability is left unused. It is chosen from the predicted distribution of the tracking error, not from a rule of thumb, and it explains why the battery entries in Table 12.1 sit well below the battery fleet's power rating.

12.5.4 Feedback correction

Reserve is what makes closed-loop correction possible. Take an instant during a 50 MW service in which telemetry reports

$$P_{\mathrm{T}} = 22\,\mathrm{MW}, \quad P_{\mathrm{E}} = 17\,\mathrm{MW}, \quad P_{\mathrm{B}} = 8\,\mathrm{MW}, \qquad P^{\mathrm{del}} = 47\,\mathrm{MW}, \tag{12.14}$$

so the tracking error is $e_t = 50 - 47 = 3\,\mathrm{MW}$. The battery controller raises its output from 8 to 11 MW and the fleet delivers 50 MW. The correction is drawn from the reserve of equation (12.13), and it is available only because the schedule did not assign the battery's full capability in advance.

This is ordinary closed-loop tracking: measure the output, compute the error, actuate. What makes it work in a VPP is that the actuation is routed to the channel in Table 12.3 whose delay matches the correction's timescale. Routing a 3 MW correction to thermostats would deliver it minutes later, by which time the error would have changed.

12.5.5 Protecting the state of charge

A correction loop that always draws on the battery will empty it. Each correction consumes energy, state of charge falls toward $\mathrm{SOC}^{\min}$, and when the floor is reached the fast channel saturates: the error it had been absorbing reappears, with no resource left to take it. The remedy is a second, slower loop that gradually adjusts the thermostat and EV schedules so that the battery's contribution can be reduced and its state of charge recovers. The fast loop keeps the tracking error near zero; the slow loop keeps the fast loop's resource available.

Figure 12.7 compares the two arrangements over a four-hour service. The upper axis is the battery fleet's contribution P_{B} in megawatts against hours into the service; the lower axis is the fleet's state of charge over the same period, with a dashed line at the floor $\mathrm{SOC}^{\min} = 0.20$. In the unlayered case the battery contribution grows as the slower classes decay, state of charge falls steeply, and the fleet reaches its floor just before three hours, at which point P_{B} must drop to zero. In the layered case the battery takes an early correction, the slower loop then relieves it by re-optimizing the thermostat and EV schedules, the battery contribution falls back, and state of charge ends the service at about 0.35, well above the floor.

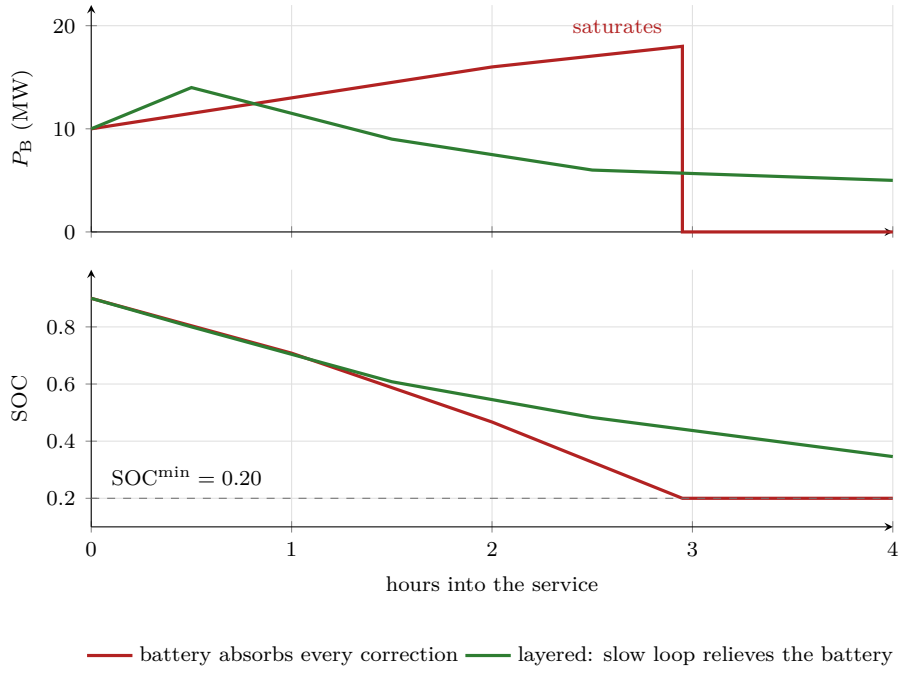


Figure 12.7: Layered control protects the fast resource. Upper panel: battery contribution. Lower panel: fleet state of charge, computed from the upper panel as $\mathrm{SOC}_{t+1} = \mathrm{SOC}_t - P_{\mathrm{B}}\,\Delta t/(60\ \mathrm{MWh})$ for a fleet of 60 MWh usable energy. Absorbing every correction in the battery drives it to its floor before three hours; relieving it with a slower adjustment of the thermostat and EV schedules ends the service near $\mathrm{SOC} = 0.35$. Trajectories are illustrative.

12.6 Objective, trade-offs and controlled recovery

The last body section writes the objective that balances tracking against the costs of achieving it, states the conflict that objective formalizes, shows that the end of a service is as much a shaping decision as its start, and assembles the complete formulation.

12.6.1 The weighted objective

Tracking is one term among several. The objective a shaping optimizer actually minimizes is

$$J = \sum_{t=1}^{H} \Big[\lambda_1 \big(P_t^{\mathrm{del}} - P_t^{\mathrm{tgt}}\big)^2 + \lambda_2\, C_{\mathrm{comfort},t} + \lambda_3\, C_{\mathrm{battery},t} + \lambda_4\, C_{\mathrm{EV},t} \Big], \tag{12.15}$$

where $C_{\mathrm{comfort},t}$ penalizes excursions of indoor temperature within or beyond the comfort band, $C_{\mathrm{battery},t}$ penalizes cycling and depth of discharge, and $C_{\mathrm{EV},t}$ penalizes charging schedules that leave little margin against departure requirements. The weights λ_i are the operating policy in numerical form. Raising λ_1 buys tighter tracking at the cost of more customer impact and more battery wear; raising λ_2 protects comfort and accepts larger tracking error. Each weight is a statement about what the program is willing to spend to hold its trajectory, and choosing them is a business decision expressed in the only language the optimizer understands.

12.6.2 Tracking against customer experience

The conflict the weights mediate is concrete. The grid asks for 50 MW. The fleet can supply 40 MW using flexibility that customers will scarcely notice: shallow setpoint changes, charging delays well inside the available slack, moderate battery discharge. The remaining 10 MW exists, but only by taking indoor temperatures further into the comfort band, pushing EVs closer to their deadlines, and discharging batteries deeper.

Whether to take it is not a physics question. It depends on program rules, on the enrollment agreements that say what may be asked of a customer and how often, on the penalty for under-delivery, and on how important the event is: a reliability emergency and an economic peak do not justify the same customer impact. Equation (12.15) does not resolve the trade-off, and it is not supposed to. It *formalizes* it, so that the choice is made once, explicitly, in the weights, rather than implicitly and differently in every event.

12.6.3 Shaped release

The end of a service is a shaping decision too. Ending a 100 MW event by releasing the whole fleet at once takes the response from 100 MW to zero in one step. Every curtailed device begins recovering simultaneously, which is the synchronization of Chapter 5 and produces the largest possible rebound peak. A shaped release instead steps the response down, $100 \to 80 \to 60 \to 40 \to 20 \to 0$, so that device states recover in overlapping groups rather than together.

Figure 12.8 compares the two endings. In both panels the horizontal axis is minutes relative to the nominal end of the event at time zero. The upper axis is the delivered reduction in megawatts: gray drops from 100 to zero at the end, green steps down in five increments of 20 MW at ten-minute intervals. The lower axis is the resulting fleet load above baseline, which is the rebound. The abrupt release produces a peak of about 45 MW roughly ten minutes after the end; the shaped release spreads the same recovery over a longer interval and peaks at about 18 MW.

The shaped release gains twice. It lowers the post-event peak, which is what the rebound of Chapter 2 threatens to create, and it delivers extra reduction while it is doing so: the stepped tail is worth $(80 + 60 + 40 + 20)\,\mathrm{MW} \times 10\ \mathrm{min} = 33\,\mathrm{MWh}$. The cost is that part of the fleet stays curtailed longer, which is a comfort cost that equation (12.15) already prices through λ_2.

Principle 12.3. *The ending trajectory of a service is as much a part of the shape as the beginning. A release schedule is a control decision, and releasing a fleet all at once is the worst available choice.*

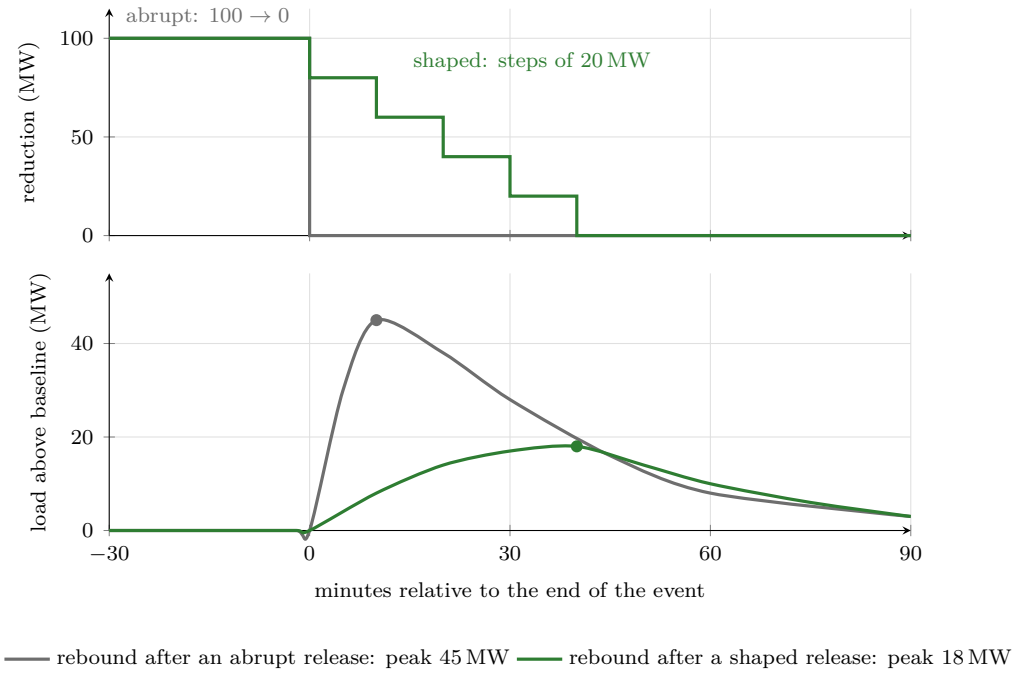


Figure 12.8: Two ways to end the same 100 MW event. Upper panel: the delivered reduction. Lower panel: the resulting load above baseline. Stepping the release down in five increments lowers the rebound peak from about 45 to about 18 MW and, because part of the fleet is still curtailed while the rest recovers, delivers an additional 33 MWh of reduction in the stepped tail. Curves are illustrative.

12.6.4 A complete formulation

Everything in the chapter can now be written in one place. Let $P^{\text{base}}(t)$ be the forecast uncontrolled load and $P_{\text{des}}(t)$ the load profile the utility wants to see. The required fleet response is their difference,

$$P_t^{\text{tgt}} = P_t^{\text{base}} - P_{\text{des},t}, \tag{12.16}$$

which is the one step that turns a desired *load* shape into a desired *response* shape. The DERMS then solves

$$\begin{aligned} \min_{P_{\text{T}},P_{\text{E}},P_{\text{B}}} \quad & \sum_{t=1}^{H} \left(P_{\text{T},t} + P_{\text{E},t} + P_{\text{B},t} - P_t^{\text{tgt}}\right)^2 \\ \text{subject to} \quad & T^{\min} \le T_{i,t} \le T^{\max} && \text{(comfort band, every building } i\text{)} \\ & \text{SOC}^{\min} \le \text{SOC}_{j,t} \le \text{SOC}^{\max} && \text{(battery limits, every unit } j\text{)} \\ & \text{SOC}_{k,T_k^{\text{dep}}} \ge \text{SOC}_k^{\text{req}} && \text{(departure requirement, every EV } k\text{)} \end{aligned} \tag{12.17}$$

together with the state dynamics of each class from Chapters 5 to 7. In production the squared term carries the weight λ_1 and the cost terms of equation (12.15) are added; the reserve of equation (12.13) appears as a bound that keeps P_{B} below the battery's availability; and the problem is re-solved at each level of Table 12.4 as new telemetry arrives. That is dynamic load shaping in a single statement: a least-squares fit of a trajectory, subject to the physics and the customer requirements of every device in the fleet.

12.6.5 The operational insight

There is one idea underneath all of the mechanisms in this chapter, and it is worth stating plainly: the VPP does not need any individual DER to behave like a generator. It needs the

portfolio to behave predictably.

Individual thermostats cycle on and off and their owners override events. Individual EVs leave when their drivers leave. Individual batteries hit their state-of-charge limits. None of this is suppressed by better hardware or better commands, and none of it has to be. With enough devices, enough diversity among them, and a schedule that sequences cohorts, substitutes classes, pre-conditions states, holds a reserve and corrects against telemetry, the aggregate trajectory is stable even though every component of it is not. That is the deepest operational idea behind virtual power plants, and dynamic load shaping is where it is used most fully: the chapter's mechanisms exist precisely because the parts are unreliable.

12.6.6 Forecasting, optimization and monitoring as one loop

Dynamic load shaping sits exactly where the three strands of this book meet. Forecasting supplies what is available: the flexibility of each class, hour by hour and location by location, together with the rebound each cohort will produce and the probability distribution of the response. Optimization chooses what to do with it: which cohorts, which device types, in which order, with how much pre-conditioning and how much reserve. Monitoring closes the loop: measured response against target, re-solved at each level of the hierarchy as the evening develops.

None of the three is useful alone. A forecast with no optimizer is a report; an optimizer with no forecast schedules resources that are not there; either one without monitoring is an open-loop plan against a system whose state is changing. The complete loop is the job, and it is the reason a VPP is an operations problem rather than a procurement problem.

12.7 What to retain

Definition 12.1 (Dynamic load shaping)**.** Dynamic load shaping is trajectory control using heterogeneous DER flexibility. It converts messy, time-limited individual responses into an aggregate that follows a target trajectory P_t^{tgt}, by coordinating pre-conditioning, cohort rotation, resource substitution, feedback correction and controlled recovery, rather than by issuing one large curtailment command.

Each mechanism in the definition appears in the chapter as a section, and Table 12.6 lists the chapter's quantities and objects with the subsection that defines each.

12.8 Exercises

Exercise 12.1. The thermostat fleet of Figure 12.1 has instantaneous capability

$$P_{\mathrm{T}}(t) = 50\, e^{-t/2.18}\,\mathrm{MW}.$$

Compute its average over each of the four hours of the service and compare the result with the thermostat column of Table 12.1. In which hours is the schedule feasible, and with what margin?

Exercise 12.2. Each cohort in Table 12.2 delivers +20 MW during its half-hour turn and then rebounds, contributing −8 MW in the first half hour after release and −4 MW in the second. Compute the gross contribution each cohort must provide in its turn so that the net response is 20 MW in all four intervals. Does the schedule remain feasible for cohorts of 15 000 thermostats, given that one cohort's maximum gross contribution is 20 MW?

Table 12.6: The quantities and mechanisms of dynamic load shaping.

Quantity or object	Definition	Section
P_t^{tgt}	target trajectory, a time series not a number	12.1.2
P_t^{del}	delivered fleet response, $\sum_k \Delta P_{k,t}$	12.2.3
tracking objective	$\min \sum_t (P_t^{\mathrm{del}} - P_t^{\mathrm{tgt}})^2$	12.1.2
portfolio schedule	per-class contributions that sum to P_t^{tgt}	12.1.4
cohort rotation	n cohorts of duration τ_0 serve up to $n\tau_0$	12.2.2
gross versus net	gross = net + recovery of released cohorts	12.2.4
pre-conditioning	choosing x_{t_0} before the service	12.2.5
e_t	tracking error $r_t - y_t$	12.3.1
actuator properties	delay, duration, uncertainty, cost per class	12.3.2
control hierarchy	day ahead, hour ahead, minutes, seconds	12.3.4
P_{net}	demand minus renewable generation	12.4.2
$R^{\max}$	bound on the ramp rate of the residual	12.4.3
$F(r, t, d)$	flexibility indexed by region, time and type	12.4.5
R_t	reserve, $P^{\mathrm{avail}} - P^{\mathrm{sched}}$	12.5.3
layered control	fast correction plus slow relief of the fast resource	12.5.5
J	weighted tracking, comfort, battery and EV costs	12.6.1
shaped release	stepped ending that limits the rebound peak	12.6.3

Exercise 12.3. From Figure 12.6, compute the largest hourly change in net load with and without the fleet, and verify that the shaped profile satisfies equation (12.10) with $R^{\max} = 50\,\mathrm{MW/h}$. Estimate how much energy the fleet moved into the early afternoon, and state in one sentence why the two profiles have nearly the same total energy.

Exercise 12.4. A 3 MW tracking error, as in equation (12.14), persists for a full hour and is corrected entirely from the battery reserve of equation (12.13). For the 60 MWh fleet of Figure 12.7, by how much does state of charge fall? Starting at $\mathrm{SOC} = 0.90$ with $\mathrm{SOC}^{\min} = 0.20$, for how many hours can the correction be sustained, and at what point must the slow loop of Section 12.5.5 intervene if the service lasts four hours?

12.9 Connection to the TERRA framework

Dynamic load shaping is what a conservation-graph model is for: a target on a node flow, a portfolio of components whose bounds move with their own states, and a least-squares fit of the simulated flow to the target over a horizon.

- **The target is a boundary condition on one node flow.** P_t^{tgt} in equation (12.16) is a prescribed time series for the aggregate flow on the grid edge, obtained by differencing a forecast load and a desired load. Shaping is then the inverse problem of finding component controls whose forward simulation reproduces that flow, and equation (12.1) is its least-squares residual.

- **Rebound is the same component relaxing, not a separate term.** The gross-versus-net accounting of equation (12.5) is not a correction factor bolted onto a reduction model. A released cohort is a set of thermal components recovering toward setpoint under their own dynamics, and its contribution to the node sum is negative for as long as that takes. A framework that simulates the components forward produces the 8 MW automatically; one that stores reductions as numbers cannot.

- **Pre-conditioning is a choice of initial condition.** Section 12.2.5 manipulates x_{t_0}, the state from which the service's forward simulation starts. Because the available flow is a function of the state, moving the initial condition moves the feasible set during the service; the pre-cooling hour and the service hour are one optimization over the same trajectory, not two decisions.

- **Horizon feasibility is a bound on accumulators.** The comfort band, the state-of-charge limits and the departure requirement in equation (12.17) are bounds on stored quantities that couple the steps of equation (12.12). Greedy dispatch fails for a structural reason: it optimizes a dynamic model one step at a time, and the accumulators carry the consequences forward.

- **Reserve and feedback are posterior quantities.** The reserve R_t of equation (12.13) is sized from the predicted distribution of the tracking error, and the correction of equation (12.14) is applied after telemetry has updated the component states. Both require inference of unmeasured states from partial measurements, which is exactly the posterior computation a probabilistic graph model supports; a deterministic model can schedule but cannot size a reserve.

- **Location is node identity.** $F(r, t, d)$ in Section 12.4.5 is flexibility attached to particular nodes of the graph. The 100 MW total and the 0.3 MW behind one transformer are flows on different edges and are not interchangeable, which is a topological fact the graph represents directly and an aggregate megawatt figure erases.

In summary, dynamic load shaping is the operating mode in which the whole modeling apparatus of the earlier chapters is actually needed at once: component dynamics to predict rebound, bounds to keep the horizon feasible, forward simulation to evaluate a candidate schedule, and posterior inference to size reserve and close the loop. The next chapter takes up the constraint that Sections 12.4.4 and 12.4.5 could only point at, namely what the distribution network itself permits.

Chapter 13

Distribution-grid constraints: why location changes DER value

13.1 System-wide and local flexibility

Every chapter so far has treated flexibility as a quantity with a magnitude, a direction and a duration. A 50 MW virtual power plant (VPP) was 50 MW, and the only questions were how long it could hold that figure and with what confidence. That treatment is correct as long as the only thing being balanced is the system as a whole: at the level of the bulk grid, a megawatt reduced in one town and a megawatt reduced in the next town are interchangeable. Once the distribution network becomes the thing being managed, they are not. The question of the chapter is

Where, electrically, is the flexibility, and does it reach the constraint that actually binds?

Jargon: feeder, substation, apparent power, reverse flow, hosting capacity, OPF

A *substation* is the facility where transmission voltage is stepped down to distribution voltage; a *feeder* is one of the medium-voltage circuits that leaves a substation and carries power out to a district, branching as it goes. *Apparent power* $S = \sqrt{P^2 + Q^2}$, measured in kilovolt-amperes (kVA), is the quantity that transformers and conductors are rated for, because heating depends on current and current depends on S, not on real power P alone; Q is reactive power, in kilovars. *Reverse flow* is power traveling from the customer end of a feeder back toward the substation, which happens when local generation exceeds local consumption. *Hosting capacity* is the amount of additional distributed generation or load a given part of the network can accept before some limit is violated. *Optimal power flow* (OPF) is the classical problem of choosing generator outputs to minimize cost subject to the network's physical equations and limits.

Key idea

Flexibility has a location as well as a magnitude and a duration. The flexibility that counts against a local constraint is only the flexibility that is electrically downstream of it.

The six body sections that follow develop that statement in order: first the distinction between system-wide and local flexibility and the sensitivity that formalizes it; then the physical limits of distribution equipment and the voltage problem that rooftop solar creates; then the dispatch problem that results, which is a form of optimal power flow; then what location does to aggregation and to the order in which resources are called; then a worked transformer example carried through to a probabilistic constraint; and finally the dependence of all of it on

network topology, the system that supplies the topology, and the planning decision that the data supports.

This section separates two quantities that the earlier chapters treated as one, and then introduces the matrix that relates a change in DER power to a change in the network.

13.1.1 Two different quantities

Suppose a utility has enrolled 100 MW of flexible DER capacity across its service territory, and that one specific distribution transformer is overloaded by 2 MW on a hot evening. The 100 MW figure is nearly irrelevant to that problem. What matters is how much flexible capacity sits behind that transformer, and if the answer is 0.5 MW, then 99.5 MW of the fleet can do nothing at all about the overload. It can be dispatched, it can be paid, and the transformer will still be 2 MW over its rating.

Written as an inequality between two quantities rather than as a remark,

$$F_{\text{sys}}(t) \neq F_{\text{loc}}(t, r), \tag{13.1}$$

where r denotes an electrical region. The left side is what the aggregator advertises; the right side is what a local constraint can draw on. The two can differ by two orders of magnitude, and nothing in the left-hand number reveals that.

13.1.2 The network as a tree

The reason for the gap is the structure of a distribution network. Power leaves a substation on a small number of feeders; each feeder branches into laterals; each lateral feeds service transformers; each service transformer feeds a handful of customers and whatever DERs those customers own. The result is close to a tree: in normal operation most distribution feeders are operated radially, so there is exactly one path from the substation to any customer.

Figure 13.1 draws such a tree. The top box is the substation; the two boxes below it are feeders; the three boxes on the next level are service transformers; the bottom boxes are groups of customers with DERs, labeled A, C and B from left to right, each with the flexible capacity it owns. The transformer T_1 is drawn in red because it is the constrained element, and the dashed outline encloses everything electrically downstream of it. Reading the figure, the overload at T_1 can be relieved only by group A, because group A is the only DER group inside the dashed outline. Group C is on the same feeder but behind a different transformer, and group B is on a different feeder altogether. Both can change their power without changing the loading of T_1 at all.

The consequence for notation is that the flexibility of Chapter 1 acquires a second argument. Instead of a single function of time,

$$F(t) \longrightarrow F(t, r), \tag{13.2}$$

the flexibility of a fleet is a function of time *and* of the electrical region in which it can be used. This is the region index already present in the notation table; the rest of the chapter is about what happens when it is taken seriously.

Principle 13.1. *Flexibility is indexed by electrical location as well as by time and duration. A megawatt is useful for a constraint only if it is electrically positioned to affect it.*

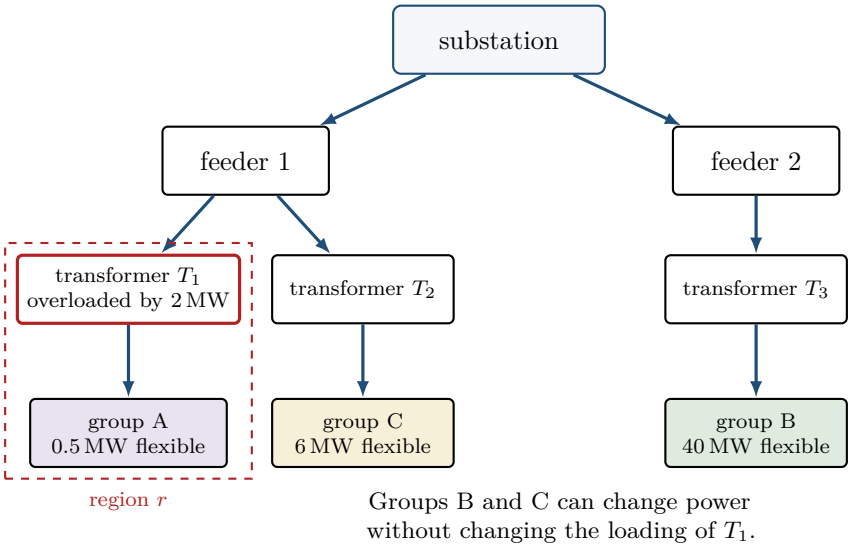


Figure 13.1: A radial distribution network. Boxes are network elements; arrows are the direction of normal power flow. The dashed outline is the electrical region r downstream of the constrained transformer T_1; only the 0.5 MW of group A lies inside it, out of a 46.5 MW total for the three groups shown. Capacities are illustrative.

13.1.3 A feeder example

A concrete case makes the arithmetic explicit. A feeder has a thermal limit of 10 MW, and the day-ahead forecast of its loading for the evening peak is 11.2 MW. The constraint the operator must enforce is

$$P_{\text{feeder}}(t) \le P^{\max}_{\text{feeder}} = 10\,\text{MW}, \tag{13.3}$$

so the corrective action required on that feeder is

$$11.2 - 10 = 1.2\,\text{MW}. \tag{13.4}$$

Whether the problem is solvable now depends on one number and one number only: the flexibility available on that feeder. If the DERs behind the feeder head can jointly reduce net load by 1.5 MW, the constraint can be met with 0.3 MW to spare. If they can reduce only 0.8 MW, it cannot be met and the operator must fall back on something else, such as switching part of the load to an adjacent feeder or accepting a period of overload. The territory-wide fleet size never enters the calculation.

Two features of equation (13.4) are worth noting. The first is that the required action is a difference of two forecasts, the load forecast and the limit, and so inherits the error of the load forecast; Section 13.5.3 returns to this. The second is that 1.2 MW is a requirement at a particular hour, not an energy requirement, so a resource that can supply 1.2 MW for ten minutes does not solve a three-hour overload. Duration matters exactly as it did in Chapter 4.

13.1.4 DER control as a network input

The loading of the feeder is not an exogenous quantity. It is the sum of everything connected behind it, which for the purposes of this book is

$$P_{\text{feeder}}(t) = P^{\text{base}}(t) + P_{\text{EV}}(t) + P_{\text{bat}}(t) - P_{\text{PV}}(t), \tag{13.5}$$

with P^{base} the uncontrolled consumption of the connected customers, P_{EV} the electric-vehicle charging load, P_{bat} the battery power (positive when charging), and P_{PV} the local solar

generation. Three of those four terms are controllable, which means the operator has direct authority over the quantity being constrained. Feeder loading can be reduced by slowing EV charging, by discharging batteries, or by reducing heating and cooling load; it can be increased by charging batteries or by accelerating EV charging.

This is a change of role rather than a change of equation. In Chapter 2 the DER was an instrument for reducing a system peak or following a price. Here the same control variables appear in a physical network constraint, so a DER command is a network-control action in the same sense that a capacitor switch or a voltage regulator tap is. The distinction between a market resource and a network asset becomes a matter of which equation the control appears in, not of what the device is.

13.1.5 The sensitivity matrix

Figure 13.1 made the point qualitatively: 1 MW from group A relieves the constraint at T_1 and 1 MW from group B does almost nothing. The quantitative statement replaces the tree picture with a linearization. Collect the network quantities of interest, voltages at nodes and loadings on branches, in a vector z, and collect the DER power changes in a vector ΔP with one entry per DER or per DER group. To first order,

$$\Delta z = \mathcal{H}\,\Delta P, \tag{13.6}$$

where $\mathcal{H}$ is the sensitivity matrix: the entry $\mathcal{H}_{ki}$ is the change in network quantity k caused by a unit change in the power of DER i, evaluated at the present operating point. The matrix is written $\mathcal{H}$ here because the plain symbol H is reserved for the optimization horizon of Chapter 3.

Equation (13.6) is the formal content of "location matters". For a radial network and a loading constraint, a row of $\mathcal{H}$ is close to an indicator function: the entry is approximately 1 for DERs downstream of the constrained element and approximately 0 for everything else, with small deviations caused by losses. For a voltage constraint the rows are not indicators at all; they decay with electrical distance, so a DER can influence a voltage some way off, weakly. In both cases the useful statement is the same. Megawatts of nameplate are not the unit in which network value is measured; the product of nameplate and sensitivity is.

13.2 Equipment and voltage limits

The constraints that bind on a distribution network are of four kinds, and this section takes them in turn: transformer apparent power, equipment temperature, node voltage and conductor current. Table 13.1 summarizes them before they are developed individually. The second column names the quantity that is limited, the third the form of the limit, and the fourth says whether the constraint depends on history as well as on the present instant. The entries in that last column are the reason a distribution constraint is not simply a bound on a power.

13.2.1 Transformer apparent power

A distribution transformer carries a nameplate rating $S_{\text{tr}}^{\max}$ in kVA, and the operational constraint is

$$S(t) = \sqrt{P(t)^2 + Q(t)^2} \le S_{\text{tr}}^{\max}. \tag{13.7}$$

The rating is in apparent power because the transformer's losses and heating follow the current, and the current follows S. A load of 450 kW at a power factor of 0.95 therefore presents $S = 474$ kVA to a 500 kVA transformer, not 450 kVA, and only 26 kVA of headroom remains.

Uncontrolled EV charging is the canonical way to exceed such a rating. Ten vehicles on Level 2 chargers behind one 500 kVA transformer add about 72 kW if they all start at once,

Table 13.1: The four limits that bind on a distribution network.

Constraint	Quantity limited	Form	Depends on history?
Transformer loading	apparent power $S(t)$	$S(t) \le S_{\text{tr}}^{\max}$	indirectly, through temperature
Equipment temperature	hot-spot temperature T_{eq}	$T_{\text{eq}} \le T_{\text{eq}}^{\max}$	yes; first-order thermal lag
Node voltage	voltage magnitude V_i	$V^{\min} \le V_i \le V^{\max}$	no; algebraic in the present state
Conductor current	current magnitude $\lvert I_\ell(t)\rvert$	$\lvert I_\ell(t)\rvert \le I_\ell^{\max}$	indirectly, through conductor temperature

which is enough to drive a peak of 550 kVA on a transformer rated for 500. The overload is not caused by a shortage of energy or of generation; it is caused by the coincidence of a dozen charging decisions. Figure 13.2 shows the two profiles. The horizontal axis is the hour of the evening, the vertical axis is apparent power at the transformer in kVA, the dashed line is the 500 kVA rating, and the shaded wedge is the interval during which the uncontrolled profile sits above it. The managed profile carries the same charging energy but is held below the rating by deferring part of it past the peak, so the shaded wedge disappears. This is the clearest single use case for a DER management system acting on a distribution constraint: the asset is not replaced, and the charging still happens.

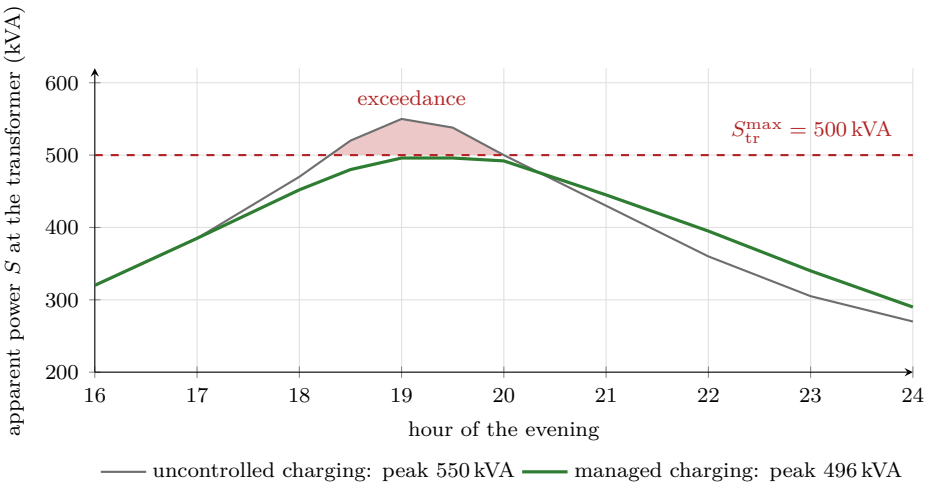


Figure 13.2: Apparent power at a 500 kVA service transformer on an evening with ten EVs charging. The uncontrolled profile exceeds the rating from about 18:20 to 20:00; the managed profile delivers the same charging energy later and stays below it. Profiles are illustrative.

13.2.2 Thermal limits have memory

Equation (13.7) is an approximation of the real limit. What destroys a transformer is not apparent power but temperature: the insulation ages at a rate that depends on the hot-spot temperature inside the winding, and that temperature responds to loading with a lag of hours. A first-order model is

$$T_{\text{eq},t+1} = f\left(T_{\text{eq},t},\, I_t,\, T_{\text{amb},t}\right), \tag{13.8}$$

with T_{eq} the equipment temperature, I_t the current and T_{amb} the ambient temperature. The constraint that the operator genuinely cares about is $T_{\text{eq}} \le T_{\text{eq}}^{\max}$, and because of the lag a short

overload may never reach the limit while a prolonged one at the same level certainly will.

Figure 13.3 shows the two cases. The horizontal axis is the time since the overload began, the vertical axis is the hot-spot temperature, the dashed horizontal line is the limit, and the two curves are the response to a four-hour overload and to a one-hour overload at the same loading level. The curves coincide for the first hour, because the equipment cannot tell how long the overload will last; thereafter the short case relaxes back toward its initial temperature, peaking at about 92°C, while the prolonged case continues to rise and crosses 110°C after about 2.6 h. The same loading is acceptable in one case and not in the other.

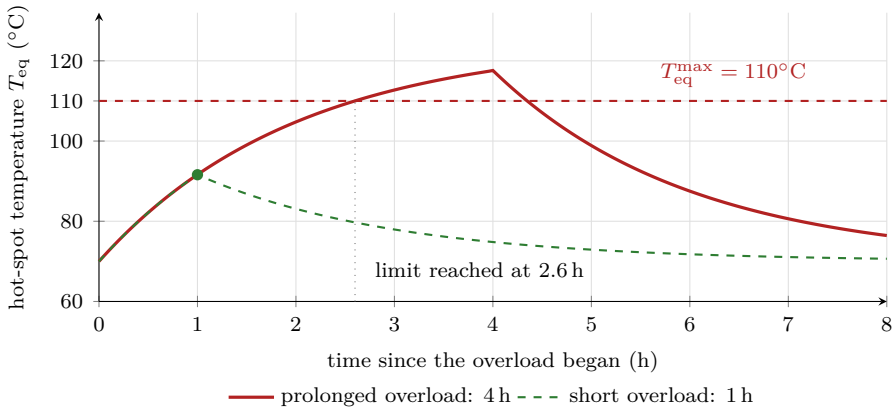


Figure 13.3: Equipment temperature under a short and a prolonged overload at the same loading. Both curves follow $T_{eq}(t) = T_\infty + (T_0 - T_\infty)\, e^{-t/\tau}$ while the overload lasts and relax back toward 70°C with the same τ afterward, with $T_0 = 70°C$, $T_\infty = 125°C$ and $\tau = 2\,h$. Parameters are illustrative.

Equation (13.8) is another state variable of the form met throughout this book, $x_{t+1} = f(x_t, u_t, w_t)$, with the equipment temperature as the state, the DER-influenced current as the control channel and the ambient temperature as the disturbance. It has the same structure as the building temperature of Chapter 5: a stored quantity, a band it must stay inside, and a time constant that creates flexibility. The practical consequence is that DER dispatch can be used to manage an equipment thermal state rather than an instantaneous power, which is what dynamic line-rating schemes do on transmission conductors and what a DERMS with a thermal model of a transformer can do on distribution.

13.2.3 Voltage limits

Voltage must be held inside a band at every node of the network,

$$V^{\min} \le V_i \le V^{\max}, \tag{13.9}$$

typically within a few percent of nominal; 0.95 to 1.05 per unit is a common operating band. The band exists because customer equipment is designed for nominal voltage, and because sustained deviation damages motors and electronics.

Unlike the thermal constraints, voltage is algebraic: it is determined by the present injections, not by their history. Every load and every DER injection moves it. For a node i supplied through a path of resistance R_{ij} and reactance X_{ij} from a stiffer node j, the standard linearization is

$$\Delta V \approx R_{ij}\, \Delta P + X_{ij}\, \Delta Q, \tag{13.10}$$

so increasing the real power injected at a node raises its voltage and increasing the real power consumed lowers it, with reactive power acting through the reactance in the same way. On

distribution circuits R and X are comparable in magnitude, which is why real-power control affects distribution voltage appreciably; on transmission, where $X \gg R$, it largely does not. Controlling P and, where the inverter allows it, Q is therefore a way of keeping equation (13.9) satisfied.

13.2.4 Reverse flow and voltage rise

The most common voltage problem on a modern feeder is caused by rooftop solar at midday. When local generation exceeds local consumption the net load is negative,

$$P_{\text{net}}(t) = P_{\text{load}}(t) - P_{\text{PV}}(t) < 0, \tag{13.11}$$

and power flows from the customer end of the feeder back toward the substation. By equation (13.10) with $\Delta P > 0$ injected, the local voltage rises, and it rises most at the far end of the feeder where R_{ij} is largest.

Figure 13.4 shows the effect. The horizontal axis is distance along a feeder from the substation, the vertical axis is voltage in per unit, and the two dashed horizontal lines are the limits of equation (13.9). The lower curve is an evening peak with no solar: voltage falls with distance, the familiar voltage drop. The upper curve is midday with a 3 MW local surplus and no DER response: voltage rises with distance and leaves the band beyond about 6.7 km, in the shaded region. The middle curve is the same midday hour with 2 MW of the surplus absorbed by flexible load; the residual reverse flow is one third as large, the rise is one third as steep, and the whole feeder stays inside the band. The same feeder is therefore at risk from both directions on the same day, and the binding limit changes sign between afternoon and evening.

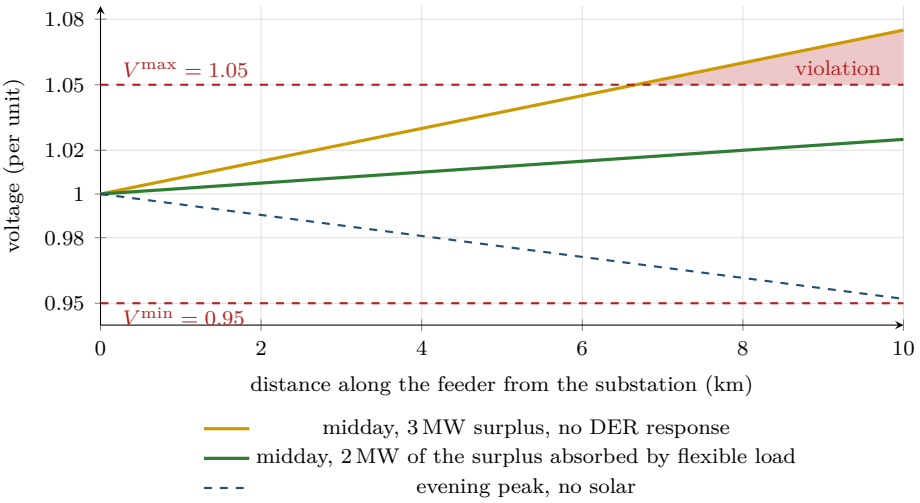


Figure 13.4: Voltage profile along a 10 km feeder. The curves are the linear profiles $V(d) = 1.000 + 0.0075\,d$ (midday, 3 MW of reverse flow), $V(d) = 1.000 + 0.0025\,d$ (midday, 2 MW absorbed so 1 MW remains) and $V(d) = 1.000 - 0.0048\,d$ (evening peak), with d in kilometers; the rise is taken proportional to the reverse flow, and the slopes are illustrative values consistent with equation (13.10) for a uniformly loaded feeder.

The DERMS response to the midday case is the opposite of the evening case. Instead of reducing load it increases it: it advances EV charging, charges batteries, and runs flexible thermal load, all of which consume the local surplus where it is produced. Only if that is insufficient does it curtail the solar inverters. Flexible demand, used this way, is a voltage-control resource.

13.2.5 Absorbing a local surplus

Example 13.1 (Absorbing a 3 MW midday surplus)**.** A feeder segment has a midday PV surplus of 3 MW. The DERs behind that segment can absorb 2 MW by advancing EV charging and 1.5 MW by charging batteries, so 3.5 MW of upward-load flexibility is available against a 3 MW requirement. The DERMS dispatches the full 2 MW of EV charging and 1 MW of battery charging:

$$2 + 1 = 3\,\text{MW}, \tag{13.12}$$

which cancels the surplus exactly, removes the reverse flow, and leaves 0.5 MW of battery capability unused as a margin. No solar is curtailed and no energy is wasted: the EV energy would have been consumed in the evening anyway, and the battery energy is available for the evening peak.

This is the clearest case in the book of increased consumption being a grid service. It reverses the instinct carried over from demand response, where the service was always a reduction, and it is one reason the direction-aware flexibility pair $F^{\uparrow}$ and $F^{\downarrow}$ of Chapter 6 must be tracked separately. A fleet with abundant $F^{\uparrow}$ and no $F^{\downarrow}$ is useless at midday on this feeder and valuable at 19:00, and the same fleet in the opposite state is valuable at midday and useless in the evening.

13.2.6 Conductor current

The last of the four limits is the current carried by each line segment,

$$|I_\ell(t)| \le I_\ell^{\max} \quad \text{for every branch } \ell, \tag{13.13}$$

which is the congestion constraint of distribution. The absolute value matters because the limit binds in both directions: a conductor is as easily overloaded by reverse flow at midday as by load at 19:00. With the constraint written this way, the dispatch problem becomes

$$\min_{u_t} J(x_t, u_t) \quad \text{subject to} \quad |I_\ell(x_t, u_t)| \le I_\ell^{\max}, \tag{13.14}$$

where the current on each branch is a function of the state and the control, computed through the network model. The DERMS is now respecting physical network limits in addition to the device and customer limits of Chapters 5 to 7.

13.3 Network-aware dispatch

With the limits in hand, this section states the dispatch problem they imply, identifies it as a relative of optimal power flow, and then develops three of its consequences: phase imbalance, hosting capacity and the ranking of resources by sensitivity.

13.3.1 From a power target to a constrained dispatch

The dispatch problem of Chapter 3 was a target-matching problem,

$$\sum_i u_i = P^{\text{tgt}}, \tag{13.15}$$

subject to each device's own feasible set $\mathcal{U}_i(x_i)$. The network-aware version keeps that equation and adds the physics of the circuit:

$$\begin{aligned} &\sum_i u_i = P^{\text{tgt}} \\ \text{subject to} \quad & g(V, \theta, P, Q) = 0, \\ & V^{\min} \le V \le V^{\max}, \\ & |I_\ell| \le I_\ell^{\max} \quad \text{for all } \ell, \end{aligned} \tag{13.16}$$

where $g(V, \theta, P, Q) = 0$ is the set of power-flow equations relating voltage magnitudes V, voltage angles θ and the real and reactive injections. The feasible set is no longer a product of per-device sets; it is coupled through the network, so the choice made for one device changes what is feasible for another even when the two devices share no owner, no type and no customer.

13.3.2 Relation to optimal power flow

Equation (13.16) is recognizably a member of the optimal power flow family. Classical OPF chooses the outputs of generators to minimize production cost subject to exactly these network equations and limits, and it has been solved in transmission control centers for decades. A network-aware DERMS solves a problem with the same mathematical shape and different decision variables: instead of a few dozen generators with smooth cost curves it has thousands of thermostats, EVs and batteries, each with a state, a customer constraint and a feasible set that changes hour by hour.

The differences are practical rather than structural. The DER problem has many more variables; the variables are coupled across time by stored energy, so it is a multi-period problem rather than a single snapshot; the feasible sets depend on states that are only partially measured; and the distribution network it is posed on is unbalanced, which the next subsection takes up. But the useful intuition is that network-aware DER dispatch is optimal power flow with distributed, state-constrained resources in place of generators.

13.3.3 Unbalanced phases

Transmission analysis usually assumes the three phases are balanced, which allows a three-phase circuit to be modeled by a single-phase equivalent. Distribution circuits are not balanced. Individual customers are connected to one phase, laterals serve different numbers of customers, and the result is

$$P_a \neq P_b \neq P_c, \tag{13.17}$$

with voltages and currents differing from phase to phase on the same pole.

The consequence for DER dispatch is that electrical location includes phase. A single-phase residential EV charger loads exactly one of the three phases, so if phase b is the one near its limit, only the DERs on phase b can help, and dispatching an equal amount on all three phases would relieve one third of the problem while needlessly disturbing the other two. Location is therefore a three-part label, not a geographic one: which feeder, which transformer, and which phase.

13.3.4 Hosting capacity

Hosting capacity answers a planning question: how much additional DER can be connected to a given part of the network before one of the limits of Table 13.1 is violated? Historically the question was asked about rooftop solar, and the answer was computed once per feeder as a fixed number of megawatts, used to decide whether an interconnection application could be approved.

Managed flexibility changes the answer. If batteries behind the feeder can absorb a midday surplus, as in Section 13.2.5, the voltage rise that limited solar penetration no longer occurs, and more solar can be connected without violating anything. If EV charging is managed, as in Section 13.2.1, the thermal overload that limited EV penetration no longer occurs. Active control does not change the conductors or the transformers; it changes the set of operating points that the existing equipment is asked to carry, and therefore the feasible operating envelope.

13.3.5 Static and dynamic hosting capacity

The two views can be written side by side. The static calculation produces a constant,

$$P_{\mathrm{PV}}^{\max} = 5\,\mathrm{MW}, \tag{13.18}$$

usually derived from a worst-case combination of minimum load and maximum irradiance. The honest statement is a function,

$$P_{\mathrm{PV}}^{\max} = f(\text{load},\ \text{weather},\ \text{topology},\ \text{DER flexibility}), \tag{13.19}$$

whose value at a given hour may be well above the static figure and, in rare combinations, below it.

The practical implication is a choice about where to put conservatism. A static number is simple to administer and is set by the worst hour of the year, so it refuses connections that would have been safe in almost every hour. A dynamic number uses the network as efficiently as the present conditions allow, at the cost of requiring a model, live data and an operating process that can act when the capacity shrinks. The same trade-off between a fixed conservative commitment and a state-dependent one appeared for fleet capacity in Chapter 4; here it applies to the network.

13.3.6 Locational value

Consider two identical 1 MW batteries. The first is installed on a feeder with ample capacity, where no constraint has bound in years. The second is installed behind a transformer that is overloaded on thirty evenings a year. The two have the same nameplate, the same energy capacity and the same round-trip efficiency, and the second is worth considerably more to the utility, because it can relieve a constraint that the first cannot reach.

Principle 13.2. *The grid value of a DER is set by the sensitivity of a binding constraint to its power, not by its rating. Two identical devices in different electrical locations are different resources.*

This is the physical basis of the locational pricing and locational-value analyses that appear in distribution planning, and it is also why an aggregator that enrolls customers wherever they happen to sign up accumulates a fleet whose network value is largely accidental.

13.3.7 Sensitivity as a dispatch priority

The quantity that makes locational value computable is the derivative already introduced in equation (13.6). Let L be the loading of the constrained element, for example the apparent power of transformer T_1 in Figure 13.1. For each DER i define

$$\frac{\partial L}{\partial P_i}, \tag{13.20}$$

the change in that loading per unit change in the DER's power. A value of 1 means a megawatt of response moves the loading by a megawatt; a value near 0 means the DER is irrelevant to this particular constraint, however large it is.

Table 13.2 lists the three DER groups of Figure 13.1 with their sensitivities to two different constraints: the loading L_{T_1} of the constrained transformer and the loading L_{F_1} of feeder 1. Reading the rows, group A is useful for both constraints, because it is downstream of both; group C is useless for the transformer and fully effective for the feeder, because it is downstream of the feeder head but not of T_1; group B is useless for both. The 40 MW in group B is thus 40 MW of system-wide capacity and zero MW of capacity for either of these two constraints, which is equation (13.1) in tabular form. A network-aware optimizer ranks candidates by the product of sensitivity and available power and dispatches the top of that list, which is a different ordering from the cost-based one.

Table 13.2: Sensitivity of two constraints to the three DER groups of Figure 13.1. Values are rounded; losses would make the nonzero entries slightly below 1.

DER group	Electrical location	$\partial L_{T_1}/\partial P_i$	$\partial L_{F_1}/\partial P_i$	Useful for
A, 0.5 MW	behind T_1, feeder 1	1.00	1.00	both constraints
C, 6 MW	behind T_2, feeder 1	0.00	1.00	the feeder only
B, 40 MW	feeder 2	0.00	0.00	neither

13.4 Location in aggregation and in the merit order

Taking location seriously changes two things that earlier chapters built: the aggregate the DERMS reports, and the order in which it calls resources. This section treats both, together with the feasible set that results from combining network and customer limits.

13.4.1 Region-indexed fleets

A VPP reported as a single number cannot answer a local question. The aggregation of Chapter 4 must therefore be replaced by a set of regional aggregates,

$$F_{\text{VPP}}(t) \longrightarrow \{F_1(t),\ F_2(t),\ \ldots,\ F_R(t)\}, \tag{13.21}$$

one per electrical region, with the system total as their sum. Figure 13.5 draws the decomposition for a 100 MW fleet across four feeders. The top box is the fleet total; the four boxes below are the per-feeder flexibilities, which sum to the total; the right-hand box is drawn in red because it is the feeder with a constraint. The figure makes the arithmetic of Section 13.1.1 visible: a constraint on feeder D can draw on 5 MW, not on 100 MW, and no amount of capacity on feeders A to C changes that.

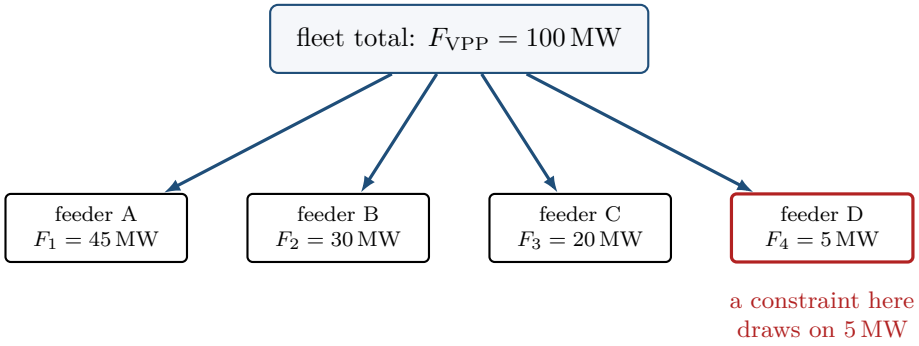


Figure 13.5: A 100 MW fleet decomposed by electrical region. The regional figures sum to the total, but only one of them is available to any given local constraint. Capacities are illustrative.

13.4.2 Spatial resolution as a modeling choice

How finely should the regions be defined? One region for the whole service territory is the aggregate of Chapter 4 and misses every local constraint. One region per household is faithful and expensive: it requires a model of every service transformer and every lateral, telemetry to match, and an optimization with as many spatial indices as customers. Neither extreme is the right answer.

The resolution that is right is the resolution the decision needs. If the decisions being made are transformer-level load management, the model must resolve service transformers. If they are

feeder-level congestion management, feeders suffice. This is the same principle as the temporal and model-order choices of Chapters 3 and 4, applied to space: model detail should match the decision, and detail beyond it costs data, computation and maintenance without changing any action.

13.4.3 Grid and customer constraints intersect

Network constraints cannot be analyzed separately from the device and customer constraints of the earlier chapters, because the same megawatt has to satisfy both. Consider a transformer that needs 1 MW of EV curtailment for the next three hours. Downstream of it, EVs can collectively reduce charging by 1.5 MW at this instant, which looks sufficient. But most of those vehicles are close to their departure deadlines, so the curtailment that can be *sustained* for three hours without missing a deadline is only 0.6 MW. The instantaneous figure satisfies the network and violates the customers; the sustainable figure satisfies the customers and not the network.

The set the optimizer may actually use is therefore an intersection,

$$\mathcal{F}_{\text{total}} = \mathcal{F}_{\text{grid}} \cap \mathcal{F}_{\text{DER}} \cap \mathcal{F}_{\text{cust}}, \tag{13.22}$$

drawn in Figure 13.6. The three circles are the three feasible sets, labeled in their non-overlapping lobes, and the region where all three overlap is $\mathcal{F}_{\text{total}}$. The figure is a reminder that each set can be generous while the intersection is narrow or, as in the example above, empty, and that a report of any one set alone overstates what is dispatchable. When the intersection is empty the operator has three options and no others: shorten the requirement, recruit flexibility elsewhere in the same region, or accept the violation.

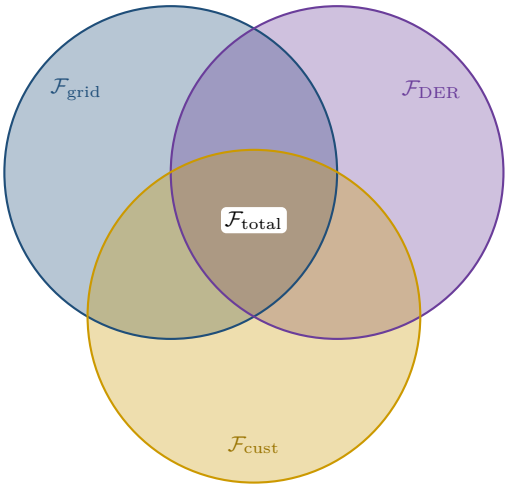


Figure 13.6: The dispatchable set as an intersection, equation (13.22). Each of the three sets can be generous while their intersection is narrow. In the example of Section 13.4.3 the 1.5 MW instantaneous curtailment lies in $\mathcal{F}_{\text{grid}} \cap \mathcal{F}_{\text{DER}}$ but outside $\mathcal{F}_{\text{cust}}$, and the 0.6 MW sustainable curtailment lies in $\mathcal{F}_{\text{DER}} \cap \mathcal{F}_{\text{cust}}$ but outside $\mathcal{F}_{\text{grid}}$, so the intersection is empty for a three-hour requirement.

13.4.4 Location-dependent merit order

Earlier chapters established a rough preference ordering among resource classes: use the resource that is cheapest, least intrusive and fastest first, which for most fleets means EVs, then thermostats, then batteries. Network constraints overturn that ordering whenever they bind locally.

Suppose the system-wide optimizer would prefer EVs, and the overloaded feeder contains few EVs but many batteries. The local requirement then forces battery dispatch first, not because batteries became cheaper but because they are the only resource with a nonzero entry in the relevant row of $\mathcal{H}$. The merit order is therefore state-dependent, as Chapter 4 already showed through stored energy, and location-dependent as well. A fixed resource ranking, written once into a dispatch rule, will be wrong for any constraint whose region does not match the ranking's assumptions.

13.5 A worked example with uncertainty

This section runs one transformer constraint from forecast to dispatch and then adds forecast error, which converts a deterministic requirement into a probabilistic one.

13.5.1 The required local response

A service transformer has a limit $P_{\text{tr}}^{\max} = 4\,\text{MW}$, and the forecast of its uncontrolled loading for the evening is 3.6, 4.4, 4.7 and 4.1 MW at 17:00, 18:00, 19:00 and 20:00. The local response the DERMS must secure in each hour is whatever is needed to bring the loading back to the limit,

$$F(t) = \max\bigl(0,\ P_{\text{load}}(t) - P_{\text{tr}}^{\max}\bigr), \tag{13.23}$$

which gives 0, 0.4, 0.7 and 0.1 MW for the four hours. Table 13.3 sets the three columns side by side. The required response is a trajectory, not a single number, and it totals 1.2 MWh of shifted energy over three hours.

Table 13.3: Required local response at a 4 MW transformer, from equation (13.23).

Hour	Forecast uncontrolled load (MW)	Limit (MW)	Required response F (MW)
17:00	3.6	4.0	0.0
18:00	4.4	4.0	0.4
19:00	4.7	4.0	0.7
20:00	4.1	4.0	0.1
Total	—	—	1.2 MWh over three hours

Figure 13.7 draws the same information. The horizontal axis is the hour, the vertical axis is transformer loading in megawatts, the dashed line is the 4 MW limit, the gray curve is the uncontrolled forecast, and the shaded area between the curve and the limit is the exceedance that must be removed. The three vertical arrows are the hourly requirements of Table 13.3, and the green dashed curve is the managed trajectory, which rides along the limit during the three constrained hours and carries the deferred energy afterward. The managed curve shows the shape of the answer: a target trajectory for a local group of DERs, with its level set by network capacity rather than by a price or a system peak.

13.5.2 Allocating the response

The local flexibility behind this transformer consists of EVs able to reduce charging by 0.5 MW and a battery able to supply 0.4 MW, a total of 0.9 MW. The binding hour is 19:00, where 0.7 MW is required. One feasible allocation is

$$0.4\ (\text{EV curtailment}) + 0.3\ (\text{battery discharge}) = 0.7\,\text{MW}, \tag{13.24}$$

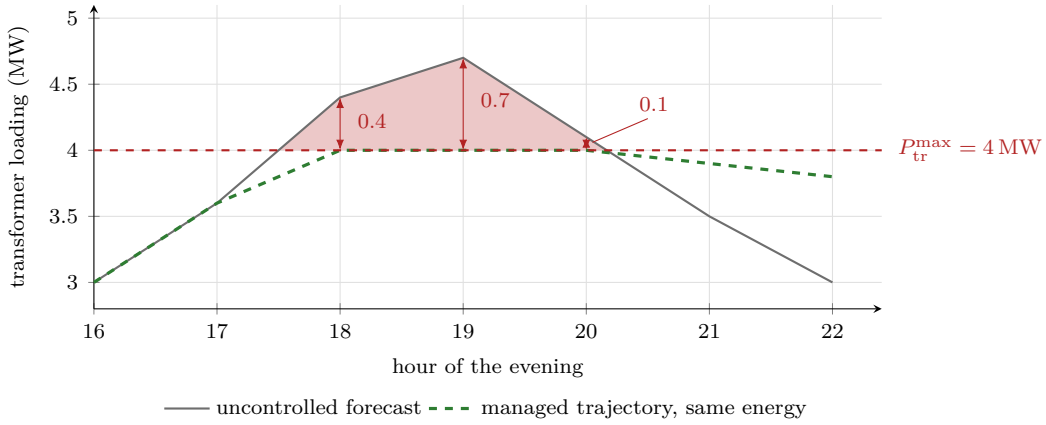


Figure 13.7: The required-flexibility trajectory at a 4 MW transformer. The shaded area is the forecast exceedance; the arrows are the hourly requirements 0.4, 0.7 and 0.1 MW; the managed trajectory holds the loading at the limit and defers the same energy to later hours. Values from Table 13.3.

which holds $P_{\text{tr}}(19{:}00) \le 4\,\text{MW}$ while leaving 0.1 MW of EV capability and 0.1 MW of battery capability in reserve. The easier hours, 18:00 and 20:00, can be covered by EV curtailment alone.

The choice between EV curtailment and battery discharge is not arbitrary. EV curtailment defers energy the customer will need before departure and so consumes the slack of Chapter 7; battery discharge consumes stored energy and so lowers the state of charge available for the remaining hours. Allocating the easy hours to EVs and the hard hour to a mix preserves battery energy for the possibility that the 19:00 forecast is low, which is precisely the situation the next subsection examines. This allocation, repeated every few minutes against updated forecasts, is the whole of network-aware DER dispatch at one transformer.

13.5.3 Forecast uncertainty

Neither number in equation (13.24) is known exactly. The transformer load forecast for 19:00 is 4.7 ± 0.4 MW and the deliverable local flexibility is 0.8 ± 0.2 MW, where the second figure differs from the 0.9 MW nameplate because some vehicles may have unplugged and the battery state of charge is itself estimated. The quantity that must stay below the limit is the difference, and the risk of a violation is

$$\Pr(P_{\text{load}} - F > P_{\text{tr}}^{\max}). \tag{13.25}$$

Treating the two as independent normal variables, $P_{\text{load}} \sim \mathcal{N}(4.7, 0.4^2)$ and $F \sim \mathcal{N}(0.8, 0.2^2)$, their difference is normal with

$$\begin{aligned} \mathbb{E}[P_{\text{load}} - F] &= 4.7 - 0.8 = 3.9\,\text{MW}, \\ \sigma = \sqrt{\text{Var}(P_{\text{load}} - F)} &= \sqrt{0.4^2 + 0.2^2} = \sqrt{0.20} = 0.447\,\text{MW}, \end{aligned} \tag{13.26}$$

so the exceedance probability is

$$\Pr(P_{\text{load}} - F > 4) = 1 - \Phi\!\left(\frac{4 - 3.9}{0.447}\right) = 1 - \Phi(0.224) \approx 0.41. \tag{13.27}$$

The deterministic plan of Section 13.5.2 looked comfortable, with 0.7 MW required against 0.9 MW available. Accounting for forecast error, the transformer is over its limit in roughly two evenings out of five. The mean of the difference is only 0.1 MW below the limit, which is a quarter of a standard deviation, so the outcome is close to a coin toss.

Figure 13.8 shows the two densities involved. The horizontal axis is the net transformer loading $P_{\text{load}} - F$ in megawatts, the vertical axis is probability density, and the vertical dashed line is the 4 MW limit. The wide left curve is the present forecast, centered at 3.9 MW, with the shaded area to the right of the limit equal to the 0.41 computed above. The second curve is the same distribution shifted so that only 1 % of its mass lies beyond the limit, which requires an expected flexibility of 1.74 MW rather than 0.8 MW. The gap between the two centers is the cost of the guarantee.

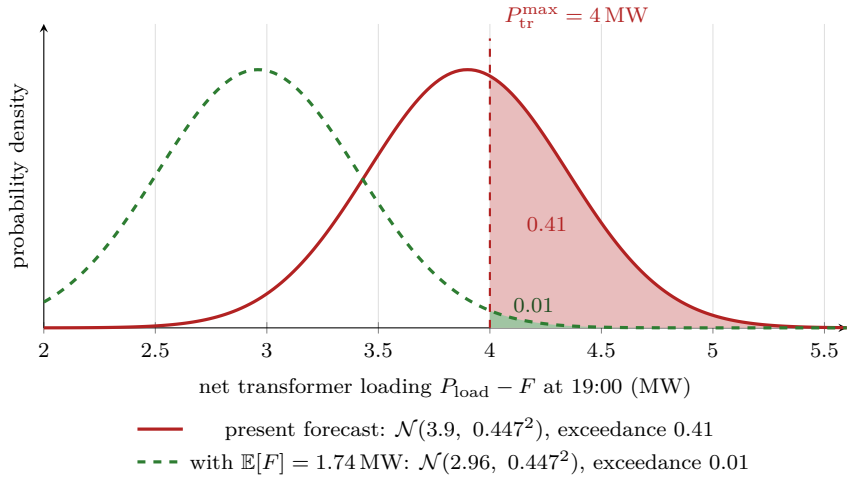


Figure 13.8: Distribution of the net transformer loading at 19:00 from equation (13.26), and the shifted distribution that satisfies the chance constraint of equation (13.28) at $1 - \epsilon = 0.99$. Shaded areas are the exceedance probabilities.

13.5.4 A chance-constrained transformer limit

The honest form of the constraint is therefore probabilistic,

$$\Pr\big(P_{\text{tr}}(t) \le P_{\text{tr}}^{\max}\big) \ge 1 - \epsilon, \qquad 1 - \epsilon = 0.99, \tag{13.28}$$

a chance constraint of the kind used in Chapter 4 for firm capacity. Under the normal model its deterministic equivalent is

$$\mathbb{E}[P_{\text{load}}] - \mathbb{E}[F] + z_{0.99}\,\sigma \le P_{\text{tr}}^{\max}, \qquad z_{0.99} = 2.326, \tag{13.29}$$

which with the figures of equation (13.26) requires

$$\mathbb{E}[F] \ge 4.7 - \big(4 - 2.326 \times 0.447\big) = 1.74\,\text{MW}. \tag{13.30}$$

The deterministic calculation asked for 0.7 MW; the 99 % constraint asks for 1.74 MW, two and a half times as much. Since only 0.9 MW is available behind this transformer, the constraint cannot be met at 0.99 confidence with the resources on hand. That is a useful answer rather than a failure of the method: it says that this transformer needs either more local flexibility, a better load forecast, or reinforcement, and it says so before the evening rather than after it.

Two remarks follow. First, the leverage is in the variance as much as in the mean: halving the load-forecast standard deviation from 0.4 to 0.2 MW reduces σ to 0.283 MW and the required expected flexibility to 1.36 MW, so forecasting is a substitute for capacity. Second, the independence assumption in equation (13.26) is optimistic. A hot evening raises the load forecast error and simultaneously reduces thermostat flexibility, which makes P_{load} and F negatively

correlated and increases $\mathrm{Var}(P_{\mathrm{load}} - F)$ above the sum of the variances. The correlation lesson of Chapter 3 applies here with its sign working against the operator.

The structural point is the one to carry forward. A chance constraint appeared earlier in this book to protect a market commitment, where the penalty for violation was financial. Here the same mathematics protects a physical asset, where the penalty is insulation aging or a failed transformer, and the confidence level is chosen by an engineering standard rather than by a contract.

13.6 Topology, the ADMS, and planning

Everything above assumed the network was known and fixed. It is neither, and this section covers the consequences: the dependence of the sensitivity matrix on switching, the utility system that supplies the topology, and the operational and planning decisions that the whole apparatus supports.

13.6.1 Topology changes the sensitivity matrix

Distribution networks are reconfigurable. Switches are opened and closed to isolate faults, to balance loading between feeders, and to allow maintenance, and after a switching operation part of a feeder may be supplied from a different substation entirely. The tree of Figure 13.1 is therefore one of many trees that the same physical conductors can form.

Every quantity derived from the tree changes with it. A DER that was downstream of transformer T_1 may, after a switch closes, be downstream of T_3 instead, so its entry in the row of $\mathcal{H}$ belonging to T_1 drops from 1 to 0 and its entry in the row belonging to T_3 rises from 0 to 1. The sensitivity matrix is a function of topology,

$$\mathcal{H} = \mathcal{H}\big(\mathrm{topology}(t)\big), \tag{13.31}$$

and a DERMS that caches it will dispatch against a network that no longer exists. Since switching is performed by distribution operators and not by the aggregator, the current topology has to arrive from another system.

13.6.2 The DERMS and the ADMS

That requirement sharpens the boundary drawn in Chapter 3. A DERMS knows DER capability, state and control: how much each device can do and how to command it. An advanced distribution management system (ADMS) knows the network: its topology, its switch positions, its voltages and loadings, and the limits of its equipment. Network-aware dispatch requires both, which means the two systems must exchange information.

Figure 13.9 shows the exchange. The three boxes are the ADMS, the DERMS and the DER fleet. The arrows along the top are the command direction: the ADMS passes the network constraints, the current topology and the sensitivities to the DERMS, which converts them into dispatch commands for individual devices. The arrows along the bottom are the feedback direction: devices report telemetry to the DERMS, which reports achieved response and per-region capability back to the ADMS. Reading the figure as a loop, the ADMS states what the network needs, the DERMS states what the fleet can do about it, and neither system can answer the combined question alone.

13.6.3 Why a utility cares

The reason this problem is now worth solving is the load it is being asked to absorb. EV adoption adds large, clustered evening loads behind service transformers sized decades ago for

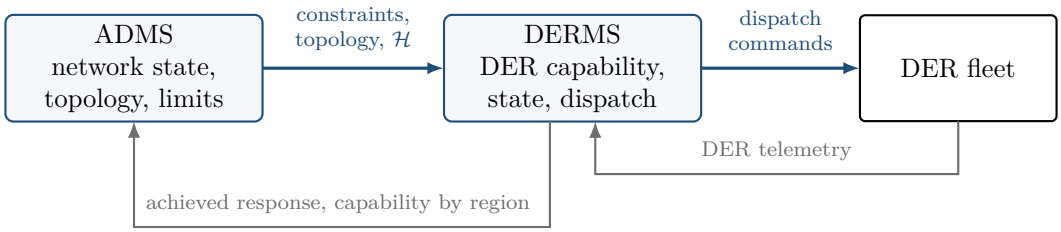


Figure 13.9: Information exchange between the two utility systems. The arrows along the top carry constraints and commands outward; the two return paths below carry telemetry and realized response back. Network-aware DER dispatch is the composition of the two systems, not a function of either.

lighting and air conditioning. Heat pumps add winter peaks where a region previously had none. Rooftop solar creates the midday reverse flow of Section 13.2.4. Home batteries add both. Each is a local phenomenon, concentrated in particular neighborhoods rather than spread evenly, which is exactly the pattern that produces local constraints while leaving system-wide totals comfortable.

The traditional response is reinforcement: a larger transformer, a reconductored feeder, a new substation. That response works and is sometimes necessary. The alternative is to manage some of these constraints operationally, using the flexibility of the same DERs that caused them, and so to defer or reduce the capital program. Whether that is the better choice is a case-by-case economic question, which the next subsection frames.

13.6.4 Flexibility and reinforcement

Flexibility is not a substitute for infrastructure in general. It is a substitute for infrastructure in specific circumstances, and the distinguishing variable is how often the constraint binds. Table 13.4 contrasts the two ends of that range. If a transformer is overloaded for twenty hours a day, relying on customer DERs means curtailing customers almost continuously, which is neither acceptable to them nor reliable for the utility; the correct answer is a bigger transformer. If the same transformer is overloaded for ten hours a year, replacing it buys a capital asset for a problem that exists 0.1 % of the time, and flexibility is attractive.

Table 13.4: When flexibility substitutes for reinforcement.

	Frequent constraint	Rare constraint
Exceedance	about 20 h per day	about 10 h per year
Customer impact of control	near-continuous curtailment	a few events per year
Reliability of the fix	depends on enrollment and overrides every day	depends on them a few times
Appropriate response	reinforce the asset	operational flexibility

Between those extremes the comparison is a genuine economic calculation: the annualized cost of the reinforcement against the cost of procuring the flexibility, both discounted over the period before load growth forces the upgrade anyway, and both adjusted for the probability that the flexibility fails to deliver, which is equation (13.28) again.

Principle 13.3. *Flexibility and reinforcement are alternatives only when the constraint is rare. The frequency with which a limit binds, not the size of the overload, decides which instrument is appropriate.*

13.6.5 Planning value of operating data

A DERMS that has been dispatching against network constraints for a few years holds a record that no planning study can otherwise produce: where constraints actually occurred, how often, how large the required response was, and how reliably the local fleet delivered it. The same four quantities are the inputs to the reinforcement decision of Section 13.6.4, so the operating history answers the planning question directly.

This is a second use for data collected for a different purpose. The measurement and verification apparatus exists to settle payments and to close the control loop; the record it leaves behind tells the planner whether a given transformer should continue to be managed operationally or be replaced, and tells the forecaster how large the error on P_{load} really was at the hours that mattered. An aggregator that discards per-region dispatch history keeps the operations and throws away the planning product.

13.6.6 The full optimization

Collecting every constraint introduced in this chapter and in the preceding ones, the dispatch problem a network-aware DERMS solves is

$$\begin{aligned} \min_{u} \quad & J(x,u) \\ \text{subject to} \quad & g_{\text{network}}(x,u) = 0, \\ & V^{\min} \le V \le V^{\max}, \\ & |I_\ell| \le I_\ell^{\max}, \qquad S_{\text{tr}} \le S_{\text{tr}}^{\max}, \qquad T_{\text{eq}} \le T_{\text{eq}}^{\max}, \\ & u_i \in \mathcal{U}_i(x_i) \quad \text{for every device } i. \end{aligned} \tag{13.32}$$

The first four lines are the network; the last line is every device and customer constraint from Chapters 5 to 7. The objective may be cost, peak, tracking error or a weighted combination, and in practice the problem is solved repeatedly on a rolling horizon in the model-predictive fashion of Chapter 3, with the state re-estimated from telemetry at each step.

Equation (13.32) is worth reading as a summary of the book so far. Every term in it was introduced separately: the device feasible sets in Chapters 1 and 2, the state dependence in Chapters 5 to 7, the aggregation in Chapter 4, and the network rows in this chapter. Nothing in the list is optional, and leaving any of it out produces a dispatch that is infeasible in some respect that the optimizer cannot see.

13.6.7 The conceptual progression

The sequence of questions the book has answered can be written in four lines:

1. Can this DER respond? (Chapter 1)
2. Can this fleet respond? (Chapter 4)
3. How should the fleet respond over time? (Chapter 3)
4. Which DERs should respond, at what time, and at what electrical location? (this chapter)

Each question contains the previous one, and the last is the operational problem in its full form. Nothing in the fourth question invalidates the answers to the first three; it adds an index to them.

13.7 What to retain

Definition 13.1 (Locational flexibility)**.** DER flexibility is locational. A network-aware DERMS must know not only how much flexibility exists, when it is available and for how long it can be sustained, but also *where electrically* it sits, because the grid contributes a further feasible region and the dispatchable resource is the intersection $\mathcal{F}_{\text{total}} = \mathcal{F}_{\text{grid}} \cap \mathcal{F}_{\text{DER}} \cap \mathcal{F}_{\text{cust}}$.

A system-wide megawatt total is not merely incomplete as a description of a fleet; it can be actively misleading, because it suggests a capability against local constraints that the fleet does not have. Table 13.5 lists the chapter's quantities with the subsection that defines each.

Table 13.5: The quantities that describe a network-constrained fleet.

Quantity	Definition	Section
$F(t, r)$	flexibility at time t usable in electrical region r	13.1.2
P_{feeder}	$P^{\text{base}} + P_{\text{EV}} + P_{\text{bat}} - P_{\text{PV}}$	13.1.4
$\mathcal{H}$	sensitivity matrix, $\Delta z = \mathcal{H}\,\Delta P$	13.1.5
$S_{\text{tr}}^{\max}$	transformer apparent-power rating, $S \le S_{\text{tr}}^{\max}$	13.2.1
T_{eq}	equipment hot-spot temperature, a state with a bound	13.2.2
ΔV	$R_{ij}\,\Delta P + X_{ij}\,\Delta Q$	13.2.3
hosting capacity	largest added DER with no limit violated	13.3.4
$\partial L / \partial P_i$	sensitivity of a constrained loading to DER i	13.3.7
$\mathcal{F}_{\text{total}}$	$\mathcal{F}_{\text{grid}} \cap \mathcal{F}_{\text{DER}} \cap \mathcal{F}_{\text{cust}}$	13.4.3
chance constraint	$\Pr(P_{\text{tr}} \le P_{\text{tr}}^{\max}) \ge 0.99$	13.5.4

13.8 Exercises

Exercise 13.1. A feeder with a 10 MW limit is forecast to carry 11.2 MW. Behind the feeder head are three DER groups with available power 0.6, 0.9 and 1.4 MW and sensitivities $\partial L_F/\partial P$ of 1.0, 1.0 and 0.4 respectively, the last group being connected through a parallel path. Compute the effective relief each group can provide and decide whether equation (13.3) can be satisfied. If the 1.4 MW group had sensitivity 1.0, how much margin would there be?

Exercise 13.2. A 500 kVA transformer serves 450 kW of load at a power factor of 0.95 lagging. Compute Q and S. Ten EVs, each drawing 7.2 kW at unity power factor, then begin charging. Compute the new S and state whether equation (13.7) is violated. How much of the 72 kW can be accommodated before S reaches 500 kVA, and by how much would the answer change if the EV inverters absorbed 40 kvar of reactive power?

Exercise 13.3. Using Figure 13.4 and the assumption stated in its caption, that the voltage rise is proportional to the reverse power flow, how much of the 3 MW midday surplus must be absorbed locally to hold $V \le 1.05$ at 10 km? Compare the answer with the 2 MW drawn in the figure, and state what the additional absorption buys.

Exercise 13.4. Repeat the risk calculation of equations (13.26) to (13.30) with a load-forecast standard deviation of 0.25 MW instead of 0.4 MW. What is $\Pr(P_{\text{load}} - F > 4\,\text{MW})$, and what expected flexibility does the 0.99 chance constraint then require? Repeat once more with P_{load} and F correlated at $\rho = -0.5$, and explain in one sentence why a negative correlation is the physically plausible sign.

13.9 Connection to the TERRA framework

This chapter is the one in which the graph stops being a metaphor. Every object introduced here is already a primitive of a conservation-graph framework.

- **The network is the graph, and loadings are edge flows.** Figure 13.1 is a conservation graph: each box is a node at which power balances, and each arrow is an edge whose flow is the loading of a transformer or a feeder section. Equation (13.5) is the conservation statement at the feeder node, and the limits of Table 13.1 are bounds on edge flows. The distinction between F_{sys} and F_{loc} in equation (13.1) is the difference between summing injections over the whole graph and summing them over the subtree below one edge.

- **The sensitivity matrix is the Jacobian of the network map.** A framework that assembles residuals and their Jacobians to run Newton's method on $g(V, \theta, P, Q) = 0$ has $\mathcal{H} = \partial z / \partial P$ of equation (13.6) as a by-product of the solve, evaluated at the current operating point rather than tabulated in advance. Equation (13.31) is then automatic: a switching operation changes the graph, reassembly produces a new Jacobian, and the sensitivities of Table 13.2 change with it.

- **Equipment temperature is a component state, not a flow bound.** Equation (13.8) is an accumulator with a first-order lag, integrated forward in time exactly like the building of Chapter 5 or the battery of Chapter 6. Figure 13.3 is a forward simulation of that component under two control histories, and the constraint $T_{\text{eq}} \leq T_{\text{eq}}^{\max}$ is a bound on a state, which is why the same apparent power is admissible for an hour and not for four.

- **Phase imbalance is several conservation domains at one node.** Equation (13.17) says the three phases balance separately, so a node carries three coupled conservation statements rather than one. A multi-conservation graph represents this without new machinery: the phases are three domains at the same node, and a single-phase charger is an edge in exactly one of them.

- **Hosting capacity and $F(t, r)$ are queries on the model.** The static number of equation (13.18) is a worst-case bound; the function of equation (13.19) is a query evaluated at the present state, namely the largest added injection for which a forward solve violates no bound. Regional flexibility, equation (13.21), is the same kind of query restricted to a subtree, and the intersection $\mathcal{F}_{\text{total}}$ of Figure 13.6 is what the query returns when network, device and customer bounds are all imposed at once.

- **The chance constraint is a quantile of a predicted edge flow.** Equation (13.28) asks for a quantile of the transformer flow under the joint posterior over load and available flexibility. Computing it requires propagating that posterior through the network map, which is why the independence assumption of equation (13.26) is a modeling choice rather than a convenience: the correlation between P_{load} and F is a property of the joint distribution that inference over the graph produces.

In summary, a network-aware DERMS is a conservation graph with bounded flows, stateful components on both the device and the equipment side, and a posterior over the quantities that are not measured; location is simply the position of a component in that graph, and locational value is the sensitivity of a binding bound to it. The next chapter leaves the distribution network and asks what the same fleet is worth in the wholesale markets and grid-service products that sit above it.

Chapter 14

Wholesale markets and grid services: what a VPP actually provides

14.1 Products, not megawatts

Chapter 4 established the virtual power plant as a single resource with a capacity, a duration, a location and a reliability, and Chapters 5 to 7 built the devices behind it. This chapter is where the fleet stops being a control problem and becomes an economic grid resource. The question it answers is

what service is the grid actually buying from the VPP?

Key idea

A VPP does not sell distributed energy resources, and it does not sell megawatts. It sells defined grid services, and each service values a different combination of power, duration, response speed, advance notice, reliability and location.

Suppose a VPP can provide 50 MW. The statement is incomplete until several further questions are answered. For five minutes, or for four hours? With ten seconds of warning, or scheduled for tomorrow evening? Anywhere on the system, or only on one distribution feeder? Delivered with 99 % reliability, or on average? Each combination of answers is a *different product*, bought by a different buyer at a different price, and a fleet that can supply one of them may be unable to supply another. A grid service is therefore a tuple,

$$S = (P,\ \tau,\ r,\ t_{\text{notice}},\ \alpha,\ \ell), \tag{14.1}$$

whose six entries are the power change, the duration for which it must be held, the ramp or response speed, the advance notice given, the reliability with which it must be delivered, and the location at which it must appear. Equation (14.1) is the refinement of the commitment fields of Chapter 4: it adds the response speed r and the notice t_{notice}, which are precisely the two entries that separate the fast balancing products from the slow ones. Figure 14.1 draws the six entries around the service they define, and Table 14.1 gives each one's unit together with typical values for two services that sit at opposite ends of the range.

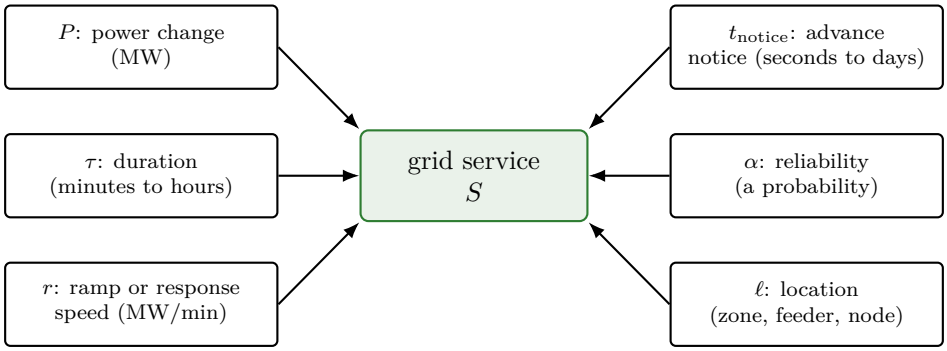


Figure 14.1: The six entries of a grid-service specification, equation (14.1). Two services with the same P and different τ, r, t_{notice}, α or ℓ are different products.

Table 14.1: The entries of the service tuple, with typical values for a day-ahead energy transaction and for a fast balancing product. The values are representative rather than taken from any one market.

Entry	What it fixes	Energy	Regulation
P	size of the net power change	20 MW	$\pm$20 MW about a baseline
τ	how long it must be held	3 h	continuous
r	how fast it must get there	within the hour	seconds
t_{notice}	how much warning is given	day-ahead	none; follow the signal
α	with what probability it is delivered	moderate	very high
ℓ	where it must appear	system-wide	system-wide

Jargon: wholesale market, capacity, reserve, regulation, co-optimization, value stacking

A *wholesale market* is the market in which electricity and grid services are traded in bulk between suppliers and load-serving entities, run by a system operator; retail customers do not transact in it, but an aggregator acting for them can. *Capacity* is a product that pays for *being available* at some future time rather than for energy delivered. *Reserve* (operating reserve) is capability deliberately held unused so that it can respond to an unexpected event such as a generator tripping. *Regulation* is continuous, second-to-second adjustment of power around a baseline to follow a signal sent by the system operator. *Co-optimization* is choosing how to allocate one physical capability across several competing services in a single optimization instead of one service at a time. *Value stacking* is earning revenue from several services with the same device over time.

Principle 14.1. *A megawatt is not a product. It becomes one only when the duration, response speed, notice, reliability and location are fixed alongside it.*

The chapter develops, in order, the wholesale products a fleet can sell, the demand-response and locational services that sit beside them, how several services are earned from one device without selling the same flexibility twice, how the forecasting and reliability requirements differ from product to product, and how much to commit when the available flexibility is uncertain.

14.2 Wholesale products

This section works through the products a system operator buys, each on the same template: what is bought, the quantity that matters, which DERs suit it, and how delivery is verified.

14.2.1 Energy and temporal arbitrage

The simplest product is energy. What is bought is a change in net consumption $P(t)$ over a period, and the quantity that matters is its integral,

$$E = \int P(t)\,\mathrm{d}t, \tag{14.2}$$

so 20 MW held for three hours is $20 \times 3 = 60$ MWh. Value comes from *when* the energy moves: reducing consumption when energy is expensive and increasing it when energy is cheap, charging batteries in low-price periods and discharging them in high-price periods. This is *temporal arbitrage*, the battery objective of Chapter 6 written as a market product.

Every DER class can provide it, because every class can shift energy in time: the thermostat by pre-cooling, the EV by moving its charging block, the battery by charging and discharging. What differs is how much energy each can move and at what cost to the customer. Verification is the easiest of any product, because energy is what revenue meters measure: the delivered quantity is the metered integral against a baseline, with the caveat that the baseline itself must be estimated.

14.2.2 Peak reduction

A utility may not care about energy at all. What it cares about is the single largest value of system load over a season, $\max_t P_{\text{system}}(t)$, because that number sizes transformers, lines and capacity purchases. If the system peak falls from 5 000 MW to 4 900 MW, the VPP has delivered 100 MW of peak reduction. The energy involved can be modest: two hours at 100 MW is 200 MWh, a tiny fraction of a day's consumption. Avoiding the peak is nonetheless extremely

valuable, because the equipment and the capacity purchases avoided are sized by the maximum and not by the average.

Peak reduction is the classic use of demand response. It suits thermostats particularly well: summer system peaks coincide with the hours in which air conditioning runs hardest, so the flexible load is largest exactly when it is needed, and a two-hour reduction is within the comfort band of most buildings. The quantity that matters is MW at the critical hours, not MWh. Verification requires a baseline at those hours and, because a program is usually paid on the seasonal peak, also requires the reduction to be achieved on the day that turns out to contain it, which is a forecasting problem rather than a control one.

14.2.3 Capacity, which is availability with a duration

A capacity product pays for being there when the system needs it. The question it asks is whether the resource can reliably be present, and the quantity that matters is not the mean of the available flexibility but a low quantile of it. Write F for the available flexibility at the hour of interest and F_q for its q-quantile, so that $\Pr(F \geq F_{0.05}) = 0.95$. A fleet may have

$$\mathbb{E}[F] = 100\,\text{MW} \quad \text{while} \quad F_{0.05} = 75\,\text{MW}, \tag{14.3}$$

and for a capacity-type obligation the meaningful number is the second one. The industry states the same thing as a "95 %-reliable value of 75 MW". This is the firm capacity of Chapter 4 under another name, and it ties the product directly to probabilistic forecasting: without a distribution for F there is no way to compute $F_{0.05}$, and the only honest alternative is to commit the mean and accept the shortfalls.

Capacity is about availability, not about energy, but a capacity obligation almost always implies a duration, and that is where fleets are caught out. Consider a battery resource with 100 MW of inverter capacity and 100 MWh of usable energy. At full output it sustains

$$\tau = \frac{100\,\text{MWh}}{100\,\text{MW}} = 1 \text{ hour}, \tag{14.4}$$

so if the capacity obligation requires four hours of sustained output, this is not a 100 MW four-hour resource. It is a 100 MW one-hour resource, or, if the four hours are mandatory, a 25 MW four-hour resource. The deliverable capacity is the function $C(\tau)$ of Chapter 4, drawn there as Figure 4.3: a capacity product selects one point on that curve, and the point it selects determines which fleets qualify. Thermostats lose out at long τ because the comfort band empties, as the decay of Figure 1.3 showed; EVs lose out when their plug-in window does not cover the hours in question.

14.2.4 Operating reserves

Operating reserve is capability held so that the system can respond to something unexpected. If a 500 MW generator trips, the balance between generation and load is broken within seconds and something must make up the difference. What is bought is not energy and not a scheduled power change; it is a *standby capability*, and the quantity that matters is the megawatts that can be delivered within the product's response time and sustained for its minimum duration.

A VPP contributes reserve if it can quickly reduce load, increase battery discharge, or interrupt and shift EV charging. Batteries suit the fastest reserve products because an inverter responds in under a second; thermostats and EV chargers suit the slower ones, where ten minutes of response time is allowed. The defining operational consequence is that the reserve must be *held back beforehand*: a battery already discharging at its full rating has no reserve to offer, and an EV charger already curtailed to zero cannot be curtailed again. Verification is unusual in that most reserve is never called, so performance is established by periodic tests and by telemetry proving that the capability was available, rather than by metered delivery.

14.2.5 The opportunity cost of holding reserve

Holding capability back has a price. Suppose a battery has 20 MW available in the hour. It can put all 20 MW into energy arbitrage, in which case the reserve it can offer is

$$P^{\text{en}} = 20\,\text{MW}, \qquad P^{\text{res}} = 0, \tag{14.5}$$

or it can split the capability,

$$P^{\text{en}} = 10\,\text{MW}, \qquad P^{\text{res}} = 10\,\text{MW}, \tag{14.6}$$

and collect the arbitrage margin on half of it and the reserve payment on the other half. Nothing in the physics prefers one split; the decision is economic, and the optimizer makes it by comparing the marginal value of a megawatt in each use.

This is the first appearance in the chapter of a theme that dominates the rest of it: services compete for the same physical capability, so they must be chosen together rather than one at a time. Choosing them together is co-optimization, and Section 14.4.3 writes it down. The opportunity cost of reserve is also the reason a reserve payment can exceed the arbitrage margin and still be worth taking only part of the time: the payment must compensate not merely for standing ready but for the energy revenue forgone while standing ready.

14.2.6 Regulation and fast balancing

Regulation is the continuous version of reserve. The system operator sends a signal every few seconds, and the resource adjusts its power to follow it: a target trajectory $P^{\text{tgt}}(t)$ around a baseline, usually with zero mean, which the resource must track. What is bought is the width of the band the resource makes available, so a resource offering ±20 MW of regulation is paid for 20 MW of capability in each direction. The quantities that matter are the band width and the *accuracy* with which the signal is followed; most operators pay a performance score that multiplies the capability payment. Figure 14.2 shows a few minutes of such a signal and a battery following it.

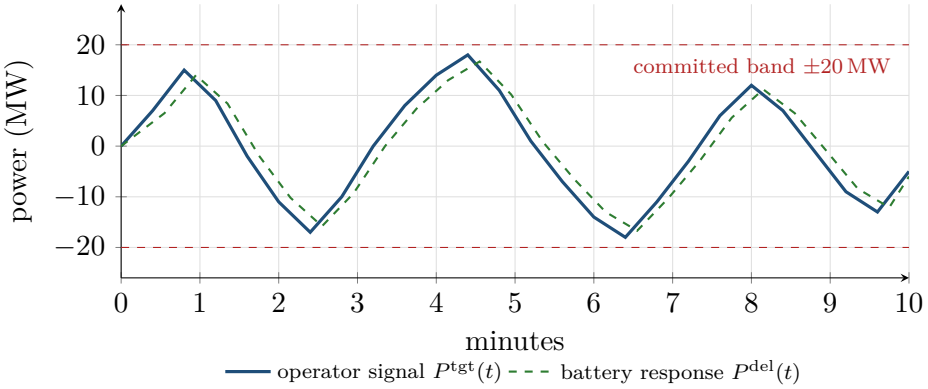


Figure 14.2: Ten minutes of a regulation service. The solid blue curve is the operator's signal, the dashed green curve the battery's delivered power, slightly lagged and slightly attenuated; the dashed red lines are the edges of the committed ±20 MW band. Both traces are illustrative.

The two traces differ in two ways that matter commercially. The response lags the signal by a few seconds, and its amplitude is a little smaller than the signal's; both reduce the performance score. Batteries are especially suited to this product because an inverter can reverse direction in milliseconds and track a trajectory precisely. Thermostat fleets can provide it, by switching groups of compressors on and off, but the response is coarse, delayed by the deadband, and vulnerable to the synchronization of Figure 5.9; it is possible rather than natural.

The requirements are therefore fast response, accurate tracking, and sufficient energy headroom, which is the subject of the next subsection.

14.2.7 Energy headroom under a biased signal

Regulation is nominally a power service, but it consumes energy whenever the signal is not balanced. A battery following a regulation signal around $P = 0$ changes its state of charge according to the accumulator equation of Chapter 6. If the signal is energy-neutral, so that $\int P^{\mathrm{tgt}}(t)\,\mathrm{d}t \approx 0$ over the hour, the state of charge oscillates about its starting value and nothing further is needed. If the signal is biased, so that $\int P^{\mathrm{tgt}}(t)\,\mathrm{d}t \neq 0$, the state of charge drifts, and it drifts at a rate set by the bias alone. Figure 14.3 integrates both cases.

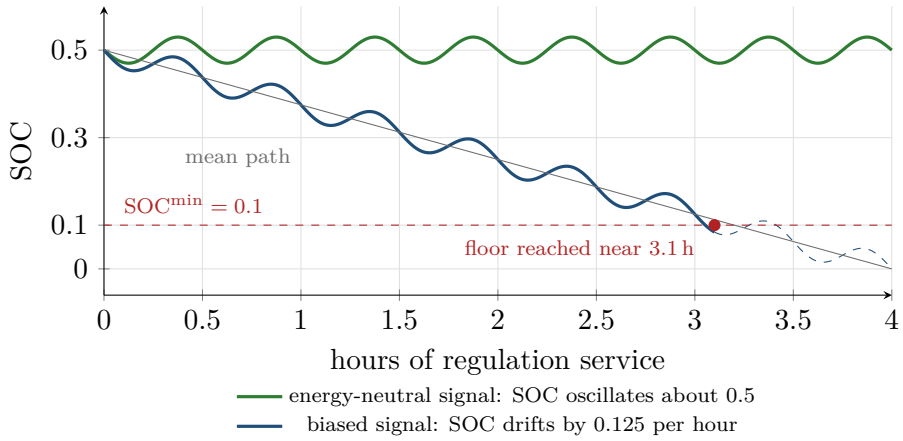


Figure 14.3: State of charge of a battery following a regulation signal for four hours. Green: the signal $r(t) = 0.75\cos(720°t)$ in per-unit of the committed band, which is energy-neutral. Blue: the same signal with a mean of $+0.25$, integrated with $\mathrm{SOC}(t) = 0.5 - k\int r$ and $k = 0.5$ per hour, so the mean drift is 0.125 per hour. The gray line is the mean path, which reaches the floor at 3.2 h; the oscillation carries the actual trajectory there near 3.1 h. Curves are illustrative.

The green curve stays within three percentage points of its starting value for the whole four hours. The blue curve falls steadily and touches $\mathrm{SOC}^{\min}$ shortly after three hours, and at that moment the battery loses one direction: it can still absorb power, so it can follow a downward signal, but it can no longer discharge, so every upward request is missed and the performance score collapses. Recovering requires buying energy back, which means leaving the regulation market or accepting a tracking error. Even the fastest, most "power-like" service therefore requires state management, and the quantity that must be forecast is not the band width but the cumulative bias of the signal.

14.3 Demand response, load shaping, and locational services

Beside the wholesale products sit services defined by a trajectory or by a place rather than by a quantity of power. This section covers called demand-response events, continuous load shaping, ramping, congestion relief, and the distribution-level services a utility buys directly.

14.3.1 Traditional demand response as a called event

Traditional demand response, the subject of Chapter 2, is the discrete member of the family. What is bought is a reduction in demand when called, over a specified window such as 17:00 to 19:00. The quantities that matter are the MW reduction, the event duration, the advance notice, the number of events allowed per season, and the customer availability behind them;

the last is what distinguishes the product from a generator contract, because the resource is only there when enough customers have not opted out.

Any class can serve it. Thermostats are the workhorse because a two-hour reduction fits inside a comfort band and requires no customer action; EV charging curtailment works when the cars are plugged in during the window. Verification is by baseline, with the rebound of Figure 2.3 deciding whether the energy was genuinely avoided or merely postponed. The product is less continuous than regulation, and in exchange it is far less demanding: tens of events a year, each with hours of notice, is a schedule a residential fleet can sustain.

14.3.2 Dynamic load shaping

Dynamic load shaping goes beyond the called event. Instead of a reduction on demand, what is bought is adherence to a target trajectory $P^{\mathrm{tgt}}(t)$ that changes through the day: the fleet is asked to be normal in the morning, to *increase* load at midday, to reduce it in the evening, and to increase it again late at night when EV charging is cheap and the system is lightly loaded. Figure 14.4 draws such a trajectory against the load the same fleet would have drawn unmanaged.

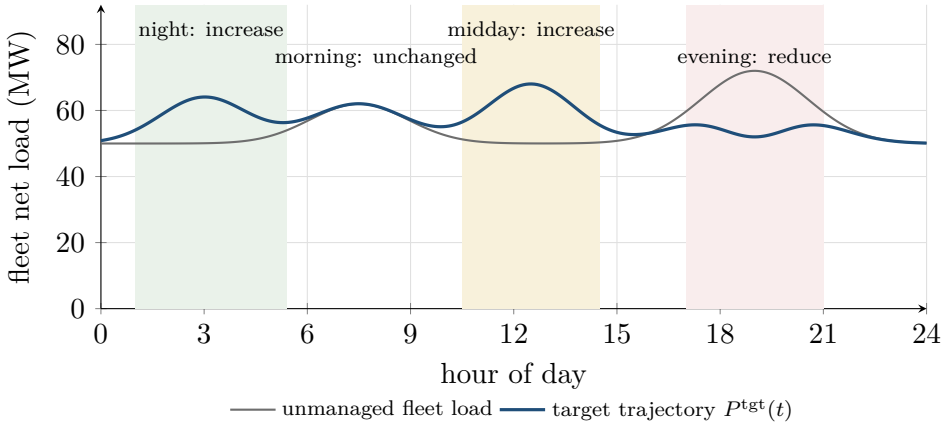


Figure 14.4: A load-shaping target through one day. The gray curve is the unmanaged fleet load, $50 + 12e^{-((t-7.5)/2)^2} + 22e^{-((t-19)/2.2)^2}$ MW; the blue curve adds three Gaussian adjustments centered at 03:00 (+14), 12:30 (+18) and 19:00 (−20) MW. The three shaded bands mark the adjusted regimes; the morning is unshaded because the target equals the unmanaged load there. Curves are illustrative.

Three features of the figure are worth naming. The midday adjustment is *upward*, which no traditional demand-response program would ever request; the evening adjustment is downward and is the only part that resembles a classic event; and the night adjustment puts energy back into the hours in which the system has room for it. Integrating the three Gaussian terms gives +45 MWh added at night, +57 MWh at midday and −57 MWh in the evening, so the day's total rises by about 45 MWh: shaping places consumption, and it does not conserve the daily total. A fleet operated this way supports renewable integration, peak management, wholesale-price optimization and local constraints at once, and it is operated actively every hour rather than a few times a season. The quantity that matters is the trajectory, and the verification question becomes a tracking-error question, $e_t = P^{\mathrm{tgt}} - P^{\mathrm{del}}$, rather than a baseline question.

14.3.3 Renewable integration

The main reason trajectories have replaced events is solar. Solar output peaks near noon, so the net load the dispatchable fleet must serve,

$$P_{\text{net}}(t) = P_{\text{demand}}(t) - P_{\text{PV}}(t), \tag{14.7}$$

falls steeply through the middle of the day and then rises steeply after sunset, when solar output vanishes while demand is still climbing toward its evening peak. A VPP reshapes that curve by increasing EV charging and battery charging at midday and reducing EV charging and discharging batteries in the evening. Figure 14.5 shows demand, solar, the resulting net load, and the net load after the VPP acts.

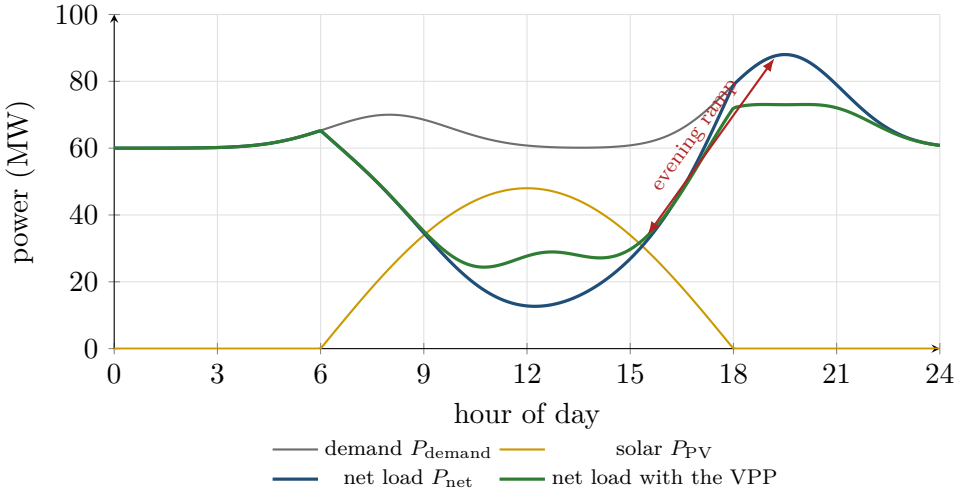


Figure 14.5: Net-load reshaping. Demand is $60 + 10e^{-((t-8)/2.5)^2} + 28e^{-((t-19.5)/2.4)^2}$ MW and solar is $\max\{0,\, 48\sin(15°(t-6))\}$ MW; the net load is their difference. The VPP adds $+16e^{-((t-12.5)/1.9)^2}$ MW of midday charging and $-15e^{-((t-19.5)/1.7)^2}$ MW of evening discharge. The red arrow marks the evening ramp of the unmanaged net load. Curves are illustrative.

The gray curve is demand, with a small morning bump and a large evening peak. The gold curve is solar, zero before sunrise and after sunset. The blue curve, their difference, is the familiar shape with a deep midday trough and a steep evening rise. The green curve is what the VPP leaves behind: the trough is partly filled by midday charging and the peak is cut by evening discharge. Two numbers summarize the effect. The midday minimum rises from about 13 MW to about 28 MW, which reduces the risk that solar has to be curtailed for want of load. The evening rise falls from about 61 MW over four and a half hours, roughly 14 MW per hour, to about 43 MW over the same interval, roughly 10 MW per hour.

14.3.4 Ramp management

The second of those numbers is itself a product. A system whose net load changes by 500 MW per minute needs resources that can move that fast, and if none are available the frequency or the voltage suffers. The service the operator wants is a bound on the rate of change,

$$\left|\frac{\mathrm{d}P_{\text{net}}}{\mathrm{d}t}\right| \le R^{\max}, \tag{14.8}$$

and a VPP provides it by supplying a trajectory $P_{\text{VPP}}(t)$ such that the effective net load $P_{\text{net}}(t) - P_{\text{VPP}}(t)$ ramps less than the raw one. What is bought is the *shape* of the response, not its size: a resource that can deliver 50 MW but takes twenty minutes to get there is useless

for a five-minute ramp, while one that delivers 20 MW in two minutes is valuable. The quantity that matters is MW per minute sustained over the ramp window.

Ramp management rewards a trajectory rather than a single number, which changes what must be forecast. A peak-reduction program needs the expected reduction at one hour; a ramp product needs the *derivative* of the fleet's deliverable power along a path, which is sensitive to the hand-off structure of Chapter 4 because a class reaching its limit mid-ramp produces exactly the ramp the service was bought to prevent. Batteries dominate here for the same reason they dominate regulation.

14.3.5 Congestion management

Every product so far has been system-wide. Congestion services are not. Suppose a transmission or distribution interface is constrained, and let P_ℓ be the flow on the constrained equipment. What is bought is a reduction in that particular flow,

$$P_\ell \le P_\ell^{\max}, \tag{14.9}$$

and only dispatch at the right location contributes to it. A 10 MW reduction behind the constraint may relieve it completely; a 100 MW reduction somewhere else does nothing at all, and a reduction on the wrong side can make the flow worse. Congestion relief is therefore *locational flexibility*, and the entry ℓ of equation (14.1) is the whole product.

This is the first service for which the aggregate of Chapter 4 is insufficient. A VPP described by a single available-megawatt figure cannot be dispatched for congestion, because the figure does not say where the megawatts are. The DERMS must retain a mapping from devices to feeders and substations, and the forecast it needs is a forecast of local load and of the current network topology, since a switching operation can move a device from one constrained interface to another. Verification is by measurement at the constrained equipment rather than at the customer meter.

14.3.6 Distribution-level services

Several further services are bought by the distribution utility rather than by a market: relieving an overloaded transformer, resolving feeder congestion, managing voltage, absorbing local solar generation that would otherwise raise voltage above limits, and deferring a planned capacity upgrade. None of these is necessarily a wholesale product, and some have no market price at all, yet their value to the utility is substantial, because the alternative is capital expenditure on equipment.

Table 14.2 sets the two families side by side. The distinction is worth keeping clear, because the two have different buyers, different settlement, different spatial resolution and different verification, and a DERMS can participate in both at the same time with the same devices. That last point is what makes Section 14.4 necessary: when one device is simultaneously enrolled in a wholesale product and in a local constraint-management program, the two commitments can conflict, and nothing but an optimizer that sees both will notice.

14.4 Value stacking and co-optimization

One device can earn from several services, but not from all of them at once. This section separates the two statements and writes down the optimization that enforces the difference.

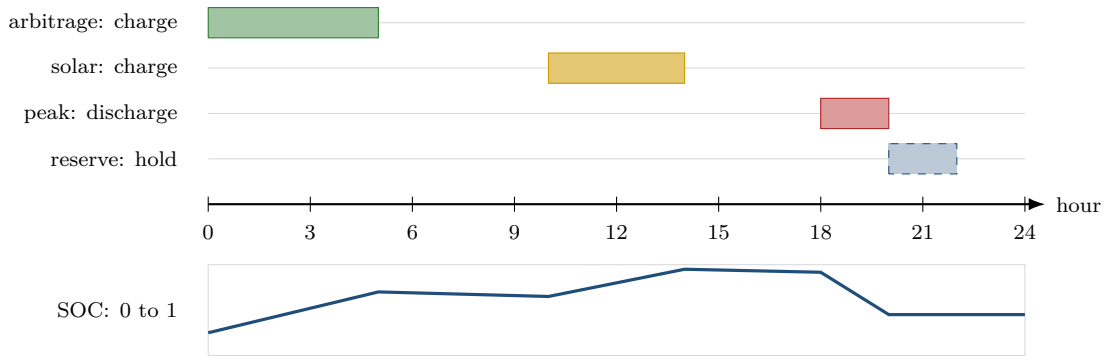
14.4.1 One device, several services

Consider a single battery over one day. At noon it charges from otherwise-surplus solar generation. At 18:00 it discharges into the evening peak and is paid for peak reduction. From

Table 14.2: Wholesale market services and distribution utility services.

	Wholesale market service	Distribution utility service
Buyer	system operator or load-serving entity	the distribution utility
Examples	energy, capacity, reserve, regulation, ramping	transformer relief, feeder congestion, voltage support, solar absorption, upgrade deferral
Location	usually system-wide or a large zone	a specific feeder, transformer or node
Settlement	market price, published after the fact	bilateral contract or tariff
Verification	metered delivery against a baseline or a signal	measurement on the distribution equipment
What it displaces	generation and capacity purchases	capital expenditure on equipment

20:00 to 22:00 it holds capability unused and is paid for operating reserve. In the hours after midnight it charges at low prices against the following day's higher ones, earning an arbitrage margin. The same physical device has provided four services in twenty-four hours, and the practice of arranging that is value stacking. Figure 14.6 lays the day out as a timeline with the battery's state of charge beneath it.



Four services, one device, one day. The reserve block moves no energy unless it is called, which is why the state of charge is flat across it; every other block changes the stored energy and therefore changes what the next block can offer.

Figure 14.6: A battery's day of stacked services. Each row is one service and each block the hours in which it is provided; the lower strip is the state of charge over the same day. Hours and levels are illustrative.

The lower strip carries the important message. The blocks are not independent: each one leaves the battery in a state that determines what the next one can offer. Charging at midday is what makes the evening discharge possible, and the evening discharge is what empties the battery before the reserve block, which is why the reserve offered from 20:00 is smaller than the battery's rating. Value stacking is a scheduling result, not an accounting trick.

14.4.2 Double counting

Services compete for the same flexibility, and the limit is physical. Suppose a battery sits at SOC = 0.60 with 20 MW of inverter capability. That 20 MW cannot be promised simultaneously to a capacity obligation, an energy schedule, a reserve product and a local congestion program, because if two of those are called in the same minute the device can serve only one of them.

Promising it anyway is *double counting*: the revenue is counted twice and the capability exists once.

The exception is the case in which the commitments are compatible. Two services that are never called together can share a capability; a reserve obligation and an energy schedule can share it if the energy schedule leaves enough headroom in the right direction; a capacity obligation for a summer afternoon and a congestion program for a winter morning do not interact at all. The optimizer's job is to tell these cases apart, which means it must know, for every pair of services, the probability that both are called at once, and must enforce the physical limit in every scenario in which they are.

Principle 14.2. *A physical capability can be sold once. Value stacking is selling the same* device *at different times; it is never selling the same* flexibility *at the same time.*

14.4.3 The co-optimization problem

Write R_k for the revenue from service k and C_{op} for the operating cost of the dispatch that produces all of them. The multi-service problem is

$$\max \sum_k R_k - C_{\text{op}} \quad \text{subject to} \quad u_t \in \mathcal{U}(x_t) \text{ and the requirements of every committed service.} \tag{14.10}$$

The first constraint is the state-dependent feasible set of Chapter 1, unchanged; the second is new, and it is what couples the services. Operating cost C_{op} carries the battery degradation of Chapter 6 and the customer-impact term of Chapter 3, so a service is worth taking only if its revenue exceeds both the opportunity cost of the capability and the wear on the devices.

Equation (14.10) is a different question from the dispatch problem of Chapter 4. Dispatch asks how to meet a given request; co-optimization asks *which service should receive the flexibility in the first place.* The answer depends on prices that are not yet known, on call probabilities, and on the state the fleet will be in when each service is called, which is why the problem is stochastic and why it is solved over a horizon rather than instant by instant.

14.4.4 Expected value of a reserve commitment

A small example makes the structure concrete. A 10 MW battery can earn \$500 for the hour by selling energy, using all 10 MW. Alternatively it can earn \$300 by selling reserve, which requires holding the 10 MW unused unless it is called. If the two are mutually exclusive, the comparison is immediate: \$500 exceeds \$300 and the battery sells energy.

The realistic case is not mutually exclusive, and it is not a comparison of certain payments. Reserve is rarely called. Let p be the probability of a call in the hour, and suppose that a call forces the battery to abandon its energy position at a cost L. Then selling reserve is worth $300 - p\,L$ in expectation, while energy is worth \$500 for certain, so reserve wins when $300 - p\,L > 500$, which it never does at these prices. But if some energy dispatch is compatible with holding reserve, so that the battery can sell P^{en} megawatts of energy and $10 - P^{\text{en}}$ of reserve, the objective becomes a function of the split and the optimum is interior whenever the reserve price per megawatt exceeds the energy margin per megawatt in part of the range. The optimization is then a stochastic economic problem: maximize expected revenue over a distribution of call events, subject to the physical constraint that a called reserve must actually be deliverable from whatever state the energy dispatch has produced. The fact that reserve is rarely called is what makes the problem worth solving rather than what makes it easy.

14.5 Requirements differ by service

Because the products differ, so does everything behind them. This section collects the differences in the forecasts, the horizons, the reliability targets, and the customer constraints that bind regardless.

14.5.1 What each service requires

Table 14.3 is the master table of the chapter. It lists the services of Sections 14.2 and 14.3 with the quantity each one cares about, the forecast it needs, the horizon on which the decision is made, and how much under-delivery it tolerates.

Table 14.3: What each service requires. The reliability column is the tolerance for under-delivery, which sets whether a mean or a low quantile should be committed.

Service	Key concern	Forecast needed	Horizon	Reliability tolerance
Energy	MWh and price	price $\hat{c}_t$ and DER availability	day-ahead to hour-ahead	high: an energy position can be bought back
Peak reduction	MW at the critical hours	which hours contain the peak; $\hat{P}$ at those hours	day-ahead	low: the peak hour happens once
Capacity	reliable availability	$\Pr(F \geq P^{\text{commit}})$	season to day-ahead	very low: penalties are severe
Reserve	held response capability	available headroom and its reliability	day-ahead to real time	very low: it is called when the system is already stressed
Regulation	response speed and tracking accuracy	very short-term response and signal bias	real time	low: scored continuously
Load shaping	the desired trajectory	the whole-day load and flexibility path	day-ahead, revised hourly	moderate: errors are averaged over the day
Congestion relief	location	local load $\hat{P}_{\text{local}}$ and network topology	day-ahead to real time	low: the constraint is a physical limit
Distribution support	local network constraints	feeder load, voltage, solar output	planning to real time	low: equipment damage is the failure mode

The table makes one negative point that is worth stating directly: the phrase "the VPP has 50 MW of capacity" is not a complete concept. It does not say which of the eight rows the 50 MW refers to, and the eight rows would give eight different numbers for the same fleet on the same day.

14.5.2 Forecasting requirements

The third column of Table 14.3 deserves expansion, because each entry is a different statistical problem. Energy-price optimization needs a forecast of the price, $\hat{c}_t$, together with the availability of the resources that would respond to it; errors cost the difference between the price realized and the price assumed. A capacity obligation needs a tail probability, $\Pr(F \geq P^{\text{commit}})$, which is a question about the shape of a distribution and cannot be answered by a point forecast at all. Regulation needs very short-term accuracy, on the scale of the signal itself, and in addition

a forecast of the cumulative bias that drives the drift of Figure 14.3. Congestion relief needs local load $\hat{P}_{\text{local}}$ and the network topology, neither of which appears anywhere in a system-wide forecast.

These are not the same model with different parameters. A model that produces excellent hourly expectations may be useless for a capacity product because it was never calibrated in the tail, and a model calibrated in the tail may be too slow to run on a four-second signal. Choosing which services to sell therefore decides which forecasting problems the platform must solve, and committing to a service whose forecast the platform cannot produce is the most common way to lose money on an otherwise sound fleet.

14.5.3 Time horizons

Decisions are taken on three horizons, and the same fleet is described differently on each. Day-ahead, the platform schedules energy, plans capacity obligations and estimates expected flexibility, working with distributions over the whole of tomorrow. Hour-ahead, it refines availability as telemetry arrives and adjusts the portfolio, which may mean buying back part of a position that now looks unsafe. In real time, it dispatches, follows regulation signals and corrects tracking errors, using the current state rather than a forecast of it.

A platform therefore needs forecasts at several horizons simultaneously, and they must be consistent: the day-ahead distribution of tomorrow's 18:00 flexibility should be the expectation, over everything not yet known, of the hour-ahead distribution that will be available at 17:00. When the horizons disagree the operator discovers it as a surprise at 17:00, with a position already committed.

14.5.4 Reliability requirements

Two services can pay the same and still deserve different commitments. Suppose service A tolerates occasional under-delivery, perhaps because the shortfall is settled at a market price that is usually close to the contract price, while service B carries severe penalties for any shortfall. For A it is reasonable to commit near $\mathbb{E}[F]$; for B the prudent commitment is a low quantile such as $F_{0.05}$. With the fleet of equation (14.3) those two rules give 100 MW and 75 MW, a difference of a quarter of the resource for the same physical fleet on the same day.

The general statement is that the economics of a service determine how the forecast uncertainty should be used. A distribution is not merely a confidence interval attached to an answer; which *feature* of the distribution becomes the answer is a function of the penalty structure. Section 14.6.6 turns that observation into an optimization.

14.5.5 Customer constraints still bind

No market price relaxes a physical or contractual limit. A valuable service does not entitle the VPP to exceed $T^{\max}$ in a customer's house, to spend the SOC^{res} a battery owner keeps for backup, or to miss the SOC^{req} an EV driver needs at departure. The feasible set

$$u_t \in \mathcal{U}(x_t) \tag{14.11}$$

is the same set as in Chapter 1, and grid economics determine only how the flexibility inside it is used, never how large it is.

This is why the chapter is placed after the device chapters rather than before them. Everything in Table 14.3 is an instruction about which point of $\mathcal{U}(x_t)$ to choose and what to be paid for choosing it. The set itself was fixed by physics and by the customer's agreement, and the next chapter examines it directly.

14.6 Commitment economics

The last body section treats the fleet as a portfolio of conditional capabilities and asks what should be promised from it. It separates the commitment decision from the dispatch decision and shows that the commitment depends on the spread of the forecast and not only on its mean.

14.6.1 The portfolio of options

Each DER in the fleet provides not a fixed capability but a *conditional* one. A thermostat offers 3 kW if the air conditioner is running, if thermal headroom exists inside the comfort band, and if the customer does not opt out. An EV offers 7 kW of curtailment if it is plugged in, if it is charging, and if there is enough slack before departure. A battery offers 5 kW if the state of charge permits and if the owner's reserve is respected. Table 14.4 collects the three, with the condition and the binding limit for each.

Table 14.4: The fleet as a portfolio of conditional capabilities. Each row is an option that exists only when its condition holds.

Device	Capability	Condition	Binding limit
Thermostat	3 kW reduction	compressor running, thermal headroom in the band, no opt-out	comfort band $[T^{\min}, T^{\max}]$, hence duration
EV charger	7 kW curtailment	plugged in, charging, slack before departure	$\mathrm{SOC}(T^{\mathrm{dep}}) \geq \mathrm{SOC}^{\mathrm{req}}$, hence energy
Battery	5 kW discharge	state of charge above the owner's reserve	$\mathrm{SOC} \geq \mathrm{SOC}^{\mathrm{res}}$, hence duration

Read down the condition column and the fleet looks less like a power plant than like a book of options whose exercise conditions are random and correlated. The VPP's commercial function is to aggregate those conditional capabilities into a service with a stated magnitude, duration and reliability, which is exactly the map $z = \phi(x)$ of Chapter 4 with a price attached to its output.

14.6.2 Revenue follows the service, not the technology

The grid pays for a response, not for a device. Whether 50 MW of evening reduction came from electric vehicles, thermostats, batteries or all three matters to the aggregator's cost structure and to its risk, and not at all to the buyer, provided the required response appeared with the required speed for the required duration at the required place. The fleet is internally heterogeneous and externally a single defined service.

That asymmetry is the commercial reason the abstraction of Chapter 4 is worth building. It also sets the aggregator's incentive clearly: it is rewarded for the cheapest internal composition that satisfies the external specification, which pushes it toward diverse portfolios whose limits do not bind at the same time.

14.6.3 Assembling an evening peak service

A utility needs 50 MW between 18:00 and 21:00. The VPP assembles it from

$$20\,\mathrm{MW}\ (\text{thermostats}) + 20\,\mathrm{MW}\ (\text{EV charging reduction}) + 10\,\mathrm{MW}\ (\text{batteries}) = 50\,\mathrm{MW}, \tag{14.12}$$

and the utility sees a 50 MW resource for three hours. The composition is the DERMS's business: which thermostat groups, which vehicles, which batteries, in what order, with what hand-off as each class approaches its limit, exactly as in the hand-off schedule of Chapter 4.

The commercial power of aggregation is visible in the arithmetic. None of the three classes could sell a 50 MW three-hour product on its own, and none of the individual customers could sell anything at all; the product exists because the aggregation exists. The three hours are also what forces the hand-off: thermostats fade as the comfort band empties, along the curve of Figure 1.3, so a 50 MW commitment held flat for three hours requires the composition to change inside it.

14.6.4 Two services in the same window

Now suppose the VPP also wants to hold 5 MW of fast reserve across the same 18:00 to 21:00 window. The total requirement is not simply 50 MW. It may be 55 MW of accessible flexibility, if the reserve must be deliverable at any moment during the peak service and the peak service is already using the first 50 MW. It may be less than 55 MW, if part of the peak-reduction dispatch can be interrupted quickly enough to serve as the reserve, which depends on the response speed r of the reserve product and on which class is carrying the peak at that moment.

Service stacking therefore couples commitments, and the coupling is a function of the overlap in time and in direction rather than a fixed adder. The practical consequence is that the two commitments must be decided together: a reserve offer accepted after the peak commitment has been made can turn a comfortable 50 MW into an infeasible 55 MW without anything changing in the fleet.

14.6.5 Commitment and dispatch

The operator faces two decisions, and they are different in kind. *Commitment* asks what should be promised: how much, of which service, for which window, at what reliability. *Dispatch* asks how to deliver what was promised: which devices, at what power, in what order. Commitment is made before the fact under uncertainty and is dominated by forecasting; dispatch is made during the fact with telemetry available and is dominated by optimization and control.

Chapters 3 to 7 were almost entirely about dispatch. The rest of this section is about commitment, which is where the uncertainty of a forecast turns into money.

14.6.6 Commitment under uncertainty

Let the available flexibility in the service window be a random variable $F \sim p(F)$, let π_C be the price paid per megawatt committed, and let π_P be the penalty per megawatt of shortfall. The operator chooses the commitment P^{commit} to maximize

$$\mathbb{E}\big[\pi_C\, P^{\text{commit}} - \pi_P\,(P^{\text{commit}} - F)^+\big], \tag{14.13}$$

where $(\cdot)^+$ is the positive part. Committing too little leaves revenue on the table, because every megawatt up to the fleet's capability earns π_C and risks nothing. Committing too much incurs penalties, because the shortfall grows with the commitment. The optimum is where the marginal penalty equals the marginal revenue,

$$\pi_P\ \Pr(F < P^{\text{commit}}) = \pi_C \qquad \text{that is} \qquad \Phi(z) = \frac{\pi_C}{\pi_P}, \quad z = \frac{P^{\text{commit}} - \mu}{\sigma}, \tag{14.14}$$

the second form holding when $F \sim \mathcal{N}(\mu, \sigma^2)$ and Φ is the standard normal distribution function. This is equation (4.19) of Chapter 4 with the revenue and penalty prices named, and Figure 4.7 there shows the shape of the objective: linear while the commitment is safely below the mean, bending over as the shortfall risk grows, falling once penalties dominate.

The feature that matters commercially is what the optimum does when the forecast changes shape rather than position. Take two forecasts with the same point estimate of 100 MW: forecast A with $\sigma = 5\,\text{MW}$ and forecast B with $\sigma = 30\,\text{MW}$. With $\pi_C = 1$ and $\pi_P = 3$, equation (14.14) gives $\Phi(z) = 1/3$ and hence $z = -0.43$, so

$$P^{\text{commit}\star} = 100 - 0.43\,\sigma = \begin{cases} 97.8\,\text{MW}, & \sigma = 5\,\text{MW}, \\ 87.1\,\text{MW}, & \sigma = 30\,\text{MW}. \end{cases} \tag{14.15}$$

Figure 14.7 plots both objectives.

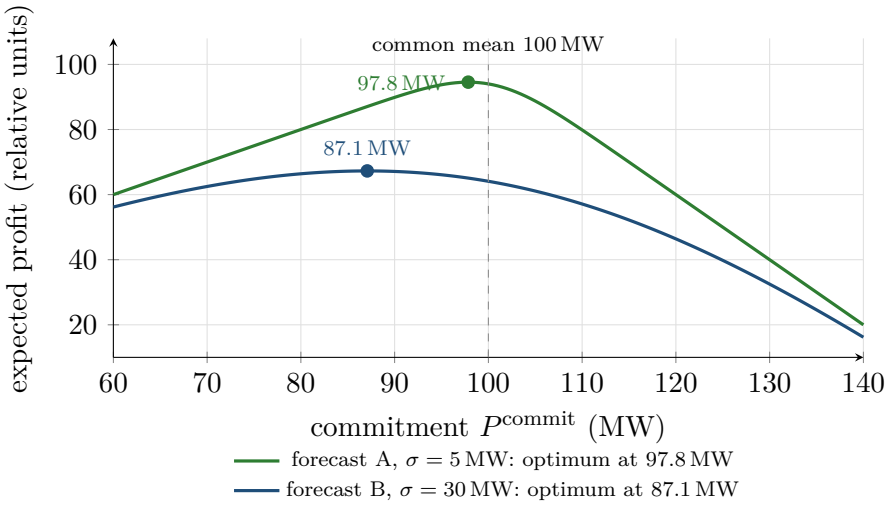


Figure 14.7: Expected profit of equation (14.13) for two forecasts with the same mean $\mu = 100\,\text{MW}$ and different spreads, with $\pi_C = 1$ per MW committed and $\pi_P = 3$ per MW short (relative units). The optimum sits $0.43\,\sigma$ below the mean in both cases, which is 2.2 MW below for forecast A and 12.9 MW below for forecast B. Data from `scripts/ch14_data.py`.

Both curves peak below the dashed line at the common mean, and the wider forecast peaks further below it and lower. Three readings follow. The prudent commitment falls as the forecast widens, from 97.8 to 87.1 MW, even though the expected flexibility is identical. The achievable profit falls as well, from 94.6 to 67.3 relative units, so the width of the forecast is itself worth money. And committing the mean instead of the optimum costs only 0.5 units under forecast A but 3.2 units under forecast B, so the penalty for ignoring the spread grows with the spread. Uncertainty changes the economic decision, not merely the confidence attached to it; this is the strongest argument in the lesson sequence for probabilistic rather than point forecasting.

Principle 14.3. *The optimal commitment depends on the spread of the forecast and not only on its mean. A narrower forecast is worth money even when it is centered in the same place.*

14.6.7 The value chain

Everything in this book now sits on one chain, drawn in Figure 14.8: device states are measured, flexibility is forecast from them, a service is committed on the strength of the forecast, an optimizer decides how to deliver it, devices are dispatched, delivery is measured and verified, and payment follows, along with the information needed to forecast better next time.

The chain is closer to commercial reality than the description a newcomer usually starts with, which is "connect the devices and control them". Connecting and controlling are two

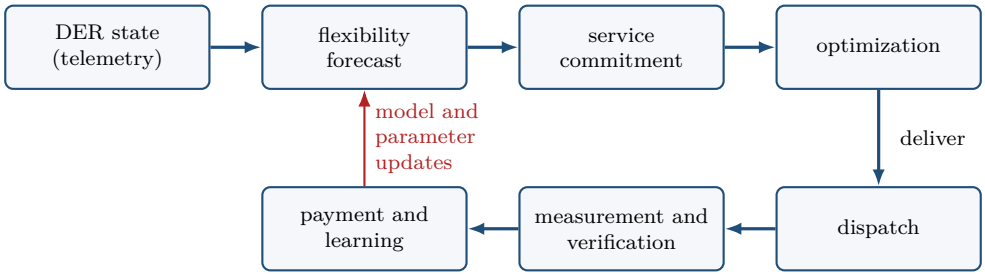


Figure 14.8: The VPP value chain. Blue arrows are the forward path from measured device states to payment; the red arrow is the feedback that makes the next forecast better, which is what turns a one-off transaction into an operating business.

boxes out of seven. The forecast decides what can be sold, the commitment decides what is sold, the verification decides what is paid, and the feedback decides whether next month is better than this one. The chapters of this book occupy the boxes: the device chapters supply the first, the DERMS chapter the fourth and fifth, the measurement-and-verification material the sixth, and the present chapter the second and third.

14.7 What to retain

Definition 14.1 (Grid service)**.** A grid service is a defined product $S = (P, \tau, r, t_{\text{notice}}, \alpha, \ell)$ that a VPP sells: a power change of a stated size, held for a stated duration, reached at a stated speed, after stated notice, delivered with stated reliability, at a stated location. A VPP does not sell distributed energy resources; it provides services, and the forecasting and optimization problems it must solve are determined by which services it sells.

The same fleet can stack several sources of value, provided the physical flexibility is not double-counted. Table 14.5 lists the chapter's quantities and the subsection that defines each.

Table 14.5: The objects that describe a grid service and a commitment.

Object	Symbol	Definition	Section
Service tuple	S	$(P, \tau, r, t_{\text{notice}}, \alpha, \ell)$	14.1
Energy	E	$\int P(t)\,\mathrm{d}t$ over the service window	14.2.1
Peak reduction	—	reduction in $\max_t P_{\text{system}}(t)$	14.2.2
Firm capability	$F_{0.05}$	the flexibility exceeded with probability 0.95	14.2.3
Sustainable duration	τ	usable energy divided by committed power	14.2.3
Held reserve	P^{res}	capability kept unused so it can respond	14.2.4
Energy–reserve split	$P^{\text{en}}, P^{\text{res}}$	allocation of one capability between two uses	14.2.5
Regulation target	$P^{\text{tgt}}(t)$	the signal a fast resource must track	14.2.6
Ramp limit	$R^{\max}$	bound on $\lvert \mathrm{d}P_{\text{net}}/\mathrm{d}t \rvert$	14.3.4
Constrained flow	P_ℓ	flow on the equipment a congestion service relieves	14.3.5
Co-optimization	—	$\max \sum_k R_k - C_{\text{op}}$ subject to $\mathcal{U}(x_t)$	14.4.3
Commitment	P^{commit}	$\arg\max \mathbb{E}[\pi_C P^{\text{commit}} - \pi_P (P^{\text{commit}} - F)^+]$	14.6.6

14.8 Exercises

Exercise 14.1. A battery resource has 100 MW of inverter capacity and 200 MWh of usable energy. Compute the largest constant power it can hold for $\tau = 1$, 2 and 4 hours, and sketch the three points on the $C(\tau)$ axes of Figure 4.3. Which of the following obligations can it accept: 100 MW for two hours, 60 MW for four hours, 80 MW for three hours? How much usable energy would the third one require?

Exercise 14.2. A 10 MW battery can sell energy at a margin of \$50 per MW for the hour, or reserve at \$30 per MW held, and the two may be split in any proportion. Compute the revenue for the splits $(10, 0)$, $(5, 5)$ and $(0, 10)$ megawatts of energy and reserve. What reserve price per megawatt would make the all-reserve allocation optimal? Now suppose holding reserve additionally requires keeping SOC above 0.4 while the energy dispatch would have taken it to 0.2: explain in one sentence why the split is then no longer a purely linear comparison.

Exercise 14.3. Using equation (14.14) with $\pi_C = 1$ and $\pi_P = 3$, compute the optimal commitment for $F \sim \mathcal{N}(100, \sigma^2)$ with $\sigma = 10$ MW and with $\sigma = 50$ MW, and the expected profit at each optimum. Then repeat the first case with $\pi_P = 10$. Which moves the commitment further: raising the penalty from 3 to 10, or raising the forecast standard deviation from 10 to 50 MW?

Exercise 14.4. A system's net load rises by 500 MW over ten minutes and the operator requires $|\mathrm{d}P_{\text{net}}/\mathrm{d}t| \le 30$ MW per minute. Find a trajectory $P_{\text{VPP}}(t)$ that satisfies the requirement, and compute the energy it uses. Then suppose half of the ramp is caused by solar output falling on one constrained feeder: explain which entry of equation (14.1) changes, and why a fleet of the same size located elsewhere cannot provide the same service.

14.9 Connection to the TERRA framework

A grid service is a statement about an edge flow over a window, and every item in this chapter is a query, a bound or an expectation on the same conservation graph the earlier chapters built.

- **A service is a query on the model.** The tuple $S = (P, \tau, r, t_{\text{notice}}, \alpha, \ell)$ of equation (14.1) is a set of conditions on the flow on one edge: its magnitude, how long it is held, its rate of change, the time at which it must be decided, the probability with which it must hold, and which edge it is. Answering "can the fleet provide S" is a forward simulation from the current state followed by a bound check, which is the $F(t, \tau)$ query of Chapter 1 with four more entries attached.

- **Reserve is a tightened bound, not a flow.** Holding P^{res} moves no energy; it narrows the admissible range of a flow variable so that a future request is certain to be feasible. The opportunity cost of equations (14.5) and (14.6) is the change in the optimal objective caused by tightening that bound, which is the bound's dual value, and that is exactly the price at which selling reserve becomes worthwhile.

- **Regulation drift is accumulator conservation.** The state-of-charge trajectories of Figure 14.3 are the battery accumulator integrated against the service signal. The loss of one direction when $\text{SOC}^{\min}$ is reached is a bound becoming active in a model that already contains it, so no separate "regulation capability" model is needed: the energy headroom is read off the same state equation that produced the power.

- **Double counting is one flow variable in two revenue terms.** On a graph each physical capability is a single variable. Selling it twice means writing two service constraints

that the same variable cannot satisfy in any scenario where both are called, which is why the co-optimization of equation (14.10) must be posed over scenarios and not over expectations.

- **Locational services need the network, not the aggregate.** The constrained flow P_ℓ of equation (14.9) and the distribution services of Table 14.2 are bounds on specific edges. They cannot be evaluated on a single aggregate node, so the model must retain the feeder topology and the assignment of devices to it, even though the wholesale products never ask for either.
- **Commitment is a decision on a predictive distribution.** Equation (14.13) is an expectation over the predicted distribution of an aggregate edge flow, and equation (14.15) shows the decision moving with that distribution's spread. Posterior inference of device states and parameters is therefore not a reporting feature but the input to the revenue calculation.

In summary, the grid buys conditions on flows, and a conservation-graph model of the fleet is what turns measured device states into a defensible statement about which conditions can be met, how reliably, and at what cost. The next chapter goes back inside the feasible set $\mathcal{U}(x_t)$, to the customer and device constraints that fix how much flexibility exists to be sold in the first place.

Chapter 15

Customer and device constraints: why DERs are not conventional generators

15.1 Purpose and the intersection of constraints

Every resource considered so far was described by its physics: a thermal mass and a comfort band in Chapter 5, an accumulator and its limits in Chapter 6, a deadline and a charging rate in Chapter 7. The physics is necessary and not sufficient. A thermostat that can physically hold 25°C belongs to a household that may not accept 25°C; a battery that can physically reach a 5 % state of charge belongs to an owner who wants a reserve for the next outage; a program contract may forbid what both the device and the customer would allow. The question of this chapter is

How much of a device's physical capability is actually available to the grid, once the customer and the contract have taken their share?

Jargon: opt-out, override, event fatigue, program rules, segmentation

An *opt-out* is a customer's refusal to participate, either standing (leaving the program) or declared in advance for one event. An *override* is a manual action taken *during* an event that cancels the command, typically pressing buttons on the thermostat or unplugging the car. *Event fatigue* is the decline in willingness to accept events as their number grows. *Program rules* are the contractual limits in the enrollment agreement: how many events per season, how long each may last, how much notice is given, which months are eligible, how far a setpoint may be moved, and what opt-out rights the customer keeps. *Segmentation* is the grouping of enrolled customers into cohorts by their measured tolerance, so that cohorts can be dispatched differently.

Key idea

A DER is useful to the grid only inside the limits imposed by the customer and the device. The grid's claim on the device is a residual, not a right.

This section establishes the frame: whose purpose the device serves, how the usable set of controls is built from three separate restrictions, and the difference between a limit that cannot be crossed and a limit that is merely expensive to approach. The sections that follow catalogue the customer constraints, the device constraints, the program rules and their economics, the spread of all of these across a fleet, and finally the feasible resource that remains.

15.1.1 The grid is not the device's primary purpose

A gas turbine exists to produce electricity. That is its only job, and every operating decision about it can be made on grid terms. None of the devices in a virtual power plant is like that. A thermostat exists to keep a building comfortable, an electric vehicle to provide transportation, a home battery to supply backup power and reduce a bill, a water heater to deliver hot water. Grid service is a secondary use of equipment bought for something else.

The consequence is a hierarchy, not a trade-off between equals. The aggregator may use the portion of the device's capability that does not interfere unacceptably with the customer's primary objective, and the word "unacceptably" is defined by the customer, not by the aggregator. Table 15.1 lists the main device classes with the purpose each one serves and the part of its operation that is left over for the grid. In every row the grid's share is a *timing* or *margin* decision inside a service that still has to be delivered: cooling still happens, the car still charges, the tank is still hot. This is the same observation as the load-shifting principle of Chapter 2, stated from the customer's side.

Table 15.1: The primary purpose of each device class and the part of its operation available for grid service.

Device	Primary purpose for the customer	What the grid may use
Thermostat and HVAC	indoor temperature inside a comfortable band	short shifts of cooling or heating within the band
Electric vehicle	transportation; enough charge by departure	when the required energy is delivered inside the plugged-in window
Home battery	backup during outages; bill reduction	charge and discharge above the owner's reserve
Water heater	hot water on demand	when the tank is reheated
Rooftop PV and inverter	own consumption; bill reduction	curtailment and reactive support inside the apparent-power limit
Commercial process load	production schedule	interruption inside process tolerances

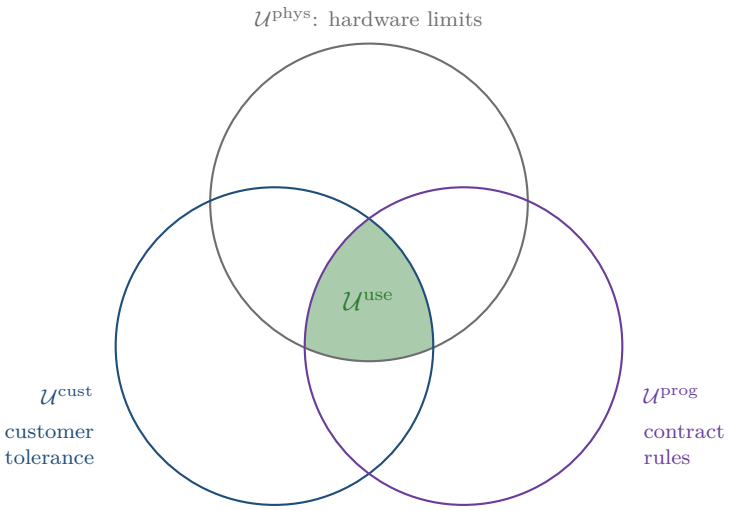
15.1.2 Usable flexibility as an intersection

Chapter 1 wrote the admissible controls of a device in state x_t as a set $\mathcal{U}(x_t)$. That set must now be split into three, because three different parties restrict it:

$$\mathcal{U}_i^{\text{use}} = \mathcal{U}_i^{\text{phys}} \cap \mathcal{U}_i^{\text{cust}} \cap \mathcal{U}_i^{\text{prog}}. \tag{15.1}$$

Here $\mathcal{U}_i^{\text{phys}}$ is what the hardware can do without damage, $\mathcal{U}_i^{\text{cust}}$ is what the customer will tolerate, and $\mathcal{U}_i^{\text{prog}}$ is what the enrollment contract permits. Only the intersection is dispatchable. Figure 15.1 draws the three sets and their intersection.

Three properties of equation (15.1) matter operationally. The intersection can be empty, in which case the device is enrolled but presently useless. It is not symmetric in time: the physical set moves with the device state, the customer set moves with the customer's plans, and the program set changes only when the contract does. And it cannot be recovered from a device data sheet, which describes $\mathcal{U}^{\text{phys}}$ alone. Every capacity number in this chapter is a statement about the intersection.



Example: the cells of a home battery permit SOC ≥ 0.05, so $\mathcal{U}^{\text{phys}}$ reaches down to 5 %. The program guarantees the owner SOC ≥ 0.30 at all times. The 25 percentage points in between are physically available and operationally unusable.

Figure 15.1: Usable flexibility is an intersection, equation (15.1). The shaded region $\mathcal{U}^{\text{use}}$ is the only part of the device's capability a dispatch may use; a control excluded by any one of the three sets is excluded altogether. Set sizes are illustrative.

15.1.3 Hard and soft constraints

The three sets contain two kinds of limit, and conflating them is a common modeling error. A *hard* constraint cannot be violated: a battery may not go below $\text{SOC}^{\min}$, an inverter may not exceed its apparent-power rating, a program may not allow an indoor temperature above $T^{\max} = 25°\text{C}$. A *soft* constraint can be violated at a cost: a household prefers 22°C and will accept 23°C for something worthwhile. Hard constraints define feasibility; soft constraints define preference. The first belongs in the constraint set, the second in the objective:

$$u_t \in \mathcal{U}^{\text{use}}(x_t) \qquad \text{against} \qquad c_{\text{comf}}(T_{\text{in}}) = \lambda\,(T_{\text{in}} - T_{\text{set}})^2. \tag{15.2}$$

Figure 15.2 shows both on one temperature axis. The horizontal axis is the indoor temperature; the vertical axis is the comfort penalty of equation (15.2) with $\lambda = 1$. The blue curve is the penalty, zero at the preferred 22°C and growing quadratically on either side; the red walls at 21°C and 25°C are the program's hard band, and the shaded regions beyond them are infeasible at any price. The two mechanisms behave differently under a high grid value: a soft constraint bends, which is what makes a penalty useful, and a hard constraint does not bend at all.

15.2 Customer constraints

This section catalogues the limits that come from the person rather than the equipment: comfort bands, manual overrides, fatigue across events, departure deadlines, and backup reserves. Each one is a restriction the device itself would not impose, and each one can move while the device is being dispatched.

15.2.1 The comfort band and the comfort penalty

Take the household of Figure 15.2. Its preferred indoor temperature is 22°C; its enrollment contract allows the aggregator to move the setpoint anywhere in $21 \leq T_{\text{in}} \leq 25°\text{C}$. The upper

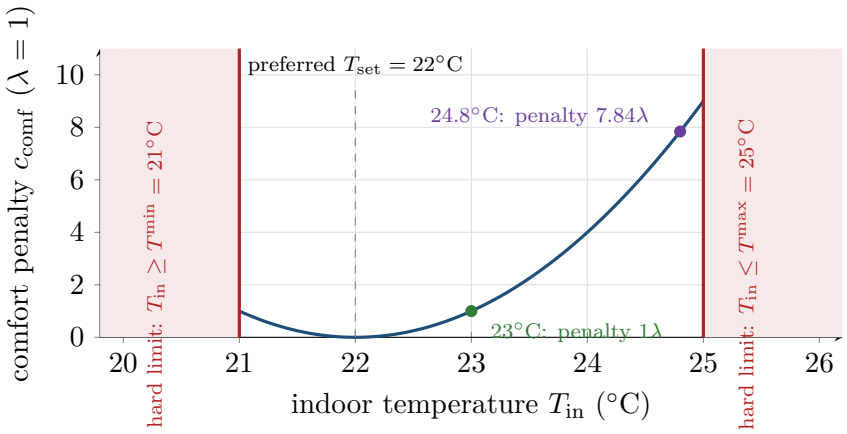


Figure 15.2: Hard and soft constraints on one axis. The quadratic penalty $\lambda(T_{\text{in}} - 22)^2$ with $\lambda = 1$ is a preference; the walls at 21°C and 25°C are the program band, and the shaded regions beyond them cannot be entered at any grid value. Penalty units are illustrative.

bound is hard: $T_{\text{in}} \leq 25°\text{C}$ is a condition of enrollment and the controller is not permitted to cross it. Inside the band the household is not indifferent, and the deviation is charged at $c_{\text{comf}} = \lambda(T_{\text{in}} - 22)^2$.

The quadratic form is a modeling choice with two useful properties. Small deviations are nearly free, which is why a 0.5°C setback is dispatched routinely, and large deviations are disproportionately expensive, which is why the controller does not simply run to the edge of the band. At 23°C the penalty is 1λ; at 24.8°C it is 7.84λ, almost eight times as much for less than twice the setback. An optimizer that trades grid value against this penalty therefore accepts 23°C for an important event and refuses 24.8°C unless the event is worth roughly eight times as much per degree. The marginal statement is the useful one: with grid value v per degree of setback, the optimizer stops at $\Delta T^* = v/(2\lambda)$, clipped at the program band.

15.2.2 Customer override

Suppose the platform commands 23°C and the occupant walks to the wall and sets 21°C. The command was accepted by the thermostat, the telemetry confirmed it, and the flexibility is gone. Worse, the building now draws *more* power than it would have without the event, because the occupant has asked for a colder house than the original setpoint.

An override is not a measurement failure or a communication fault; it is the customer exercising a right the program granted. The practical consequence is that the delivered response is a function of human behavior as well as of physics, so customer behavior belongs inside the resource model. A model that computes the thermal response of 100 000 buildings and then assumes every setpoint command is honored will overstate the fleet, and the error is not random noise around the truth: it is a bias in one direction, always toward less delivery than predicted.

15.2.3 The override probability

The natural object is the override probability p^{ovr} of Chapter 4, written as a function rather than a constant:

$$p^{\text{ovr}} = f(\Delta T,\ \text{event duration},\ T_{\text{out}},\ \text{time of day},\ \text{prior events}). \tag{15.3}$$

Each argument has an evident sign. A larger setback is noticed sooner; a longer event gives more opportunity to notice; a hotter day makes the same setback less tolerable; an evening

event finds more people at home; and a customer who has already sat through several events this month is quicker to intervene.

The dependence on ΔT has a consequence that is easy to miss when p^{ovr} is treated as a constant. The expected delivered setback of one thermostat is the product

$$\mathbb{E}[\Delta T^{\text{del}}] = \Delta T \left(1 - p^{\text{ovr}}(\Delta T)\right), \tag{15.4}$$

a product of one factor that grows with the request and one that shrinks. Figure 15.3 plots both for an illustrative logistic override model

$$p^{\text{ovr}}(\Delta T) = \left[1 + e^{-(\Delta T - 3.5)/0.7}\right]^{-1}, \tag{15.5}$$

whose midpoint 3.5°C is the setback at which half the customers intervene and whose scale 0.7°C sets how abruptly they do. The horizontal axis is the requested setback; the left vertical axis is the expected delivered setback from equation (15.4), the right vertical axis the override probability. The dashed red curve rises monotonically, from 0.03 at 1°C to 0.90 at 5°C. The solid green curve rises, peaks at $\Delta T \approx 2.75$°C with about 2.05°C delivered, and then falls: at 4°C requested only 1.32°C is expected, less than the 1.79°C expected from a 2°C request.

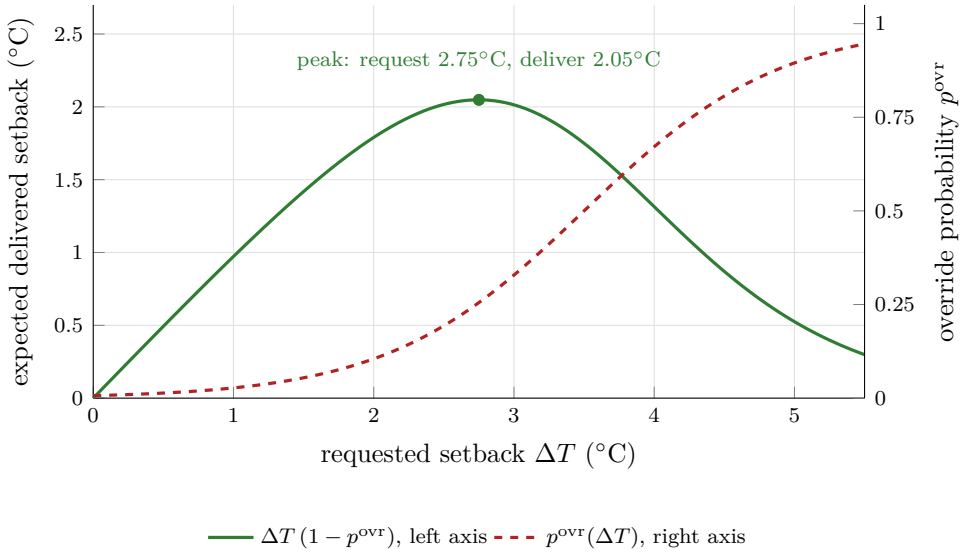


Figure 15.3: Override probability and expected delivered setback for the illustrative logistic model of equation (15.5), midpoint 3.5°C and scale 0.7°C. The dashed red curve (right axis) is p^{ovr}; the solid green curve (left axis) is $\Delta T(1 - p^{\text{ovr}})$ from equation (15.4). The curve peaks at 2.75°C and falls thereafter. Parameters are illustrative.

Principle 15.1. *Requesting more response can deliver less. Where the override probability rises with the request, the expected delivery has an interior maximum, and the dispatch that maximizes it is smaller than the dispatch the program permits.*

The practical reading of Figure 15.3 is that the aggressiveness of a dispatch is a decision variable with an optimum, not a dial to be turned up when more capacity is needed. It also changes what a forecast must contain: the quantity to forecast is the delivered response, which depends on the command, so the forecast and the dispatch cannot be computed in sequence and must be solved together.

15.2.4 Event fatigue

One event per month is tolerated well. Fifteen events per month is a different program: opt-out rates rise, overrides become habitual, and enrollment itself declines as customers leave. The quantity that degrades is not the thermal response of the buildings, which is unchanged, but the willingness of the people in them.

Fatigue makes customer flexibility an intertemporal resource. Using it today reduces how much is available tomorrow, which is formally the same structure as the depletion of a battery's stored energy in Chapter 6 and as the within-event decay of Figure 1.3 in Chapter 1, now acting across events instead of within one. The simplest usable model adds a state per customer, the number of recent events N_i, together with a recovery rate at which that state decays, and makes both p_i^{ovr} and the enrollment probability functions of it. A dispatch optimizer that ignores the state will spend the fleet's tolerance in the first hot week of the summer and find it missing in August.

15.2.5 The electric-vehicle departure deadline

For an electric vehicle the binding customer constraint is usually the one introduced in Chapter 7,

$$\mathrm{SOC}(T^{\mathrm{dep}}) \geq \mathrm{SOC}^{\mathrm{req}}, \tag{15.6}$$

and it is a hard constraint: a car that is not charged enough cannot make the trip. Charging may be delayed only while slack remains, and the amount of slack is set by the customer's schedule.

Take a 60 kWh pack at $\mathrm{SOC} = 0.40$ with $\mathrm{SOC}^{\mathrm{req}} = 0.80$, a 7.2 kW charger and $\eta = 0.92$. The energy required is $E^{\mathrm{need}} = 0.40 \times 60 = 24\,\mathrm{kWh}$, which takes $24/(0.92 \times 7.2) = 3.62\,\mathrm{h}$ of charging. Two customers own the identical car:

- departure in 10 h: urgency $U = 3.62/10 = 0.36$, slack 6.38 h, and the charging block may be placed almost anywhere in the window;
- departure in 90 minutes: $U = 3.62/1.5 = 2.41 > 1$. The requirement cannot be met even at full power. In 1.5 h the charger delivers 9.94 kWh, which brings the car to $\mathrm{SOC} = 0.57$, and the only sensible dispatch is maximum power immediately.

In the first case the vehicle is a 7.2 kW flexible load for most of the night; in the second it is a 7.2 kW inflexible load and, in a scarcity hour, an unwelcome one.

Same hardware, different resource. The two vehicles have the same state of charge, the same $P^{\max}$, the same $E^{\max}$ and the same efficiency. Every number on the data sheet agrees, and their flexibility differs by a factor that is not finite, since one of them has none. Device specifications alone cannot define virtual-power-plant capacity; the customer's schedule is part of the specification. The urgency metric of Chapter 7 is exactly the scalar that distinguishes the two cases, and the fleet consequence is visible in Figure 7.7, where the deadline distribution determines how much of the fleet can be moved at all.

15.2.6 The backup reserve, and its drift

A home battery of $E^{\max} = 13.5\,\mathrm{kWh}$ sits at $\mathrm{SOC} = 0.70$. Its cells permit discharge to 0.10. The owner bought it for outage backup and insists on keeping $\mathrm{SOC}^{\mathrm{res}} = 0.40$. The energy available to the grid is

$$E^{\mathrm{avail}} = (\mathrm{SOC} - \mathrm{SOC}^{\mathrm{res}})\, E^{\max} = (0.70 - 0.40) \times 13.5 = 4.05\,\mathrm{kWh}, \tag{15.7}$$

not $(0.70 - 0.10) \times 13.5 = 8.10\,\mathrm{kWh}$. The owner's preference, which appears nowhere in the battery's specification, has halved the resource. At a 5 kW discharge the deliverable duration falls from 1.62 h to 0.81 h, which is the $C(\tau)$ curve of Figure 4.3 shifted bodily to the left.

The reserve is not a constant. Suppose the same owner normally sets $\mathrm{SOC}^{\mathrm{res}} = 0.20$ and, on seeing a severe-weather forecast, raises it to 0.60. The usable energy falls from $(0.70 - 0.20) \times 13.5 = 6.75\,\mathrm{kWh}$ to $(0.70 - 0.60) \times 13.5 = 1.35\,\mathrm{kWh}$, a reduction of 80 % with no change in state of charge, weather-driven grid need, or hardware. The reduction arrives on precisely the days when the grid is most likely to need the fleet, because the same storm forecast drives both. Customer preferences are therefore a quantity to be observed and forecast, like weather or load, and their correlation with grid stress has the sign that hurts.

> **Principle 15.2.** *Available flexibility is not fixed by hardware. It is set by customer preferences that change over time and that are correlated with the conditions under which the flexibility is needed.*

15.3 Device constraints

The equipment imposes its own limits, which hold regardless of grid value: protection bounds, cycling limits, state-dependent ratings, and a wear cost that is real even though no meter records it.

15.3.1 Protection limits and the inverter capability curve

A battery system carries a set of bounds that exist to protect the hardware:

$$P^{\min} \le P_t \le P^{\max}, \qquad \mathrm{SOC}^{\min} \le \mathrm{SOC}_t \le \mathrm{SOC}^{\max}, \qquad T_t^{\mathrm{bat}} \le T^{\mathrm{bat,max}}. \tag{15.8}$$

These are enforced in the device's own firmware and are not negotiable by the platform: a command outside them is clipped or rejected, and a dispatch built on the assumption that it will be honored produces a tracking error the platform cannot explain.

Jargon: apparent power and the capability curve

An inverter's rating is an *apparent power* in kilovolt-amperes (kVA), not a real power in kilowatts. Real power P (kW) and reactive power Q (kvar) are two components of the same current, and the rating bounds their combination, $P^2 + Q^2 \le S_{\max}^2$. The boundary of that region drawn in the (P, Q) plane is the *capability curve*.

The inverter constraint is the one most often omitted from a dispatch written in megawatts alone:

$$P_t^2 + Q_t^2 \le S_{\max}^2. \tag{15.9}$$

Figure 15.4 draws it for a 7.6 kVA inverter shared by a 5 kW battery and a 7.6 kW photovoltaic array. The horizontal axis is real power, the vertical axis reactive power; the shaded disk is the set allowed by equation (15.9), and the dashed vertical lines at $\pm 5\,\mathrm{kW}$ are the battery's own power rating. Two operating points are marked. At $P = 5\,\mathrm{kW}$ the reactive capability is $\sqrt{7.6^2 - 5^2} = 5.72\,\mathrm{kvar}$; at $P = 7.0\,\mathrm{kW}$ of solar output it has fallen to $\sqrt{7.6^2 - 7.0^2} = 2.96\,\mathrm{kvar}$. The cross marks a request for 5 kW and 7 kvar together, which needs $\sqrt{5^2 + 7^2} = 8.60\,\mathrm{kVA}$ and is refused even though each component alone is within rating. Real and reactive service compete for one device, which is why a volt/VAR request and an energy dispatch to the same inverter cannot be planned independently.

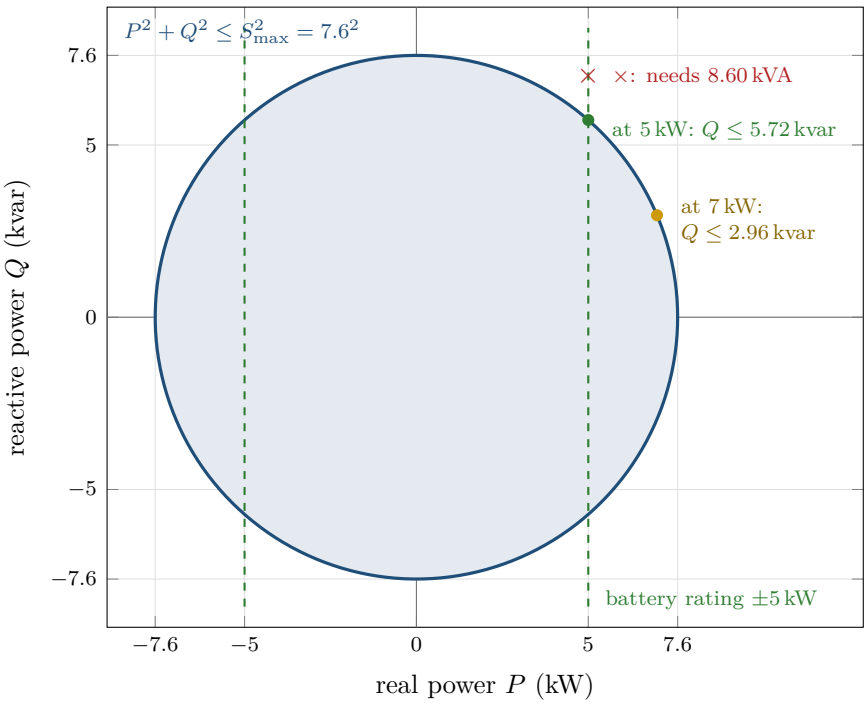


Figure 15.4: Inverter capability curve, equation (15.9), for $S_{\max} = 7.6\,\text{kVA}$. The shaded disk is the feasible set in the (P, Q) plane; the dashed green lines are the battery's ±5 kW rating. Reactive capability shrinks as real power grows, so the two services compete.

15.3.2 Minimum on and off times

A compressor must not be switched rapidly. Starting a motor draws a large inrush current and the refrigerant takes time to equalize, so rapid cycling, called *short cycling*, shortens equipment life and degrades efficiency. The protection is stated as two dwell times:

$$t_{\text{on}} \ge t_{\text{on}}^{\min}, \qquad t_{\text{off}} \ge t_{\text{off}}^{\min}, \tag{15.10}$$

typically a few minutes each for residential air conditioning. The device enforces them whether or not the platform knows about them.

Two consequences follow for dispatch. First, response speed is capped: a thermostat fleet cannot be modulated on a 30-second cycle however fast the communication path is, which bears directly on whether the fleet can offer fast-responding services. Second, the dispatch problem acquires integer structure, because equation (15.10) constrains a binary compressor state rather than a continuous power. The usual way out at fleet level is to leave individual devices alone and modulate the *number* of devices switched, which is the rotation scheme of Figure 5.11 in Chapter 5: a smooth aggregate response built from devices that each respect their own dwell times.

15.3.3 State-dependent charging limits

The charging constraint on a battery or vehicle is written $0 \le P_t \le P^{\max}$, but $P^{\max}$ is not a constant. It depends on the vehicle, the charger, the cell temperature and the state of charge:

$$0 \le P_t \le P^{\max}(\text{SOC}_t, T_t). \tag{15.11}$$

Near a high state of charge the charging current is reduced to protect the cells, a behavior called *taper*, and at low cell temperature the limit is reduced throughout.

Figure 15.5 shows an illustrative version. The horizontal axis is state of charge, the vertical axis the maximum charging power. The green curve is a warm pack: 11 kW until SOC = 0.80, then falling linearly to 2 kW at full charge. The blue dashed curve is the same pack at 0°C: limited to 5 kW from the start and tapering from SOC = 0.55 to 1 kW. The practical effect on scheduling is direct. A car at SOC = 0.90 charges at 6.5 kW rather than 11 kW, so the last few kilowatt-hours before a deadline take disproportionately long and a schedule computed with the nameplate rate will miss the deadline. Physical capability is state-dependent, exactly as $P^{\text{avail}}(x)$ was in Chapter 1, and here the state includes a temperature the platform may not measure.

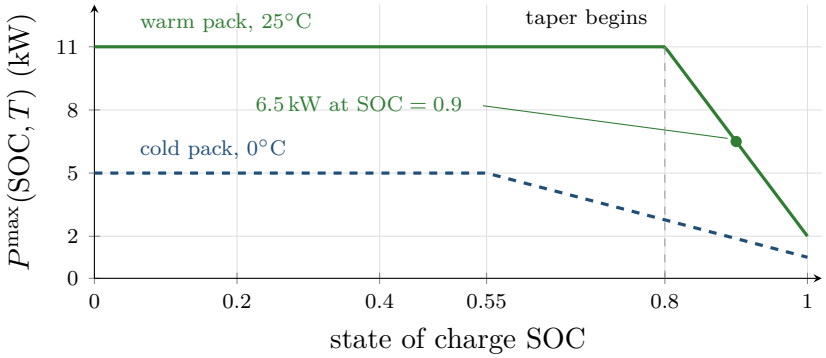


Figure 15.5: State-dependent charging limit, equation (15.11). Warm pack: 11 kW held to SOC = 0.80, then linear to 2 kW at SOC = 1. Cold pack: 5 kW held to SOC = 0.55, then linear to 1 kW. Both curves are illustrative piecewise-linear approximations of a manufacturer taper schedule.

15.3.4 Degradation as an economic constraint

Deep, fast cycling at a high temperature is physically permitted and economically undesirable, because it consumes battery life. The cost is written

$$c_{\text{deg}} = \lambda_{\text{deg}}\, g(\text{SOC},\ \text{DoD},\ P,\ T), \tag{15.12}$$

where the depth of discharge DoD is the fraction of capacity removed in a cycle, and g increases in depth, in rate and in temperature. For an illustrative order of magnitude, a 13.5 kWh pack costing $8,000 and warranted for 4,000 equivalent full cycles carries $4{,}000 \times 13.5 = 54{,}000$ kWh of throughput, or about $0.15 per kWh discharged. A 4 kWh discharge therefore costs roughly $0.59 of battery life before any efficiency loss is counted.

Two conclusions follow. The marginal cost of battery flexibility is not zero, so a dispatch that treats stored energy as free will cycle the fleet more than the owner would choose and will, over a season, transfer value from the customer to the program. And because g depends on how the energy is taken, not only on how much, two dispatches with the same energy can have different costs: a shallow cycle at moderate power is markedly cheaper than a deep one at full power, which gives the optimizer a reason to spread a requirement over many devices rather than exhaust a few.

15.4 Program rules and the economics of an event

A feasible dispatch is not automatically a worthwhile one. This section covers what the event costs once customers are paid and equipment wears, the contractual rules that bound the optimizer, and the budget of events that must be allocated across a season.

15.4.1 Compensation and the net value of an event

Suppose the virtual power plant earns \$100 from an event. Customers are compensated \$30 for their participation, and the batteries dispatched during it lose \$20 of life by equation (15.12). The net value is

$$100 - 30 - 20 = \$50. \tag{15.13}$$

The arithmetic is trivial and its implication is not: a dispatch inside every constraint in Sections 15.2 and 15.3 can still be the wrong decision. Had the event been worth \$45, every constraint would have been satisfied and the program would have lost money.

Principle 15.3. *Feasible is not optimal. The feasible set answers whether a dispatch is allowed; the objective, net of customer compensation and equipment wear, answers whether it is worth making.*

This is why the objective of Chapter 3 carries a customer term beside its tracking term, and why the commitment curve of Figure 4.7 in Chapter 4 peaks below the forecast mean. The costs that make the net value smaller than the gross value are mostly costs the grid does not see.

15.4.2 Program rules as constraints

The enrollment contract is a set of hard constraints with the same standing as the physical ones. Table 15.2 lists the usual rules, a typical value for each, and what each one bounds in the optimizer. The entries are ordinary contract terms, and each maps onto a recognizable mathematical object: a budget, a horizon, a decision deadline, an eligibility window, a bound, a source of randomness, or a recovery constraint.

Table 15.2: Typical program rules and what each constrains. Values are illustrative of residential thermostat programs.

Rule	Typical value	What it constrains
Maximum events per season	20	a budget on the number of dispatches, allocated over months
Maximum event duration	4 h	how long one response must be held; fixes τ in $C(\tau)$
Notification lead time	30 min	how late the dispatch decision may be made, hence which forecast is used
Seasonal eligibility	June–September	the days on which the resource exists at all
Allowable setpoint adjustment	± 4°C	the bound $T^{\max}$ in the comfort constraint
Opt-out rights	per event, unlimited	makes the delivered response a random variable
Minimum time between events	24 h	recovery of thermal state and of customer tolerance

The important point is that $\mathcal{U}^{\text{prog}}$ can bind before $\mathcal{U}^{\text{phys}}$ and before $\mathcal{U}^{\text{cust}}$. A fleet of buildings that could physically hold a four-hour response, with customers who would tolerate it, delivers two hours if the contract says two hours. Program design is therefore a lever on capacity of the same kind as hardware procurement, and a cheaper one.

15.4.3 The event budget

A program allows 20 thermostat events per summer. By mid-August 18 have been called and two remain. A moderately valuable event is available today. Whether to call it is not a feasibility question: two events are in hand and the day qualifies. It is an allocation question, and the rule is the familiar one for a scarce stock,

$$\text{call the event today} \iff v_{\text{today}} \geq \mathbb{E}\big[v^{(2)}_{\text{remaining}}\big], \tag{15.14}$$

where the right-hand side is the expected value of the second-best remaining opportunity in the season, that being what today's event displaces. If today's event is worth \$6,000 and the hot spell forecast for next week carries an expected \$14,000, the two events are held. The marginal value of a remaining event is a shadow price, and it rises as the stock falls and falls as the season ends with the stock unspent.

15.4.4 The analogy with stored energy

The event budget is formally the same object as a battery's stored energy, and recognizing this saves inventing new machinery for it. A battery holds E^{avail} kilowatt-hours; a program holds N^{rem} events. Both deplete when used, both have a hard floor, both refill on their own schedule, and both therefore create an intertemporal allocation problem whose solution is a shadow price on the stock. Table 15.3 sets the two side by side. The analogy is exact enough that the same dynamic-programming formulation solves both, with the number of remaining events playing the role of the state of charge.

Table 15.3: The event budget as a stored quantity.

Battery energy	Program event budget
$E^{\text{avail}} = (\text{SOC} - \text{SOC}^{\text{res}})E^{\max}$ kilowatt-hours available	$N^{\text{rem}} = N^{\max} - N^{\text{used}}$ events available
depletes with each discharge	depletes with each event called
refills by charging, at a price and over time	refills at the start of the next season
hard floor at the owner's reserve SOC^{res}	hard floor at $N^{\text{rem}} \geq 0$
shadow price: marginal value of a stored kilowatt-hour	shadow price: marginal value of one remaining event
spend when the price is highest in the remaining horizon	call when the grid value is highest in the remaining season

15.5 Heterogeneity, fairness and participation

Customer constraints are not one set of numbers but a distribution across the fleet. This section covers that spread, the cohorts it suggests, the fairness problem created by always dispatching the cheapest customers, and the uncertainty in how many devices respond at all.

15.5.1 Customer heterogeneity

Customer A accepts a 3°C setback without complaint. Customer B tolerates 1°C. Customer C overrides every event and effectively opts out. Electric-vehicle owners differ in the same way: departure times spread over several hours, required states of charge from 0.5 to 0.9, willingness to delay charging from complete to none.

The right object is therefore a distribution, not a value. Figure 15.6 shows an illustrative one for the tolerable setback across an enrolled thermostat population. The horizontal axis is the setback a customer will accept without overriding; the vertical axis is the share of the population at each value. The distribution has mass at zero, the customers who will not participate in practice; a mode at 2°C; and a tail to 5°C. Its mean is 2.2°C, and that mean is the misleading number: a dispatch that requests 2.2°C from everyone is comfortably inside the tolerance of 40 % of the fleet, at the edge for 30 %, and outside it for the remaining 30 %. The aggregate response is the sum over the distribution, not the response of the average customer.

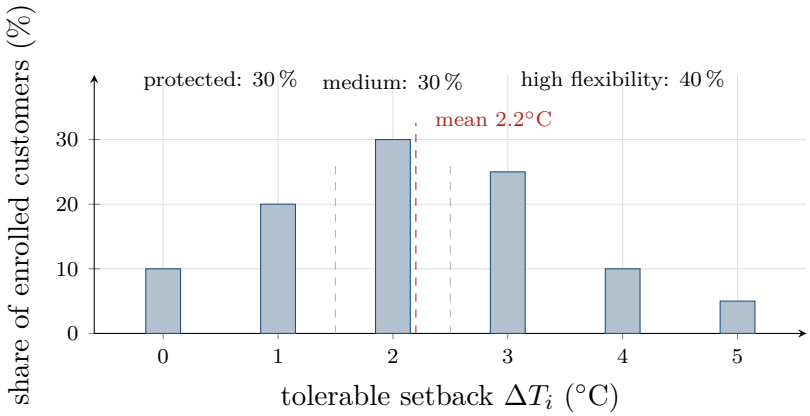


Figure 15.6: Illustrative distribution of tolerable setback across an enrolled thermostat population, with the cohorts of Table 15.4 marked. The mean, 2.2°C, describes no individual customer: 30 % of the fleet tolerates 1°C or less. Shares are illustrative.

15.5.2 Segmentation into cohorts

Because the distribution is wide, treating the fleet as homogeneous wastes the tolerant customers and antagonizes the sensitive ones. The standard response is to segment: estimate each customer's tolerance from history and assign them to a cohort, then dispatch cohorts differently. Table 15.4 gives the three cohorts implied by Figure 15.6, with the share, the tolerance and the dispatch rule for each.

Table 15.4: Segmentation of the population of Figure 15.6 into dispatch cohorts.

Cohort	Share	Tolerable ΔT	Dispatch rule
High flexibility	40 %	3–5 °C	dispatched first, and with the largest setback
Medium	30 %	2°C	dispatched for larger events, at a moderate setback
Protected	30 %	0–1 °C	dispatched last or not at all; counted as capacity only in emergencies

Segmentation is a second kind of aggregation. Chapter 3 aggregated devices by type and region to make the fleet forecastable; cohorts aggregate customers by tolerance to make the fleet dispatchable in order of increasing customer impact. Both reduce a large population to a few numbers a dispatch can use, and both depend on estimating a per-customer quantity that is never measured directly.

15.5.3 Fairness as a cost and as a constraint

An optimizer that minimizes customer impact will dispatch the high-flexibility cohort of Table 15.4 at every event, because those customers are the cheapest. Over a season the same households absorb every event while the protected cohort absorbs none, which is both inequitable and self-defeating: the tolerant customers become fatigued in the sense of Section 15.2.4 and the program loses its best resource.

The remedy is to make repeated use expensive. Add to the objective a per-customer term

$$c_i = \lambda_i \, N_i^{\text{recent}}, \tag{15.15}$$

where N_i^{recent} counts the customer's recent events, so that a heavily used customer is costlier to dispatch than an unused one of the same tolerance, and participation rotates without being explicitly scheduled. Where rotation must be guaranteed rather than encouraged, the same quantity is imposed as a hard constraint,

$$N_i \le N^{\max} \quad \text{for every } i, \qquad \text{or} \qquad |N_i - N_j| \le \delta \quad \text{for every } i, j, \tag{15.16}$$

the first a per-customer cap and the second a balance condition. Either way the optimizer is now managing customer burden alongside grid performance, and the rotation of Figure 5.11 in Chapter 5, introduced there to avoid synchronizing compressors, acquires a second justification.

15.5.4 Participation uncertainty

Ten thousand enrolled devices do not mean ten thousand available devices tomorrow. Customers opt out in advance, devices lose network connectivity, settings are changed, thermostats are put in vacation mode, vehicles are not plugged in. The number of responding devices is a random variable $N^{\text{avail}} \sim p(N)$, and that randomness propagates directly into fleet capacity.

Take 10,000 thermostats of 3 kW each. The enrolled total is

$$P^{\text{enroll}} = 10{,}000 \times 3\,\text{kW} = 30\,\text{MW}, \tag{15.17}$$

and with an expected participation of 80 % the expected delivery is $0.8 \times 30 = 24$ MW. The reliable figure is lower again, and how much lower depends on correlation. Modeling each device as available with probability $p = 0.8$ and pairwise correlation ρ, the variance of the count is

$$\operatorname{Var}\big[N^{\text{avail}}\big] = Np(1-p)\,\big[1 + (N-1)\rho\big]. \tag{15.18}$$

With $\rho = 0$ the standard deviation is 40 devices, or 0.12 MW, and the 5th percentile of delivery is 23.8 MW: independence makes the fleet almost deterministic. With $\rho = 0.1$ the bracket in equation (15.18) is 1,000.9, the standard deviation is 1,265 devices or 3.80 MW, and the 5th percentile falls to 17.8 MW. This is Figure 3.4 of Chapter 3 again: a small common cause, such as a heat wave that keeps people at home and makes setbacks less tolerable, destroys most of the averaging benefit of a large fleet. The chain

$$\underbrace{30\,\text{MW}}_{\text{enrolled}} > \underbrace{24\,\text{MW}}_{\text{expected}} > \underbrace{17.8\,\text{MW}}_{\text{reliable at } 95\,\%} \tag{15.19}$$

is the customer-side version of the nameplate, available and firm capacities of Chapter 4.

15.5.5 Learned behavior and learned reliability

None of these quantities has to be guessed. A program with operating history has, for every past event, the command sent and the response metered, which is exactly the data needed to fit the override model of equation (15.3): a regression or classifier for p^{ovr} on the setback, the

outdoor temperature, the event length, the hour and the number of recent events. Customer behavior then enters the pipeline in the same place as weather, as a forecast feeding a dispatch, rather than as a safety factor applied afterward.

The same applies to the devices. Device i may respond successfully to 95 % of commands and device j to 60 %, through firmware quirks, marginal connectivity or an unreliable gateway. Each device therefore carries a learned reliability score $r_i \in [0, 1]$, estimated from its own history of commanded and delivered responses, and r_i is observable in a way that a customer's tolerance is not: it depends on equipment rather than on a person, and it is stable enough to be useful for months.

15.5.6 Reliability-weighted dispatch

With reliability scores in hand, the quantity to use in a dispatch is not the nominal power but the effective power

$$P_i^{\text{eff}} = r_i \, P_i. \tag{15.20}$$

A 5 kW device with $r_i = 0.5$ is not equivalent to a 5 kW device with $r_i = 0.99$; they differ by a factor of two in expectation and by more than that in variance. Table 15.5 works a four-device example: 20 kW of nominal capacity provides 14.5 kW of effective capacity, and the ordering by P_i^{eff} differs from the ordering by P_i, so device D displaces device B in the dispatch order despite B and A sharing the same rating.

Table 15.5: Reliability-weighted dispatch, equation (15.20). Nominal and effective capacity of four devices.

Device	P_i (kW)	r_i	P_i^{eff} (kW)	Dispatch order
A	5.0	0.99	4.95	first
D	7.0	0.60	4.20	second
C	3.0	0.95	2.85	third
B	5.0	0.50	2.50	last
Total	20.0	–	14.50	

Two refinements are worth noting. Expectation is not the whole story: if a commitment must be met with high probability, a few reliable devices are preferable to many unreliable ones of the same expected total, because the variance of the sum is smaller. And r_i is partly actionable, since a persistently low score usually indicates a fixable gateway or firmware problem rather than an unwilling customer.

15.6 The feasible resource

The chapter's constraints are now assembled into one feasible set, and that set is compared with a conventional generator's. The section ends with the arithmetic that matters operationally: how much of a nominal fleet survives all of the constraints at once.

15.6.1 Privacy and the control boundary

A platform that needs customer constraints does not need customer lives. For control purposes the required information is operational: the power available now, the deadline, and the bounds of the feasible set. Why a household wants the car charged by 06:00 is irrelevant; that $T^{\text{dep}} = 06{:}00$ and $\text{SOC}^{\text{req}} = 0.8$ is all the optimizer uses.

The principle is to expose operational constraints and not personal detail, and it has an engineering payoff beyond its evident ethical one. A device interface specified as a feasible set,

the triple of available power, energy and deadline, is narrow, stable and vendor-independent, whereas an interface specified by customer attributes is none of those things. The same reduction appears in Chapter 4 between the fleet and the grid, and here it appears between the customer and the platform.

15.6.2 The feasible set with four arguments

The admissible control of Chapter 1 depended on the device state alone. Everything in this chapter adds an argument:

$$u_t \in \mathcal{U}\big(x_t,\ \theta_t^{\text{cust}},\ \theta^{\text{dev}},\ \theta^{\text{prog}}\big)\ \cap\ \mathcal{U}_t^{\text{net}}, \tag{15.21}$$

where x_t is the physical state, θ_t^{cust} the customer's current preferences and schedule, θ^{dev} the device's protection parameters, θ^{prog} the program rules, and $\mathcal{U}_t^{\text{net}}$ the network limits at the device's location. Only the first argument is a state in the mechanical sense; the second drifts with human decisions, the third is fixed by hardware, the fourth by contract. At fleet level the same statement reads

$$\mathcal{F}_{\text{VPP}} = \mathcal{F}_{\text{phys}} \cap \mathcal{F}_{\text{cust}} \cap \mathcal{F}_{\text{prog}} \cap \mathcal{F}_{\text{net}}, \tag{15.22}$$

which is equation (15.1) with a fourth term for the distribution network. This is the true feasible resource, and no single data source describes it: the hardware vendor knows $\mathcal{F}_{\text{phys}}$, the customer knows $\mathcal{F}_{\text{cust}}$, the contract fixes $\mathcal{F}_{\text{prog}}$ and the utility holds $\mathcal{F}_{\text{net}}$.

15.6.3 Why a DER is not a conventional generator

A conventional generator is adequately described for dispatch by a fixed interval, $P^{\min} \le P_g \le P^{\max}$, plus ramp limits. The corresponding statement for a DER is a function,

$$P^{\text{avail}} = f\big(\text{device state},\ \text{customer state},\ \text{time},\ \text{history},\ \text{program},\ \text{network}\big), \tag{15.23}$$

which is a much richer object: it varies by the minute, it depends on what the fleet was asked to do yesterday, and it can be refused by its owner. Chapter 4 made this point for the aggregate, where the feasible band of the fleet moved with state and prior dispatch; Table 15.6 makes it device by device.

Table 15.6: A conventional generator and a DER as dispatchable objects.

Property	Conventional generator	DER
Feasible set	fixed interval $[P^{\min}, P^{\max}]$	$\mathcal{U}(x_t, \theta_t^{\text{cust}}, \theta^{\text{dev}}, \theta^{\text{prog}})$, changing every step
Owner's objective	sell energy	comfort, transportation, backup, bill
Who may refuse	nobody	the customer, during the event
Depletion	fuel, on a slow scale	state of charge, comfort band and event budget, within hours
Cost of output	fuel and maintenance	compensation, degradation and customer burden
Availability	known; derated for forced outages	random; forecast from behavior and connectivity
Population	tens of units	10^4 to 10^6 units

15.6.4 The complete dispatch problem

Collecting the chapter gives the dispatch problem in the form the rest of the book uses:

$$\begin{aligned} \min_{u} \quad & J_{\text{grid}} + \lambda_1 J_{\text{comf}} + \lambda_2 J_{\text{deg}} + \lambda_3 J_{\text{burden}} \\ \text{subject to} \quad & x_{t+1} = f(x_t, u_t, w_t), \\ & u_t \in \mathcal{U}^{\text{phys}} \cap \mathcal{U}^{\text{cust}} \cap \mathcal{U}^{\text{prog}}, \\ & \text{network constraints at every location.} \end{aligned} \tag{15.24}$$

The four terms in the objective are the grid's tracking or cost objective, the comfort penalty of equation (15.2), the degradation cost of equation (15.12), and the fairness cost of equation (15.15). The weights λ_1 to λ_3 convert customer impact, equipment wear and burden into the units of grid value, and choosing them is a program-design decision rather than a mathematical one. The hard constraints are in the set; the preferences are in the objective, which is the separation of Section 15.1.3 applied at scale.

15.6.5 Worked example: nominal and usable capacity

A utility requests 20 MW for two hours from a portfolio of thermostats, electric vehicles and batteries. The nominal figures, taken from device ratings and enrollment counts, are 15 MW of thermostats, 10 MW of vehicles and 10 MW of batteries, a total of 35 MW, which appears to leave a comfortable margin. Applying the chapter's constraints class by class:

- **thermostats**: comfort bands and the override behavior of Figure 15.3 leave 10 of the 15 MW;
- **vehicles**: departure deadlines leave 5 of the 10 MW, since cars with U near one cannot be delayed and cars already unplugged cannot be reached;
- **batteries**: owner reserves, by equation (15.7), leave 6 of the 10 MW.

The usable total is $10 + 5 + 6 = 21$ MW against a 20 MW request, a margin of 1 MW or 4.8 %, as drawn in Figure 15.7. The horizontal axis groups the three classes and the total; the gray bars are nominal capacity, the green bars usable capacity, and the dashed line is the request. The request is barely feasible, and the comfortable margin suggested by the nominal figures was an illusion created by reading device ratings as capacity.

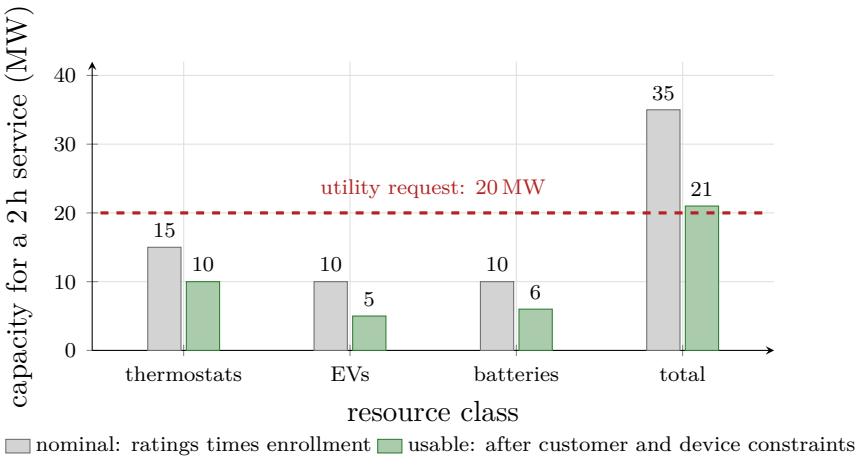


Figure 15.7: Nominal and usable capacity by class for the example of Section 15.6.5. Nominal capacity totals 35 MW; comfort bands, departure deadlines and backup reserves reduce it to 21 MW against a 20 MW request. Values are illustrative but typical in magnitude.

Principle 15.4. *Nominal capacity is not usable capacity. The reduction is not a safety factor to be applied at the end; it is the result of evaluating the customer, device, program and network constraints on the current state.*

15.7 What to retain

Definition 15.1 (Usable DER flexibility)**.** The flexibility of a DER is not owned by the grid operator. It exists inside the customer's and the device's constraints, and the usable part is the intersection $\mathcal{U}^{\text{phys}} \cap \mathcal{U}^{\text{cust}} \cap \mathcal{U}^{\text{prog}}$, further intersected with the network's limits. The task of a virtual power plant is not to maximize control but to extract useful grid flexibility while preserving the customer's primary service.

DER orchestration is therefore physics and optimization *plus* behavior, preferences, reliability and program design. The behavioral part is not a correction applied to a physical model; it has its own states, its own dynamics and its own forecasts. Table 15.7 lists the chapter's quantities with the subsection that defines each.

Table 15.7: The quantities and objects that describe a constrained DER.

Quantity	Definition	Section
$\mathcal{U}^{\text{use}}$	$\mathcal{U}^{\text{phys}} \cap \mathcal{U}^{\text{cust}} \cap \mathcal{U}^{\text{prog}}$	15.1.2
hard vs. soft limit	in the constraint set vs. in the objective	15.1.3
c_{comf}	$\lambda(T_{\text{in}} - T_{\text{set}})^2$	15.2.1
$p^{\text{ovr}}(\Delta T)$	override probability; $\mathbb{E}[\Delta T^{\text{del}}] = \Delta T(1 - p^{\text{ovr}})$	15.2.3
N_i^{recent}	recent events per customer; the fatigue state	15.2.4
SOC^{res}	owner reserve; $E^{\text{avail}} = (\text{SOC} - \text{SOC}^{\text{res}})E^{\max}$	15.2.6
$S_{\max}$	apparent-power rating; $P^2 + Q^2 \le S_{\max}^2$	15.3.1
$t_{\text{on}}^{\min}$, $t_{\text{off}}^{\min}$	minimum dwell times against short cycling	15.3.2
$P^{\max}(\text{SOC}, T)$	state-dependent charging limit (taper)	15.3.3
c_{deg}	$\lambda_{\text{deg}}\, g(\text{SOC}, \text{DoD}, P, T)$	15.3.4
net event value	gross value minus compensation minus degradation	15.4.1
N^{rem}	events remaining in the season; a stored quantity	15.4.3
c_i	$\lambda_i N_i^{\text{recent}}$, the fairness cost	15.5.3
N^{avail}	number of responding devices; variance by equation (15.18)	15.5.4
r_i, P_i^{eff}	learned reliability and $r_i P_i$	15.5.6
$\mathcal{F}_{\text{VPP}}$	$\mathcal{F}_{\text{phys}} \cap \mathcal{F}_{\text{cust}} \cap \mathcal{F}_{\text{prog}} \cap \mathcal{F}_{\text{net}}$	15.6.2

15.8 Exercises

Exercise 15.1. A household has the comfort penalty $c_{\text{comf}} = \lambda(T_{\text{in}} - 22)^2$ with λ equal to \$0.40 per squared degree per hour, inside a program band $21 \le T_{\text{in}} \le 25$°C. An event is worth \$1.10 per °C of setback per hour. What setback maximizes net value, and is it inside the band? Repeat for an event worth \$4.00 per °C per hour and state which constraint binds.

Exercise 15.2. Using the logistic override model of equation (15.5), compute p^{ovr} and the expected delivered setback at $\Delta T = 2$, 3 and 4°C. Which request delivers more, 2°C or 4°C? Then shift the midpoint to 4.5°C and recompute the maximizing request; explain in one sentence why it moves.

Exercise 15.3. A home battery has $E^{\max} = 13.5\,\text{kWh}$, SOC $= 0.70$ and a $5\,\text{kW}$ inverter. Compute the usable energy and the deliverable duration with $\text{SOC}^{\text{res}} = 0.20$ and with $\text{SOC}^{\text{res}} = 0.60$. For a fleet of 5,000 such batteries, express both cases as a capacity in MW for a two-hour service.

Exercise 15.4. For the 10,000 thermostats of Section 15.5.4, use equation (15.18) to compute the standard deviation of delivered power and the 5th percentile under $\rho = 0$, $\rho = 0.02$ and $\rho = 0.1$. How much enrolled capacity would be needed at each ρ to make 24 MW firm at the 95 % level?

15.9 Connection to the TERRA framework

The constraints of this chapter are not an appendix to a physical model. Each one is an ordinary element of a graph-based model, and naming which element it is removes most of the apparent difficulty.

- **Customer limits are bounds, not objectives.** The comfort band is a bound on a component state, the departure requirement of equation (15.6) is a bound on an accumulator at one future time, and the owner reserve of equation (15.7) is a floor on the same accumulator. All three are the same kind of object as a thermal or electrical limit, with the single difference that their values are supplied by a person and may change between steps.
- **Soft constraints live in the objective, hard ones in the feasible set.** Equation (15.24) keeps the quadratic comfort penalty, the degradation cost and the fairness cost as terms, and keeps the program band, the apparent-power circle of equation (15.9) and the dwell times of equation (15.10) as constraints. A model that mixes the two either refuses dispatches it should price or prices dispatches it should refuse.
- **Override makes the commanded flow differ from the realized flow.** The command is an input; the delivered power is an edge flow with a random multiplier. Fitting p^{ovr} and the reliability r_i is posterior inference of component parameters from the history of commanded and metered pairs, and the interior maximum of Figure 15.3 is a property of the fitted model rather than an operational rule of thumb.
- **The event budget is an accumulator with a floor.** Table 15.3 is an identity, not an analogy: N^{rem} obeys a conservation law with a decrement per event and a seasonal reset, so the same forward simulation and the same shadow price apply to it as to stored energy. The fatigue state N_i^{recent} is a second such accumulator, with decay instead of a reset.
- **Participation enters as a distribution on the aggregate flow.** Equation (15.18) propagates per-device availability, with its correlation, to the variance of a node's net flow; the firm number in equation (15.19) is a quantile of that distribution, computed jointly rather than assembled from marginals.
- **Usable capacity is a query, not a separate model.** Figure 15.7 is produced by evaluating equation (15.22) on the current states and parameters and asking for the largest sustainable flow; no capacity model exists beside the physical one, which is what keeps the nominal and usable figures consistent with each other.

In summary, a DER is a bounded component whose bounds are partly written by a person, whose delivered flow is a random function of the command, and whose stocks include an event budget and a stock of customer tolerance alongside stored energy. The next chapter turns to the architecture of the platform that has to hold all of this at once.

Chapter 16

DERMS and VPP architecture: how the whole system fits together

16.1 The layered architecture

Every earlier chapter of this book treated one part of the problem: a device, a service, a forecast, an optimization, a verification procedure. This chapter assembles them. The question is

What software and data flow are required to turn thousands of distributed devices into an operational virtual power plant?

Chapter 3 answered a coarser version of the same question with the eight-stage chain of Figure 3.1 and the core loop of Figure 3.10. Those pictures are correct and will be used here; what this chapter adds is the engineering detail that the chain leaves out, namely which object crosses each boundary, how often each stage runs, what happens when a stage fails, and how uncertainty travels from a sensor to a dispatch decision.

Jargon: layer, telemetry, vendor adapter, cadence, fallback, latency

A *layer* is a stage of the pipeline with a defined input, a defined output and no knowledge of how its neighbors are implemented; layers can be replaced independently. *Telemetry* is the stream of measurements a device reports to the platform. A *vendor adapter* is the piece of software that translates one manufacturer's application programming interface (API) into the platform's internal vocabulary, in both directions. *Cadence* is how often a layer runs: every two seconds, every five minutes, once a day. A *fallback* is the simpler model or rule used when the preferred one is unavailable. *Latency* is the delay between a physical event and the moment the platform learns of it.

Key idea

A distributed energy resource management system (DERMS) is a probabilistic closed-loop control system built as a stack of layers. Information flows up the stack and control flows down it; verification closes the loop, and the residual between what was predicted and what was delivered is the signal that improves the next cycle.

16.1.1 The ten layers

Figure 16.1 draws the stack. Each box is a layer; the text to the right of each box is the object that layer hands upward or downward; the vertical black arrows are the ordinary path through the stack; the red arrow is the fast correction from monitoring back to optimization, and the green arrow is the learning path from model updating back to forecasting. Read from the

bottom of the stack, the figure is the eight-stage chain of Figure 3.1 with the data platform, the monitoring loop and the model-updating step made explicit.

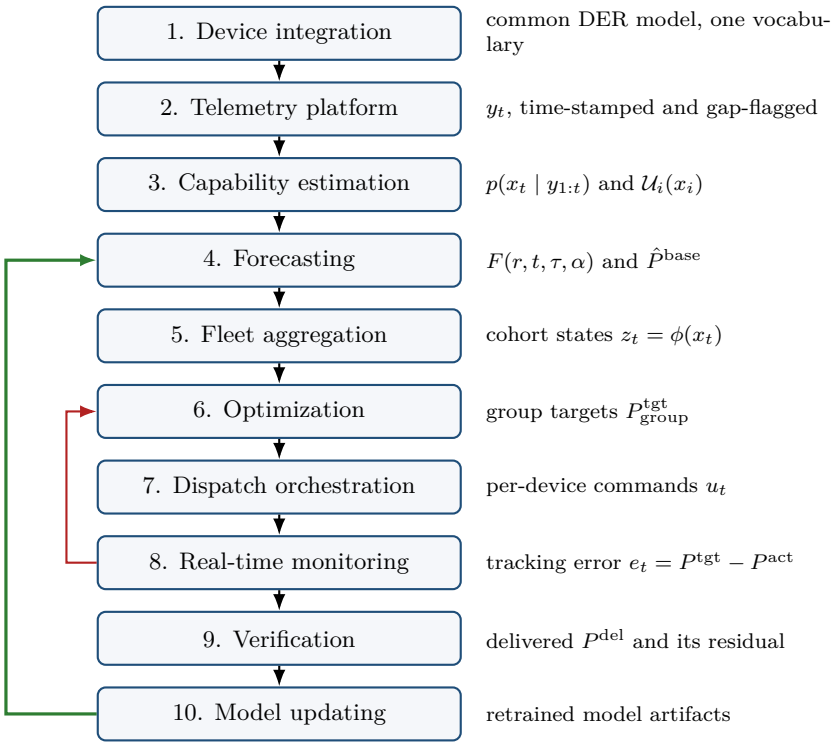


Figure 16.1: The ten-layer DERMS architecture; the physical devices sit below layer 1 and are not part of the platform. Black arrows are the ordinary path down the stack; the red arrow is the fast correction from monitoring back to optimization within an event; the green arrow is the slow learning path from model updating back to forecasting. The object beside each box is what the layer produces.

The layers, with the input each consumes and the output each produces, are as follows.

1. **Device integration.** Input: the application programming interfaces of every device manufacturer in the fleet. Output: a common DER model, in which every battery has a state of charge SOC, every thermostat an indoor temperature T_{in} and a setpoint T_{set}, and every charger a plug-in indicator a_t. This layer is the only place in the system that knows a vendor's field names.

2. **Telemetry platform.** Input: device messages arriving asynchronously, at different rates, with missing and duplicated records. Output: a time-series store of y_t in which every value carries a timestamp, a freshness flag and a quality flag.

3. **Capability estimation.** Input: the telemetry history $y_{1:t}$ and the device models. Output: a belief about the present state, $p(x_t \mid y_{1:t})$, and from it the feasible action set of each device, $\mathcal{U}_i(x_i)$. This is the layer that answers "what can the fleet do right now".

4. **Forecasting.** Input: the state beliefs, weather and calendar data, and the behavioral history of each cohort. Output: the baseline forecast $\hat{P}^{\text{base}}$ and the flexibility forecast $F(r, t, \tau, \alpha)$, with the exogenous drivers the optimizer also needs. This is the layer that answers "what will the fleet be able to do later".

5. **Fleet aggregation.** Input: per-device states and capabilities. Output: a small set of cohort-level quantities $z_t = \phi(x_t)$ – available power, available energy, sustainable duration, uncertainty and location – one tuple per cohort instead of one per device.

6. **Optimization and scheduling.** Input: the cohort states, the forecasts, the grid target $P^{\text{tgt}}(t)$ and the constraint set. Output: group targets over the horizon, not device commands.

7. **Dispatch orchestration.** Input: group targets and the current eligibility of each device. Output: concrete commands, translated by the vendor adapters and sent through the manufacturers' APIs.

8. **Real-time monitoring.** Input: the target trajectory and the metered response. Output: the tracking error $e_t = P^{\text{tgt}} - P^{\text{act}}$ and, when it is nonzero, a request for re-allocation.

9. **Verification.** Input: metered load and a baseline model. Output: the delivered quantity $P^{\text{del}} = P^{\text{base}} - P^{\text{act}}$ and the residual between what was forecast and what was delivered.

10. **Model updating.** Input: the accumulated residuals and the record of command success. Output: retrained model artifacts for layers 3 and 4, after which the loop repeats with better models.

16.1.2 Inputs, outputs and cadence

A layer is defined as much by how often it runs as by what it computes. Table 16.1 lists the ten layers with their input, output and typical cadence. The cadences span four orders of magnitude, from seconds in the monitoring layer to weeks in the model-updating layer, and the spread is the reason the architecture decomposes into layers at all: a component that must answer in two seconds cannot be the same component that retrains a gradient-boosted model on a year of history.

Table 16.1: The ten layers, their inputs, their outputs and their typical cadence. Cadences are representative of production systems rather than fixed by any standard.

Layer	Input	Output	Cadence
1. Device integration	vendor API payloads	common DER model records	on enrollment and on schema change
2. Telemetry platform	device messages	cleaned, flagged y_t	seconds to 15 minutes per device
3. Capability estimation	$y_{1:t}$, device models	$p(x_t \mid y_{1:t})$, $\mathcal{U}_i(x_i)$	every telemetry cycle, 1–5 minutes
4. Forecasting	states, weather, calendar, history	$F(r, t, \tau, \alpha)$, $\hat{P}^{\text{base}}$	5 minutes to hourly, plus a day-ahead run
5. Fleet aggregation	device states and capabilities	cohort tuples z_t	every optimization cycle
6. Optimization	z_t, forecasts, P^{tgt}, constraints	group targets over H steps	1–15 minutes, rolling horizon
7. Dispatch orchestration	group targets, eligibility	device commands u_t	on every target change
8. Real-time monitoring	P^{tgt}, metered P^{act}	e_t, re-allocation requests	2 seconds to 1 minute
9. Verification	metered load, baseline model	P^{del}, forecast residuals	per event, daily, monthly
10. Model updating	residuals, command-success records	retrained model artifacts	days to weeks

16.2 Sensing: device integration, telemetry and state

The first three layers take the system from heterogeneous device interfaces to a statement of what the fleet can do at this instant. Their difficulty is that none of the three delivers what the next layer would like to have.

16.2.1 Device integration

A real fleet mixes thermostats, electric-vehicle chargers, home and commercial batteries, water heaters and commercial building controls. Each manufacturer exposes a different API, different telemetry fields, different command formats and a different update frequency. The same physical quantity arrives under three names: one vendor reports `SOC_pct`, another `battery_level`, a third `energy_remaining` in kilowatt-hours. Table 3.1 of Chapter 3 tabulates that case, and Figure 3.2 there draws the layer: each vendor adapter translates once into the common DER model, and every layer above sees only that model.

Device integration is mostly a software and platform problem rather than a modeling problem, and it is nevertheless where a large fraction of the engineering effort of a real DERMS goes. The architectural reason to isolate it is that the alternative is intolerable: if a vendor's field names reached the optimizer, adding a manufacturer would mean editing every algorithm in the stack.

16.2.2 Telemetry and data quality

With the adapters in place, layer 2 collects telemetry. For a battery the record is $y_t = \{\text{SOC},\ P,\ \text{availability}\}$; for a thermostat, $\{T_{\text{in}},\ T_{\text{set}},\ \text{HVAC state}\}$; for an electric vehicle, $\{\text{SOC},\ \text{plugged in},\ P_{\text{EV}}\}$.

Collecting those records is not "reading the sensor". The layer must handle messages that never arrive, messages that arrive with an old timestamp, devices that report every ten seconds beside devices that report every fifteen minutes, fields that are populated on some firmware versions and null on others, and duplicate records after a communication outage is repaired. The output is therefore not a clean matrix of values but a matrix of values with per-element quality annotations, and every layer above has to be able to act on a partially annotated input.

Principle 16.1. *Telemetry is a data-quality problem before it is a measurement problem. A layer that assumes complete, fresh, uniformly sampled input will produce confident answers from stale and missing data.*

16.2.3 Observations are not the state

Even perfect telemetry is not the state. In general

$$y_t \neq x_t, \tag{16.1}$$

because the measurement map of equation (3.2) in Chapter 3, $y_t = h(x_t) + v_t$, selects only part of the state. An electric-vehicle API reports the state of charge and the plug-in flag; it does not report the driver's departure time, which is the variable that determines how long charging can be deferred. A thermostat reports the indoor temperature; it does not report the building's thermal mass C, its envelope conductance $1/R$, the occupancy or the internal heat gains Q_{int}, all of which determine how far the temperature will drift during an event. Figure 3.3 of Chapter 3 draws this structure: the state advances by f, each observation is a noisy function h of the state, and an estimator runs the chain backward.

The architectural consequence is a layer. A system that wires telemetry straight into forecasting has no place to put the inference that closes the gap between y and x, and will treat the unobserved part of the state as though it did not exist.

16.2.4 Capability estimation

Layer 3 answers one question: what is the fleet capable of doing right now? Formally it maps each device's state to that device's feasible action set,

$$x_i \longmapsto \mathcal{U}_i(x_i), \tag{16.2}$$

which is the state-dependent constraint set introduced in Chapter 1 and developed for batteries in Chapter 6. A home battery with a 5 kW inverter at SOC = 0.80 has

$$-5\,\text{kW} \le P_{\text{bat}} \le 5\,\text{kW}$$

available in both directions, with the discharge direction limited by energy rather than power only near the end of a long service. The same battery at SOC = 0.10 has $P^{\text{dis}} \approx 0$: the inverter is unchanged, the nameplate is unchanged, and the usable discharge capability is gone. Nothing in the telemetry record said so; the mapping (16.2) said so.

This layer is what converts telemetry into current feasible flexibility, and it is the first layer whose output a utility would recognize as an answer.

16.2.5 Capability and forecast are different objects

Two quantities are easily conflated. The current capability F_t is what the fleet can do now, and it is the product of state estimation on present telemetry. The forecast capability $\hat{F}_{t+\tau}$ is what the fleet will be able to do at a lead time τ, and it is the product of forward prediction under uncertain weather, uncertain customer behavior and uncertain device availability.

The first is an inference about the present and is bounded in accuracy by sensor coverage; the second is a prediction about the future and is bounded in accuracy by the predictability of people and weather. They use different inputs, have different error structures, are validated differently, and belong in different layers. A system that computes F_t and calls it a forecast will commit capacity it cannot hold, because the states that produce F_t will have moved by the time the service is called.

16.3 Forecasting and aggregation layers

Layers 4 and 5 are where the forecasting role becomes central and where the fleet stops being a list of devices. This section states what the forecasting layer must emit for the optimizer to be able to use it, and why aggregation sits between the two.

16.3.1 What the forecasting layer produces

Layer 4 produces, at minimum, a baseline forecast and a flexibility forecast per resource class:

$$\hat{P}^{\text{base}}, \qquad \hat{F}_{\text{thermostat}}, \qquad \hat{F}_{\text{EV}}, \qquad \hat{F}_{\text{battery}}.$$

Depending on the optimization problem it also supplies the exogenous drivers: the outdoor temperature forecast $\hat{T}_{\text{out}}$, the utility's system load forecast $\hat{P}_{\text{sys}}$, and the market price forecast $\hat{c}_t$. Each of these is a separate prediction problem with its own data, its own horizon and its own error structure, and Table 3.2 of Chapter 3 lists the inputs of the flexibility forecasts by device type.

16.3.2 Decision-ready forecast output

The optimizer usually needs more than a single number. A forecast delivered as $\hat{F} = 100\,\text{MW}$ cannot answer the questions the optimizer actually asks: how much can be promised with 95 % confidence, how much is available on a particular feeder, and how much of it can be sustained for four hours rather than fifteen minutes. The interface the optimizer needs is therefore a function of four arguments,

$$F(r, t, \tau, \alpha), \tag{16.3}$$

with r the location or resource group, t the delivery time, τ the required duration and α the quantile level; equivalently, the layer returns a distribution $p(F)$ per location and duration rather than its mean. This extends the three-argument object $F(r, t, d)$ of equation (3.8) in Chapter 3 with the duration and the quantile, which are exactly the two arguments a commitment decision needs.

Figure 16.2 contrasts the two interfaces. Panel (a) is the scalar interface: one box containing one number feeding the optimizer. Panel (b) is the decision-ready interface, drawn as a grid whose rows are feeders and whose columns are required durations; each cell contains three bars, the low, median and high quantiles of the flexibility available on that feeder for that duration. The scalar of panel (a) is one of the twenty-seven numbers in panel (b), summed over rows and with the quantile information discarded. An accurate model behind the interface of panel (a) is still an unusable model.

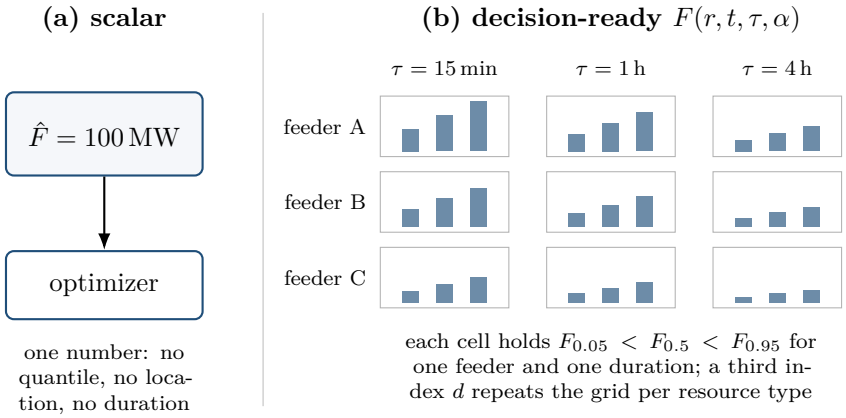


Figure 16.2: The forecast interface. (a) A scalar forecast. (b) The decision-ready object of equation (16.3): rows are feeders, columns are required durations, and the three bars in each cell are the 5th, 50th and 95th percentiles of available flexibility. Bar heights are illustrative.

16.3.3 The aggregation layer

Layer 5 turns N device-level resources into a small number of fleet representations. A fleet of 100 000 thermostats becomes perhaps 20 cohorts, grouped by region, building type, program and recent event history, and the cohorts become regional flexibility curves. The output is the reduced representation

$$z_t = \phi(x_t), \tag{16.4}$$

whose components are the available power, the available energy, the sustainable duration, the uncertainty and the location of each cohort. Equation (4.14) of Chapter 4 is the same map; what matters architecturally is that the uncertainty is a component of z_t and not an afterthought, because the variance of a cohort sum is not the sum of the device variances. Figure 3.4 of Chapter 3 quantifies that: a pairwise correlation of only $\rho = 0.1$ raises the fleet standard deviation tenfold above the independent case, so a cohort tuple that reports a mean

without a correlation-aware spread is wrong by an order of magnitude in exactly the quantity a commitment depends on.

16.3.4 Why aggregation precedes optimization

Optimizing one million individual control variables every few minutes is expensive, and it is also unnecessary. The hierarchy is

device states → cohort states → portfolio optimization → cohort targets → device commands,

so the expensive combinatorial decision is made over twenty cohorts and the cheap allocation decision is made over the devices inside each cohort. Figure 3.9 of Chapter 3 contrasts the flat and hierarchical architectures, and Section 3.6.2 there gives the scaling argument.

Two properties make the reduction safe rather than merely convenient. First, the cohorts are chosen so that devices within a cohort are interchangeable for the purpose of the decision, which is what allows the lower level to pick any devices that sum to the target. Second, the reduction keeps the quantities the upper level needs – power, energy, duration, uncertainty, location – so the portfolio optimizer is not solving a different problem from the one the devices face. When either property fails, the hierarchy produces targets that the lower level cannot meet, and the failure appears at layer 8 as a persistent tracking error.

16.4 Decision, actuation and monitoring

Layers 6 to 8 choose what to do, make it happen on real hardware, and check the result while the service is still running. The distinction that organizes them is that an optimizer does not emit device commands and a command is not a response.

16.4.1 The optimization layer

Layer 6 receives the device states x_t or, more usually, the cohort states z_t; the forecasts $\hat{F}_{t:t+H}$ over the horizon; the grid target $P^{\mathrm{tgt}}(t)$; and the constraint sets $\mathcal{U}$. It solves

$$\min_{u_{t:t+H}} \; J \qquad \text{subject to} \qquad x_{t+1} = f(x_t, u_t, w_t), \quad u_t \in \mathcal{U}(x_t), \tag{16.5}$$

with the objective decomposed as in equation (3.10) of Chapter 3 into tracking, customer, cost and risk terms. This is the receding-horizon structure of Figure 3.7 there: the plan covers H steps, only its first step is applied, and the plan is recomputed at the next cycle from a new state estimate.

16.4.2 What the optimizer outputs

The optimizer does not output 100 000 device commands. It outputs group targets, for example

$$P^{\mathrm{tgt}}_{\mathrm{A}} = 20\,\mathrm{MW}, \qquad P^{\mathrm{tgt}}_{\mathrm{B}} = 15\,\mathrm{MW}, \qquad P^{\mathrm{tgt}}_{\mathrm{C}} = 10\,\mathrm{MW}, \tag{16.6}$$

for three groups that together must deliver 45 MW. Which particular devices inside group A provide its 20 MW is decided one layer below, by logic that knows which devices are currently eligible, which were used in the last event, and which are closest to a customer limit. The separation matters because the two decisions have different cadences and different information: the optimizer runs every few minutes on forecast quantities, while the allocation runs on the latest telemetry and can change its mind about individual devices without re-solving the portfolio problem.

16.4.3 Dispatch orchestration

Layer 7 converts group targets into commands. For the thermostat cohort the command may be a setpoint offset $\Delta T_{\text{set}} = +2°\text{C}$; for an electric-vehicle group, a charging limit $P_{\text{EV}}^{\max} = 3\,\text{kW}$; for the batteries, $P^{\text{dis}} = 4\,\text{kW}$. The commands then travel outward through the vendor adapters of layer 1 into the manufacturers' APIs.

Figure 16.3 draws the control path as five boxes in series. Reading left to right: the optimizer emits a group target; the fleet-target stage fixes how much each cohort owes; the dispatch allocation chooses the individual devices and their shares; the vendor adapter encodes the request in the manufacturer's format; and the device command arrives at the hardware, where a local controller executes it. The dashed arrow returning along the bottom is the telemetry path of layers 1 and 2, which is the only evidence the platform ever gets that anything happened.

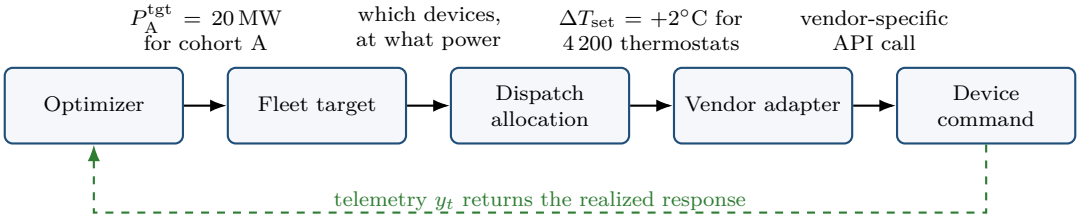


Figure 16.3: The control path from an optimized group target to a device command, with the object that crosses each boundary. Telemetry, the dashed return path, is the only evidence of what the command achieved. Device counts and setpoints are illustrative.

16.4.4 Command acknowledgment

A command has four distinguishable fates: created by the dispatch layer, sent to the vendor, received by the device, and acted on by the device. Each step loses some of the population, so in general

$$u^{\text{requested}} \neq u^{\text{realized}}, \tag{16.7}$$

which is equation (3.11) of Chapter 3 restated as an operational accounting problem. Figure 16.4 shows the ladder as four horizontal bars whose lengths are the number of commands surviving each step, for an illustrative dispatch to 20 000 thermostats. The bars shorten from 20 000 created to 17 400 acted on: 40 commands fail at the vendor interface, 860 never reach a device that is offline or unreachable, and 1 700 reach a device whose local controller refuses or whose customer has overridden it. The fleet response is therefore 87 % of what was commanded, and a platform that tracks only the "created" count has no way to know it.

16.4.5 Real-time monitoring

Layer 8 watches the actual fleet power against the target and computes

$$e_t = P^{\text{tgt}}(t) - P^{\text{act}}(t), \tag{16.8}$$

the tracking error of equation (3.12) in Chapter 3. When $e_t \neq 0$ it has three remedies: dispatch additional devices from the same cohort, change the allocation among cohorts, or use the batteries to absorb the residual error because they are the fastest and most predictable resource in the fleet. This is ordinary feedback control, and Figure 3.6 of Chapter 3 draws the loop. What makes it architecture rather than control theory is the cadence: the measure-compare-correct cycle has to be short compared with the duration of the service, which is why layer 8 runs on a timer of seconds while layer 6 runs on a timer of minutes.

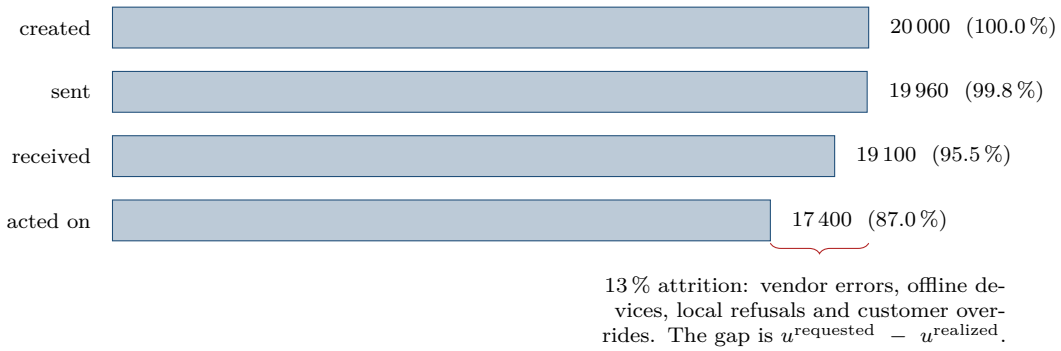


Figure 16.4: The command acknowledgment ladder for an illustrative dispatch to 20 000 thermostats. Bar length is the number of commands surviving each step. Only the last bar corresponds to a physical change in load.

16.4.6 Three loops at three cadences

The whole architecture runs as three nested loops. Figure 16.5 draws them as three concentric regions. The innermost region is the fast loop: measure the response and correct the allocation, on a cadence of seconds to minutes, using layers 8 and 7. The middle region is the slow loop: update the forecast and re-optimize the schedule, on a cadence of minutes to hours, using layers 4 to 6. The outermost region is the learning loop: verify what was delivered and retrain the models, on a cadence of days to weeks, using layers 9 and 10. Each outer loop changes the premises of the loop inside it, and each inner loop runs many times per cycle of the loop outside it.

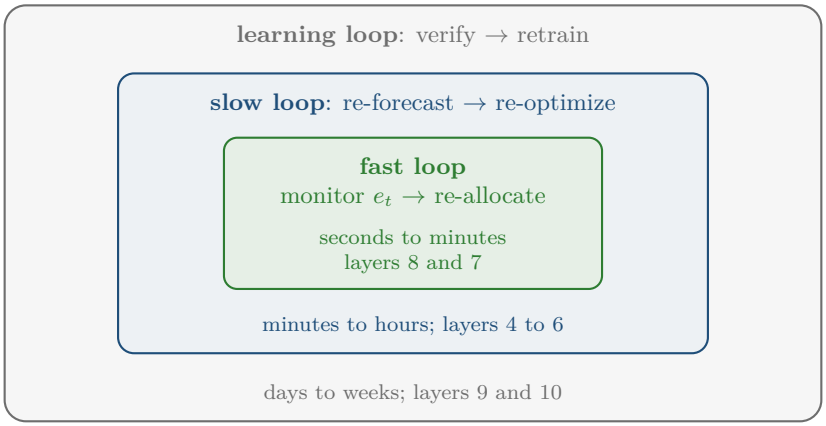


Figure 16.5: The three nested control loops of the architecture, with their cadences and the layers each uses. An outer loop changes the models and schedules that the loop inside it executes against.

16.5 Verification, learning and cadence

Layers 9 and 10 are what make the architecture improve rather than merely operate. This section covers the verification layer, the residual it produces, the learning loop it feeds, and the three planning horizons over which the whole stack runs.

16.5.1 The verification layer

Layer 9 estimates what the fleet actually delivered,

$$P^{\mathrm{del}} = P^{\mathrm{base}} - P^{\mathrm{act}}, \tag{16.9}$$

and uses it to decide whether the committed service was provided. The subtraction looks like metering, but only P^{act} is metered: the baseline P^{base} is the consumption that would have occurred without the dispatch, which never occurred and therefore has to be estimated. Measurement and verification is consequently a counterfactual estimation problem, treated at length in Chapter 8 of this book, and its error enters every performance number the platform reports.

16.5.2 Residuals as training data

Suppose the forecasting layer predicted 50 MW of reduction and verification establishes 42 MW. The 8 MW residual is not merely a performance report; it is a labeled training example. Accumulated over many events it teaches the models specific, actionable facts: that hot days produce higher override rates, so p^{ovr} should depend on T_{out}; that one device group has a systematically lower command success rate, so its contribution should be derated; that a cohort dispatched on three consecutive days delivers less on the third. Each of those is a term that layer 4 can learn and layer 6 can then exploit.

> **Principle 16.2.** *Performance data is future forecast data. Every verified event is a measurement of the fleet's response model, and an architecture that does not return verification output to the forecasting layer discards its most relevant training set.*

16.5.3 The architecture as a learning loop

The most useful one-line picture of the whole system is therefore

$$\text{observe} \to \text{estimate} \to \text{forecast} \to \text{optimize} \to \text{act} \to \text{verify} \to \text{learn}, \tag{16.10}$$

after which the loop returns to observation. Compared with the chain of equation (3.1) in Chapter 3, the single addition is the last arrow, and it is the arrow that distinguishes a platform that gets better from a platform that merely runs. The three loops of Figure 16.5 are the three speeds at which equation (16.10) is traversed.

16.5.4 Day-ahead, hour-ahead and real-time

The same ten layers run at three planning horizons, with different inputs refreshed and different decisions taken. Table 16.2 sets them side by side. Day-ahead, the platform forecasts tomorrow's availability and schedules planned dispatch, which is also when a market offer is made. Hour-ahead, it updates the weather and the availability estimate and revises the schedule. In real time, telemetry and the realized response correct the dispatch within the commitment already made. The architecture does not change between the three; the cadence, the freshness of the inputs and the reversibility of the decision do.

16.5.5 Information up, control down

The two directions through the stack carry different objects. Figure 16.6 shows four levels – device, group, fleet, grid objective – as four horizontal boxes, with the upward path labeled on the left and the downward path on the right. Upward, a device's telemetry becomes a state

Table 16.2: The same architecture at three planning horizons.

Horizon	Inputs refreshed	Decision taken	Cadence
Day-ahead	weather forecast, calendar, enrollment	planned dispatch schedule and market offer	once or twice a day
Hour-ahead	updated weather, current states, availability	revised schedule, cohort selection	every 15–60 minutes
Real-time	telemetry, realized response, e_t	re-allocation inside the committed target	seconds to minutes

and capability estimate, which becomes a cohort aggregate, which becomes the fleet capability $F(r, t, \tau, \alpha)$ offered to the grid. Downward, a grid objective becomes a fleet target, which becomes a group target, which becomes a device command. The upward direction is inference and reduction; the downward direction is decision and allocation. Each upward step loses detail deliberately and must carry uncertainty in exchange; each downward step adds detail and must respect the constraints discovered on the way up.

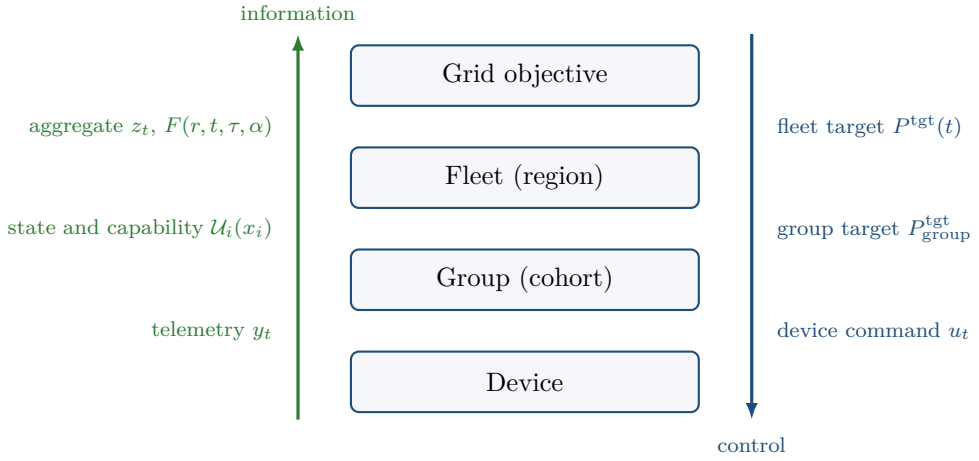


Figure 16.6: Information flows up the hierarchy and control flows down it. The labels beside each arrow are the objects that cross the corresponding boundary.

16.6 Interfaces, robustness and uncertainty

The last body section covers what surrounds the stack and what happens when parts of it fail: the external systems it exchanges data with, where the forecasting role sits inside it, how models are trained and served, what the system does when a model or a telemetry feed is unavailable, and how uncertainty travels from a sensor to a dispatch decision.

16.6.1 External systems

The DERMS is not a closed system. It receives weather data (temperature, humidity, solar irradiance), the utility's own system load forecast $\hat{P}_{\text{sys}}$, market signals π_t, network constraints from the advanced distribution management system (ADMS), and grid-service requests P^{tgt}; and it sends back schedules, dispatch records and performance reports.

Figure 16.7 places these around the stack. The dashed frame encloses the platform; inside it, the five internal boxes are the pipeline of Section 16.1.1 compressed into telemetry and state estimation, forecasting, aggregation, optimization and dispatch, with verification below. The

gold boxes along the top are the external systems, each drawn above the internal layer it feeds: the utility load forecast and the weather service feed forecasting; market prices, the ADMS constraints and the grid-service request feed optimization; and the platform's own schedules and performance reports leave from dispatch. Below the frame are the physical devices, which receive commands along the lower path, return telemetry along the left path, and provide the metered load that verification consumes. The green arrow from verification back to forecasting is the model-update path of layer 10.

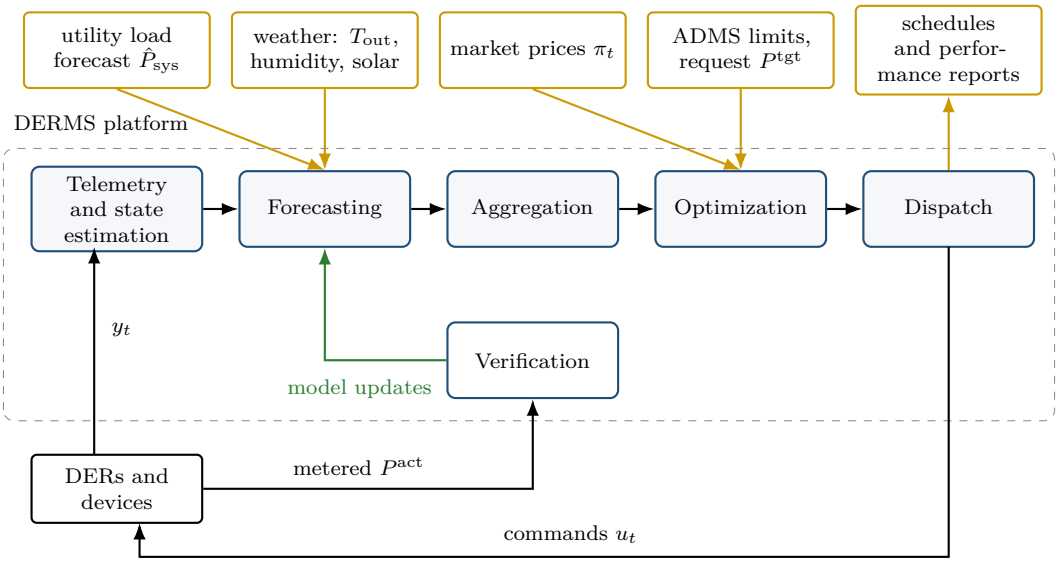


Figure 16.7: The platform and its neighbors. Gold boxes are external systems, placed above the internal layer each one feeds; the dashed frame is the platform boundary; the devices below exchange commands and telemetry with it. The green arrow is the model-update path from verification to forecasting.

Figure 3.8 of Chapter 3 places the same platform in the utility's organizational hierarchy, beneath the ADMS and above the devices; Figure 16.7 is the data-flow view of the same relationship.

16.6.2 Where the forecasting role sits, and why interfaces matter

A forecasting engineer working on this system does not sit inside the single box labeled "forecasting". Table 16.3 lists six distinct objects such an engineer owns and the layer that consumes each. The role spans the center of the architecture, from capability estimation to verification, because every layer in that span consumes a probabilistic statement about physical behavior.

The second row is where interface mismatch usually appears. A model may be accurate and still unusable. If the forecast returns $F = 100\,\text{MW}$ but the optimizer needs $F_{0.05}$, $F_{0.5}$ and $F_{0.95}$, the model has an interface problem rather than an accuracy problem. If the optimizer needs flexibility per feeder, F_r, because the ADMS constraint is local, and the forecast is system-wide, then the model is insufficient no matter how small its system-wide error is. Product requirements – which arguments the output is indexed by, and what distributional information it carries – matter as much as the choice of algorithm.

16.6.3 Online and offline models

Models in this architecture live in two regimes. An *online* model runs inside the operating loop and must answer within one cadence: the real-time flexibility estimate of layer 3 is online. An

Table 16.3: Six places the forecasting role touches the architecture.

Object	Symbol	Consumed by
Baseline forecast	$\hat{P}_t^{\text{base}}$	verification (layer 9) and the tracking term of J
Flexibility forecast	$F(r, t, \tau, \alpha)$	optimization and commitment (layer 6)
Customer-behavior model	p^{ovr}	capability estimation and forecasting (layers 3–4)
Device-availability model	$\Pr(a_t = 1)$	aggregation and derating (layer 5)
Decision coupling	$u^\star$ from $\mathbb{E}[J]$	optimization: the forecast enters the objective, not a dashboard
Verification model	residual $P^{\text{del}} - \hat{P}$	verification and model updating (layers 9–10)

offline model is trained on history outside the loop: the customer-response model that supplies p^{ovr} is offline.

Figure 16.8 draws the two lanes. The upper lane is offline: historical telemetry and verification records feed a training pipeline, which emits a model artifact. The artifact is the object that crosses between the lanes. The lower lane is online: live telemetry enters inference against that artifact, which produces the flexibility estimate, which produces the dispatch decision. The green path along the bottom returns new verification data to the training set, which is the learning loop of Section 16.5.3 drawn in terms of software artifacts rather than layers. The separation matters operationally: the online lane must be fast and must never fail, while the offline lane may be slow and may fail without stopping operations.

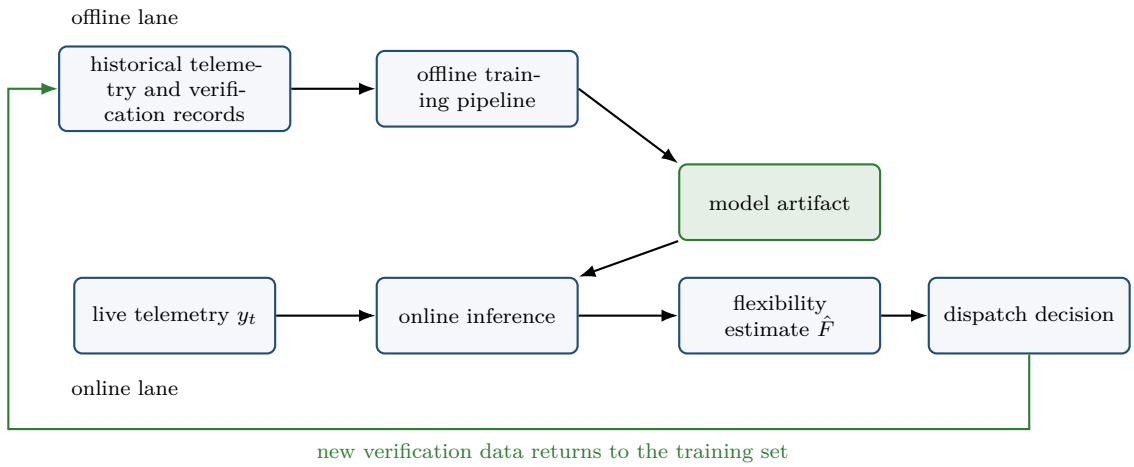


Figure 16.8: Offline training and online inference. The model artifact is the only object that crosses from the training lane to the operating lane; verification data crosses back along the green path.

16.6.4 Fallbacks and confidence-dependent dispatch

If the machine-learning forecast service is unavailable, the platform still has a committed service to deliver. Robustness is therefore designed as a chain of tiers, from the preferred model to a rule that cannot fail:

$$\text{primary model} \rightarrow \text{fallback model} \rightarrow \text{safe operating rule}.$$

Table 16.4 gives the three tiers, what each emits and what is lost by falling to it. The pattern is graceful degradation: each tier delivers a usable answer in a poorer currency, and the stack above is written to accept the poorer currency. A platform whose optimizer requires full quantiles has no fallback tier, because the second and third tiers cannot produce them.

Table 16.4: The fallback chain of the forecasting layer.

Tier	Model	Output	What is lost
Primary	probabilistic machine-learning flexibility model	$F(r, t, \tau, \alpha)$, full quantiles by location and duration	nothing
Fallback	persistence or recent-history average	point forecast with a fixed assumed spread	quantile shape and most location resolution
Safe rule	conservative default capacity per cohort	one derated number per cohort	most of the usable capacity, but not the service

Model confidence also changes the dispatch directly. Suppose $\mathbb{E}[F] = 100\,\text{MW}$. If the uncertainty is large, the operator derates and commits 70 MW; if it is small, 95 MW can be committed from the same expected value. The committed quantity is a quantile, not a mean, exactly as in the chance constraint $\Pr(P^{\text{del}} \geq P^{\text{commit}}) \geq 1 - \epsilon$ of Chapter 4. This is the architectural reason probabilistic forecasting is useful rather than merely tidy: the confidence has to travel with the prediction, because a layer downstream multiplies by it.

16.6.5 Latency and missing telemetry

Two data problems have larger practical consequences than most modeling choices. The first is latency. If thermostat telemetry arrives 15 minutes late, then what the platform calls y_t is a measurement of x_{t-15}:

$$y_t = h(x_{t-15}) + v_t, \tag{16.11}$$

and every state estimate built from it is 15 minutes stale. For a building with a thermal time constant of two hours, 15 minutes of drift is a material fraction of the comfort band, so the capability estimate of Section 16.2.4 is biased in a direction that depends on the weather. Handling this means carrying the timestamp and the freshness of every input, and propagating the state forward from the measurement time to the decision time rather than treating the latest record as current.

The second is missing telemetry. If 10 % of devices stop reporting, their capability has not necessarily vanished; what has vanished is the evidence for it. Three policies are available, and Table 16.5 states what each assumes and what it trades. Declaring the silent devices unavailable understates the fleet and forgoes capacity that exists. Inferring their state from the last known value and the model keeps the capacity but is wrong precisely when the cause of the silence also changes the state, such as a local power outage. Applying a conservative derating factor κ to the silent population is the tunable middle, and the choice of κ is an operational risk decision rather than a modeling one.

16.6.6 Uncertainty propagation through the architecture

Uncertainty accumulates along the pipeline. Measurement uncertainty becomes state uncertainty, state uncertainty becomes forecast uncertainty, and forecast uncertainty becomes dispatch uncertainty. The desirable architecture propagates it; the common one loses it at every module boundary, because each module publishes a point estimate and the next module has no way to know how good it was.

Table 16.5: Three policies for a silent fraction of the fleet.

Policy	Assumption	Utilization	Risk
Declare unavailable	the device contributes nothing	lowest	lowest; capacity is understated and may be bought elsewhere
Infer the state	last state propagated forward by the model	highest	highest; wrong when the cause of the silence also moved the state
Conservative derate	the device contributes a fraction $\kappa < 1$	intermediate, set by κ	tunable; κ is chosen from the observed reporting-gap statistics

Figure 16.9 shows both cases for an illustrative fleet whose expected capability is 100 MW. The horizontal axis is the pipeline stage and the vertical axis is capability in megawatts. The solid black line is the expected capability, which is the same at every stage. The shaded blue band is the propagated $\pm 2\sigma$ interval: it is ± 2 MW at the measurement stage, ± 6 MW after state estimation, ± 14 MW after forecasting and ± 18 MW after the dispatch decision. The dashed red lines are the band a system would declare if each module passed only a point estimate with the measurement-stage spread attached. The marked point is a realized delivery of 83 MW: inside the propagated band, where it is an expected tail outcome, and far outside the declared band, where it looks like a model failure. The two interpretations lead to opposite operational responses, and only the first is correct.

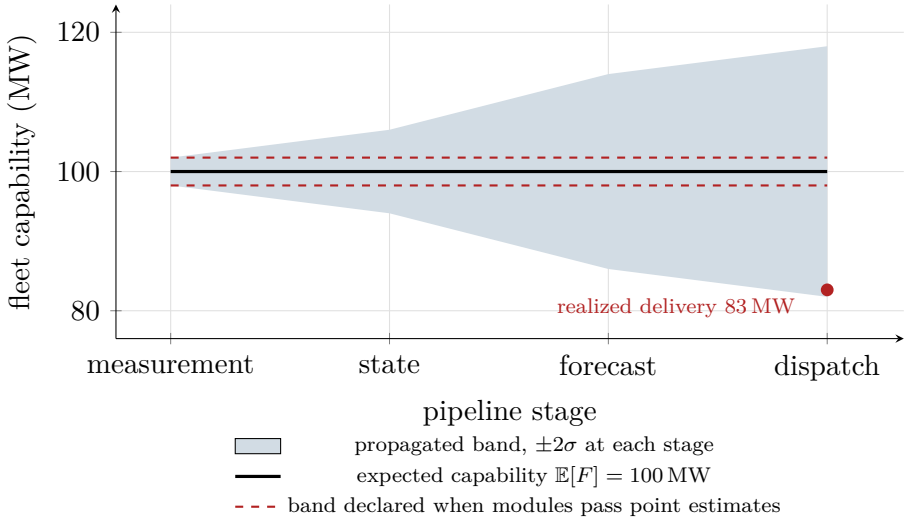


Figure 16.9: Uncertainty along the pipeline, illustrative. The shaded band is the propagated $\pm 2\sigma$ interval, widening from ± 2 MW at measurement to ± 18 MW at dispatch; the dashed lines are the band a system declares when each module passes a point estimate with the measurement spread attached. The realized outcome is a tail event under the first description and an anomaly under the second.

16.6.7 The full mathematical view

The whole architecture is one probabilistic closed-loop control system, and it can be written in six lines:

$$\begin{aligned} y_t &= h(x_t) + v_t && \text{measurements arrive,} \\ x_t &\sim p(x_t \mid y_{1:t}) && \text{the state is inferred,} \\ (x_{t+1:t+H},\, F_{t:t+H}) &\sim p(\,\cdot \mid y_{1:t}) && \text{the future is forecast,} \\ u^{\star}_{t:t+H} &= \arg\min_{u \in \mathcal{U}} \mathbb{E}\big[J(x,u)\big] && \text{a decision is made,} \\ x_{t+1} &= f(x_t, u_t, w_t) && \text{the physical system responds,} \\ y_{t+1} &= h(x_{t+1}) + v_{t+1} && \text{and the loop repeats.} \end{aligned} \tag{16.12}$$

Line one is layers 1 and 2; line two is layer 3; line three is layers 4 and 5; line four is layers 6 and 7; line five is the physical fleet, which is not a layer at all; and line six returns to layer 2. Layers 8 to 10 are the statement that the realized y_{t+1} is compared with the prediction and that the comparison updates h, f and the forecast distributions. Every layer of Figure 16.1 appears in equation (16.12), and every term of equation (16.12) is implemented by some layer.

16.7 What to retain

Definition 16.1 (DERMS architecture)**.** The architecture of a DERMS is the chain telemetry → state → forecast → aggregate → optimize → dispatch → verify, closed by verify → learn → better forecast. It is a probabilistic closed-loop control system in which information flows up a hierarchy and control flows down it, run at three cadences simultaneously.

The forecasting and optimization role lives at the point where uncertain physical capability is transformed into an operational decision, which spans state estimation, forecasting, optimization and control rather than any one of them. Table 16.6 lists the ten layers with the section of this chapter that defines each and the chapter of this book that treats it in depth.

Table 16.6: The ten layers, where each is defined in this chapter, and where it is treated in depth.

Layer	What it produces	Section	Chapter
1. Device integration	common DER model	16.2.1	3
2. Telemetry platform	flagged telemetry y_t	16.2.2	3
3. Capability estimation	$p(x_t \mid y_{1:t})$, $\mathcal{U}_i(x_i)$	16.2.4	3, 15
4. Forecasting	$F(r,t,\tau,\alpha)$, $\hat{P}^{\text{base}}$	16.3.2	10
5. Fleet aggregation	cohort states $z_t = \phi(x_t)$	16.3.3	9, 13
6. Optimization	group targets over H steps	16.4.2	11, 13
7. Dispatch orchestration	device commands u_t	16.4.3	11
8. Real-time monitoring	tracking error e_t	16.4.5	12
9. Verification	delivered P^{del} and residuals	16.5.1	8
10. Model updating	retrained model artifacts	16.5.2	8, 10

16.8 Exercises

Exercise 16.1. A home battery has 13.5 kWh usable, a 5 kW inverter, an owner reserve of $\mathrm{SOC}^{\mathrm{res}} = 0.20$ and a measured $\mathrm{SOC} = 0.80$. Write $\mathcal{U}_i(x_i)$ for a service that must be sustained for two hours, stating both the power bound and the energy bound and which one binds. Repeat at $\mathrm{SOC} = 0.10$ and say what the capability-estimation layer reports in each case.

Exercise 16.2. A forecasting layer currently returns $\hat{F} = 100$ MW with a standard deviation of 12 MW. The optimizer needs the 5th percentile per feeder for durations of 15 minutes, 1 hour and 4 hours, on three feeders. How many numbers must the interface of equation (16.3) return if three quantiles are reported per cell? State which of the two failure modes of Section 16.6.2 the scalar interface exhibits.

Exercise 16.3. Using the acknowledgment ladder of Figure 16.4, suppose each thermostat that acts delivers 0.6 kW of reduction. Compute the realized fleet reduction, the reduction the dispatch layer would have predicted from the "created" count, and the derating factor the optimizer should apply to future thermostat commands if this ladder is representative.

Exercise 16.4. Thermostat telemetry arrives 15 minutes late, as in equation (16.11), and the indoor temperature is drifting upward at 1.2°C per hour during an event. By how much is the reported T_{in} in error at the moment of the decision, and how much of a 2°C comfort band does that consume? Then state which policy of Table 16.5 you would apply to a cohort whose telemetry has stopped entirely, and why.

16.9 Connection to the TERRA framework

The ten layers of this chapter are, layer for layer, the workflow of a graph-based physical model with inference and control on top of it.

- **Layers 1 to 3 build the graph and condition it.** The common DER model is the declaration of components, their state variables and the edges they attach to; the capability map $\mathcal{U}_i(x_i)$ of equation (16.2) is the set of bounds on those edges given the component states; and $p(x_t \mid y_{1:t})$ is the posterior over the unmeasured states and parameters, inferred from whatever measurements the vendors happen to expose.

- **Layer 4 is forward integration from the posterior.** The object $F(r, t, \tau, \alpha)$ of equation (16.3) is a predicted edge flow at a future time with its distribution retained, obtained by integrating the transition map forward from the current belief with the disturbance w drawn from its own forecast. The quantile index α exists because the integration produces a distribution, not a number.

- **Layer 5 is model reduction that must carry the covariance.** The map $z_t = \phi(x_t)$ of equation (16.4) replaces a high-dimensional state by a few aggregate quantities per cohort. A framework that holds the joint posterior, including the shared weather that correlates devices, reduces correctly; one that sums marginals reports the cohort spread too small by the factor of Figure 3.4.

- **Layers 6 and 7 are decisions on the graph, decomposed.** Equation (16.5) uses the graph as its prediction model and the posterior as its initial condition. The group targets of equation (16.6) are interface flows between subgraphs: the portfolio problem is solved on the coarse graph, and each cohort then solves its own allocation subject to the flow it owes at the interface.

- **Layers 8 to 10 are conditioning on new data.** The tracking error e_t and the verification residual $P^{\text{del}} - \hat{P}$ are residuals between prediction and measurement at two different timescales. Feeding them back is updating the posterior over states in the fast loop and over model parameters in the slow one, so that h, f and the disturbance distributions all improve with operating experience.

- **The uncertainty band of Figure 16.9 is the difference between one model and five.** A single graph with one joint posterior propagates measurement uncertainty into dispatch uncertainty automatically, because every stage is a transformation of the same distribution. Five independent modules exchanging point estimates cannot, which is why the declared band in that figure is both narrower and wrong.

In summary, the architecture of a DERMS is a conservation graph of physical components, a posterior over its unmeasured states, a forward prediction from that posterior, a decision taken against the prediction, and a conditioning step that returns the realized outcome to the model. The ten layers are the engineering form of those five operations, and the three cadences of Figure 16.5 are how often each is performed. This chapter closes the arc of the book: the lesson sequence continues with one aggregator's commercial platform, which is outside the scope of these chapters.

Chapter 17

A commercial aggregation platform: how one platform maps onto the architecture

17.1 The platform and its value chain

Every preceding chapter of this book developed one part of an aggregation problem, and Chapter 16 assembled those parts into a ten-layer architecture. That architecture was written without reference to any particular company. This chapter takes one commercial aggregation platform and asks how much of the architecture appears in it, in what order, and what the platform contains that the architecture does not predict. The question of the chapter is

Which parts of the generic DERMS architecture does a commercial aggregation platform actually implement, and what does it add?

The facts in this chapter are as stated in the public product material of one large residential DER aggregator at the time of writing, 2026; the company is not named here, and none of its claims has been independently verified. Every number, date and capability below is therefore introduced as something the aggregator states, reports or describes, with the year attached, and none of it is a finding of this book. Where a stated capability is matched against a formulation from an earlier chapter, the formulation is the book's and the capability is the company's; the correspondence is an interpretation, not a claim about the implementation, which is proprietary.

Jargon: Edge DERMS, device ecosystem, enrollment

An *Edge DERMS* is a DER management system operated by an aggregator on devices that the customer owns, at the grid edge, as opposed to a system that a utility runs on its own equipment. A *device ecosystem* is the set of manufacturers and service providers whose products the platform can read and command, each through its own application programming interface. *Enrollment* is the customer-program step of signing a customer, obtaining consent to control a device, and establishing connectivity to it; a fleet's capacity is bounded by it before any physics applies.

Key idea

The ten-layer architecture of Chapter 16 is close to the published description of a working platform, with one stage the architecture omits: enrollment. A virtual power plant is a control problem downstream of a customer-acquisition problem, and the second sets the size of the first.

17.1.1 The value chain

The aggregator's product language, as stated on its public pages in 2026, centers on building and managing virtual power plants, optimizing load shape, operating across DER types, and using near-real-time information to manage performance. Read as a sequence of operations, that language describes the chain

enroll DERs, observe DER behavior, forecast load and flexibility, group resources, optimize dispatch, control DERs, measure performance.

Figure 17.1 draws the seven stages as a row of boxes. Each box is one stage of the chain; the label beneath each box is the chapter of this book that treats that stage in detail. The figure is not a reconstruction of the company's software; it is the published sequence of operations with the book's chapters attached to it.

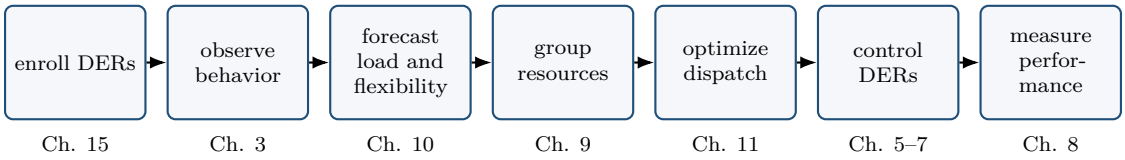


Figure 17.1: The seven-stage value chain implied by the aggregator's own product language as published in 2026. Beneath each stage is the chapter of this book that develops it. The first stage, enrollment, has no counterpart in the ten-layer architecture of Figure 16.1.

Compared with Figure 16.1, the chain is the same pipeline with two differences. The architecture's layers 1 and 2, device integration and the telemetry platform, are compressed into the single stage "observe DER behavior"; and a stage appears at the front that the architecture does not have at all. The generic architecture began with devices already connected. A commercial platform begins earlier, with a customer who has not yet agreed to anything.

17.1.2 Device integration as a line of business

The aggregator does not manufacture distributed energy resources. According to its product material in 2026 it integrates with device manufacturers and service providers across thermostats, electric vehicles and electric-vehicle supply equipment, batteries, solar inverters, water heaters and other DER classes, and it states that the platform manages more than 2.5 million DERs for more than 160 utility clients. The integration work is therefore not a detail of the engineering; it is part of what the company sells.

Figure 17.2 shows the structure. The upper part of the figure has a column of vendor classes on the left, one box per class; a single box in the middle labeled "vendor adapters"; and the platform on the right. Every arrow enters the platform through the adapter box, which is the only part of the system that knows a manufacturer's field names, as in layer 1 of Section 16.1.1. The lower part of the figure is the enrollment front end discussed in the next subsection. What the upper part asserts is that the value of the platform lies partly in abstracting away heterogeneous hardware: downstream of the adapter box, a thermostat is a comfort band and an indoor temperature regardless of who built it, which is precisely the common DER model of Chapter 3.

17.1.3 Enrollment as part of the platform problem

A virtual power plant cannot be operated without customers enrolled in it. The aggregator's material emphasizes customer acquisition, enrollment, ongoing engagement and retention as platform functions rather than as marketing activities, and in 2026 it reports more than 1.4

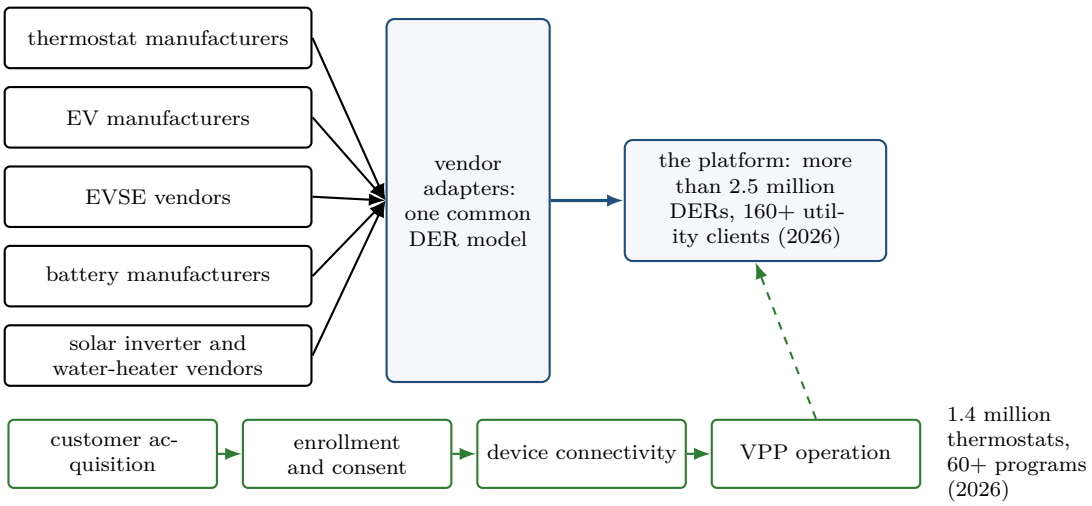


Figure 17.2: Upper part: the device ecosystem as described on the aggregator's public pages in 2026. Vendor classes on the left reach the platform only through adapters that translate each manufacturer's interface into one common DER model. Lower part: the enrollment front end, which precedes every stage of Figure 17.1. The counts are as stated by the company in 2026.

million enrolled thermostats across more than 60 thermostat programs. The front end is therefore a chain of its own, drawn in the lower part of Figure 17.2: customer acquisition, enrollment and consent, device connectivity, and only then virtual-power-plant operation. The dashed arrow in that figure carries the enrolled, connected devices into the platform.

The consequence for the engineering is quantitative. Chapter 4 defined the enrolled nominal flexibility P^{enroll} and the committed capacity P^{commit}, and the coverage ratio $P^{\text{enroll}}/P^{\text{commit}}$ as the overbooking margin. Enrollment sets the numerator. A forecasting improvement raises the fraction of P^{enroll} that can be committed with confidence; an enrollment campaign raises P^{enroll} itself. The two are substitutes at the margin, and a platform that owns both can trade one against the other. That trade has no representation anywhere in Chapter 16, which treated the fleet as given.

Enrollment also introduces attrition. Devices leave the fleet when customers move, replace equipment, or opt out of a program, so P^{enroll} is not a constant but a stock with an inflow and an outflow. The customer constraints of Chapter 15 govern the outflow: a program that calls too many events, or holds devices too long during each one, loses enrollment and therefore capacity. This is the mechanism by which a comfort constraint becomes a capacity constraint on a timescale of months.

17.2 Telemetry and DER-specific forecasting

This section covers what the platform observes once a device is enrolled and connected, and the forecasting capability the company states it built on top of that observation.

17.2.1 Grid-edge telemetry

Once a device is enrolled and connected, the platform receives information from the grid edge that a utility would often not otherwise possess. The aggregator describes its DER load forecasting as using device state, device power consumption, weather, and real-time customer behavior to predict continuously the load from grid-edge devices. In the notation of this book,

the telemetry y_t of Chapter 16 feeds a forecast of the controllable load,

$$\hat{P}_{\mathrm{DER}}(t+\tau) = f\big(y_{1:t},\ \hat{w}_{t+1:t+\tau},\ \mathcal{D}\big), \tag{17.1}$$

where $\hat{w}$ is the forecast of the exogenous drivers, chiefly weather, and $\mathcal{D}$ is the behavioral history of the enrolled population. The distinguishing feature of equation (17.1) is not its functional form but its inputs: a utility forecasting system has the metered feeder load and the weather, and does not have y_t at device level.

17.2.2 System load and controllable-device load

The aggregator states that it began building its own DER load forecasting capability in 2018, because conventional utility forecasting did not adequately predict the shape, magnitude and timing of the load activity of connected devices. The distinction is between two different target variables. A traditional utility forecast predicts system or substation load: the sum of everything behind a meter, dominated by thousands of uncontrollable loads, and consequently smooth. A DER forecast predicts the load specifically associated with controllable grid-edge devices: a much smaller quantity, concentrated in short intervals, and driven by the switching of individual appliances.

Figure 17.3 contrasts the two. The upper panel plots an illustrative substation load against the hour of the day; the lower panel plots, on its own scale, the load of the controllable devices inside that same substation. The solid curve in the lower panel is the device load; the dashed curve in the lower panel is what would be obtained by scaling the system curve of the upper panel to the device total, which is the best a system-load forecast can do about device load. The device curve has three narrow peaks, at the morning ramp, at midday and in the early evening; the scaled system curve has one broad hump. The shapes differ in width, the peak heights differ by a factor of about two and a half, and the timing of the morning peak is absent from the scaled curve entirely. Those three discrepancies are the shape, magnitude and timing that the company states conventional forecasting did not resolve.

The reason the two differ is developed in Chapter 10. A system load is an average over a very large, diverse population, and averaging suppresses the switching of individual devices; a device load is the sum over a small, deliberately homogeneous subpopulation whose members respond to the same price, the same weather and the same commands, so their switching does not average away. Forecasting the quantity one can influence rather than the quantity one happens to measure is the practical content of the distinction, and it is the same argument that separated flexibility forecasting from load forecasting in Chapter 10.

17.2.3 Fifteen-minute re-issue

The aggregator states that these forecasts are updated using current device state, current power consumption and external feeds such as weather, with forecast information updated in real time and represented at 15-minute intervals. A forecast that is re-issued every quarter hour is a different object from a forecast that is produced once a day: it is a sequence of overlapping predictions,

$$\hat{P}_{t+1:t+H} = f\big(P^{\mathrm{act}}_{t},\ x_t,\ \hat{w}_{t+1:t+H},\ \mathcal{D}\big), \tag{17.2}$$

re-evaluated at every t with the newest measurement P^{act}_{t} and the newest state estimate x_t substituted in. This is the receding horizon of Chapter 11 applied to the forecast rather than to the plan.

Figure 17.4 shows the effect. The horizontal axis is time in minutes; the gray curve is the device load as it is realized; the three shaded bands are forecasts issued fifteen minutes apart, each starting at the realized value at its issue time and widening with the horizon. The bands are narrow where they begin and wide at the far end, because uncertainty grows with lead time;

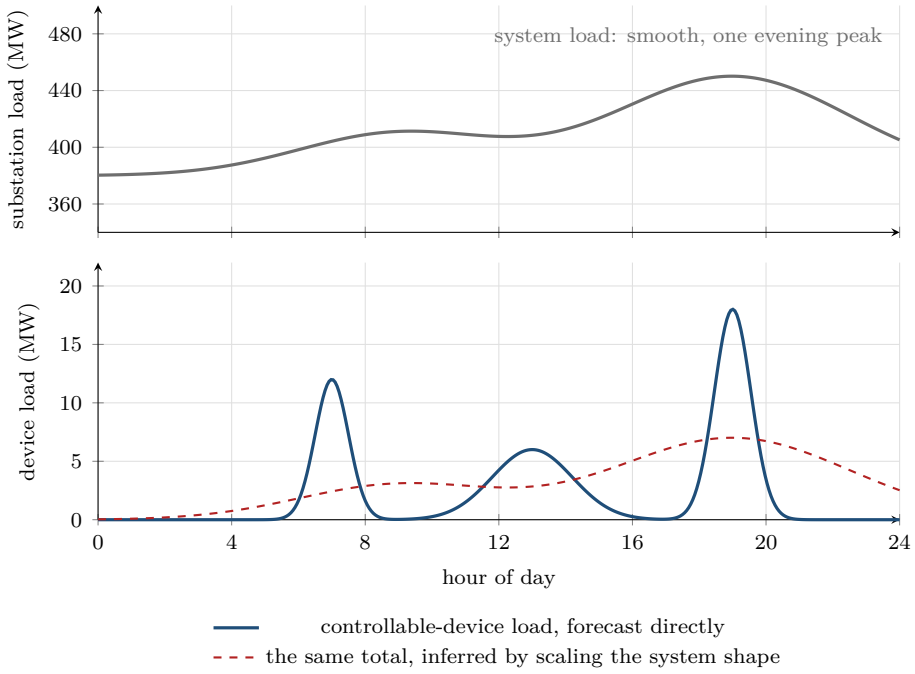


Figure 17.3: Why a system-load forecast does not resolve device load. Upper panel: an illustrative substation load, $380 + 70\exp[-(h-19)^2/(2\cdot 3.5^2)] + 30\exp[-(h-9)^2/(2\cdot 3^2)]$ in MW. Lower panel: an illustrative controllable-device load, $18\exp[-(h-19)^2/(2\cdot 0.55^2)] + 12\exp[-(h-7)^2/(2\cdot 0.5^2)] + 6\exp[-(h-13)^2/(2\cdot 1.2^2)]$, and the dashed curve obtained by scaling the system shape by 0.10. All values are illustrative; the distinction between the two target variables is the one the aggregator states it addressed from 2018 onward.

successive bands start from a better-known present and therefore sit closer to what eventually happens. The practical consequence is that the optimizer never consumes a single forecast. It consumes the most recent one, and the difference between consecutive issues is itself information: a forecast that moves by 3 MW between two issues tells the dispatch layer that its last plan was built on a premise that has already changed.

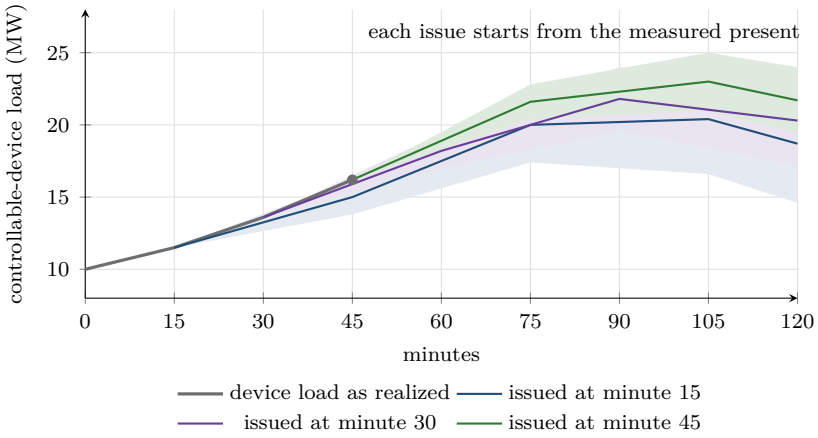


Figure 17.4: A forecast re-issued every fifteen minutes. The gray curve is the realized device load; each shaded band is one issue of the forecast, anchored at the realized value and widening with the horizon. Values are illustrative; the 15-minute representation and the real-time update are as described by the aggregator in 2026.

17.3 From demand response to dynamic load shaping

The platform's control capability, as published, spans two generations: event-based demand response and continuous load shaping. This section follows that progression, including the named algorithm the company uses for it.

Jargon: snapback, packetized energy management

Snapback is an operator's term for the post-event rebound of Chapter 2: the load above baseline that appears when curtailed devices recover the state they gave up. *Packetized energy management* is a coordination technique in which each device asks a coordinator for permission to consume a short "packet" of energy, and the coordinator grants or defers the request; the aggregate is shaped by the grant rate rather than by a command sent to each device.

17.3.1 The demand-response pipeline

The aggregator's demand-response product, as described on its product pages in 2026, runs a four-step pipeline: forecast the potential load shed, develop an event strategy, dispatch the event, and observe near-real-time results. Those four steps are Chapters 10, 11, 2 and 8 respectively, and the chain of Figure 8.4 is the accounting that closes the fourth step: a request, a forecast, a dispatch and a verified delivery are four different numbers, and a demand-response product has to report all four.

Event-based demand response remains part of the platform, but the published material no longer treats it as the end state. The forecast step is the one that constrains the rest: an event strategy is a choice among feasible trajectories, and the set of feasible trajectories is whatever the flexibility forecast says it is. The progression described in the next subsection is, in

these terms, a progression from choosing one trajectory per event to choosing a new trajectory continuously.

17.3.2 Dynamic load shaping

The aggregator describes a capability it calls Dynamic Load Shaping, which it states uses machine learning together with optimization to coordinate DER dispatch automatically against grid objectives, and which it states can coordinate multiple DER types rather than treating each class independently. Decomposed, the description contains exactly three of this book's components: a forecast, an optimization, and a control action, run as a loop rather than as a one-off event. That loop is Figure 11.8, and the trajectory-tracking formulation it implies is the subject of Chapter 12.

Two features of the description are worth separating. The first is automation: the dispatch is chosen by an optimizer against a stated objective rather than by an operator selecting an event. The second is cross-resource coordination: thermostats, vehicles and batteries are dispatched against one objective, which is the portfolio problem of Chapter 12 and in particular the schedule of Figure 12.2, where no resource's contribution is constant in time and the sum is what matters. A platform that dispatches each class separately cannot produce that schedule, because the thermostat column and the battery column have to be chosen jointly to hold the total flat.

17.3.3 The staggered cohort test

The aggregator states that in its dynamic-load-shaping tests it dispatched staggered thermostat groups to maintain a consistent load reduction over three hours while limiting the control duration of any individual device to two hours, and that this reduced the immediate post-event snapback. The three numbers in that statement correspond exactly to the three mechanisms of Chapter 12.

The first is cohort rotation. A single thermostat group cannot hold a reduction for three hours, because its comfort band is consumed; Figure 5.11 shows four groups rotated through half-hour curtailments so that each group's decaying contribution overlaps the others and the stacked total stays near the target. The second is the per-device duration limit. Capping any one device at two hours is a customer constraint of the kind catalogued in Chapter 15, and it is also what forces the rotation: if a device could be held for the full three hours, one group would suffice. The third is the reduced snapback. Releasing cohorts at different times spreads the recovery, which is the stepped release of Figure 12.8, where ending a 100 MW event in five increments rather than at once lowers the rebound peak from about 45 MW to about 18 MW.

The staggered-cohort exercise constructed in Chapter 12 is therefore close to a problem the company states it has run in the field. The constraint structure is the interesting part: three hours of aggregate service built from devices that each tolerate two hours of control implies a lower bound on the number of cohorts, and that bound is a function of the thermal time constant of the buildings rather than of anything in the software.

17.3.4 Firm multi-hour dispatch

The aggregator describes a capability for firm multi-hour dispatch. The stated motivation is that traditional thermostat demand response produces a load reduction that declines over the course of an event; the stated mechanism is that devices are organized into groups and given optimized dispatch schedules so that the fleet produces an approximately flat multi-hour load shape.

Figure 17.5 draws the contrast it is built on. The horizontal axis is hours since the start of a three-hour service and the vertical axis is the fleet's load reduction in megawatts. The declining curve is a single group dispatched at once: it begins at 100 MW and decays, passing 45 MW

after two hours. The nearly flat curve is the same fleet organized into groups with staggered schedules, holding about 80 MW for the whole three hours; the dashed line is the 80 MW target. The single-group dispatch overshoots the target for the first thirty-four minutes and is below it for the remaining two hours and twenty-six minutes. Both curves are drawn from the fleet's own capability, so the flat one is not more resource; it is the same resource spent at a constant rate instead of all at once.



Figure 17.5: The idea behind Firm Load Dispatch. The declining curve is $100\,e^{-t/2.505}$ MW, chosen so that it passes through 100 MW at the start and 45 MW after two hours; the flat curve is $80 + 2.2\sin(540t)$ MW, an approximately constant 80 MW held for three hours. Values are illustrative and the decay is a fitted exponential, not a measurement; the motivation and the grouping mechanism are as described by the aggregator in 2026.

Firm Load Dispatch is thus Chapter 12 stated in operational terms. The shaping problem there was to track a target trajectory with a fleet whose per-device capability decays; the named product is the special case in which the target is flat and the devices are thermostats. The word "firm" is the point: a flat, multi-hour reduction that can be promised in advance is a capacity product in the sense of Chapter 14, whereas a reduction that decays from 100 to 45 MW is an energy product that happens to start large.

17.3.5 Packetized energy in the coordination layer

The aggregator states that packetized-energy-management technology is used to improve DER-management algorithms and to shape the net load of a virtual power plant dynamically. In the architecture of Chapter 16 this belongs in layer 7, dispatch orchestration: it is a way of turning a group target into device-level actions, not a way of computing the group target.

The mechanism differs from a broadcast command in a way that matters for the fleet dynamics. A broadcast setpoint change sends the same instruction to every device at the same instant, which is exactly the synchronization failure of Chapter 5: devices that were cycling independently begin to cycle together, and the aggregate develops the oscillation of Figure 5.9. A request-and-grant scheme has the opposite property. Each device asks at a time determined by its own state, and the coordinator shapes the aggregate by choosing what fraction of requests to grant, so the devices remain desynchronized by construction and the control variable is a scalar grant rate rather than a vector of commands. The company's published position is simply that it has been adding coordination algorithms over time rather than issuing static demand-response commands; the architectural reading is that the actuation layer and the optimization layer evolve separately, which is why Chapter 16 made them different layers.

17.4 Objectives, targets and resource classes

This section maps the optimization objectives the company states onto the formulations of earlier chapters, identifies where the target comes from, and takes the three resource classes in turn.

17.4.1 The four stated objectives

The aggregator's material, as published in 2026, names four optimization objectives: wholesale price optimization, which shifts consumption toward lower-price periods; customer-rate optimization, which shifts charging according to the customer's tariff; sustained load reduction, which dispatches DER groups at a fine enough granularity to maintain a response over longer events; and distribution load management, which shifts load to stay below distribution limits. Table 17.1 lists the four against the formulation each corresponds to and the chapter in which that formulation was derived. Every one of the four was written down earlier in this book before the company's material was consulted, which is the strongest available evidence that the formulations are the standard ones rather than inventions of this book.

Table 17.1: The four optimization objectives the aggregator states, the formulation each corresponds to, and where this book derives it. The objective names and descriptions are the company's, as published in 2026; the formulations are this book's.

Stated objective	Formulation	Where derived
Wholesale price optimization	$\min \sum_t \pi_t P_{\mathrm{VPP}}(t)\,\Delta t$, with π_t the wholesale energy price	Ch. 11, 14
Customer-rate optimization	$\min \sum_t c_t^{\mathrm{cust}} P_{\mathrm{VPP}}(t)\,\Delta t$, with c_t^{cust} the retail tariff the customer faces	Ch. 7, 15
Sustained load reduction	$\min \sum_t \left(P_{\mathrm{VPP}}(t) - P^{\mathrm{tgt}}(t)\right)^2$ over a long event, with per-device duration limits	Ch. 12
Distribution load management	$P_{\mathrm{feeder}}(t) \le P_{\mathrm{feeder}}^{\max}$ for every constrained feeder	Ch. 13

The company also states that the objective is not simply to maximize load shed, but to choose DER schedules that satisfy a particular grid objective, where the objective itself changes according to wholesale cost, distribution constraints, customer rates, peak reduction and sustained capacity. That is a multi-objective problem of the form developed in Chapter 11,

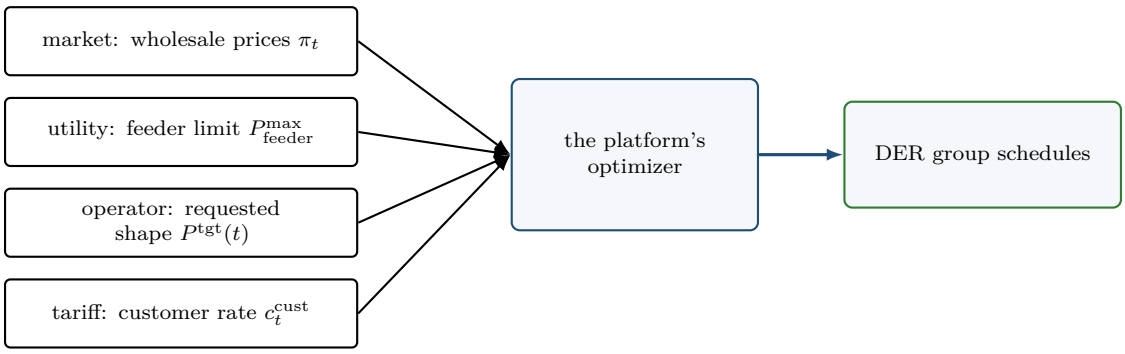
$$J = \sum_k \lambda_k J_k, \tag{17.3}$$

with the weights λ_k reset as conditions change, subject to the feeder constraint of the fourth row of Table 17.1, which is a constraint rather than a term. The engineering content of equation (17.3) is in the weights: four of the five named objectives are priced in dollars and one, sustained capacity, is a reliability requirement, so the weights are not all commensurable and some of them encode a policy decision rather than a market price.

17.4.2 External sources of the target

The objective that the optimizer pursues is in general supplied from outside the platform. A market system supplies the wholesale prices π_t; a utility defines a distribution limit $P_{\mathrm{feeder}}^{\max}$; an operator requests a target trajectory $P^{\mathrm{tgt}}(t)$; a tariff defines c_t^{cust}. Figure 17.6 draws the four sources as boxes on the left, the platform's optimizer as a single box in the middle, and the DER group schedules it emits as a box on the right. Each arrow entering the optimizer carries

one of the four objects just listed. The figure makes the position of the platform explicit: it is the intelligence and orchestration layer between a signal it does not choose and devices it does not own.



The platform chooses neither the objective nor the devices; it chooses the schedules that connect them.

Figure 17.6: Where the grid objective comes from. Four external sources on the left supply the price, the limit, the requested shape and the tariff; the optimizer converts whichever of them is active into group schedules. The position of the platform as an orchestration layer is as described in the aggregator's material in 2026.

One implication is organizational and one is technical. The organizational implication is that the platform's product is a schedule, not a megawatt-hour: it sells the conversion of an objective into device actions. The technical implication is that P^{tgt} in every formulation of Chapters 11 and 12 is an exogenous input that can change during execution, which is why the monitoring layer of Chapter 16 exists and why the horizon is receding rather than fixed.

17.4.3 Electric vehicles

The aggregator markets electric-vehicle managed charging around shifting charging toward lower-cost periods while meeting the driver's needs, and states that it manages relationships with multiple vehicle and supply-equipment manufacturers. The first half of that description is the managed-charging problem of Chapter 7: minimize $\sum_t c_t P_t \Delta t$ subject to the availability gate and the departure requirement $\text{SOC}(T^{\text{dep}}) \geq \text{SOC}^{\text{req}}$. The problem posed there is therefore representative of the product rather than a simplification of it.

The second half, the vendor relationships, is the reason the first half is harder in practice than on paper. Different manufacturers expose different quantities: some report state of charge, some report only whether the vehicle is plugged in; some accept a target finish time, some accept only a start and stop command; some report with a latency of seconds and some of many minutes. The urgency metric U_i of Chapter 7 requires E_i^{rem} and T_i^{rem}, and a fleet in which a third of the vehicles do not report state of charge forces either an estimate of E_i^{rem} from the charging history or a policy that does not use it. That is a modeling decision created by the device ecosystem of Figure 17.2, not by the physics.

17.4.4 Batteries in the portfolio

In 2026 the aggregator and one of its utility customers added a residential battery program to existing thermostat and electric-vehicle virtual-power-plant programs at the same utility. The stated fact is the program; the consequence for the optimization is that the portfolio at that utility now contains all three of the resource classes developed in Chapters 5 to 7.

A heterogeneous portfolio is a more interesting optimization than any of its parts, for the reason set out in Chapter 12: the three classes have different and partly complementary limitations. Thermostats deliver a large reduction immediately that decays as the comfort band is consumed. Vehicles deliver little in the early evening, when many are still driving, and more later as they plug in. Batteries deliver a reduction that is constant in power but strictly limited in energy, and can be held in reserve. Figure 12.2 is the schedule that exploits exactly this complementarity: the thermostat contribution falls, the vehicle contribution rises, the battery contribution is withheld early so that energy remains late, and the sum is held at the target in every hour. Adding batteries to a thermostat and vehicle program is therefore not simply more megawatts; it adds the one resource whose contribution the optimizer can place anywhere inside the event.

17.4.5 Bidirectional vehicles

In July 2026, according to a joint announcement, the aggregator joined Eversource, National Grid, Sunrun and The Mobility House in a residential vehicle-to-grid test in Massachusetts. The modeling consequence is a single change of bounds. The managed-charging model of Chapter 7 restricts the vehicle to

$$0 \le P_{\mathrm{EV}}(t) \le a_t \, P^{\mathrm{ch}}, \tag{17.4}$$

whereas a bidirectional vehicle obeys

$$-\, a_t \, P^{\mathrm{V2G}}_{\max} \le P_{\mathrm{EV}}(t) \le a_t \, P^{\mathrm{ch}}, \tag{17.5}$$

with the departure requirement $\mathrm{SOC}(T^{\mathrm{dep}}) \ge \mathrm{SOC}^{\mathrm{req}}$ unchanged.

Relaxing one bound changes the character of the fleet. Under equation (17.4) the only decision is when to consume energy that will be consumed anyway, so the fleet can reduce net load at most to zero. Under equation (17.5) the fleet can inject, so the problem becomes the distributed-storage optimization of Chapter 6 with a departure deadline attached, and the quantity that limits it is the energy above the driver's requirement rather than the charging that has not yet happened. The operational burden rises with it: degradation, compensation, interconnection and customer acceptance all enter, which is why Chapter 7 observed that most programs begin with managed charging. A pilot is the appropriate instrument for that transition, and the announcement describes a pilot.

17.5 Scale and the platform architecture

Two published aspects of the platform remain: the numbers that describe its size, and the sequence of stages its material supports. This section takes them together, because the first is the reason the second is organized hierarchically.

17.5.1 The stated numbers

Table 17.2 collects the quantities the company has published, each with the year in which it was stated. The 2024 figures and the 2026 figures are not restatements of one another and should not be read as a time series: they are different quantities measured on different bases, and the table reports each as stated rather than reconciling them.

The first row is the one that constrains the engineering. Two thousand megawatts of dispatchable flexibility is comparable to a large conventional generating station, and it is assembled from devices whose individual contribution is between a fraction of a kilowatt and a few kilowatts. A fleet of that composition cannot be optimized device by device at a dispatch cadence of minutes, for the reasons given in Chapter 9: the decision variable count is in the

Table 17.2: Scale as stated by the aggregator, with the year of the statement. Every figure is from the company's own public material; none is independently verified here, and the rows are not a consistent time series.

Quantity	As stated	Year stated
Dispatchable flexibility managed	more than 2 000 MW	2024
Utilities served	more than 70	2024
Largest individual programs	several above 100 000 enrolled devices	2024
DERs under management	more than 2.5 million	2026
Utility clients	more than 160	2026
Enrolled thermostats	more than 1.4 million	2026
Thermostat programs	more than 60	2026
Forecast representation	15-minute intervals, updated in real time	2026

millions, the telemetry is incomplete, and the marginal value of resolving one device is negligible. The response is the hierarchy of Figure 9.4, in which each level summarizes the level below it by a few aggregate quantities and the optimization at every level stays tractable. The published scale is therefore the reason aggregation and clustering are not optional.

17.5.2 The platform architecture

The company's public material supports a particular sequence of stages, and Figure 17.7 draws it. The figure has three parts. The left column is the data path into the platform: customer enrollment, the connected devices it produces, and the telemetry those devices report, with weather and other external feeds joining the same trunk from below. The center column is the model path: forecasts, the fleet and group model built from them, and the optimization that runs against that model, with the grid objective entering from the right as the external input of Section 17.4.2. The bottom row is the action path: group dispatch, the devices acting, and the measurement of what they delivered. The heavy arrow from the measurement box returns around the outside of the figure to the forecasts, which is the update-and-re-optimize step.

Figure 17.7 and Figure 16.1 describe the same operations in two different geometries, and the comparison is informative. The ten-layer stack is drawn as a vertical pipeline because its organizing idea is that information flows up and control flows down through replaceable layers. The platform figure is drawn as a cycle because its organizing idea is that the enrollment path, the model path and the action path are three different kinds of work that meet at the optimizer. Neither drawing contains anything the other lacks; the stack emphasizes interfaces and the cycle emphasizes the loop, and the three cadences of Figure 16.5 are invisible in both. A platform diagram with no times on it is a statement about data flow only.

17.6 The forecasting and optimization role

The final section locates a specific engineering role inside the platform just described, states the problems that role owns, and works two of them far enough to show what they require.

17.6.1 Where the role sits

The role under discussion is a senior forecasting and optimization engineer at such a platform. Its position in the pipeline is the pair of stages in the middle: DER data arrives, a forecast produces the available and expected response, and an optimization converts that response into

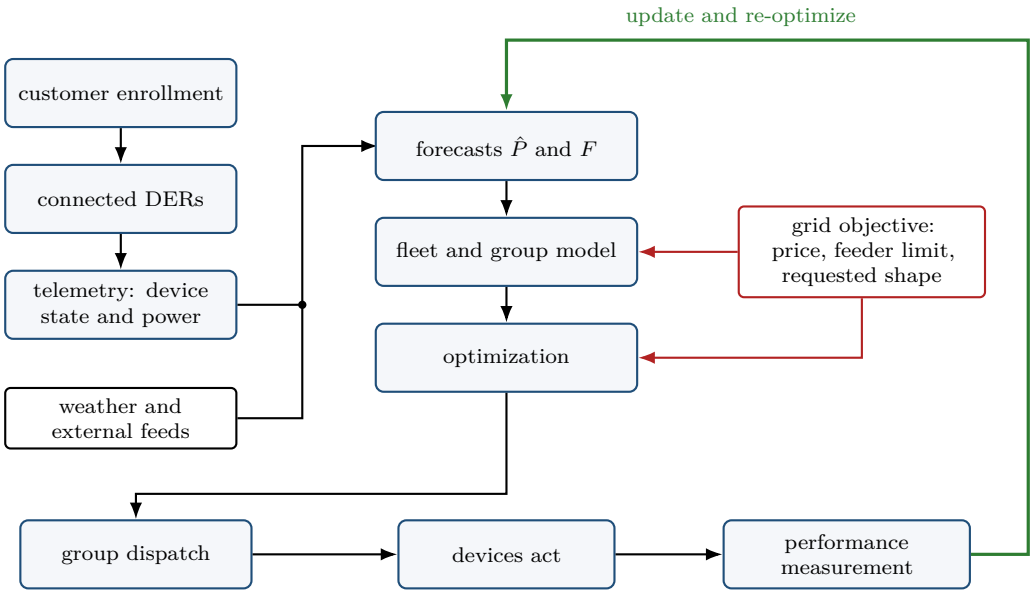


Figure 17.7: The platform architecture as the sequence of stages the aggregator's public material supports: enrollment and telemetry on the left, the forecast, model and optimization in the center with the grid objective entering from the right, and dispatch, action and measurement along the bottom, closed by the update path. Internal implementation details are proprietary and are not represented here.

a dispatch schedule. Figure 17.8 draws the pipeline vertically. The four boxes, from top to bottom, are the telemetry and state of the fleet, the forecast of its expected behavior and capacity, the optimization that produces dispatch schedules, and the dispatch and measurement that follow. The brace on the left marks the two stages that the forecasting track of these lessons covers; the brace on the right marks the two that the aggregation and operations track covers; the dashed box marks the two middle stages, which are the role. The arrow on the far left is the forecast update that closes the loop.

The description that the figure rules out is "a data scientist who predicts something". A prediction role ends when the prediction is accurate. This role ends when the dispatch schedule built on the prediction performs, which makes the relevant error metric a function of the decision rather than of the forecast, as Chapter 10 argued when it defined the forecast target from the optimizer's requirements. The work is closer to physical-system forecasting feeding operational decision-making than to statistical prediction.

17.6.2 The four problem statements

Table 17.3 states the four problems such an engineer faces, one per resource class plus one for the mixture, with the objects each requires and the chapter that supplies them. Read down the "question" column, the four are not four instances of one problem. Each asks for a different quantity: a sustained reduction per cohort, a schedule, a reservation decision, and a composition.

None of the four is a pure forecasting problem and none is a pure optimization problem. The thermostat question asks for a forecast whose target, sustained reduction per cohort, is defined by the optimization that will consume it; the battery question asks for an optimization whose objective depends on a forecast of a future need that has not been requested yet. The coupling is the reason the two stages are one role in Figure 17.8.

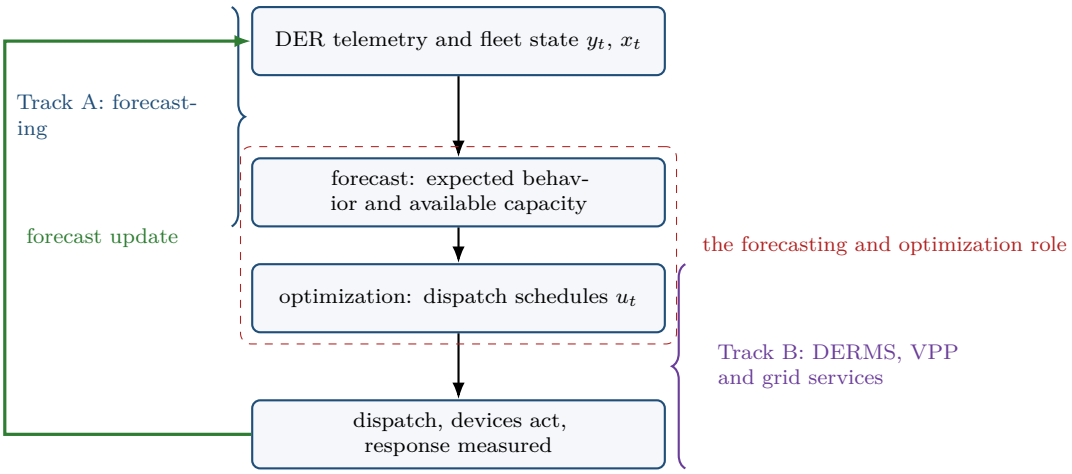


Figure 17.8: Where the two lesson tracks meet. The vertical pipeline runs from fleet state through forecast and optimization to dispatch and measurement; the braces mark the stages each track develops, and the dashed box marks the two stages the role owns. The role is a forecast-to-decision role, not a prediction role.

Table 17.3: The four problems the forecasting and optimization role owns, as set out in the 2026 lesson material, with the objects each requires.

Resource	Question	Objects required	Chapter
Thermostats	Given tomorrow's weather and the current fleet state, how much sustained reduction can each cohort deliver?	T_{in}, $T^{\min}$, $T^{\max}$, RC, cohort assignment, $F(t,\tau)$	5, 10
EVs	Given expected charging demand, departure constraints and prices, how should charging be scheduled?	$a_{i,t}$, SOC_i, T_i^{dep}, $\text{SOC}_i^{\text{req}}$, π_t	7
Batteries	Given the distribution of states of charge and future grid needs, how much energy should be preserved rather than dispatched now?	SOC_i, $\text{SOC}_i^{\text{res}}$, $E_i^{\max}$, P_i^{dis}, future P^{tgt}	6
Mixed fleets	What combination produces the requested shape at minimum customer impact and acceptable risk?	all of the above, plus λ_k and a risk level	12, 11

17.6.3 Moving up the control-sophistication curve

The company's older public material emphasizes demand-response events; its newer material emphasizes automatic optimization, dynamic load shapes, coordination across DER types, and grid objectives applied in near real time. Read as a sequence, the published emphasis moves from scheduling events to continuously optimizing a heterogeneous set of flexible resources. That is the same transition this book made between Chapter 2 and Chapter 12, and it is a change in the mathematical object being chosen: an event is a start time, an end time and a group, while a shape is a trajectory with a feasibility condition at every step.

The transition is the most plausible reason such a role exists. Event scheduling can be operated from a user interface by a program manager; continuous optimization of a heterogeneous fleet against an exogenous objective cannot, because the feasible set has to be recomputed from device states at every cycle. A platform in the middle of that transition needs the forecast of the feasible set to be an engineered artifact rather than a rule of thumb, and that artifact is the interface of Chapter 16.

17.6.4 Domain modeling and feasibility

A purely statistical model that predicts next period's flexibility F_{t+1} from recent history is useful and easy to build. A stronger treatment asks why flexibility changes at all, and the answer is always a state variable: for a thermostat, the position of the indoor temperature inside the comfort band, as in the thermal state of charge $\mathrm{SOC}^{\mathrm{th}}$ of Chapter 5; for a battery, the state of charge relative to the owner's reserve, as in Chapter 6; for a vehicle, the state of charge, the departure time and whether the car is plugged in, as in Chapter 7.

Those state variables determine feasibility, which a statistical model of the aggregate cannot recover. A fleet whose thermostats sit at the top of the comfort band has no downward flexibility whatever the history says, and a regression on recent flexibility will not know it; Figure 1.3 is the shape such a model gets wrong, because the decay it shows is a function of accumulated state and not of elapsed time. The role therefore benefits from four things at once: physical modeling, time-series methods, uncertainty quantification and optimization. A candidate strong in only the second is a candidate who will forecast the wrong variable accurately.

17.6.5 Operational reality

The word "optimization" in the role title does not mean the derivation of new solution methods. The objective is to make a few million real devices satisfy a utility's requirements reliably, and in that setting the following matter at least as much as the solver:

- **noisy telemetry**, so the state entering the forecast is itself an estimate (Chapter 3);
- **stale state**, so a reported indoor temperature may be fifteen minutes old and the device has drifted since (Section 16.2.2);
- **device heterogeneity**, so two thermostats of different make expose different fields and accept different commands (Section 17.1.2);
- **customer opt-outs**, which remove capacity during the event that was counted before it (Chapter 15);
- **forecast error**, which has to be carried as a distribution rather than discarded (Chapter 10);
- **incomplete response**, so the commanded, acknowledged and delivered quantities are three different numbers (Figure 8.4);
- **changing targets**, since P^{tgt} is exogenous and may be revised mid-event (Section 17.4.2);
- **computational scale**, which forces the hierarchy of Figure 9.4 rather than a single large program.

Each item on that list degrades a quantity the optimizer treats as known. The practical consequence is that the useful forecast is not the most accurate one but the one whose stated uncertainty is honest, since a dispatch built on an over-confident forecast fails in the field while a dispatch built on a wide but correct one merely commits less.

17.6.6 Two problems worked briefly

The first problem is the thermostat problem of Table 17.3 with numbers attached. A fleet of 100 000 thermostats is enrolled; the utility wants 80 MW from 17:00 to 20:00; a standard simultaneous dispatch produces about 100 MW initially and falls to about 45 MW after two hours, which is the declining curve of Figure 17.5. The required reduction is below the initial capability and above the two-hour capability, so the question is one of allocation over time rather than of total resource. A workable sequence is: forecast each cohort's response $F_k(t)$ from weather and state; model the thermal depletion and the rebound that follow a curtailment; cluster the fleet into cohorts with similar time constants and similar remaining band; optimize a staggered schedule $u_{k,t}$; track $P^{\text{act}} - P^{\text{tgt}}$ during the event; and correct with feedback. The optimization at step four is

$$\begin{aligned} \min_{u_{k,t}} \ & \sum_t \Big(\sum_{k=1}^{K} F_k(t, u_{k,\cdot}) - 80\,\text{MW} \Big)^2 + \lambda \sum_k \sum_t u_{k,t} \\ \text{subject to} \quad & \sum_t u_{k,t}\,\Delta t \le 2\ \text{h for every cohort } k, \qquad u_{k,t} \in \{0,1\}, \end{aligned} \tag{17.6}$$

where $u_{k,t}$ indicates that cohort k is curtailed at step t, F_k depends on the cohort's own curtailment history through its thermal state, and λ prices customer impact. The duration constraint is what makes the problem combinatorial and the staggering necessary; with K cohorts each limited to two hours of a three-hour service, the cohorts must overlap, and equation (17.6) chooses the overlap. This is a direct application of the dynamic load shaping of Chapter 12 and of Firm Load Dispatch as described in Section 17.3.4.

The second problem is the vehicle problem. A fleet of 40 000 electric vehicles must each reach a required charge by departure, and hourly prices are known. The deterministic form is

$$\begin{aligned} & \min_{P_{i,t}} \ \sum_t \pi_t \sum_{i=1}^{40\,000} P_{i,t}\,\Delta t \\ & \text{subject to} \quad \text{SOC}_{i,t+1} = \text{SOC}_{i,t} + \frac{\eta_i P_{i,t} \Delta t}{E_i}, \qquad 0 \le P_{i,t} \le a_{i,t}\,P_{i,\max}, \qquad \text{SOC}_{i,T_i^{\text{dep}}} \ge \text{SOC}_i^{\text{req}}, \end{aligned} \tag{17.7}$$

a linear program that separates across vehicles when no shared constraint binds, and whose solution is then the urgency ordering of Chapter 7. The problem becomes interesting when $a_{i,t}$, T_i^{dep} and $\text{SOC}_i^{\text{req}}$ are uncertain, because the terminal constraint must then hold with a stated probability, $\Pr(\text{SOC}_{i,T_i^{\text{dep}}} \ge \text{SOC}_i^{\text{req}}) \ge 1 - \epsilon$, and the vehicles couple through any feeder limit of Table 17.1. The first form is a textbook exercise; the second is the problem the platform actually has.

17.6.7 The description to retain, and where the tracks meet

Two common descriptions of the company are too narrow and too vague respectively. "A demand-response company" describes one product of several and the oldest one. "A VPP software company" is accurate but says nothing about what the software does. The description this chapter supports is that the aggregator operates a distributed cyber-physical control platform that turns customer-owned DERs into forecastable and dispatchable grid flexibility, and that the direction of the published development is toward making that aggregate flexibility

behave more like a controllable power-system resource. The two adjectives carry the content: forecastable is Chapter 10, dispatchable is Chapters 11 and 12, and the move from the first to the second is the role of Figure 17.8.

That figure is also the answer to where the two lesson tracks meet. The forecasting track supplies the expected behavior and capacity of the fleet with its uncertainty; the aggregation and operations track supplies the fleet model, the constraints and the grid objective; the role converts the two into a dispatch schedule, and the measured response returns to the forecast. The convergence is not a change of subject at the boundary between the tracks. It is one loop in which the forecast exists to make a decision and the decision generates the data that improves the forecast.

17.7 What to retain

Definition 17.1 (The platform as a loop)**.** The aggregator's platform, as described in its own public material in 2026, is the loop: connect DERs, understand their state, forecast their behavior, aggregate their flexibility, optimize dispatch, control the devices, measure the response, and update. Forecasting is load-bearing in that loop because the optimizer cannot safely decide what to dispatch unless it knows what the fleet is likely to be capable of delivering.

The loop is the architecture of Chapter 16 with one stage added in front of it, enrollment, and one named algorithm inside it, Firm Load Dispatch. Table 17.4 lists each stated capability of the platform against the chapter of this book that explains it and the section of this chapter that discusses it.

17.8 Exercises

Exercise 17.1. Using the declining curve of Figure 17.5, $P(t) = 100\,e^{-t/2.505}$ MW, compute the energy delivered over the three-hour service and compare it with the energy delivered by a flat 80 MW for three hours. Which is larger, and what does the comparison say about whether Firm Load Dispatch creates additional capability or redistributes existing capability? State the one assumption about the fleet that the comparison requires.

Exercise 17.2. A service must last three hours and no individual device may be controlled for more than two hours, as in the test described in Section 17.3.3. Assuming every cohort is identical and delivers a constant reduction while controlled, find the smallest number of equal cohorts and a schedule for them that holds the aggregate constant for the full three hours. Then explain why the real fleet of Figure 5.11 needs more cohorts than your answer.

Exercise 17.3. A utility states a distribution limit $P^{\max}_{\text{feeder}} = 8$ MW and, separately, a wholesale price series with a minimum at 03:00. A fleet of vehicles behind that feeder would charge at 11 MW if all charged at the price minimum. Write the combined problem using the first and fourth rows of Table 17.1, say which of the two is a constraint and which is an objective, and explain what the optimizer does with the 3 MW that cannot be served at 03:00.

Exercise 17.4. Table 17.2 reports more than 2 000 MW of dispatchable flexibility across more than 70 utilities as stated in 2024, and more than 2.5 million DERs as stated in 2026. Explain why dividing one by the other does not give the average flexibility per device, listing the quantities that would be needed to do the calculation correctly, and identify which stage of Figure 17.1 each missing quantity would come from.

Table 17.4: Each platform capability, as stated by the aggregator in 2026, against the chapter of this book that explains it. The capabilities are the company's; the mapping is this chapter's interpretation.

Stated capability	Chapter of this book	Section
Device integration across six DER classes	3, device integration and the common model	17.1.2
Customer acquisition and enrollment	15, customer and program constraints	17.1.3
Grid-edge telemetry from device state and power	3, sensing the fleet	17.2.1
DER load forecasting distinct from system load	10, flexibility forecasting	17.2.2
Real-time forecasts at 15-minute intervals	10 and 11, receding horizon	17.2.3
Demand-response events with verified results	2 and 8, events and verification	17.3.1
Dynamic Load Shaping across DER types	12, dynamic load shaping	17.3.2
Staggered cohorts over a three-hour service	5 and 12, rotation and release	17.3.3
Firm Load Dispatch: a flat multi-hour shape	12 and 14, shaping and capacity products	17.3.4
Packetized coordination of devices	5 and 16, desynchronization and actuation	17.3.5
Four optimization objectives	11, 13, 14, dispatch and services	17.4.1
External grid objectives and targets	13 and 14, limits and products	17.4.2
EV managed charging and V2G pilots	7, managed charging	17.4.3, 17.4.5
Residential battery programs	6, batteries as resources	17.4.4
Operation at the scale of millions of devices	9, fleet aggregation	17.5.1

17.9 Connection to the TERRA framework

Mapping a commercial platform onto a graph framework is a test of whether the framework's objects are the operational ones.

- **The device ecosystem is the component library.** The vendor adapters of Figure 17.2 translate heterogeneous interfaces into components with declared states, edges and bounds: a thermostat with T_{in} inside $[T^{\min}, T^{\max}]$, a battery with SOC inside $[\text{SOC}^{\text{res}}, \text{SOC}^{\max}]$, a charger with a plug-in indicator a_t gating its edge. Everything above the adapter layer operates on the graph, not on the vendors.
- **Enrollment is graph construction, not graph solving.** The front end of Section 17.1.3 decides which components exist and at which nodes they attach. A framework in which components can be added and removed at run time represents a fleet whose membership changes daily; one in which the graph is fixed at build time does not.
- **DER-specific forecasting is forward simulation from a posterior, not regression on a sum.** The distinction of Figure 17.3 is the difference between predicting an aggregate edge flow statistically and integrating the component states forward. The second resolves shape, magnitude and timing because the switching is in the model; the first cannot, because the model has averaged it away.
- **Fifteen-minute re-issue is repeated conditioning.** Each issue in Figure 17.4 is the same forward integration started from a posterior that has absorbed one more measurement. The narrowing of successive bands at a fixed wall-clock time is the information gain from that conditioning, which is why the difference between two consecutive issues is a usable signal.
- **Firm Load Dispatch is a trajectory optimization with per-component duration bounds.** Equation (17.6) tracks a target edge flow at a node while every component carries a bound on its own cumulative control, and F_k depends on the component's accumulated thermal state. The flat aggregate of Figure 17.5 is produced by the coupling between the node's conservation equation and those per-component accumulators.
- **The external objective is a boundary condition.** The price, the feeder limit and the requested shape of Figure 17.6 enter the graph as a cost on an edge flow, a bound on an edge flow and a reference for an edge flow respectively. All three are imposed at the interface node, which is why one optimizer serves four objectives without changing the component models.

In summary, a commercial aggregation platform is a conservation graph whose membership is set by a customer program, whose component states are inferred from vendor telemetry, whose forecasts are forward integrations of that inference, and whose dispatch is a trajectory optimization against a boundary condition supplied from outside. The loop of this chapter and the ten layers of Chapter 16 are two descriptions of those same operations. The next chapter stops describing and operates: one worked case, from a utility's request to a verified delivery.

Chapter 18

A worked case: design and operate a virtual power plant for a utility

18.1 The case and the service definition

This chapter is the capstone of the sequence: a single operating problem, carried from the first reading of the request to the verification report after the event, using nothing that has not already been established in Chapters 1 to 16. The problem is the one a utility actually poses – deliver a stated reduction over a stated window – and the work consists of discovering which constraint binds, creating capability where it is short, and operating the plan under forecast error. This first section states the request, the portfolio that must serve it, and the two things that must be written down before any optimization: the service and the target trajectory.

Jargon: service definition, binding period, pre-conditioning

A *service definition* is the triple (power, duration, reliability) that a commitment must satisfy; a request quoted only in megawatts is incomplete. The *binding period* (informally, the hard hour) is the interval of the horizon in which the tightest constraint is active; it is usually not the first interval. *Pre-conditioning* is deliberately moving a resource's state before the event – cooling buildings below setpoint, advancing vehicle charging, charging a battery – so that more flexibility exists during it.

18.1.1 The request and the portfolio

A utility asks an aggregator for 80 MW of load reduction from 17:00 to 20:00. The aggregator operates three resource groups, listed in Table 18.1 with the limit that binds each and the chapter in which that limit was derived: a thermostat fleet forecast at 50 MW of flexibility at the start of the window, an electric-vehicle fleet at 35 MW, and a battery fleet rated at 25 MW.

The three nominal figures sum to $50 + 35 + 25 = 110$ MW against a request of 80 MW, so the request appears to leave a comfortable margin. That conclusion is premature, and the rest of the chapter is an account of why. Every number in Table 18.1 is an instantaneous figure, and the request is not instantaneous.

Key idea

A portfolio is sized by the sustainable profile of its resources in the most constrained interval of the service, not by the sum of their instantaneous capabilities at the start of it.

Table 18.1: The portfolio available for the case. The nominal figures are the flexibility each group offers at 17:00; the binding limit is the constraint that eventually decides what the group can actually do.

Group	Nominal at 17:00	Binding limit	State variable	Chapter
Thermostats	50 MW	comfort band $T^{\min} \le T_{\text{in}} \le T^{\max}$	indoor temperature T_{in}	5
Vehicles	35 MW	departure requirement $\text{SOC}(T^{\text{dep}}) \ge \text{SOC}^{\text{req}}$	state of charge and charging slack	7
Batteries	25 MW	usable energy $E_{\text{B}}^{\max} = 45\,\text{MWh}$ and $\text{SOC}^{\min} \le \text{SOC} \le \text{SOC}^{\max}$	state of charge SOC	6
Portfolio	110 MW	against a request of 80 MW: an apparent margin of 30 MW		

18.1.2 The service is a sustained profile, not a peak

The first question to ask of the request is what the service actually is. Eighty megawatts from 17:00 to 20:00 is a three-hour obligation, so the energy that must be displaced is

$$E^{\text{service}} = \int_{17:00}^{20:00} P^{\text{tgt}}(t)\,\mathrm{d}t = 80 \times 3 = 240 \text{ MWh}. \tag{18.1}$$

A resource that can produce 80 MW for ten minutes and a resource that can produce 80 MW for three hours are the same number in a capacity table and different products in operation. The service definition that the aggregator must answer therefore has three parts, in the sense of Chapter 14: the power, 80 MW; the duration, three hours; and the reliability with which both must hold, which the request as stated leaves unspecified and which Section 18.4.1 shows to be the decisive term.

The distinction is the same one drawn in Chapter 4 between nameplate capacity and the deliverable capacity $C(\tau)$ for a required duration τ: for every resource in Table 18.1, $C(3\,\text{h})$ is smaller than $C(0)$, and for the battery it is smaller by a factor of three.

18.1.3 The target trajectory and the tracking objective

The request is turned into a trajectory,

$$P^{\text{tgt}}(t) = 80\,\text{MW}, \qquad 17{:}00 \le t \le 20{:}00, \tag{18.2}$$

and the dispatch problem is to choose the three group powers $P_{\text{T}}(t)$, $P_{\text{E}}(t)$ and $P_{\text{B}}(t)$ – the reductions delivered by thermostats, vehicles and batteries – so that

$$P_{\text{T}}(t) + P_{\text{E}}(t) + P_{\text{B}}(t) \approx P^{\text{tgt}}(t) \tag{18.3}$$

over the whole window. The approximation sign is made precise by a tracking cost, as in Chapter 11,

$$J^{\text{track}} = \sum_{t} \big[P_{\text{T}}(t) + P_{\text{E}}(t) + P_{\text{B}}(t) - 80\big]^2, \tag{18.4}$$

summed over the dispatch intervals of the window. Equation (18.4) penalizes over-delivery and under-delivery equally, which is a modeling choice rather than a fact: a commitment settled with a shortfall penalty is asymmetric, and Chapter 11 writes the asymmetric form. The symmetric cost is used here because it keeps the feasibility question – can the sum reach 80 at all, in every hour – separate from the pricing question. Feasibility is what the next two sections establish, and it turns out to be the whole difficulty.

18.2 Resource models

Each group is now written as a profile over the three hours rather than as a single number. The three profiles decay, for three different physical reasons, and this section ends by showing that the decay is one phenomenon seen through three models.

18.2.1 Thermostats: a profile that decays as the band is consumed

The thermostat fleet offers $F_{\mathrm{T}}(17{:}00) = 50\,\mathrm{MW}$, but Chapter 5 showed that this figure cannot be held. Curtailing cooling lets the indoor temperature rise toward the upper edge of the comfort band; once a house reaches $T^{\max}$ its compressor must be allowed to run, and that house leaves the pool of available reduction. The fleet's sustainable reduction therefore falls hour by hour as the band is spent, exactly as in Figure 1.3 of Chapter 1. The forecast sustainable profile for this case is

$$F_{\mathrm{T}} = 50\,\mathrm{MW}\ (17\text{–}18), \quad 35\,\mathrm{MW}\ (18\text{–}19), \quad 20\,\mathrm{MW}\ (19\text{–}20).$$

The quantity that depletes is thermal energy: the comfort band is a store, the curtailment is a withdrawal from it, and the profile is the discharge curve of that store. A 50 MW thermostat fleet is not a 50 MW generator, and the difference is not a derating factor but a shape.

18.2.2 Vehicles: a profile that decays as slack disappears

The vehicle fleet offers 35 MW of curtailable charging at 17:00, and that figure also falls through the evening, to

$$F_{\mathrm{E}} = 35\,\mathrm{MW}\ (17\text{–}18), \quad 30\,\mathrm{MW}\ (18\text{–}19), \quad 20\,\mathrm{MW}\ (19\text{–}20),$$

for an entirely different reason. Nothing about a vehicle gets hot. What shrinks is scheduling slack: as the evening proceeds, each car has less time left before its departure and the same energy still to deliver, so its urgency U_i of Chapter 7 rises, and a car whose urgency approaches one can no longer be curtailed at all without missing its deadline. The fleet loses flexibility because its deadlines approach, which is a property of the clock and the customers' plans, not of the weather.

18.2.3 Batteries: a power limit and an energy budget

The battery fleet has a power rating $P_{\mathrm{B}}^{\max} = 25\,\mathrm{MW}$ and usable fleet energy $E_{\mathrm{B}}^{\max} = 45\,\mathrm{MWh}$. Those two numbers are not independent: delivering the rating for the whole window would require $25 \times 3 = 75\,\mathrm{MWh}$, which the fleet does not have. At full output the fleet runs for

$$\tau^{\max} = \frac{E_{\mathrm{B}}^{\max}}{P_{\mathrm{B}}^{\max}} = \frac{45}{25} = 1.8\ \mathrm{h}, \tag{18.5}$$

so the deliverable capacity of Chapter 4 is

$$C(\tau) = \min\Big(P_{\mathrm{B}}^{\max},\ \frac{E_{\mathrm{B}}^{\max}}{\tau}\Big), \tag{18.6}$$

drawn in Figure 18.1. The horizontal axis of that figure is the required duration τ in hours and the vertical axis is the power the fleet can hold for that duration; the shaded set is the pairs (τ, P) the fleet can actually serve, bounded above by the 25 MW rating for short durations and by the energy hyperbola $45/\tau$ for long ones. The two bounds cross at $\tau = 1.8\,\mathrm{h}$. The blue point is the obligation the battery will eventually be given in this case, 25 MW for one hour, which lies on the power-limited edge of the set; the red point is 25 MW for three hours, which lies outside it. Because the energy is a budget spread over the horizon, a decision to use the battery in the first hour is a decision not to have it in the third, and that coupling is what makes the case a scheduling problem rather than three independent hourly problems.

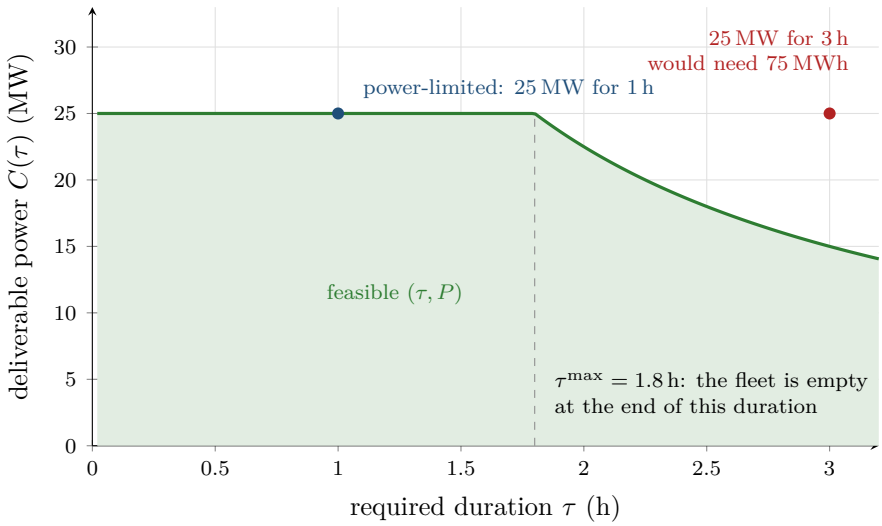


Figure 18.1: The battery fleet's capability set, equation (18.6) with $P_{\mathrm{B}}^{\max} = 25\,\mathrm{MW}$ and $E_{\mathrm{B}}^{\max} = 45\,\mathrm{MWh}$. The boundary is the 25 MW rating for durations below 1.8 h and the energy hyperbola $45/\tau$ above it. The one-hour obligation lies on the boundary; the three-hour one lies outside the set.

18.2.4 State depletion as the common mechanism

The three profiles are collected in Figure 18.2 and tabulated in Table 18.2. In the figure, the left-hand axis is power in megawatts and the three step curves are the sustainable reductions of the thermostat fleet (purple), the vehicle fleet (blue) and the battery power rating (green, dashed and constant, because a power rating does not decay). The right-hand axis is energy in megawatt-hours and the rising green line is the energy the battery would have drawn if it ran at its rating from 17:00; the dashed red horizontal is the 45 MWh budget, and the line crosses it 1.8 hours into the window. Two curves fall and one does not, and the one that does not is limited by the line that crosses the budget.

Table 18.2 adds the arithmetic the figure leaves implicit: the hourly sum of the three capabilities, the target, and the margin between them. The margin is +30 MW in the first hour, +10 MW in the second and −15 MW in the third. The last row is the whole problem of the chapter, and it was invisible in Table 18.1.

Table 18.2: Hourly capability of the portfolio against the target. The margin is the sum of the three capabilities minus 80 MW. The final row shows that the portfolio as forecast cannot serve the third hour.

Hour	F_{T} (MW)	F_{E} (MW)	$P_{\mathrm{B}}^{\max}$ (MW)	Sum (MW)	P^{tgt} (MW)	Margin (MW)
17:00–18:00	50	35	25	110	80	+30
18:00–19:00	35	30	25	90	80	+10
19:00–20:00	20	20	25	65	80	−15
Energy (MWh)	105	85	45 (budget)	235	240	−5

The energy row of Table 18.2 is a second, independent warning. Integrating the forecast profiles gives 105 MWh from thermostats and 85 MWh from vehicles, and the battery can add at most its 45 MWh budget, for 235 MWh against the 240 MWh the service requires by equation (18.1). The portfolio is short on energy as well as short in the third hour, and both shortfalls are properties of the state trajectories rather than of the ratings.

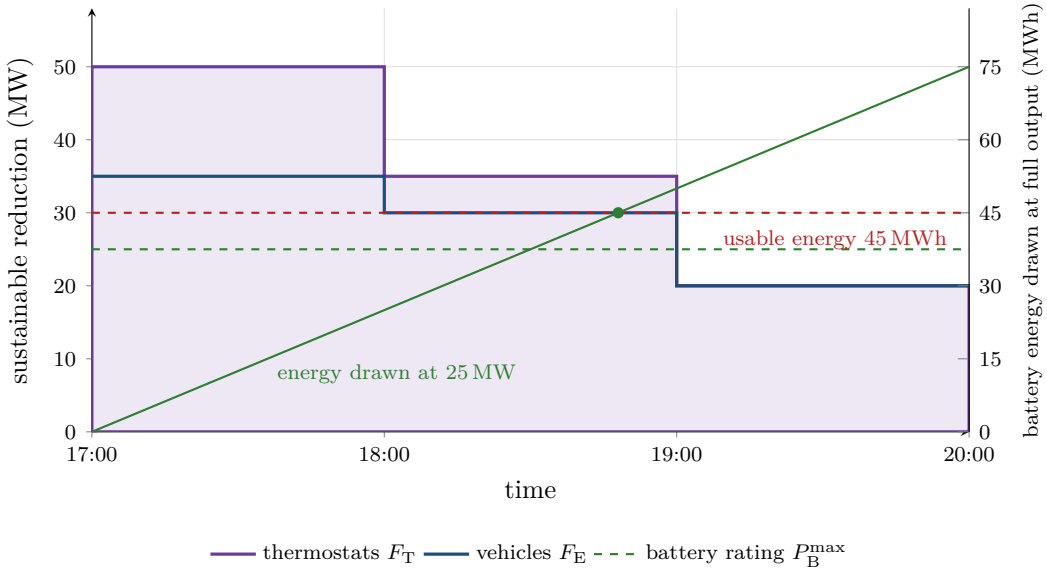


Figure 18.2: The three resource profiles over the service window. Left axis: the sustainable reduction of each group, as step functions over the three hours. Right axis: cumulative battery energy if the fleet ran at its 25 MW rating, which reaches the 45 MWh budget at 18:48. The thermostat and vehicle profiles are the forecasts of Sections 18.2.1 and 18.2.2.

Table 18.3 states the common mechanism. In all three groups a stored quantity is drawn down by the act of delivering the service, and the flexibility available later is a function of how much of that store remains. The physics differs – thermal energy in a building's mass, unserved charging energy before a deadline, electrochemical energy in a cell – and the mathematics does not: each group has a state, a bound on that state, and a flexibility that is a function of the distance to the bound.

Table 18.3: State depletion in the three groups: one phenomenon, three physical mechanisms. The last column names the quantity whose exhaustion ends the group's contribution.

Group	State	Why flexibility decays	Store being spent
Thermostats	T_{in} relative to $T^{\max}$	curtailment raises T_{in}; houses at $T^{\max}$ must be released	thermal energy of the building mass, $C(T^{\max} - T_{\mathrm{in}})$
Vehicles	E^{rem}, T^{rem}, hence U	time to departure shrinks while the energy owed does not	scheduling slack, $T^{\mathrm{rem}} - E^{\mathrm{rem}}/P^{\max}$
Batteries	SOC	discharging lowers SOC toward $\mathrm{SOC}^{\min}$	usable energy, $E_{\mathrm{B}}^{\max}$

Principle 18.1. *Every flexible resource is a store with a bound. Its instantaneous capability is a power limit; its capability over a horizon is set by how much of the store remains, so the dispatch problem is the allocation of stores across the horizon.*

18.3 Finding a feasible plan

The plan is found by four attempts, each of which fails against a different constraint before the fifth succeeds. This section follows them in order, with the numbers of every attempt, because the sequence of failures is what identifies the moves that make the service possible. Figure 18.4 collects four of the schedules as stacked bars and Table 18.7 summarizes the whole sequence.

18.3.1 Greedy dispatch and its failure in the last hour

The natural first attempt is to dispatch the cheapest resource fully, hour by hour, and call on the battery only for whatever remains. In the first hour the thermostat fleet can give 50 MW, so 30 MW of vehicle curtailment completes the target and the battery is untouched. In the second hour the capabilities have fallen to 35 and 30, so the battery supplies 15 MW. In the third hour thermostats and vehicles together give only 40 MW, so the battery is asked for 40 MW – against a rating of 25. Table 18.4 records the schedule and the violation.

Table 18.4: The greedy schedule. Each hour is satisfied in turn using the most available resource first; the third hour then asks the battery for 40 MW, which exceeds its 25 MW rating, and would in any case require 55 MWh of a 45 MWh budget.

Hour	P_{T} (MW)	P_{E} (MW)	P_{B} (MW)	Total (MW)	Energy (MWh)	Verdict
17:00–18:00	50	30	0	80	0	feasible, but spends the thermal store fastest
18:00–19:00	35	30	15	80	15	feasible
19:00–20:00	20	20	40	80	40	infeasible: $40 > P_{\mathrm{B}}^{\max} = 25$ MW
Total	105	80	55	240	55	infeasible: 55 > 45 MWh

Greedy dispatch fails for a structural reason, not an arithmetic one. It treats each hour as a separate problem, and the resources are coupled across hours through their states: every megawatt-hour of thermal store spent at 17:30 is unavailable at 19:30, and the battery energy left for the third hour is whatever the first two did not take. A dispatch rule that does not look ahead will always exhaust the flexible resources in the easy hours and arrive at the hard one with the inflexible remainder. This is the horizon argument of Chapter 11 and Figure 11.3, where a myopic battery plan reaches its reserve an hour before the service ends.

18.3.2 Working backward from the binding period

The correct starting point is the most constrained interval. From the last row of Table 18.2, the hour 19:00–20:00 offers

$$F_{\mathrm{T}} + F_{\mathrm{E}} + P_{\mathrm{B}}^{\max} = 20 + 20 + 25 = 65\,\mathrm{MW} < 80\,\mathrm{MW} = P^{\mathrm{tgt}}, \tag{18.7}$$

a shortfall of 15 MW against the target even if every resource is used at its limit and nothing is spent earlier. With the forecasts as given, the request is infeasible. No optimizer can repair this, because the infeasibility is not in the choice of dispatch but in the forecast profiles themselves: the feasible set of the third hour does not contain the target.

That observation reframes the problem. If the hard hour cannot be served with the profiles as forecast, then the profiles must be changed, and the only way to change them is to act on

the resource states before and during the earlier hours. The plan must be built backward: fix what the hard hour needs, then ask what the earlier hours are permitted to spend.

> **Principle 18.2.** *Identify the binding period first and design the plan backward from it. The earlier intervals are not opportunities to deliver cheaply; they are the budget from which the binding interval's capability is bought.*

18.3.3 Creating capability in the hard hour

Two moves raise the third-hour capability, one per deferrable resource, and Figure 18.3 shows what they do to the profiles.

Pre-cooling the buildings. The thermostat fleet's third-hour figure of 20 MW assumes the buildings enter the window at their setpoints and are curtailed from 17:00 onward. If instead the fleet is cooled below setpoint before 17:00, every house starts the window further from $T^{\max}$, which is additional thermal store, and the same is achieved in part by curtailing less aggressively during the first hour so that less of the band is spent early. Chapter 5 and Figure 1.8 of Chapter 1 give the mechanism; Chapter 12 gives the fleet-level version. Suppose pre-conditioning raises the third-hour thermostat capability from 20 to 30 MW. That is an extra 10 MW held for one hour, so it is 10 MWh of additional thermal storage made available by cooling earlier, paid for with electricity consumed before 17:00 when the system is less stressed. The hard hour becomes

$$30 + 20 + 25 = 75\,\text{MW},$$

still 5 MW short.

Advancing vehicle charging. The second move is the mirror image of the first. The instinct before an event is to delay vehicle charging so that more charging remains to be curtailed; the effect in the third hour is the opposite, because a car whose charging has been postponed is a car whose urgency is high at 19:00 and which therefore cannot be curtailed. Advancing some charging earlier in the day reduces the energy that must still be delivered in the evening, lowers urgency across the fleet, and leaves more cars that are charging but not urgent at 19:00. Suppose the third-hour vehicle capability rises from 20 to 25 MW. That is about 5 MW of curtailable charging, or roughly $5000/7.2 \approx 700$ vehicles on 7.2 kW chargers. The hard hour becomes

$$30 + 25 + 25 = 80\,\text{MW},$$

which meets the target exactly. The resolution did not come from a better optimizer; it came from managing resource state before the event.

18.3.4 Two further binding constraints

A third-hour plan of $30 + 25 + 25$ is only one hour of a three-hour schedule, and the preservation that makes it possible restricts the earlier hours. Holding 30 MW of thermostat capability until 19:00 means not drawing the full 50 MW at 17:00: the fleet must be worked at about 35 MW in the first hour so that enough of the comfort band survives. The resulting schedule is the first block of Table 18.5: $35 + 35 + 10$, then $30 + 30 + 20$, then $30 + 25 + 25$. Every hour reaches 80 MW, which is what a feasibility check on the power balance alone would report.

The battery energy, however, is $10 + 20 + 25 = 55$ MWh against a 45 MWh budget. The schedule is infeasible by 10 MWh, and it is infeasible in a way that no single hour's power balance reveals, because the violated constraint couples all three hours. Interview cases, like

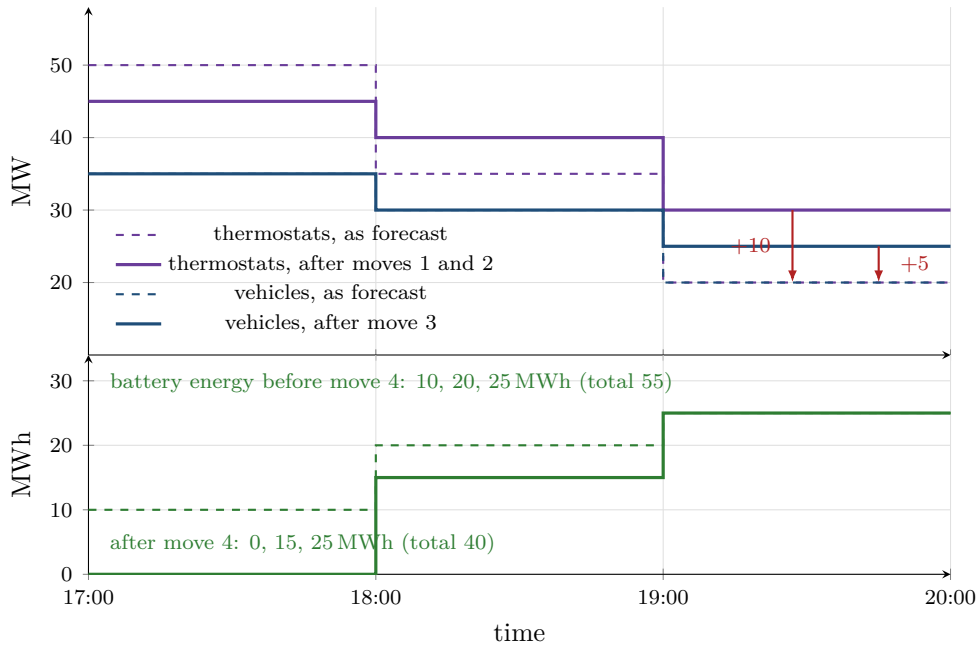


Figure 18.3: The four enabling moves as changes to the hourly profiles. Top: the thermostat and vehicle capabilities as forecast (dashed) and after pre-conditioning, cohort rotation and charging advancement (solid); the red arrows mark the +10 MW and +5 MW created in the hard hour. Bottom: the battery energy spent per hour before and after the energy is reserved for the hard hour. All values are the case's figures from Sections 18.3.3 to 18.3.5.

operating problems, contain several interacting constraints, and satisfying the one that was binding a moment ago often exposes the next.

The obvious repair is to take the 10 MWh out of the first hour, where the battery was least needed. Setting $P_{\mathrm{B}} = 0$ in the first hour leaves the second block of Table 18.5: the battery energy now sums to $0 + 20 + 25 = 45$ MWh and fits, but the first hour must then find 80 MW from thermostats and vehicles alone, whose preserved capabilities are $35 + 35 = 70$ MW. The first hour is short by 10 MW. Pushing the thermostat fleet to 45 MW in that hour would close the gap and would also consume the thermal store that the third hour depends on. The problem is a genuine optimization: capability in the hard hour and capability in the first hour are drawn from the same store, and the question is how to divide it.

18.3.5 Cohort rotation and the feasible schedule

The division is possible because the thermostat fleet is not a single lump. Chapter 5 introduced rotation and Chapter 12 developed its accounting: divide the fleet into cohorts A, B and C, curtail one cohort at a time, and release each cohort before its houses reach $T^{\max}$. Each building then spends only part of the window curtailed and recovers in between, so the fleet as a whole stays inside the comfort band far longer than a fleet curtailed continuously. The price is paid in the first hour: with only part of the fleet curtailed at any instant, and with the cohorts that have just been released drawing their recovery power, the net reduction available at 17:00 falls from 50 to about 45 MW. That net figure is net of rebound in the sense of Figure 12.3, where the gross reduction of the curtailed cohort and the recovery of the released one are added with their signs.

The optimized thermostat profile becomes

$$F_{\mathrm{T}} = 45\,\mathrm{MW}\ (17\text{–}18), \quad 40\,\mathrm{MW}\ (18\text{–}19), \quad 30\,\mathrm{MW}\ (19\text{–}20),$$

Table 18.5: The two intermediate schedules; the battery column is cumulative energy over the window. Iteration 2 pre-conditions the fleets and satisfies every hourly balance but exceeds the 45 MWh battery budget; iteration 3 removes the first-hour battery energy and then cannot reach 80 MW in the first hour.

Iter.	Hour	P_T (MW)	P_E (MW)	P_B (MW)	Total (MW)	Battery (MWh)	Verdict
2	17–18	35	35	10	80	10	balance met
2	18–19	30	30	20	80	30	balance met
2	19–20	30	25	25	80	55	energy budget exceeded by 10 MWh
3	17–18	35	35	0	70	0	short by 10 MW
3	18–19	30	30	20	80	20	balance met
3	19–20	30	25	25	80	45	energy budget exactly met

a flatter profile with a lower first hour and a far higher third hour, and with a larger integral – 115 MWh against the 105 MWh of the unrotated forecast – because rotation wastes less of the band on houses that are already at their limit. Table 18.6 gives the resulting schedule, with each group's hourly capability in parentheses so that the binding entries are visible.

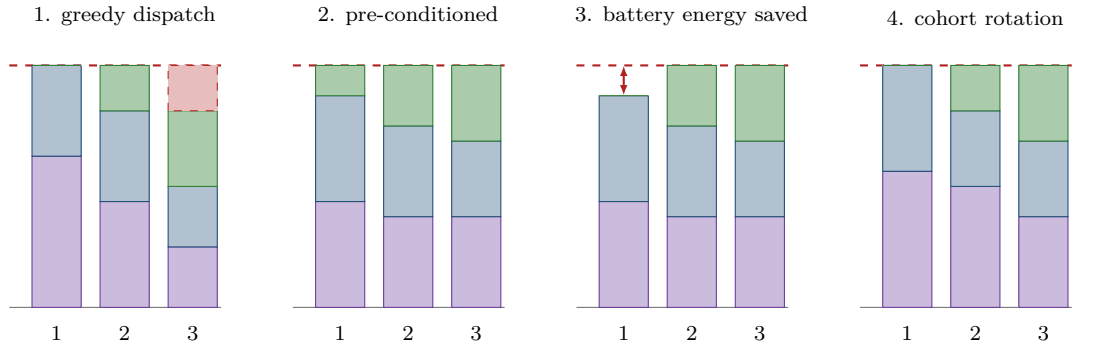
Table 18.6: The feasible schedule. The thermostat and vehicle entries give the dispatched power with that hour's capability, after the four moves, in parentheses; an entry equal to its capability is binding. Every hour reaches 80 MW, the battery energy totals 40 MWh of a 45 MWh budget, and the three groups together deliver exactly the 240 MWh of equation (18.1).

Hour	P_T (MW)	P_E (MW)	P_B (MW)	Total (MW)	Cumulative battery (MWh)
17:00–18:00	45 (45)	35 (35)	0	80	0
18:00–19:00	40 (40)	25 (30)	15	80	15
19:00–20:00	30 (30)	25 (25)	25	80	40
Energy (MWh)	115	85	40	240	40 of 45

Three checks confirm the schedule. Every hourly total is 80 MW, so equation (18.3) holds and the tracking cost (18.4) is zero in expectation. The battery energy is $0+15+25 = 40$ MWh, inside the 45 MWh budget with 5 MWh to spare. And the energy delivered is $115+85+40 = 240$ MWh, which equals the service energy of equation (18.1) – as it must, since every hour is met exactly. The thermostat and vehicle entries are binding in five of six cells, which is a warning the next section takes up: a schedule with almost no slack is a schedule with almost no tolerance for forecast error.

Figure 18.4 draws the four schedules side by side. Each panel has three stacked bars, one per hour, with the thermostat contribution at the bottom in purple, the vehicle contribution above it in blue and the battery contribution on top in green; the dashed red line is the 80 MW target. The first panel's third bar carries a red block above the green one, the part of the greedy schedule's 40 MW battery call that exceeds the rating. The second panel reaches the target in every hour but spends 55 MWh of battery energy. The third panel fits the energy budget and falls 10 MW short in the first hour, marked by the red arrow. The fourth panel reaches the

target in every hour within every limit.



Hour 3 asks 40 MW of a 25 MW battery; the red block cannot be delivered.

Every hour reaches 80 MW, but the battery energy is $10 + 20 + 25 = 55$ MWh against a 45 MWh budget.

The battery energy now fits, but hour 1 reaches only 70 MW.

Every hour reaches 80 MW and the battery energy is $0 + 15 + 25 = 40$ MWh. Feasible.

Purple: thermostats P_{T}. Blue: vehicles P_{E}. Green: battery P_{B}. Red, dashed outline: a block that violates a limit. Hours 1, 2 and 3 are 17:00–18:00, 18:00–19:00 and 19:00–20:00.

Figure 18.4: The four candidate schedules of Section 18.3 as stacked hourly bars against the 80 MW target (dashed). Panels 1 to 3 each violate a different constraint – a power rating, an energy budget, and the first hour's capability – and panel 4, the rotated schedule of Table 18.6, violates none.

18.3.6 What made the solution possible

Table 18.7 summarizes the five readings of the problem, from the nameplate sum to the feasible schedule, with the constraint that fails in each. Read down the third column, the sequence is a tour of the book: a power rating, an energy budget, a state trajectory, and a fleet-level control pattern.

Four moves made the difference, and none of them is an optimization technique. Preconditioning created thermal store before the window. Cohort rotation changed the shape of the thermostat profile, trading 5 MW in the first hour for 10 MW in the third. Advancing vehicle charging reduced the evening energy obligation and so lowered urgency. Saving battery energy treated the 45 MWh as a budget to be allocated across the horizon rather than as a resource to be spent when convenient. All four are state-management decisions taken before or early in the event, and all four were taught by earlier chapters; Table 18.13 at the end of the chapter lists them with those chapters.

18.4 Uncertainty, reserves and real-time operation

The schedule of Table 18.6 is a plan built from point forecasts. This section attaches uncertainty to those forecasts, holds a reserve against it, and then operates the plan through a tracking error at 17:05, a weather revision at 17:30 and a behavioral revision shortly after.

18.4.1 Forecast uncertainty and the zero-margin plan

The first-hour capabilities are not known exactly. The forecasts, with their standard deviations, are

$$F_{\mathrm{T}} = 45 \pm 8\,\mathrm{MW}, \qquad F_{\mathrm{E}} = 35 \pm 5\,\mathrm{MW}, \qquad F_{\mathrm{B}} = 25 \pm 1\,\mathrm{MW},$$

Table 18.7: The sequence of attempts and the constraint that binds in each. Each row changes one thing about the previous row, and the constraint that fails changes with it.

#	What is assumed or changed	Constraint that fails	Verdict
0	nameplate sum, 110 ≥ 80 MW	third-hour capability 65 < 80 MW	infeasible
1	greedy hour-by-hour dispatch	battery power 40 > 25 MW in hour 3	infeasible
2	pre-cooling and charging advancement, capability preserved for hour 3	battery energy 55 > 45 MWh	infeasible
3	first-hour battery energy removed	first-hour capability 70 < 80 MW	infeasible
4	thermostat cohort rotation	none	feasible

the thermostat band being much the widest because it depends on weather, occupancy and customer behavior, and the battery band the narrowest because an inverter does what it is told. Treat the delivery error of each group as additive with these standard deviations and independent across groups. Then any schedule whose entries sum to 80 MW delivers

$$P^{\text{del}} \sim \mathcal{N}(80,\ \sigma^2), \qquad \sigma = \sqrt{8^2 + 5^2 + 1^2} = \sqrt{90} = 9.5\,\text{MW}, \tag{18.8}$$

and consequently $\Pr(P^{\text{del}} \geq 80\,\text{MW}) = 0.5$. A plan whose expected delivery equals the target meets the target in half of all events. Figure 18.5 draws the distribution: the horizontal axis is the delivered reduction in the first hour, the curve is the density of equation (18.8), the vertical red line is the 80 MW target, and the shaded area to its left is the half of the probability mass that constitutes a shortfall. The dashed marker is the fifth percentile, $80 - 1.645 \times 9.5 = 64.4$ MW: one event in twenty delivers 64 MW or less.

If the utility's service definition includes a reliability level – the third term of Section 18.1.2 – then the plan must carry a margin. For $\Pr(P^{\text{del}} \geq 80) \geq 0.95$ the expected delivery must be $80 + 1.645 \times 9.5 = 95.6$ MW, the green arrow in Figure 18.5. The first hour can supply that from its 105 MW of capability. The hard hour cannot: its capability after all four moves is exactly 80 MW, so a 0.95 commitment in that hour would have to be no larger than $80 - 15.6 = 64.4$ MW. The honest reading of the case is therefore that an 80 MW commitment over three hours is a roughly even-odds commitment in its last hour, and that a firm 80 MW requires either more enrolled resource or a shorter window.

Principle 18.3. *A schedule whose expected delivery equals the target is a coin flip. Reliability is bought with margin, the margin is largest where the forecast is widest, and the binding period is usually the interval where margin is hardest to find.*

The independence assumption behind $\sigma = 9.5$ MW is optimistic. The thermostat and vehicle errors share drivers – the same weather, the same evening, the same customers – and Chapter 9 showed in Figure 3.4 that a common-mode component does not average away. With a correlation of 0.3 between the thermostat and vehicle errors the variance becomes $90 + 2 \times 0.3 \times 8 \times 5 = 114$ and σ rises to 10.7 MW, which widens the required margin to 17.6 MW.

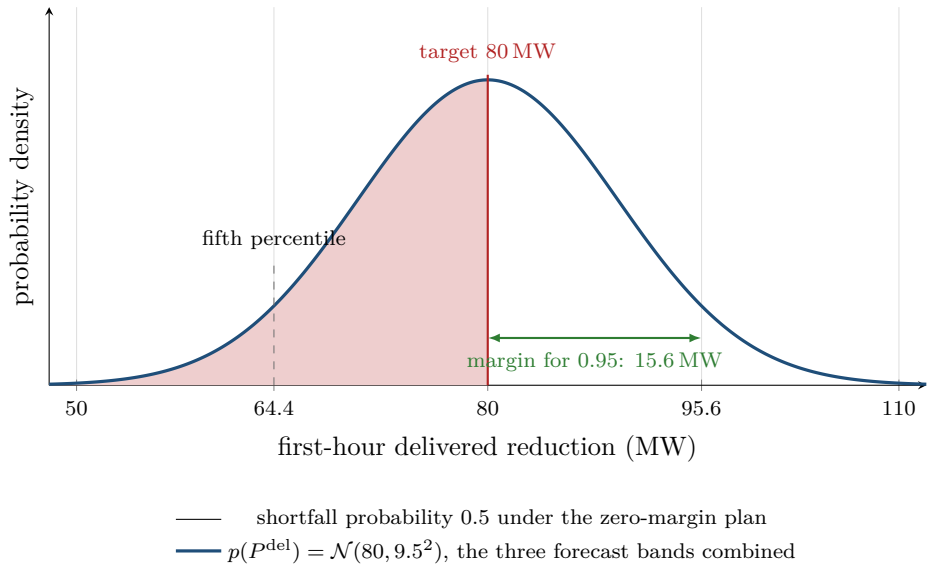


Figure 18.5: The delivery distribution of a schedule that sums to the target. The density is $p(P^{\text{del}}) = \exp[-(P^{\text{del}} - 80)^2/(2 \cdot 9.5^2)]/(9.5\sqrt{2\pi})$, from the three forecast standard deviations of Section 18.4.1 combined independently. Half the mass lies below the target; reaching a 0.95 delivery probability requires 15.6 MW of additional expected response.

18.4.2 Reserves

Since a point-forecast schedule delivers the target half the time, part of the portfolio is held back to correct the realization rather than committed to the plan. Define, for the battery,

$$P_{\text{B}}^{\text{res}}(t) = P_{\text{B}}^{\text{avail}}(t) - P_{\text{B}}^{\text{sch}}(t), \tag{18.9}$$

the difference between the power the battery could produce and the power the schedule asks of it. This is the operating reserve of Chapter 14, held inside a fleet rather than on a system: unused capability kept for an event that has not happened yet.

The first hour is rescheduled to create it. Instead of $45 + 35 + 0$, the plan becomes $43 + 32 + 5 = 80$ MW, shown in Table 18.8. Two megawatts of thermostat capability and three of vehicle capability are left unused, and the battery is started at 5 MW so that it is already online. The reserve is then $25 - 5 = 20$ MW of battery power, which is $20/9.5 = 2.1$ standard deviations of the delivery distribution, so a correction of the size the forecast error is likely to require is available with probability $\Phi(2.1) \approx 0.98$.

The last row of Table 18.8 is the catch. Starting the battery at 5 MW in the first hour spends the 5 MWh that Table 18.6 had left over, so the battery energy is now allocated in full: $5 + 15 + 25 = 45$ MWh. The 20 MW of reserve is a power reserve with no energy behind it. It can correct an error for a few minutes; sustaining it would borrow energy from a later hour, which is exactly what happens next.

18.4.3 The event begins: a tracking error at 17:05

Five minutes into the event the telemetry reports

$$P_{\text{T}} = 39\,\text{MW}, \qquad P_{\text{E}} = 31\,\text{MW}, \qquad P_{\text{B}} = 5\,\text{MW}, \qquad \text{total } 75\,\text{MW},$$

against a target of 80, so the tracking error of Chapter 3 is $e_t = P^{\text{tgt}} - P^{\text{del}} = 5$ MW. The thermostat fleet is 4 MW below its scheduled 43 and the vehicles 1 MW below their 32. Before anything is done about it, the size of the error is worth placing: 5 MW is 0.53σ of the delivery

Table 18.8: The first-hour plan with a reserve. The scheduled column sums to the target; the reserve column is equation (18.9) applied to each group. Only the battery's reserve is fast and certain enough to use inside the hour, and the energy row shows that the power reserve has no energy reserve behind it.

Group	Capability (MW)	σ (MW)	Scheduled (MW)	Reserve (MW)	Usability of the reserve
Thermostats	45	8	43	2	slow and uncertain; using it also spends the thermal store
Vehicles	35	5	32	3	available, but raises urgency later
Batteries	25	1	5	20	immediate and precise
Portfolio	105	9.5	80	25	20 MW of it is usable within seconds
Battery energy	45 MWh	–	45 MWh	0 MWh	power reserve, no energy reserve

distribution of equation (18.8). An error this size is not a failure of anything; it is the forecast band, and a well-calibrated plan should produce errors of this magnitude in about half of all events.

The correction is to raise the battery from 5 to 10 MW, which restores $39 + 31 + 10 = 80$ MW. This is the right instrument: the battery is the fastest and most predictable member of the portfolio, its own forecast band is ± 1 MW, and it is the one resource whose response does not depend on a customer. Chapter 11 draws the same assignment in Figure 11.6, where the battery takes whatever remains between the realized thermostat response and the target.

Figure 18.6 shows the first hour. The upper axis plots the delivered total against the 80 MW target, with the dip to 75 MW at minute five and the recovery immediately after; the lower axis plots the battery power, a step from 5 to 10 MW at the same moment. The two dashed vertical lines mark the events of this section and the next: the replan at 17:05 and the weather update at 17:30. The annotation on the lower axis is the consequence: holding 10 MW for the remaining 55 minutes draws

$$5 \times \tfrac{5}{60} + 10 \times \tfrac{55}{60} = 0.42 + 9.17 = 9.6 \text{ MWh},$$

against the 5 MWh the schedule had budgeted for the hour. The tracking error has been closed in power and transferred into an energy problem later in the horizon.

18.4.4 Re-optimizing the remaining horizon

The correction changed a state, so the plan built on the old state is no longer the right plan. Of the 45 MWh budget, 9.6 have been used and 35.4 remain; the hard hour still requires 25 MWh if the battery is to hold 25 MW through it, so the second hour may use at most $35.4 - 25 = 10.4$ MWh instead of the 15 it was given.

The second hour is therefore re-solved with a battery allocation of about 10 MWh. The slack needed to absorb the difference is already in Table 18.6: the second hour's vehicle entry,

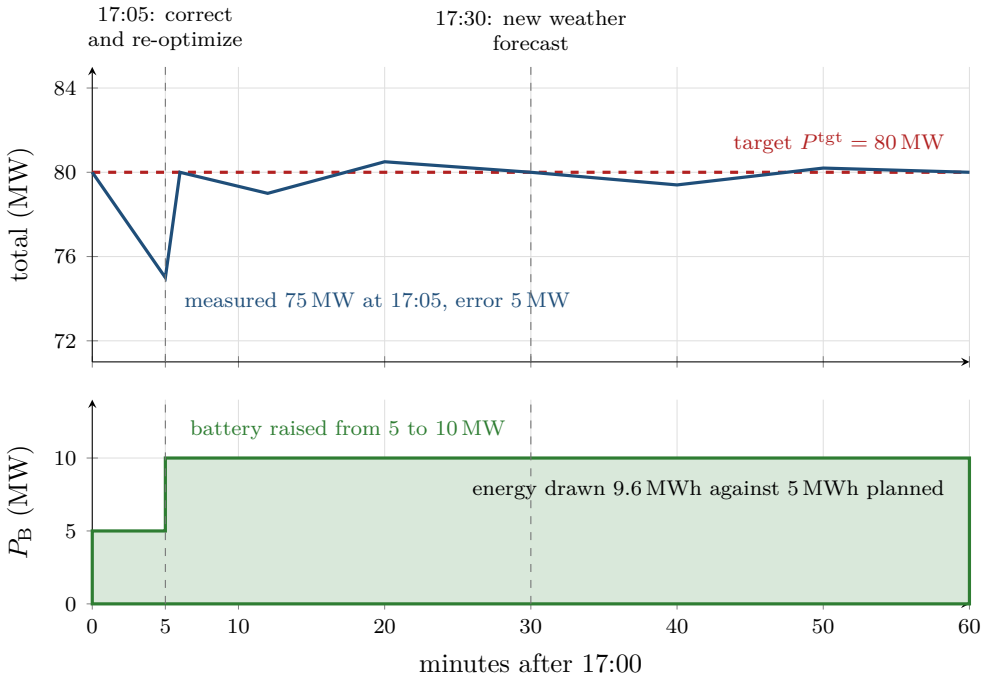


Figure 18.6: The first hour of the event. Upper axis: the delivered total against the target, with the 5 MW tracking error at 17:05 and its correction. Lower axis: the battery power, stepped from 5 to 10 MW to close the error, which draws 9.6 MWh instead of the 5 MWh budgeted for the hour. The dashed verticals mark the replan and the weather update.

25 MW of a 30 MW capability, is the one cell in that table that was not binding. Raising it to its capability gives

$$40 + 30 + 10 = 80\,\text{MW} \quad (18{:}00\text{–}19{:}00),$$

with battery energy $9.6 + 10 + 25 = 44.6$ MWh, inside the budget with 0.4 MWh to spare, and with the hard hour's 25 MW preserved. The replan is a receding-horizon step in the sense of Chapter 11 and Figure 11.7: measure the state, re-solve the remaining horizon from it, apply the next action, and discard the rest of the plan. It also illustrates why slack is kept deliberately. The replan succeeded because one cell of the schedule was not binding; had every cell been at its capability, the 4.6 MWh of battery energy spent on the correction would have had to be taken out of the hard hour.

18.4.5 A hotter forecast at 17:30

Half an hour into the event the weather forecast is revised upward. The instinct is to welcome it, because hotter weather means more air conditioning and therefore more load to curtail. That reading is incomplete, and the correct one has two opposing parts, both derived in Chapter 5:

- **Instantaneous capability rises.** A higher T_{out} raises the compressors' duty cycle, so more HVAC load is running at any moment and more of it can be shed. The thermostat fleet's MW figure for the next interval goes up.
- **Sustainable duration falls.** A higher T_{out} also means a larger temperature difference across the envelope, so T_{in} rises faster once cooling is withheld. Each house reaches $T^{\max}$ sooner, so the fleet's reduction cannot be held as long. The figure for the third hour goes down.

The two effects act on different parts of the same profile, and their net effect on the service depends on where the binding period is. Here the binding period is the last hour, so a hotter forecast is, on balance, bad news: it offers more power in the interval that already had margin

and less in the interval that had none. The operational response is to revise the whole profile $F_{\mathrm{T}}(t)$ – raising the near term, lowering the far term – and re-solve, not to scale a single capability number up. A platform whose forecast interface returns one megawatt figure per cohort, rather than a profile with a duration index, cannot represent this revision at all; that is the interface argument of Chapter 16.

18.4.6 A higher override rate

The second revision is behavioral. The plan assumed an override probability $p^{\mathrm{ovr}} = 0.05$, the fraction of customers who cancel the event on their thermostat; the observed rate during the first hour is 0.12. Overrides remove houses from the responding population, so the thermostat forecast must be rescaled by the ratio of responding fractions,

$$\hat{F}_{\mathrm{T}}^{\mathrm{new}}(t) = \hat{F}_{\mathrm{T}}(t)\,\frac{1 - p^{\mathrm{ovr\,obs}}}{1 - p^{\mathrm{ovr\,fcst}}} = \hat{F}_{\mathrm{T}}(t) \times \frac{0.88}{0.95} = 0.926\,\hat{F}_{\mathrm{T}}(t), \tag{18.10}$$

which turns the rotated profile $45, 40, 30$ MW into $41.7, 37.1, 27.8$ MW. The consequence falls where the plan has no margin. The hard hour becomes

$$27.8 + 25 + 25 = 77.8\,\mathrm{MW},$$

a shortfall of 2.2 MW about ninety minutes before the hour begins. Every other entry in the hard hour is already at its capability, so there is nothing to reallocate: the 0.4 MWh of battery slack is worth 0.4 MW for the hour, the vehicles are at 25 of 25, and the thermostat figure is the one that just fell. The available responses are to recruit additional cohorts into the last hour at the cost of a larger rebound afterward (Chapter 12), to buy the difference elsewhere, or to tell the utility at 17:40 that the last hour will deliver about 78 MW. Giving quantified notice two hours ahead of a 2.2 MW miss is an operational result, not a failure; discovering the miss at 19:55 would be the failure. The cycle just completed – observe the override rate, update the forecast, re-optimize, and report what the update implies – is the learning loop of Chapter 16 running inside a single event.

18.5 Attribution and verification

After the event two questions remain: where in the chain a shortfall arose, and how much the fleet is credited with having delivered. The two are different questions, and the second has an uncertainty of its own.

18.5.1 Four causes of a short response

Suppose the thermostat fleet was believed to offer 45 MW and the measured response was 40 MW. Four mechanisms produce that same number, and they call for four different remedies. Table 18.9 states each with the arithmetic it implies and the measurement that distinguishes it.

The useful question after an event is therefore not why the forecast missed but where in the chain the discrepancy entered. Figure 8.4 of Chapter 8 is the diagnostic picture: a forecast capability, a requested dispatch, a device-level response and a verified grid response are four separate numbers, and the gaps between consecutive pairs have different causes and different owners. The first gap is an operator decision, the reserve of Section 18.4.2; the second is a control-path loss; the third may be a real loss or an artifact of the baseline. Reporting a single delivered figure destroys precisely the information needed to decide what to change before the next event.

Table 18.9: Four causes of a 5 MW gap between a 45 MW forecast and a 40 MW measured response, with the evidence that identifies each and the chapter that treats it. The remedies have nothing in common.

Cause	Mechanism	Arithmetic	Distinguishing evidence	Ch.
Forecast error	the capability was never there: only 40 MW was available	$F = 40$, all of it dispatched and delivered	per-device responses match their models, the aggregate does not match the forecast	10
Control path	45 MW was available, but commands reached only 90 % of devices	$0.90 \times 45 = 40.5$ MW	acknowledgment counts below dispatched counts	16
Customer behavior	overrides ran at 0.12 instead of 0.05	$45 \times 0.88/0.95 = 41.7$ MW, equation (18.10)	setpoint-restored events in telemetry, by cohort	15
Baseline error	the fleet delivered 45 MW and the counterfactual was estimated too low	measurement and verification reports 40 MW of a true 45	device telemetry and verified response disagree; a control group settles it	8

18.5.2 Probabilistic measurement and verification

The verification for this event runs as follows. The metered aggregate load of the enrolled population during the window averaged 900 MW. The estimated counterfactual baseline – what the population would have drawn without the event – was 982 MW. The verified reduction is therefore

$$\Delta P = P^{\text{base}} - P^{\text{act}} = 982 - 900 = 82\,\text{MW}, \tag{18.11}$$

against a target of 80, and the event appears successful by 2 MW.

The appearance is less firm than the arithmetic suggests, because P^{base} was estimated, not measured. With a baseline standard deviation of 10 MW, which is typical of a regression baseline on a population this size, the verified reduction is itself a random variable, $\Delta P \sim \mathcal{N}(82, 10^2)$, and the quantity the contract cares about is

$$\Pr(\Delta P \geq 80\,\text{MW}) = \Phi\Big(\frac{82 - 80}{10}\Big) = \Phi(0.2) \approx 0.58. \tag{18.12}$$

Figure 18.7 draws this. The horizontal axis is the verified reduction, the curve is the density of ΔP, the dashed vertical is the point estimate of 82 MW, and the red vertical is the 80 MW contract threshold; the shaded region to the right of the threshold has area 0.58. The event was delivered, and the verification supports that conclusion with odds of about four to three rather than with certainty. This is the probabilistic measurement and verification of Chapter 8, whose Figure 8.5 is the same construction at device-fleet scale.

The reported figure should accordingly be a probability, not a point. “The event delivered 82 ± 10 MW, meeting the 80 MW commitment with probability 0.58” is a statement the operator can act on: it says the margin is thin relative to the measurement, and that either a larger

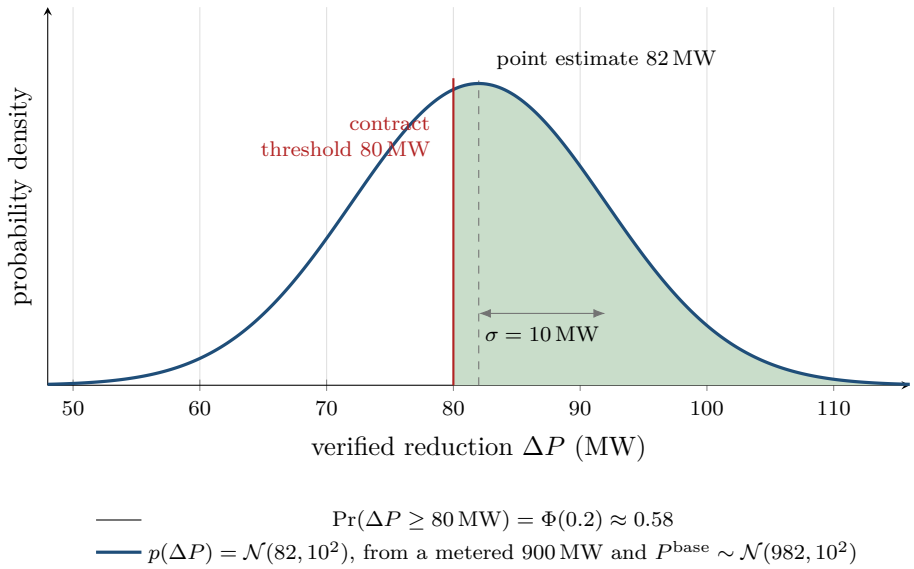


Figure 18.7: Probabilistic verification of the event. The density is $p(\Delta P) = \exp[-(\Delta P - 82)^2/(2 \cdot 10^2)]/(10\sqrt{2\pi})$, implied by equation (18.11) and a baseline standard deviation of 10 MW. The shaded area right of the 80 MW threshold is 0.58: the event is verified as delivered with moderate, not high, confidence.

margin or a better baseline is needed before the next settlement. "The event delivered 82 MW" is not.

One further qualification belongs with the number. Equation (18.11) is an average over the whole three-hour window, and the shortfall predicted in Section 18.4.6 falls in the last hour alone. An event-average verification can show a surplus while one interval was short, which matters because contracts are sometimes settled interval by interval; the hourly decomposition should be reported alongside the average, with its own baseline uncertainty, which is wider because each hour contains less data than the window.

18.5.3 From attribution to the next event

The residual between what was predicted and what was verified is the input to the slowest of the three loops of Chapter 16, drawn in Figure 16.5. Three quantities from this event should be written back into the models: the realized override rate of 0.12, which belongs in the behavioral model of the cohorts that overrode; the realized thermostat profile, which belongs in the flexibility forecast as evidence about how quickly the comfort band is spent at this outdoor temperature; and the verified response together with the device-level response, whose difference is evidence about the baseline rather than about the fleet. Chapter 8 makes the last point sharply: a residual that is the same sign on every event is a baseline bias, and no amount of better dispatch will remove it.

18.6 Presenting the reasoning

The case has a second audience: whoever asks for the reasoning out loud, in an interview, a design review or a conversation with the utility. This section states the answer in the order it should be given, and then answers the standard follow-up questions, each with the table or equation that supports it.

18.6.1 The structure to state first

The answer begins with structure, not with equations. Five sentences cover the case, in the order in which the work actually proceeds.

Key idea

1. Define the service precisely: power, duration and reliability – here 80 MW, three hours, and a reliability level the request leaves unstated.
2. Forecast *sustainable* flexibility per resource type over the window, as a profile, not instantaneous capacity.
3. Track the state variables that constrain future flexibility: thermal state, charging slack and departure times, state of charge.
4. Optimize over the whole horizon, working backward from the binding period, so that valuable resources are not exhausted early.
5. During the event, use telemetry and verification to update the forecasts and re-optimize, with the fastest resource absorbing tracking error.

Mathematics follows if it is asked for. Stating the structure first is not a presentational preference: the four moves of Section 18.3.6 are visible from the structure and invisible from any single equation.

18.6.2 What to forecast, and what to optimize

The forecasting question has a specific answer, and it is not system load. Four objects are needed:

- thermostat flexibility as a function of time and required duration, $F_{\mathrm{T}}(t, \tau)$, since the 50 MW of Section 18.2.1 is only true for τ near zero;
- vehicle flexibility $F_{\mathrm{E}}(t, \tau)$, driven by plug-in counts, energy needs and departure times, hence by the urgency distribution of the fleet;
- battery state of charge $\mathrm{SOC}(t)$ and the capability $C(\tau)$ of equation (18.6) that follows from it;
- the portfolio distribution $p\big(F_{\text{fleet}}, t{:}t+H\big)$, jointly across resources and time, because the margin calculation of equation (18.8) needs a covariance and not three separate bands.

The optimization question has an equally specific answer. The full problem behind Table 18.6 is

$$\min_{u} \sum_{t} \Big[\lambda_1 \big(P_{\mathrm{VPP}}(t) - P^{\mathrm{tgt}}(t)\big)^2 + \lambda_2 C^{\mathrm{comfort}}(t) + \lambda_3 C^{\mathrm{battery}}(t) + \lambda_4 C^{\mathrm{EV}}(t) \Big] \tag{18.13}$$

subject to, for every device i and every step t,

$$T^{\min} \le T_i(t) \le T^{\max}, \qquad \mathrm{SOC}_i(T_i^{\mathrm{dep}}) \ge \mathrm{SOC}_i^{\mathrm{req}}, \qquad \mathrm{SOC}^{\min} \le \mathrm{SOC}_i(t) \le \mathrm{SOC}^{\max},$$

together with the state-transition map of each device and, when the service carries a reliability level, the chance constraint

$$\Pr\big(P^{\mathrm{del}}(t) \ge P^{\mathrm{tgt}}(t)\big) \ge 0.95. \tag{18.14}$$

The four cost terms are the tracking error of equation (18.4), the customers' comfort, battery degradation and the inconvenience imposed on drivers; the weights λ_k express how much of each the operator will trade for a megawatt. The constraint list is where the case's difficulty lives: remove the comfort band and the energy budget and the problem is trivial.

18.6.3 Model choice and evaluation

Asked which models to use, the answer does not begin with an architecture. It begins with four questions: what exactly is being forecast, over what horizon, with what data available at prediction time, and what the operational loss of an error is. Only then does a model ladder apply, and Table 18.10 gives it in the order a project should climb it. Each rung must earn its complexity against the rung below, measured on the operational metric rather than on fit.

Table 18.10: The model ladder for flexibility forecasting. A rung is justified only when it beats the rung below it on the operational metrics of Table 18.11.

Rung	What it adds	What it costs
Persistence	a defensible floor: yesterday's or last week's same hour	nothing; fails on weather changes
Historical average by hour and weather bin	seasonality and a crude weather response	needs a season of events
Linear regression on weather, calendar and recent telemetry	interpretable coefficients, easy monitoring	misses saturation and interactions
Gradient-boosted trees	nonlinearity, interactions, missing-value handling	no extrapolation beyond observed conditions
Recurrent or sequence models	temporal dependence and state-like memory	data-hungry; opaque failure modes
Physics-informed or hybrid models	the RC dynamics and the energy budgets as hard structure, so forecasts respect Table 18.3	most engineering effort; the usual right answer for profiles

Evaluation has three tiers, in Table 18.11, and a model can pass the first and fail the third. A forecast that is unbiased on average but overconfident in its bands produces exactly the zero-margin plan of Section 18.4.1; a forecast whose errors are largest in the binding period is worse for this service than one with a higher overall root-mean-square error whose errors fall in the first hour. The best forecast is the one that supports the better decision.

18.6.4 Uncertainty and scale

Uncertainty enters the decision in one of four ways, all of which start from a distribution $p(F)$ rather than a number. The first is to dispatch against a quantile: plan on the fifth percentile of capability instead of the mean, which in the first hour of this case means planning on 64.4 MW rather than 80 MW. The second is the chance constraint (18.14), which is the same idea written as a requirement on the outcome. The third is scenarios: sample futures, as in Figure 10.8 of Chapter 10, and optimize the expected cost across them, which captures the fact that a dispatch choice now changes the capability later. The fourth is robust optimization against the worst case in an uncertainty set, appropriate when the penalty for a shortfall is severe and asymmetric.

Scale is the other standard question, and the answer is explicitly not to put a million devices into one solver. The hierarchy of Chapters 9, 13 and 16 is: devices are grouped into cohorts by state, location and behavior; each cohort is reduced to a small flexibility model – available power, available energy, sustainable duration, uncertainty, location; a portfolio optimizer works

Table 18.11: Three tiers of evaluation for a flexibility forecast. The third tier is the one the service is paid on.

Tier	Metrics	Failure it catches
Point accuracy	mean absolute error, root-mean-square error, bias by hour and by cohort	a systematically optimistic capability forecast
Probabilistic quality	interval coverage, quantile (pinball) loss, calibration of the predicted distribution	bands too narrow, which is what makes a chance constraint unsafe
Operational value	target-tracking error, frequency and size of delivery shortfall, customer impact (overrides, comfort excursions), revenue net of penalties	a model that is accurate on average and wrong in the binding period

on the cohorts and produces cohort targets; and local dispatch logic allocates each cohort's target among its devices. The case as solved is already in that form: the three groups of Table 18.1 are cohort-level objects, the thermostat cohorts A, B and C of Section 18.3.5 are a finer partition of one of them, and the optimization has nine variables rather than a million. Reduction is what makes the problem solvable, and the conditions under which the reduction is faithful are what Chapter 9 establishes.

18.6.5 Failure modes

A list of failure modes is more useful when it is organized by where in the system the failure occurs, because the mitigation lives at that place. Table 18.12 groups them into six families and names the instance of each that this case produced.

Table 18.12: Failure modes by family, with the instance each family produced in this case. The mitigation belongs to the layer named by the family.

Family	Failure modes	Instance in this case
Data	bad or missing telemetry, stale state, latency, clock skew	the 17:05 decision rests on telemetry that is minutes old
Forecast	weather error, drift, bias, overconfident bands	the hotter forecast at 17:30 and the 9.5 MW band
Physical and network	state exhaustion (battery depletion, vehicle deadlines), synchronized rebound, distribution constraints	the 45 MWh budget and the rebound priced into the rotated profile
Human	overrides, opt-outs, enrollment churn, comfort complaints	the override rate of 0.12 against a forecast 0.05
Control	tracking error, command delivery failures, actuation limits, control-loop interaction	the 5 MW error at 17:05
Verification	baseline bias, baseline uncertainty, misattribution between the four causes	$\Pr(\Delta P \geq 80) = 0.58$ on a nominal 82 MW

Two families deserve emphasis because they are the ones most often left out. Synchronized rebound belongs to the physical family and is created by the control system itself: ending an event for every device at once produces a load spike that can exceed the reduction, which is why the rotation of Section 18.3.5 staggers release. Overconfidence belongs to the forecast family and is the most dangerous single failure in the list, because it is invisible in point metrics and shows up only as commitments that cannot be kept.

18.6.6 The full framework

Every part of the case fits in one loop. At each step the platform observes telemetry, estimates the state, forecasts what the fleet will be able to do, optimizes against that forecast, acts, and measures the result:

$$\begin{aligned}
\text{observe:} &\quad y_t, && (18.15)\\
\text{estimate:} &\quad p(x_t \mid y_{1:t}), && (18.16)\\
\text{forecast:} &\quad p\big(F_{t:t+H} \mid y_{1:t}\big), && (18.17)\\
\text{optimize:} &\quad u^* = \arg\min_{u}\ \mathbb{E}\big[J(x,u)\big] \text{ s.t. } x_{t+1} = f(x_t, u_t, w_t), && (18.18)\\
\text{act:} &\quad u_t = u_t^*, && (18.19)\\
\text{measure:} &\quad y_{t+1}, \quad \text{and repeat.} && (18.20)
\end{aligned}$$

The optimize line shows only the dynamics; the full constraint list is the one written under equation (18.13), and the expectation is taken over the forecast distribution produced by the line above it.

The case exercised every line. The estimate step is the state depletion of Table 18.3; the forecast step is the profiles of Figure 18.2 with the bands of Section 18.4.1; the optimize step is equation (18.13) solved backward from the binding period; the act step is the schedule of Table 18.6; and the measure step closes the loop at 17:05, at 17:30 and again in the verification of Section 18.5.2.

18.6.7 What the question is testing

A case of this kind is not a test of one discipline. It is simultaneously a forecasting problem (the profiles and their bands), an optimization problem (equation (18.13) with a binding period), a controls problem (receding-horizon replanning and tracking error), a physical-system problem (thermal stores, deadlines, state of charge) and a production-software problem (telemetry, latency, acknowledgment, cadence). An answer that supplies only a model architecture misses the energy budget, the binding period and the verification; an answer that supplies only optimal-control equations misses the override rate, the baseline uncertainty and the fact that the request was infeasible as forecast. The interesting work, and the work this book has described, is at the intersection.

18.7 What to retain

Definition 18.1 (Designing and operating a virtual power plant for a service)**.** Serving a stated request from a portfolio of distributed resources is the problem of allocating depletable states over a horizon. Define the service as power, duration and reliability; forecast each resource as a sustainable profile with a band; find the binding period and plan backward from it; create capability there by managing state before the event; hold a reserve sized by the forecast band; and re-optimize from measurements as the event proceeds.

The four state-management moves that made this case feasible are collected in Table 18.13, with the chapter that established each. None of them is an optimization method, and together they moved the binding hour from 65 MW of capability to 80 MW.

Table 18.13: The four enabling moves, their effect in the case, and where each was established.

Move	What it changes	Effect in the case	Ch.	Section
Pre-conditioning	lowers T_{in} before 17:00, adding thermal store	hard-hour thermostats $20 \rightarrow 30$ MW	1, 5	18.3.3
Cohort rotation	curtails part of the fleet at a time, so each house stays inside the band longer	profile $50/35/20 \rightarrow 45/40/30$ MW	5, 12	18.3.5
Charging advancement	removes evening energy obligation, lowering urgency	hard-hour vehicles $20 \rightarrow 25$ MW	7	18.3.3
Battery energy budgeting	treats 45 MWh as a budget allocated across the horizon	battery energy $10/20/25 \rightarrow 0/15/25$ MWh	6, 11	18.3.4

The track as a whole reduces to one picture, Figure 18.8. The inner panel is the chain this book followed: a distributed energy resource has state-dependent flexibility; flexibility aggregates over a fleet; the aggregate is forecast; the forecast is optimized into a schedule; the schedule shapes load dynamically; and the shaped load is a grid service. The six boxes around the outside are the operating loop that executes that chain repeatedly – observe, forecast, optimize, dispatch, verify, learn – with each pass re-planning the chain from new measurements. Every section of this chapter is a point on one of the two: the resource models and the schedule are the chain, and the 17:05 correction, the 17:30 revision and the verification are the loop.

18.8 Exercises

Exercise 18.1. The utility asks for 90 MW instead of 80 MW, over the same window, with the same portfolio and the same four moves. Recompute the binding hour, state the shortfall, and compute the battery energy a 90 MW schedule would require if thermostats and vehicles are dispatched at their capabilities in all three hours. Is the request feasible, and if not, what is the largest flat commitment that is?

Exercise 18.2. A hotter day raises the thermostat fleet's instantaneous capability but shortens its sustainable duration, giving the rotated profile $52, 42, 26$ MW in place of $45, 40, 30$ MW. Recompute the binding hour, decide whether the 80 MW schedule of Table 18.6 survives, and say which of the two effects of Section 18.4.5 dominates for this service.

Exercise 18.3. Pre-cooling is not permitted, so the rotated thermostat profile is $45, 40, 22$ MW. Keeping the vehicle and battery figures of Table 18.6, find the largest flat three-hour commitment the portfolio can serve, give the schedule that achieves it, and verify that it respects the 45 MWh battery budget.

Exercise 18.4. With the delivery standard deviation of 9.5 MW from equation (18.8), compute the margin required for delivery probabilities of 0.95 and 0.99, and hence the largest commitment the binding hour can carry at each level given its 80 MW of capability. Then repeat the 0.95 case with a correlation of 0.3 between the thermostat and vehicle errors, and comment on which of the two assumptions – the reliability level or the independence – costs more megawatts.

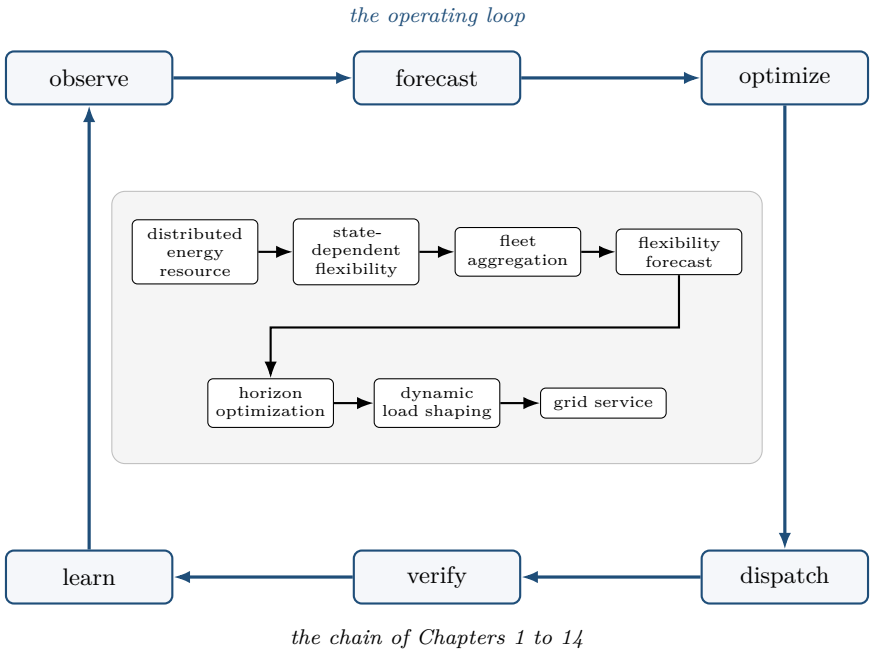


Figure 18.8: The Track B mental model: the seven-stage chain from a device to a grid service, inside the six-stage loop that operates it. The chain is a modeling statement, read left to right and then wrapped; the loop is an operating statement, executed repeatedly at the three cadences of Figure 16.5.

18.9 Connection to the TERRA framework

The whole case is one small graph, carried forward in time and conditioned on measurements as they arrive.

- **The case is one node with three component populations.** Equation (18.3) is a conservation statement at a single node: the reductions delivered by the thermostat, vehicle and battery populations must sum to the flow the utility requested. Each population is a set of components with its own state – T_{in}, charging slack, SOC – attached to that node by an edge whose bound is the group's capability. The nine numbers of Table 18.6 are three edge flows at three times.

- **The capability profiles are bounds that depend on component state.** The decaying profiles of Figure 18.2 are not exogenous data; they are the edge bounds induced by the component states, which is why they change when the states are managed. Pre-conditioning and rotation are interventions on states, and the profile change of Figure 18.3 is their consequence.

- **Each iteration is a forward simulation.** Every row of Table 18.7 is a candidate control trajectory integrated forward through the component dynamics and tested against the bounds. The violated constraint reported in the third column is the output of that integration: a power bound in iteration 1, an accumulator bound in iteration 2, a state-dependent bound in iteration 3.

- **The binding-period test is a feasibility query.** Equation (18.7) asks whether the target lies inside the feasible set of flows at one time given the states reachable at that time. Answering it before choosing any schedule is the correct order of operations, and it is a query about the graph, not about an optimizer.

- **Reserves are posterior quantiles.** The band $\sigma = 9.5\,\mathrm{MW}$ of equation (18.8) is a predictive distribution obtained by propagating the posterior over component states and the disturbance forecast through the model. The reserve of Table 18.8 is sized from a quantile of that distribution, and the chance constraint (18.14) is a statement about the same quantile; the correlation correction is the off-diagonal term of the covariance that a joint posterior carries and three separate bands do not.
- **The 17:05 correction is conditioning.** The measurement $(39, 31, 5)$ updates the posterior over the component states; the battery step is an action taken under the updated belief; and the re-optimization of Section 18.4.4 is a new forward solve from it. The override revision of equation (18.10) and the verification of equation (18.12) are the same operation on slower timescales, acting on a behavioral parameter and on a baseline model rather than on a state.

In summary, the case is a conservation node with three populations of bounded, depletable components; a posterior over their states; a forward prediction of what they can deliver; a decision taken against that prediction and checked in its binding period; and a sequence of conditioning steps as telemetry, weather and behavior arrive. That is also a description of every chapter that preceded this one, which is the reason the capstone needed no new machinery. The sequence ends here: what the reader carries forward is the habit of asking, of any request, which state depletes, when it binds, and what must be done beforehand.