SCALE WITHOUT SURPRISE

Visual Companion

Managing Demand, Risk, and Scale

Timothy O’Brien

A shorter route into the book

The main ideas, selected explanations, practical decisions, and chapter diagrams. Follow the links when you need the full argument, worked examples, and qualifications.

Draft edition. Preliminary chapters remain preliminary here.

https://withoutsurprise.com/briefs

Copyright © 2026 Timothy O’Brien. All rights reserved.

Contents

Seasonal Variation 3

The Long Tail and the Spike 7

A Portfolio of Customers 10

The Relationship Discipline 13

Reserve Capacity 16

Managing Risks and Anomalies 19

Commitments Between Organizations 27

Multiple Markets 31

The Monthly Capacity Review 37

What Scale Brings 39

Markets for Compute 42

CHAPTER 14 / 3 MINUTE BRIEF

Seasonal Variation

Read the full chapter → / Read this brief online →

The main idea

When someone asks for a better capacity forecast, I'd first want to understand what's wrong with the one they have. Perhaps the model handles seasonality poorly, or perhaps the organization keeps changing the conditions the model was supposed to describe. If the launch moved, the promotion expanded, or a new feature changed the kind of work each customer performs, those are different problems from failing to recognize a recurring pattern in the measurements.

I want a good model, and I want to know how well it performs, but I also want access to the people who can explain what the business intends to do. Otherwise we can spend a lot of time improving our estimate of what would have happened if the plan had stayed the same. The model may be doing an excellent job within its assumptions while the capacity team prepares for the wrong workload.

There is real mathematical work here, and we'll do some of it. A forecast earns trust when it is scored against what was known on the day it was made, in the unit of the decision it served, with the business's information kept beside the model.

Explore this section in the full chapter →

Start with a comparison we can understand

A **seasonal-naive forecast** gives us a useful starting point: it repeats the latest observation from the matching season, so next Monday's forecast is the most recently observed Monday. Hyndman and Athanasopoulos treat it as a standard benchmark in *Forecasting: Principles and Practice*, with a blunt rule: a new method that can't beat these simple alternatives isn't worth considering.[1]

Explore this section in the full chapter →

Recreate what you could have known

The evaluation has to recreate what we knew when each forecast would have been made: choose a date in the past, forecast using only the information available then, compare the result with what happened, then move the date forward and repeat. Each starting date is a **forecast origin**; the repeated comparison, which can test tomorrow's forecast or one made weeks ahead, is **rolling-origin evaluation**.[2]

Our companion starts after four weeks of history, advances the origin weekly, and compares both methods at one, seven, eight, fourteen, and forty-two days. Eight days corresponds to the constructed launch decision; forty-two days corresponds to the initial supply estimate. I'd want to see the latter before relying on the model for purchasing: predicting tomorrow well says little about a decision whose supply consequence arrives six weeks later. That horizon also has the fewest completed cases, seven, and their overlapping six-week windows make them far from independent.

Explore this section in the full chapter →

[1] Rob J. Hyndman and George Athanasopoulos, *Forecasting: Principles and Practice*, third edition, section 5.2, "Some simple forecasting methods," https://otexts.com/fpp3/simple-methods.html, accessed October 3, 2026. Defines the seasonal-naive method, including its general form for horizons beyond one season, and says a new method that does not beat these benchmarks is not worth considering. The four-week matching-weekday mean and numerical fixture are constructed for this book.

[2] Hyndman and Athanasopoulos, *Forecasting: Principles and Practice*, section 5.10, "Time series cross-validation," https://otexts.com/fpp3/tscv.html, accessed October 3, 2026. Describes evaluation on a rolling forecasting origin, including multi-step horizons. The companion's weekly origins and evidence-availability dates are this book's adaptation.

Put the conversation next to the model

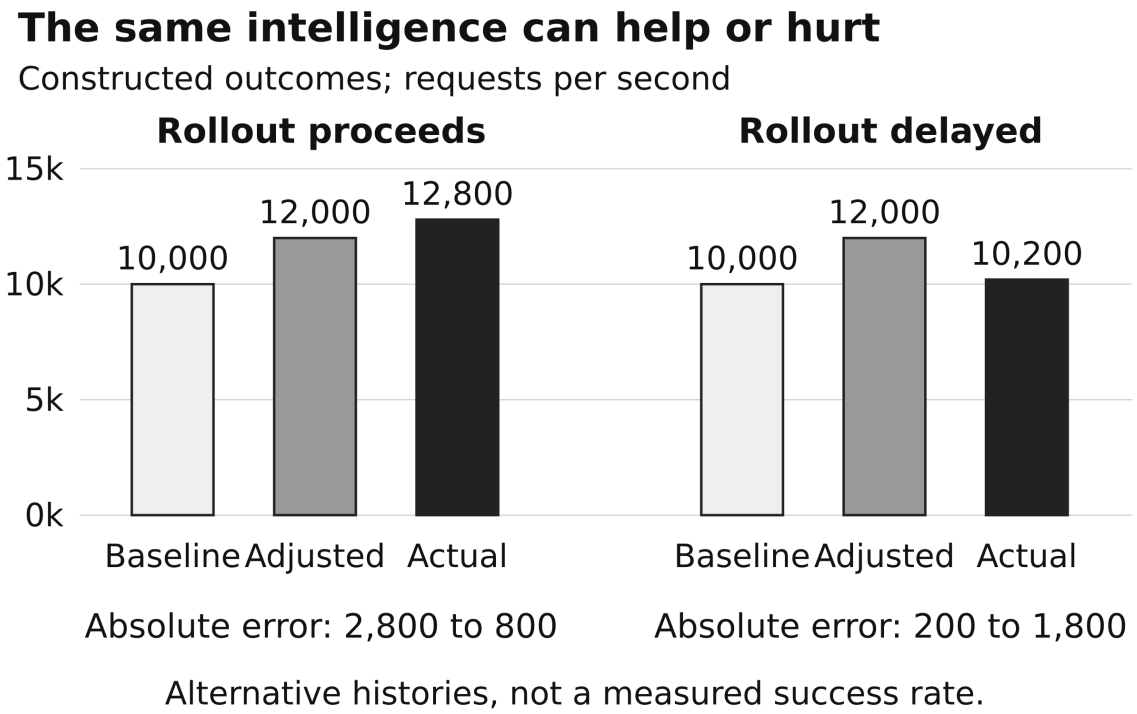


Diagram 14.1. The same event adjustment reduces error when the rollout proceeds and increases error when it is delayed. The bars show two alternative constructed outcomes.

Explore this section in the full chapter →

What the uncertainty changes

A **prediction interval** gives a range with a stated probability under a model and its assumptions. We can check whether a claimed 90 percent interval contains roughly 90 percent of later outcomes at the horizon we care about. A low-demand and a high-demand business scenario give us a different kind of range.[3] Both can help a decision, but our staged 12,000 and full-rollout 14,000 cases don't become the 50th and 95th percentiles merely because those labels would make the chart easier to present. We haven't established those probabilities.

Explore this section in the full chapter →

The discussion becomes useful when we connect those mechanisms to options. A delayed rollout may let us defer the purchase. Delayed supply may mean staging longer, if the business and batch owner can accept the consequences. A changed request mix requires someone qualified to reassess the service rate. Where we have defensible probability estimates, they can inform those choices. Where we don't, the alternatives and their conditions still deserve to be examined.

I sometimes call forecasting magic, partly because of how much judgment can disappear behind the final number. I'd rather let the reader see that judgment: why we chose the historical comparison, what we learned from the business, and where we still have to make an assumption before the decision deadline. That doesn't eliminate guessing. It gives us something specific to revisit when the world behaves differently.

Explore this section in the full chapter →

[3] Hyndman and Athanasopoulos, *Forecasting: Principles and Practice*, section 5.5, "Distributional forecasts and prediction intervals," https://otexts.com/fpp3/prediction-intervals.html, accessed October 3, 2026. Supports the distinction between point forecasts, forecast distributions, and prediction intervals; this book's rollout scenarios are not calibrated intervals.

CHAPTER 15 / 3 MINUTE BRIEF

The Long Tail and the Spike

Read the full chapter → / Read this brief online →

The main idea

If the largest twenty projects account for most of your capacity, you're going to spend a lot of time with those teams: their plans, the technologies they've chosen, and which assumptions are likely to change. That is a sensible use of a capacity team's attention, particularly when a decision in any one of those projects could affect everyone sharing the same constrained resources.

The difficulty is what that understanding allows you to say about the rest. Suppose the twenty have individual forecasts and regular conversations with the capacity team, while another thousand projects appear under "Other" in the review. We might know those smaller projects' current consumption quite well without knowing what could make them change together, and that gap matters even when the aggregate number looks reassuringly small.

Closing that gap doesn't take a thousand more customer meetings. Small projects rarely threaten the plan one at a time; they threaten it when something they share, such as a platform default, a common rollout, or a business event, moves many of them at once. That is the level at which to give the remainder an owner and an investigation threshold.

Explore this section in the full chapter →

Measure concentration before naming its distribution

If we want to make a power-law claim, there's more statistical work to do. One concentration figure can't settle it: some power-law exponent will reproduce any single point on the curve, and so will some log-normal. Clauset, Shalizi and Newman showed that the familiar test, a straight line on a log-log plot, can badly misestimate the exponent and says nothing about whether the data follow a power law at all. Their procedure estimates where the power law starts and its exponent, tests the fit, and compares alternatives; the log-normal often can't be ruled out.[4] I'd keep that work separate from ranking the portfolio, which has to hold up whatever the distribution turns out to be.

Explore this section in the full chapter →

A small change can spend the remaining headroom

Assume we have 1,030 units of usable supply under the scenario's operating conditions. Against baseline demand of 1,000, that leaves 30 units for forecast error and change. The small projects average 200 / 1,000 = 0.2 units each, so an additional 0.02 per project could look unremarkable to anyone reviewing individual consumers.

Now suppose that increase comes from a shared configuration change reaching all thousand small projects. Its aggregate effect is 1,000 × 0.02 = 20 units: 10 percent of tail demand, 2 percent of total demand, and two-thirds of our available headroom.

Scenario	Large projects	Small projects	Total demand	Supply minus demand
Baseline	800	200	1,000	30
Small projects each add 0.02	800	220	1,020	10
Large projects also grow 2%	816	220	1,036	-6

[4]Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review* 51(4), 661–703 (2009), https://doi.org/10.1137/070710111; preprint https://arxiv.org/abs/0706.1062, checked October 3, 2026. Supports the weakness of least-squares fits on logarithmic plots, the maximum-likelihood, goodness-of-fit and likelihood-ratio procedure, the difficulty of ruling out the log-normal, and exponents typically between 2 and 3. It says nothing about the distribution of any computing workload.

A close forecast for the large projects doesn't protect this plan: 2 percent growth in the head, on top of the tail change, exhausts the margin. Failures and bursts beyond these stated conditions need their own test even when the last column is positive.

Explore this section in the full chapter →

Put an investigation threshold next to the headroom

Response time should influence the threshold. Quickly obtainable, qualified capacity may give us room to learn from a small change. A supply process taking months requires earlier action, and the test should then count everything that could land within the lead time, from scheduled shared changes and expected arrivals to the cohort's recent bias. That's why a low spending rank doesn't settle whether a workload needs investigation: its contribution may be small in the budget and large relative to the room we have left before we can respond.

Once we understand the proposed mechanism, we can discuss an action with someone able to carry it out: stage the common change, obtain supply, qualify an optimization, or accept the exposure through the appropriate business owner. If the application teams can't influence the shared change, telling a thousand of them to use less would leave the relevant decision untouched.

Explore this section in the full chapter →

CHAPTER 16 / 3 MINUTE BRIEF

A Portfolio of Customers

Read the full chapter → / Read this brief online →

The main idea

One of the choices a capacity team has to make is how much work to ask of its customers. A major project that consumes a substantial part of a constrained pool may need recurring conversations with both the engineering team and the people who understand the business, and asking those questions is part of supporting it. Asking the same of every application on a shared platform would spend everyone's time, theirs and ours, on conversations that change nothing.

For a thousand small consumers, I may get much further with a platform owner who understands adoption, can change common defaults, and knows how to reach the members when something unusual happens. Those customers don't need to become capacity planners; the team providing the service may be best placed to understand what they will ask of it.

An experiment needs something else again: close attention for a while, as the organization learns what the system does and decides what it might become. I wouldn't describe that as less mature than a stable forecasting relationship; it suits a different kind of uncertainty. Specialist attention is the scarce resource in all three cases, and this chapter's process spends it deliberately, deciding who gets an individual forecast, who shares a group forecast, and what moves a consumer between them.

Explore this section in the full chapter →

3. Assign a planning treatment

Treatment	What the team maintains	Relationship needed
Individually managed projects	Workload model, business drivers, scenarios, commitments, and forecast revisions	Named technical and business owners
Cohorts of smaller projects	Membership, aggregate baseline, variation, shared drivers, and project-level exceptions	A cohort owner who can reach member teams
Temporary exceptions	A time-bounded investigation of a new, changing, or consequential workload	The person who can resolve the specific uncertainty

Explore this section in the full chapter →

4. Gather intelligence at the appropriate level

- **Record:** event `E-ROLLOUT`, revision 1, known June 25, three days before the forecast it informs.
- **Change and source:** product proposes wider customer access; the product lead supplied the plan.
- **Boundary:** additional demand on the web worker pool, measured as requests per second at the existing request mix.
- **Alternatives:** a full rollout adds an assumed 4,000; a staged rollout adds 2,000. Timing and scope remain negotiable.
- **Overlap:** zero of this increment is assumed to be present in the historical baseline used for this decision.
- **Dependencies:** the rollout proceeds on the stated schedule, the measured request mix remains representative, and the transferred capacity is qualified before expansion.
- **Review and expiry:** review by June 30; no forecast made after July 7 may use this revision.

Explore this section in the full chapter →

6. Promote, investigate, and release

A project or cohort needs closer attention when its growth, forecast error, contribution, or consequence crosses the threshold we've agreed to use. A small workload can qualify before its spend is large, by growing quickly, supporting a critical transaction, or needing a resource with a long lead time, and a suspected shared driver can qualify a cohort before any member is large. Sometimes, in the monthly meetings, a very small project that nobody thinks is a big deal starts giving signals that it is bigger than even its own team thinks. It may not be asking for capacity today, but it is going to need it. In each case I'd want a person responsible for resolving the uncertainty and a deadline tied to the decision, rather than an indefinite designation that the project is being watched.

Explore this section in the full chapter →

There also needs to be a way out of the special treatment. A system can stabilize enough to join an existing cohort, or an investigation can resolve the uncertainty that justified individual attention. We should agree what would bring it back, then release the extra effort. I'd set the release condition below the promotion trigger and require it to hold for two consecutive reviews, so a consumer near the line doesn't change treatment every month, and I'd return its history to the cohort baseline and its open events to the cohort owner. Otherwise every temporary exception becomes permanent, and the central team's limited attention goes to maintaining arrangements nobody has reconsidered.

Explore this section in the full chapter →

CHAPTER 17 / 3 MINUTE BRIEF

The Relationship Discipline

Read the full chapter → / Read this brief online →

The main idea

Whoever can warn us about a change in demand is probably doing something other than waiting for the capacity team's questionnaire. They may be trying to close a sale, decide which customers get a feature first, or work out whether a migration is going to finish at all. Your deadline is another deadline in that person's day, and before answering, they have a perfectly reasonable concern: what are you going to do with the answer?

That concern usually has a history. If what they told us last time was later held against them, or reporting a change earned criticism instead of help, we have taught them that keeping quiet is easier. We can call this a communication problem, but we helped create it.

Most of these people do not report to the capacity team, and nobody can be ordered to mention a cancellation they are still arguing about. What we control is what happens to the information once it reaches us, and that decides whether we hear about the next change in time.

Explore this section in the full chapter →

Make early information worth sharing

I would begin by asking what decision the customer is trying to make, then work backward to the information that could improve our answer. An earliest plausible date, the work likely to change, a credible range, and somebody who can update it may be enough to preserve an option. A central template can help us remember these questions; it should not turn every empty field into a reason to send the request back. If I cannot explain how a field affects this decision, I should be careful about making someone else spend an afternoon filling it in.

The return trip matters just as much. Show the customer the supply range or conditional cost estimate their information made possible, including the part we still cannot answer. After the event, come back with what happened and what we learned, even if their adjustment made the forecast worse. Otherwise, from their perspective, information disappears into a process that sends another questionnaire next month.

Explore this section in the full chapter →

Work through the specialists

A capacity team needs access to specialists who have lived with their systems, although hiring one person who has suffered through a difficult database does not somehow give us mastery of all databases. The tools in the earlier chapters can make it tempting to imagine a central team measuring everything independently. At enterprise scale, I would build partnerships with the people who understand the streaming platforms, databases, AI infrastructure, and applications, then hire central staff who know enough to ask them worthwhile questions. Those are already demanding jobs.

I want the specialists to disagree with me when my question is incomplete. If I ask how much throughput a database can support, the useful answer may begin with a discussion of maintenance that I had not included. A test with maintenance paused can tell us something about foreground throughput, but if the customer needs both activities to happen, we have another experiment to discuss. Similarly, an AI team may have measured how many requests meet a latency target without yet evaluating answer quality. I would rather learn that distinction from the team than discover, after making a commitment, that I had combined two results that never held at the same time.

Explore this section in the full chapter →

When the answer has a sponsor

Claim	Question that makes it inspectable
“Efficiency improved.”	Is the useful outcome unchanged? Were workload mix, latency, quality, durability, maintenance, and recovery conditions comparable?
“This is the demand forecast.”	Is it a forecast, sales ambition, funding request, or commitment? Who supplied the assumptions, and are contrary scenarios preserved?
“We saved money.”	Was cost removed, deferred, reallocated, or avoided in a hypothetical future? Are migration, parallel running, remaining commitments, and shared costs included?

Explore this section in the full chapter →

CHAPTER 18 / 3 MINUTE BRIEF

Reserve Capacity

Read the full chapter → / Read this brief online →

The main idea

Show me a large amount of idle capacity and I want to understand it. I may also want to leave it alone, at least until someone can explain what happens if we take it away. That can be an uncomfortable answer in a cost review, especially when the chart appears to have done the investigation for us: here are the resources, here is the low utilization, and here is the proposed saving.

Perhaps the resources are waiting for a recovery, a customer deadline, or maintenance whose timing is difficult to predict. Perhaps the event that justified them has already passed and nobody has revisited the hold. Both explanations are possible, which is why I begin with the work the reserve protects and the conditions under which it becomes useful. Calling it insurance does not establish that it is adequate; calling it waste does not establish that we can release it.

Explore this section in the full chapter →

Available for what?

Return to the constructed portfolio's web service, whose generic failure groups do not specify a provider's regional or zonal layout. It has 64 qualified workers in four groups of sixteen, at a synthetic 250 requests per second per worker for the specified workload mix and quality. That gives us 16,000 in normal operation and 12,000 after losing one group, against a baseline peak of 10,000.

As Volume I, Chapter 2 showed, a report comparing only normal capacity with demand shows 6,000 requests per second available, but 4,000 of that room disappears with the failed group, leaving 2,000 under the condition we have promised to support. If we offer the whole 6,000 to a new customer, we have spent part of the allowance for failure without deciding to change that promise.

Explore this section in the full chapter →

Borrowed capacity changes the failure calculation

Configuration	Normal (requests/s)	After losing the largest group (requests/s)	Peak-use ceiling
64 workers, four groups of 16	16,000	12,000	75%
72 workers, four groups of 18	18,000	13,500	75%
72 workers, one group of 24 and three of 16	18,000	12,000	66.7%
80 workers, four groups of 20	20,000	15,000	75%
80 workers, five groups of 16	20,000	16,000	80%

The constructed portfolio assumes these rates were tested in each configuration and that a lost group's traffic spreads evenly over the survivors; adding workers to a real system does not establish a proportional increase in throughput. The eighty-worker rows are future configurations that need new supply and the batch workers' return; the last also needs a qualified fifth group.

Explore this section in the full chapter →

Borrowing is an agreement with an end

Volume I, Chapter 4's proposal borrows the workers for thirty days, beginning when they are ready on day two. That covers day 2 up to, but not including, day 32. New supply is estimated to be qualified and ready on day 42, leaving ten days the launch calculation has not covered.

Until day 32 the service survives a group loss at 13,500 requests per second. When the workers go back, it returns to 12,000: exactly the staged demand, with nothing left for forecast error, and 800 short if demand reaches 12,800, the peak in Volume I, Chapter 4's first constructed outcome. That exposure lasts ten days, or thirty-eight if supply slips to day 70. Keeping the workers would close it by changing somebody else's commitment.

Explore this section in the full chapter →

I want the next decision scheduled while alternatives remain available: renew the agreement, qualify another source, constrain demand, or bring the unmet obligation to someone who can change it. Work back from day 32 by the lead time of the slowest alternative we still want. Reminding everybody on day 32 leaves them solving the problem on the day the solution is required.

Explore this section in the full chapter →

Decide what would permit release

A reserve can outlive its original purpose remarkably easily if nobody defined an ending. A launch finishes, a migration changes, a customer departs, or the recovery design improves, while the allocation looks exactly as it did before. When authorizing a hold, agree what evidence would permit its release so that the next review can ask a concrete question.

Explore this section in the full chapter →

CHAPTER 19 / 4 MINUTE BRIEF

Managing Risks and Anomalies

Read the full chapter → / Read this brief online →

The main idea

Earlier in my career, I had a small business building a search engine. We had a primary system and a backup, which was considerably less reassuring after lightning struck the building and took out both. That was effectively the end of the business. I went back to school.[5]

The story takes a few sentences now. What mattered then was that we had prepared for a narrower problem than the one that occurred: having a backup had not given us a recovery we could perform after losing the place that contained it.

That experience is behind the way I use the word **anomaly**: an event outside the ordinary operating assumptions of the business, one that changes what the capacity plan can promise. No capacity team can predict these events, and I would not ask one to try. I would ask it to know, before one arrives, which decisions will be its own, what each would cost, and how long each would take.

Explore this section in the full chapter →

[5]Tim O'Brien, *Leaving the Cloud Isn't for Everyone*, October 31, 2025, "The Risk Premium You're Paying Cloud to Carry." The passage is the author's recollection.

Force: the kinds of events that hit the business

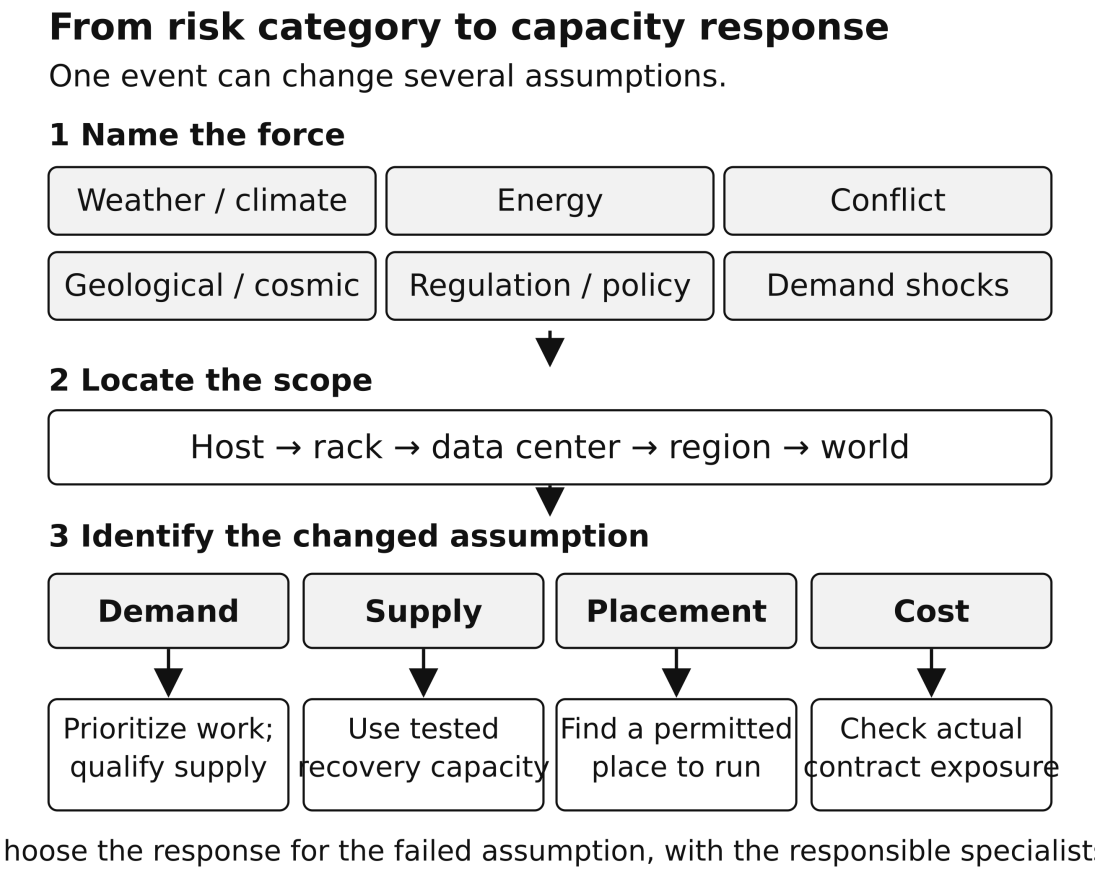


Diagram 19.1. Classify an event by force and scope, then identify which capacity assumptions it changes. The same event can affect demand, supply, permitted placement, and cost; each calls for a different response with the responsible specialists.

Explore this section in the full chapter →

Weather, climate, and the map you inherit

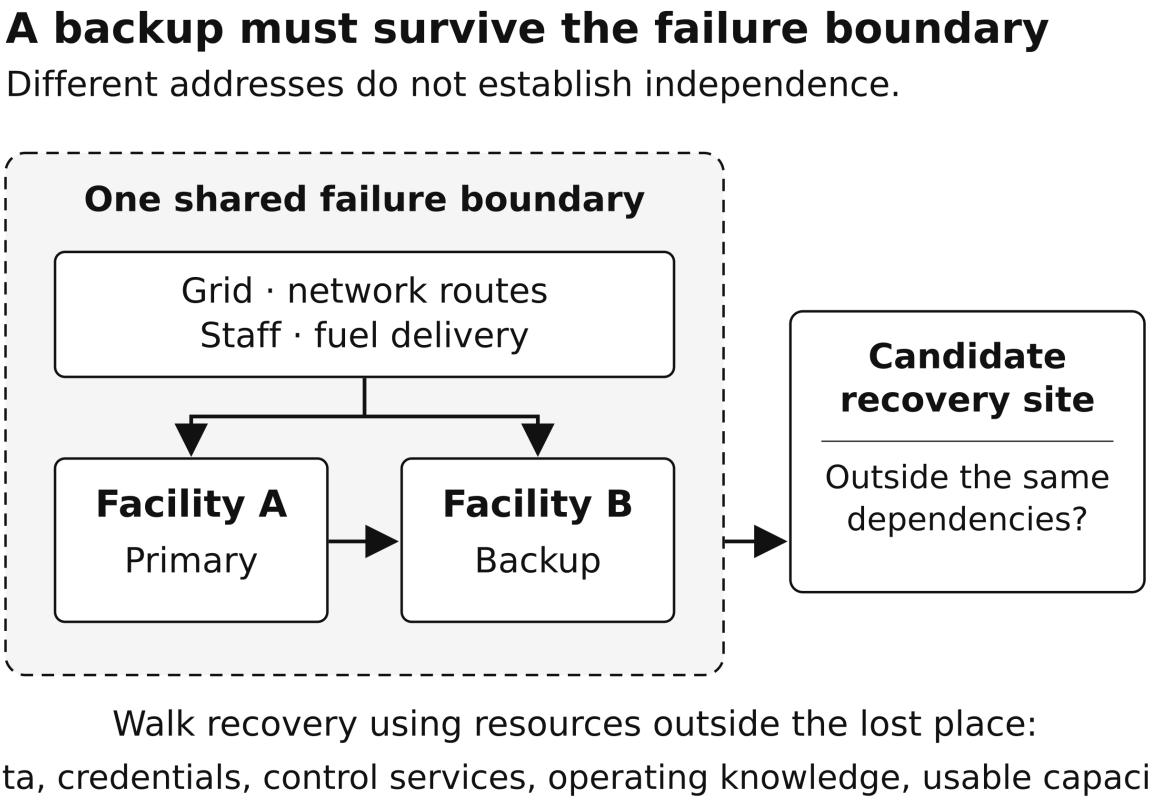


Diagram 19.2. Different addresses do not establish independence. This schematic places two facilities inside one shared failure boundary; a candidate recovery site still needs its dependencies checked and its recovery resources qualified.

Explore this section in the full chapter →

Solar storms: warning and preparation

Solar flares and coronal mass ejections are different phenomena; for the power grid, the one that matters is an Earth-directed coronal mass ejection, or CME. Forecasters can watch a CME leave the Sun. The fastest reach Earth in 15 to 18 hours, and slower ones take several days. Spacecraft near the L1 point, about 1.5 million kilometers toward the Sun, measure the disturbance as it passes, which often gives 15 to 60 minutes of warning.[6] That last interval is far too short to provision and qualify replacement infrastructure. Compare the warning with the actual recovery time, and hold in advance whatever the promised response needs.

Explore this section in the full chapter →

[6]NOAA SWPC, Coronal Mass Ejections, checked October 4, 2026: the fastest Earth-directed CMEs arrive "in as little as 15-18 hours," slower ones in several days; DSCOVR at L1 "can often provide 15 to 60 minutes advanced warning." Typical intervals, not guarantees.

Meteors and the loss of a place

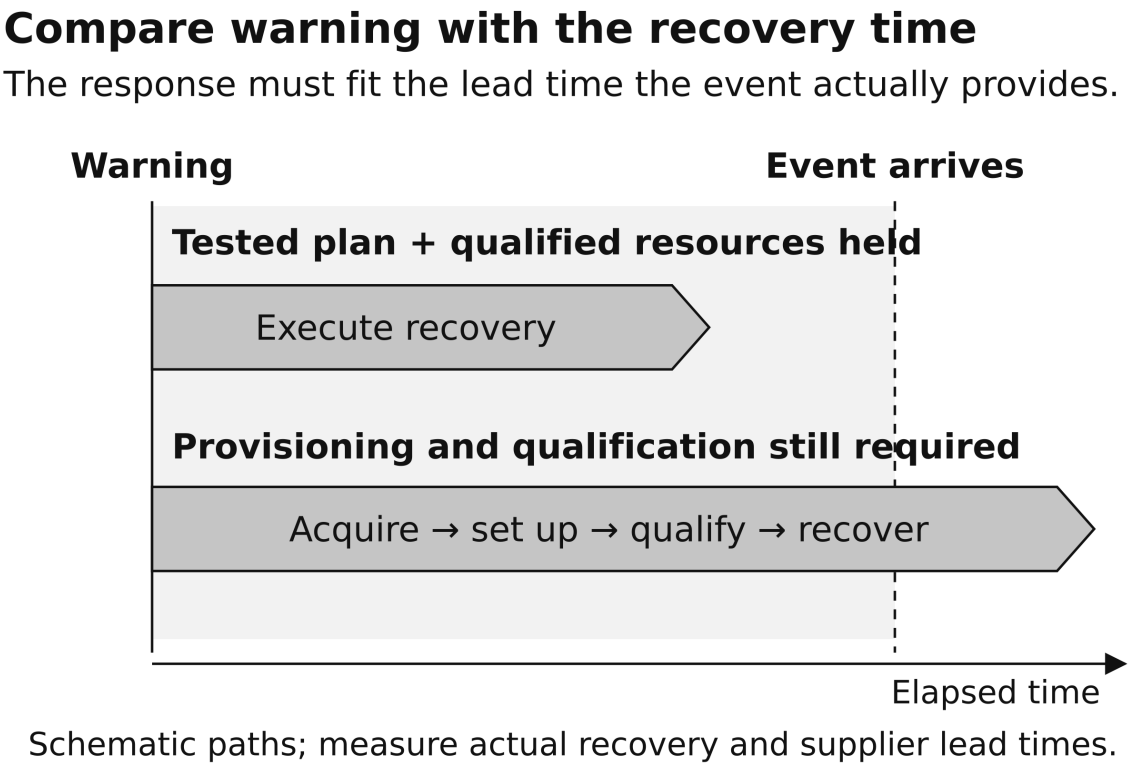


Diagram 19.3. Compare the warning available with the work still needed to recover. These schematic paths illustrate why resources that still need acquisition and qualification cannot be treated as ready capacity; actual response and supplier lead times must be established for the promised recovery.

Explore this section in the full chapter →

COVID, and the anomaly that changes demand

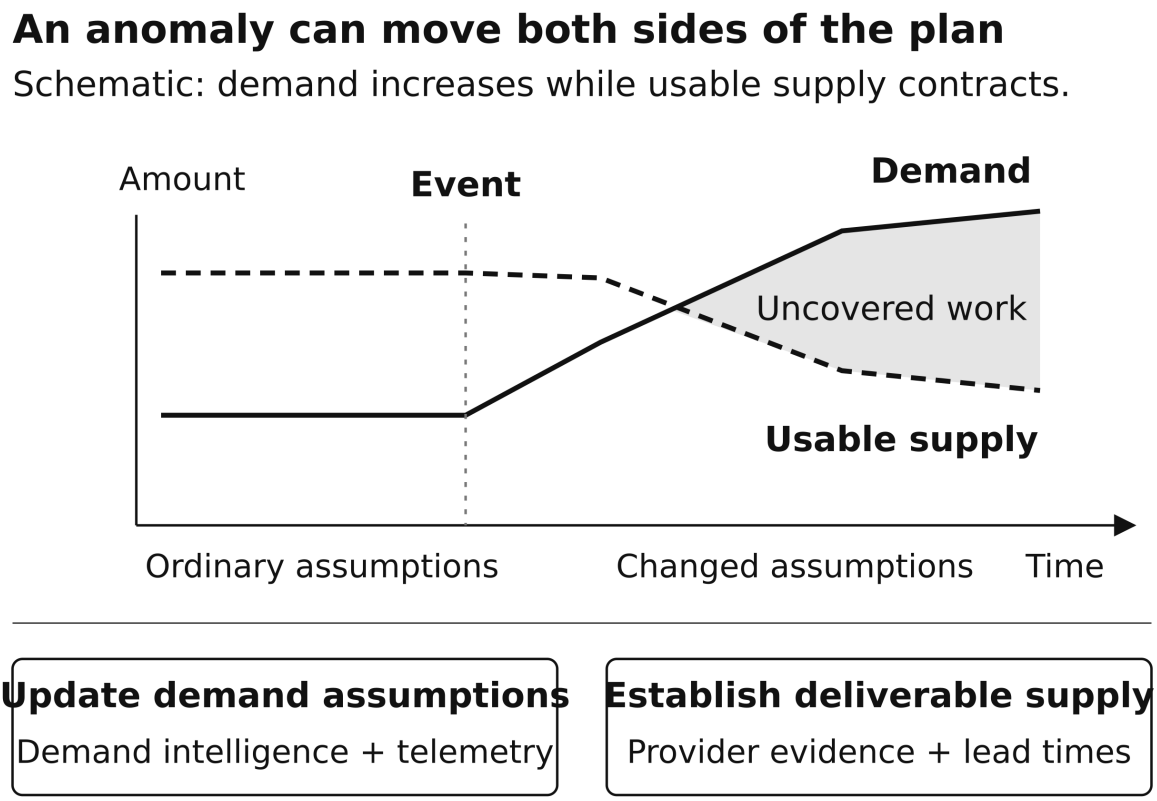


Diagram 19.4. An anomaly can change demand and supply together. This schematic, rather than a measured historical series, shows why updated demand assumptions and evidence of deliverable supply are separate parts of the response.

Explore this section in the full chapter →

The headroom you hold is the failover you can do

This is Chapter 18's failure-group arithmetic with regions as the groups. Take N equal regions sharing the work evenly, each able to sustain a qualified rate C, and let $u_{\max}$ be the highest fraction of C a survivor may carry after one region is lost. Normal utilization u must then satisfy

$$u \le \frac{N-1}{N}\, u_{\max}.$$

With two regions and $u_{\max} = 0.8$, the ceiling is 40 percent: two regions at $0.4C$ send $0.8C$ to the survivor, leaving 20 percent of its qualified capacity free. Letting the survivor run to its full tested limit raises the ceiling to 50 percent. Three regions with the same margin can each run at about 53.3 percent ($3 \times 0.533C \div 2 \approx 0.8C$); at two-thirds each, a loss leaves the survivors no headroom. These are upper bounds under stated assumptions, not utilization targets. Uneven routing, recovery work, a changed request mix, and shared failures each need their own calculation, and no percentage guarantees zero downtime.

Explore this section in the full chapter →

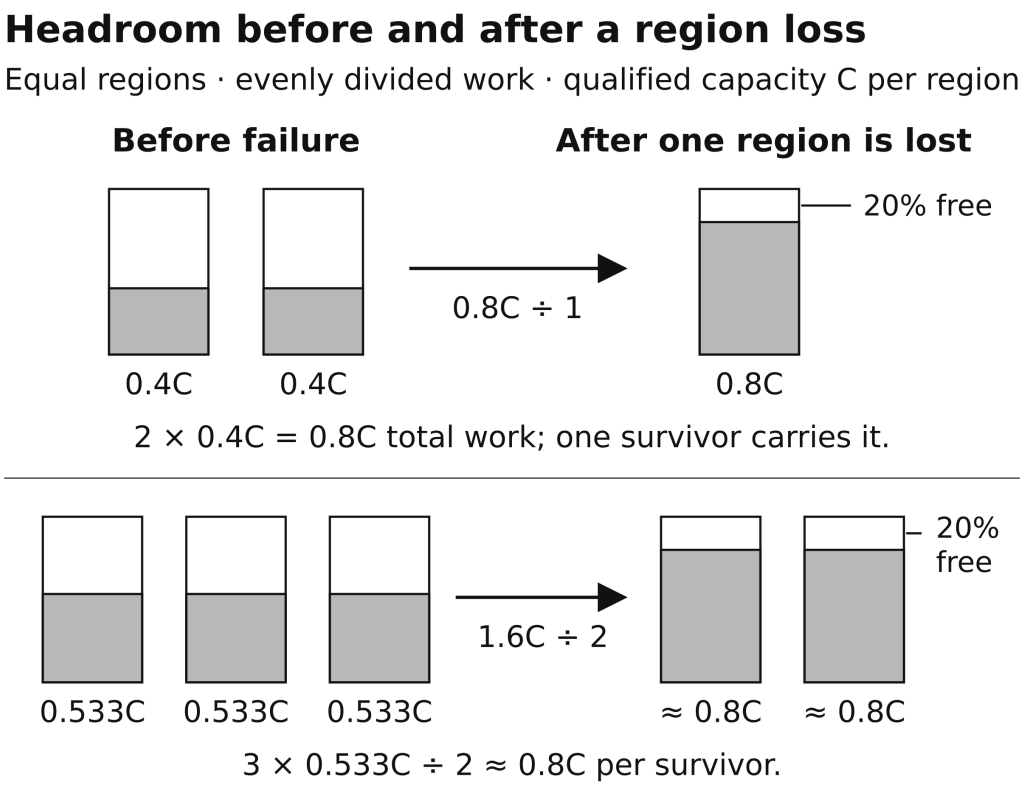


Diagram 19.5. Each outlined bar is a region's qualified capacity C; its gray portion is the work it carries. Under the stated equal-region assumptions, both arrangements leave about 20 percent free after one region is lost. Routing, state, recovery work, and shared failures still need qualification.

Explore this section in the full chapter →

What the capacity team is for, when it happens

The organization should assign ownership of business continuity and disaster recovery explicitly. In the arrangement described here, the capacity team contributes without owning the whole plan: when someone asks “can we handle this?”, we answer alongside the service operators, risk specialists, and business owners whose requirements define what handling it means.

That answer cannot be assembled during an event. Keep a short record for each scenario the organization has decided to plan for:

- **Event and scale**, with the date of the evidence.
- **Affected workloads**, their criticality, and the dependencies they share (Volume I, Chapter 1).
- **Surviving capacity**, and what it is qualified to do (Chapter 18).
- **Recovery deadline**, and the lead time for anything not already held.
- **Permitted destinations**, including legal eligibility (Chapter 21).
- **Provider commitments**, and what they entitle us to in a shortage (Chapter 20).
- **Decision owner and trigger**, and the evidence that would change the response.
- **Accepted gaps**: who accepted them, and when they will be reviewed.

I would write the first record for the scenario the organization is most confident it has covered, because confidence is what stops people asking what two sites share. Its findings go to the next monthly review (Chapter 22) and the next provider conversation.

Explore this section in the full chapter →

CHAPTER 20 / 3 MINUTE BRIEF

Commitments Between Organizations

Read the full chapter → / Read this brief online →

The main idea

"We expect to have it in six weeks" is useful information. It becomes a different kind of information when a project plan turns it into a date and a customer announcement turns that date into a promise. By the time the estimate returns to the capacity team, several people may be relying on a level of certainty that the supplier never offered.

Nobody needs to have intended that transformation. Each person used the best information available to move their own work forward while the original answer's conditions became harder to see. We need those conditions to survive the journey, whether the arrangement is between companies or between teams whose shared company logo conceals different customers, budgets, and obligations.

Explore this section in the full chapter →

A discount and a resource are different things

Azure and AWS arrangements, as of October 2026:

Arrangement	What we owe	What we have when supply is short
Spend commitment: Azure savings plan, AWS Savings Plans[7][8]	A fixed amount per hour for one or three years	No capacity; unused hours expire but are still paid for

[7] Microsoft Learn, "What are savings plans?" and "Automatically renew your savings plan", checked October 4, 2026. Hourly commitment for one or three years; unused hours expire; no cancellation or refund; current savings plan prices apply; renewal turned off by default, with pay-as-you-go rates after expiry. The $10/$7 illustration is hypothetical, not an Azure offer.

[8] Amazon Web Services, Savings Plans User Guide, "What are Savings Plans?", "Compute Savings Plans and Reserved Instances", "Understanding how Savings Plans apply to your usage", and "Returning a purchased Savings Plan", checked October 3, 2026 (return conditions checked October 4, 2026). Hourly commitment with prices fixed for the term; no capacity reservation; no cancellation; unused hours not carried over; the seven-day return conditions.

Arrangement	What we owe	What we have when supply is short
Consumption commitment, such as Microsoft's MACC[9]	A total over the term, perhaps with milestones; any shortfall is charged	No capacity; only eligible spending reduces the balance
Instance commitment: Azure Reserved VM Instance, AWS regional Reserved Instance[10][11]	The reservation price for one or three years, upfront or monthly	No capacity; the discount covers only what we can deploy
Capacity reservation: Azure on-demand capacity reservation, AWS On-Demand Capacity Reservation[12][13]	The on-demand rate for the reserved quantity, used or not; discounts can apply	The reserved size and quantity in one region or zone; Azure's SLA remedy is a service credit

[9]Microsoft Learn, "Track your Microsoft Azure Consumption Commitment (MACC)", checked October 4, 2026. Commitment, milestones, shortfall charged as Azure prepayment credit, and alerts 90, 60, and 30 days before expiry. The page does not address capacity availability; each agreement's terms govern.

[10]Microsoft Learn, "Prepay for Azure virtual machines to save money" and "Transition guide for retired Azure Reserved VM Instances", checked October 3, 2026 (transition guide checked October 4, 2026). Reserved VM Instances are a billing discount, not a capacity guarantee, and can be combined with capacity reservations; capacity priority requires single-subscription scope; purchases and renewals for several older series ended July 1, 2026, with existing reservations running to term. Vendor documentation; terms change.

[11]Amazon Web Services, Amazon EC2 User Guide, "Reserve compute capacity with EC2 On-Demand Capacity Reservations", "Concepts for Amazon EC2 Capacity Reservations", "View the state of a Capacity Reservation", "Capacity Reservation pricing and billing", and "Regional and zonal Reserved Instances (scope)", checked October 3, 2026 (future-dated terms and eligible families checked October 4, 2026). Billing and discounts for Capacity Reservations; zonal versus regional Reserved Instances; future-dated lead times, minimum commitment, eligible families, and states. Product limits change.

[12]Microsoft Learn, "On-demand capacity reservation in Azure" and "Create a capacity reservation", checked October 3, 2026 (overview checked October 4, 2026). Region or zone holds billed at the VM rate whether used or not, with discounts applicable; no SLA for capacity priority; the SLA's worked service-credit example and customer-calculated claims; ND-series and NCads H100 v5 unsupported. Microsoft's SLA text governs, and supported sizes change.

[13]Amazon Web Services, Amazon EC2 User Guide, "Reserve compute capacity with EC2 On-Demand Capacity Reservations", "Concepts for Amazon EC2 Capacity Reservations", "View the state of a Capacity Reservation", "Capacity Reservation pricing and billing", and "Regional and zonal Reserved Instances (scope)", checked October 3, 2026 (future-dated terms and eligible families checked October 4, 2026). Billing and discounts for Capacity Reservations; zonal versus regional Reserved Instances; future-dated lead times, minimum commitment, eligible families, and states. Product limits change.

[14]Amazon Web Services, Amazon EC2 User Guide, "Capacity Blocks for ML" and "Capacity Blocks pricing and billing", checked October 4, 2026. Up to eight weeks ahead and 64 instances per block; no cancellation; price set at purchase by supply and demand, charged upfront; no Savings Plans or Reserved Instance discounts. Supported types and Regions change.

[15]Amazon Web Services, Amazon EC2 User Guide, "Reserve compute capacity with EC2 On-Demand Capacity Reservations", "Concepts for Amazon EC2 Capacity Reservations", "View the state of a Capacity Reservation", "Capacity Reservation pricing and billing", and "Regional and zonal Reserved Instances (scope)", checked October 3, 2026 (future-dated terms and eligible families checked October 4, 2026). Billing and discounts for Capacity Reservations; zonal versus regional Reserved Instances; future-dated lead times, minimum commitment, eligible families, and states. Product limits change.

Dated capacity: AWS Capacity Blocks for ML, future-dated Capacity Reservations[14][15]	Blocks: the full price upfront, no cancellation. Future-dated: the on-demand rate for a minimum term, usually 14 days	The reserved instances for the booked period; a future-dated request can be delayed or declined before it starts

Explore this section in the full chapter →

The internal promise is still a promise

I would keep the agreement short enough that the people relying on it can read it, but specific enough to establish the resource or service, quantity and unit, placement, readiness date, duration, quality or failure condition, and payer. It also needs the authority to change the arrangement and the demand assumption behind it. If the plan works only for a staged rollout, that condition is doing more work than another list of administrative approvals.

Explore this section in the full chapter →

Evaluate the price of certainty

In a simplified pricing illustration, variable purchase costs $10 per unit-period. A take-or-pay agreement offers $7 per unit-period for 100 units over twelve periods, with no resale, cancellation, or other charges. Assume use never exceeds 100 units in any period, so there are no overage charges. We would commit to $100 \times 12 \times \$7 = \$8{,}400$, the same amount that 840 unit-periods would cost at the variable price.

There are 1,200 possible unit-periods, so break-even use is $840 \div 1{,}200 = 70$ percent, which is simply $\$7 \div \$10$. At 600 unit-periods, variable purchase would cost $6,000, so the agreement costs $2,400 more; at all 1,200, variable purchase costs $12,000 and the agreement costs $3,600 less. That $3,600 is the most it can gain, while demand that disappears entirely would lose the whole $8,400. The $7 rate has not changed between the cases; the amount of useful demand has changed its value to us.

Explore this section in the full chapter →

Carry the conditions into the review

Agreement gives us a set of conditions to watch. Product changes the rollout, a new request mix changes the service rate, a supplier revises delivery, or qualification reveals less

throughput than expected, while the borrowed workers approach their return date regardless. Any one of these changes can reopen the combined decision; Chapter 22 turns them into triggers agreed in advance.

Someone has to own those connections, not just the inputs. In the constructed portfolio, the capacity lead maintains the joined view and brings the choice back before the available options expire; the people with authority over the promises still decide which ones can change.

Explore this section in the full chapter →

CHAPTER 21 / 4 MINUTE BRIEF

Multiple Markets

Read the full chapter → / Read this brief online →

The main idea

Once an organization serves several markets, some of its placement rules become conditions that no commercial agreement can change, and two locations with the same machines stop being interchangeable. The earlier examples kept those privacy, sector, and infrastructure rules in the background. Scale makes them harder to coordinate, but a small business using an overseas provider meets them too. Before moving work between locations, the capacity team needs to know which ones its specialists have cleared, and on what conditions.

The example organization from this book has run a European operation for years, under its own budget and its own agreement with Provider B, which is why Volume I, Chapter 1's ledger leaves it out. That bill put the web workers on Provider A, the serving layers behind the data platforms and AI assistant on Provider B, and the customer and order records in the organization's own data center. European customers are served from Provider B's European regions, chosen partly because its data-residence terms satisfied legal review, while Provider A keeps serving the Americas. The same React application, the same tier structure, the same databases behind it, and a topology that looks arbitrary until somebody asks why the European workers cannot simply fail over to Virginia. The duplicated stack shows up in the bill long before anyone mentions the law.

Explore this section in the full chapter →

Location is a capacity property

Once placement is constrained, geography becomes part of what "capacity" means. The tested worker capacity is unchanged, but the locations are no longer interchangeable: a region that could absorb the web tier's failover traffic may be unable to hold the data the tier serves. Chapter 18 asked what a reserve is available for, and each candidate location an-

swers differently. Capacity counts toward a workload only where it is permitted, available, and tested.

Then price the permitted arrangements, including quoted rates, chargeable data transfers, and the standby or duplicated services each topology requires. Volume I, Chapter 1's "where it runs" view was drawn with two providers and a data center; I would add permitted locations and their conditions to it before comparing costs, so that the cheapest row is also one we could operate.

Explore this section in the full chapter →

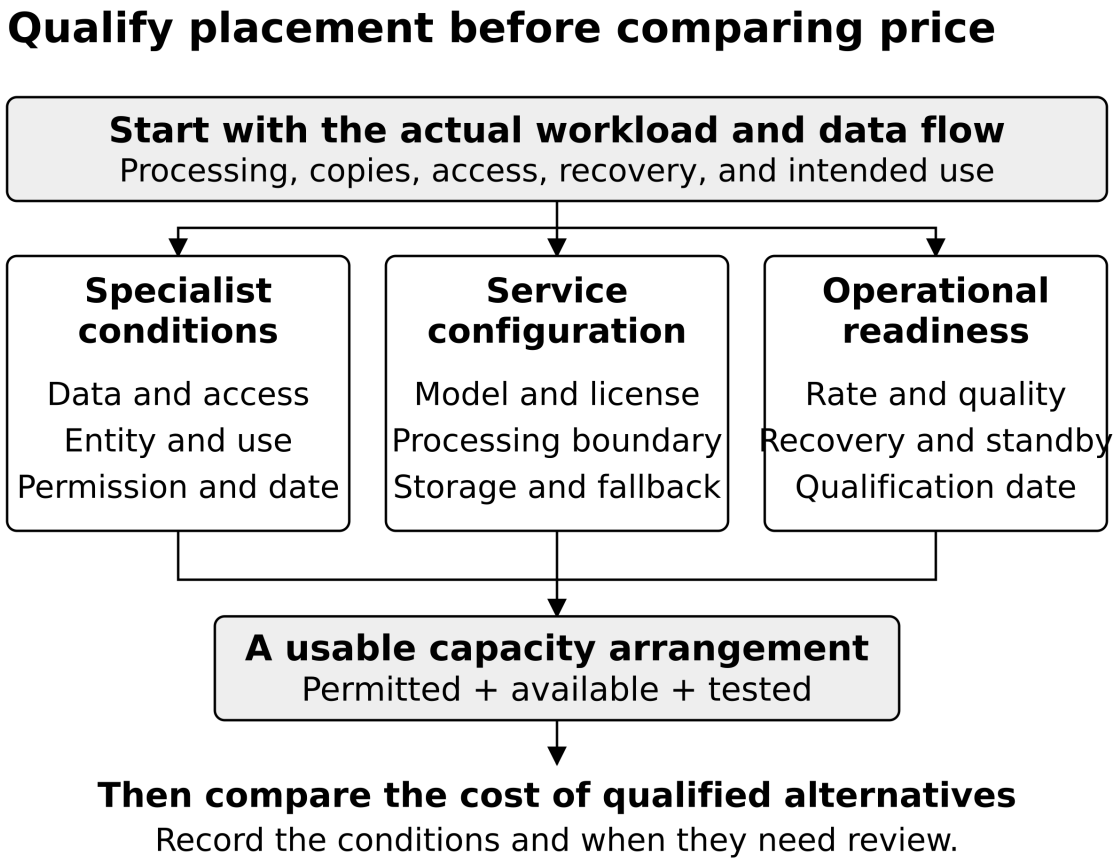


Diagram 21.1. Placement becomes usable capacity only when specialist conditions, service configuration, and operational readiness agree. Compare prices among those qualified arrangements, and retain each condition's review date.

Explore this section in the full chapter →

Consider segmentation by market

Segmentation is an option to evaluate, not a requirement to infer from a country boundary. I would identify which workloads need separate placement or operations, then look for what still connects them: identity, telemetry, support tooling, backups, and the deployment pipeline can each carry data or access across the line. Separation can limit the effect of a local rule change; it cannot guarantee independence. Where lawful sharing is possible, weigh its savings against the cost of separating later.

Explore this section in the full chapter →

Segmentation costs utilization

Separated systems will usually not reach the efficiency of one pooled system. That is a fact of life, and it does not contradict the earlier chapters: deliberately unused capacity can protect a service or satisfy a constraint.

The mechanism is the minimum configuration that availability and recovery requirements impose: replicas, standby capacity, test environments, and the people and systems to operate them. If two deployments must stay separate and each needs its own standby, idle capacity in one cannot meet the other's obligation. Show those duplicated minimums and why they exist. Judging each market against an unconstrained global pool would mistake a placement requirement for an efficiency failure.

Explore this section in the full chapter →

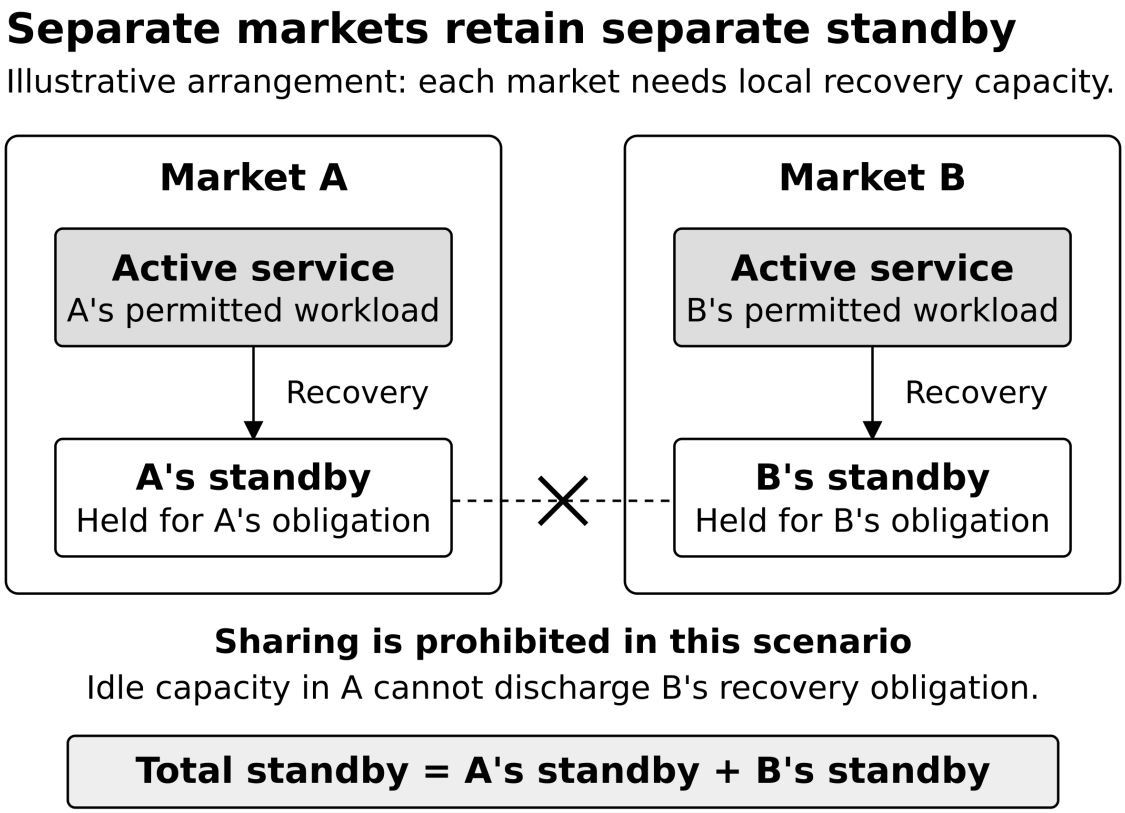


Diagram 21.2. In this illustrative arrangement, sharing is prohibited and each market needs its own standby. The unused capacity protects a local recovery obligation; it cannot be counted again as the other market's reserve.

Explore this section in the full chapter →

The monthly review across markets

Suppose the organization starts measuring the reuse of AI infrastructure across markets, and maximizing reuse becomes a success criterion. Capacity that looks idle on one market's chart may be unavailable to another, and a criterion written as if the deployment were unified will mark correctly segmented systems as failures. So the review keeps the constraint register beside the efficiency numbers, with specialists confirming the conditions behind any cross-market comparison. When a rule or interpretation changes, note its effective date, the options it removes or adds, and the time needed to respond, so nobody discovers an expired condition during a failed recovery or a purchase review.

Explore this section in the full chapter →

One practice or two

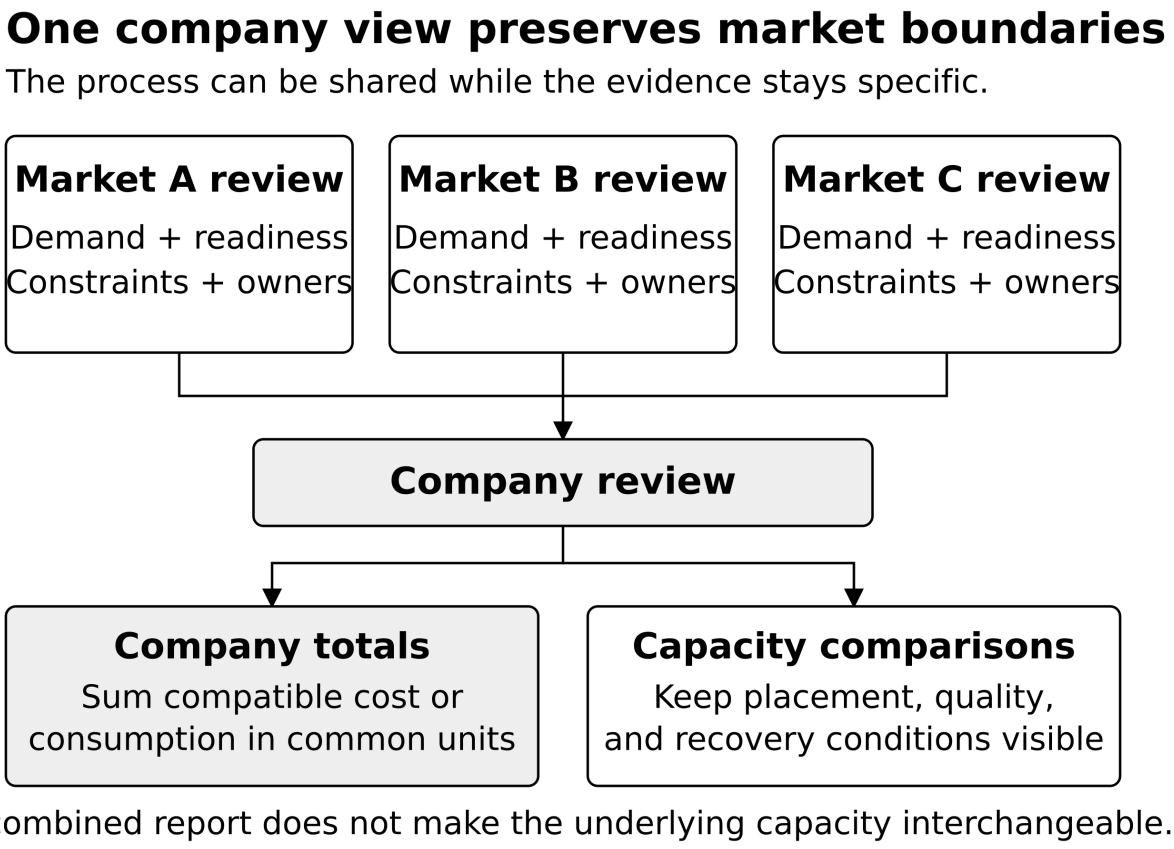


Diagram 21.3. Separate market reviews can feed one company view. Compatible cost or consumption units can be summed; capacity comparisons must preserve the placement, quality, and recovery conditions that make each market's numbers meaningful.

Explore this section in the full chapter →

CHAPTER 22 / 3 MINUTE BRIEF

The Monthly Capacity Review

Read the full chapter → / Read this brief online →

The main idea

The staged launch happened, and the peak in our constructed example was 12,800 requests per second. We had a historical baseline of 10,000 and a business-adjusted forecast of 12,000, so both estimates were low. Assuming the tested request mix and quality conditions still apply, the temporary arrangement supports 13,500 after losing one modeled failure group. That leaves 700 requests per second of degraded headroom, less than half the 1,500 we planned.

We could make this a meeting about a successful launch or a meeting about a forecast that missed. I would rather begin with what the result means for the next choice. The room left is smaller than the amount by which the forecast just missed, another rollout stage is still planned, and the borrowed workers have a return date that our satisfaction with the launch will not move.

Explore this section in the full chapter →

Keep the original forecast

To learn anything from the comparison, we need the baseline as it existed before the conversation, the adjustment, the dates when each became available, and the later observation. Chapter 14 showed what happens otherwise: Product's July 8 revision, which raised the first stage's increment to 2,800 and the full rollout's to 4,500, would make the June 28 forecast look perfect if it leaked in. The revision belongs in the next decision instead.

Explore this section in the full chapter →

Review the decision, not only the prediction

Forecast error tells us something important, but it cannot carry the entire assessment. A conservative estimate might be numerically worse while preserving an option the business values. An accurate average can conceal an unacceptable peak, and an improved demand forecast does not make supply in the wrong location useful. We can even satisfy one customer's resource requirement while imposing a delay on another customer that nobody in the review has examined.

Explore this section in the full chapter →

The agenda

The pack and the meeting share one order, and the time goes where the deadlines are:

1. **Corrections**, first and briefly, starting with the capacity team's own, so that nobody argues a decision on a number we have already withdrawn.
2. **Decisions due before the next review**, in deadline order, each with its record. In an hour, this should take at least half. Each ends with an owner's decision, a stated refusal, or a date by which one will exist.
3. **Exceptions that crossed a threshold agreed in advance**: forecast bias at the decision horizons (Chapter 14); room under the stated failure condition and its trend, where Chapter 19's slow drift in failover headroom appears; and arrangements expiring inside their lead time (Chapters 18 and 20).
4. **The portfolio view**, a few minutes of totals and trends, keeping attribution, enterprise cash, and committed spend apart (Volume I, Chapters 1 and 13), with AI consumption beside the accepted work it produced (Volume I, Chapters 7 and 9).
5. **Carried items**: investigations past their date, and any measurement we keep requesting and never obtain, with what is preventing it and which decision the delay now blocks.
6. **Read back the record**: decisions, conditions, refusals, triggers, and the next update date for each estimate.

Explore this section in the full chapter →

CHAPTER 23 / 3 MINUTE BRIEF

What Scale Brings

Read the full chapter → / Read this brief online →

The main idea

Near its end, Volume I, Chapter 1 multiplied its bill by ten: the same five views, $4,480,000 over thirty days, roughly $54.5 million a year, and a promise that the conversations around those views change at that size. The change I would watch most closely is where a decision lands. Even the smaller ledger has agreements that carry several owners' consumption, but at that size a capacity lead and a finance partner can keep the connections in their heads. At ten times the bill, the same agreements can carry minimum spending terms with shortfall charges in the millions, and a sensible choice by one owner can create an obligation in a budget that owner never sees.

Explore this section in the full chapter →

The portfolio at ten times the bill

Take Volume I, Chapter 1's ledger and multiply every line by ten, holding the workload categories, locations, unit rates, and funding arrangements fixed. This is a bookkeeping assumption for the comparison. It does not predict that a real service can grow tenfold without changing its architecture, performance, or commercial terms. The thirty-day table reads:

Work supported	Thirty-day modeled cost
Data platforms	$1,200,000
AI support assistant	$1,000,000
Transaction-facing web workers	$768,000
Shared platform hosting many small applications	$600,000
Shared network, observation, and support	$396,000
Legacy system being retired	$240,000
New system replacing it	$180,000
Deferrable batch workers	$96,000
Total	**$4,480,000**

On a 365-day basis, $4,480,000 ÷ 30 × 365 = $54,506,667, or about $54.5 million a year. Twelve identical thirty-day periods would instead total $53.76 million. The web pool keeps its shape at ten times the size: 640 workers at the same $40 per worker-day, now 640 × $40 × 30 = $768,000, still arranged in failure groups and still reading from serving layers whose own lines have grown tenfold. The eight compatible batch workers become eighty, and the transfer that moved $9,600 of attribution in Volume I, Chapter 1 moves $96,000 here. The larger transfer deserves the same explanation of displaced work and responsibility.

Explore this section in the full chapter →

What the number buys, and what it costs

A larger budget can improve our negotiating position, but annual spending is not the same as an annual commitment. We should decide how much to promise after examining the stable portion of demand. Volume I, Chapter 1's constructed bill was about 63 percent committed or fixed; multiplying its lines by ten preserves that proportion. It does not prove that larger organizations necessarily commit a larger share. If we choose a higher minimum to obtain a better rate, show the reduced freedom to shrink alongside the saving.

Explore this section in the full chapter →

The obligations grow their own management

Volume I, Chapter 1 described obligations as promises that outlast the reasons we made them. At the small scale, a capable finance partner and a diligent engineer can enumerate the commitments, subscriptions, minimums, and refresh cycles together. At $54.5 million, the enumeration itself is a job: an obligation register, a renewal calendar, usage-versus-commitment tracking, and a reconciliation between what the agreements promise and what the workloads actually consume. Organizations that skip this do not avoid the obligations; they discover them at renewal time, one at a time, in the order the providers raise them.

Explore this section in the full chapter →

Where I would start

The five views still answer the same questions at either size: what the money does, who answers for it, where it runs, how its charges behave, and who funds it. What changes is that each view needs someone to maintain it. The work view needs owners for its largest lines; attribution needs an allocation process with someone accountable for whatever stays unallocated; the location view needs an owner for each agreement; the charge view becomes the register; and the funding view needs program end dates that Finance can see coming. A reader with a smaller budget can use the same questions without adopting an organization built for this chapter's spending level.

Explore this section in the full chapter →

CHAPTER 24 / 3 MINUTE BRIEF

Markets for Compute

Read the full chapter → / Read this brief online →

The main idea

This book has treated capacity as decisions made inside one organization: what we measure, promise, reserve, and commit to pay. This closing chapter looks outward. Compute already has markets; what 2026 added was an attempt to list futures on the rental price of a GPU, so that buyers and sellers could pass price risk to someone else. As of October 4, 2026, none of those contracts was trading on a U.S. exchange. The work they ask of a buyer, defining exactly what is priced and comparing our own price with an outside one, can start now.

Treat the chapter as advanced reading and a direction of travel; nothing here changes the practice of the previous twenty-three chapters.

Explore this section in the full chapter →

A market appears

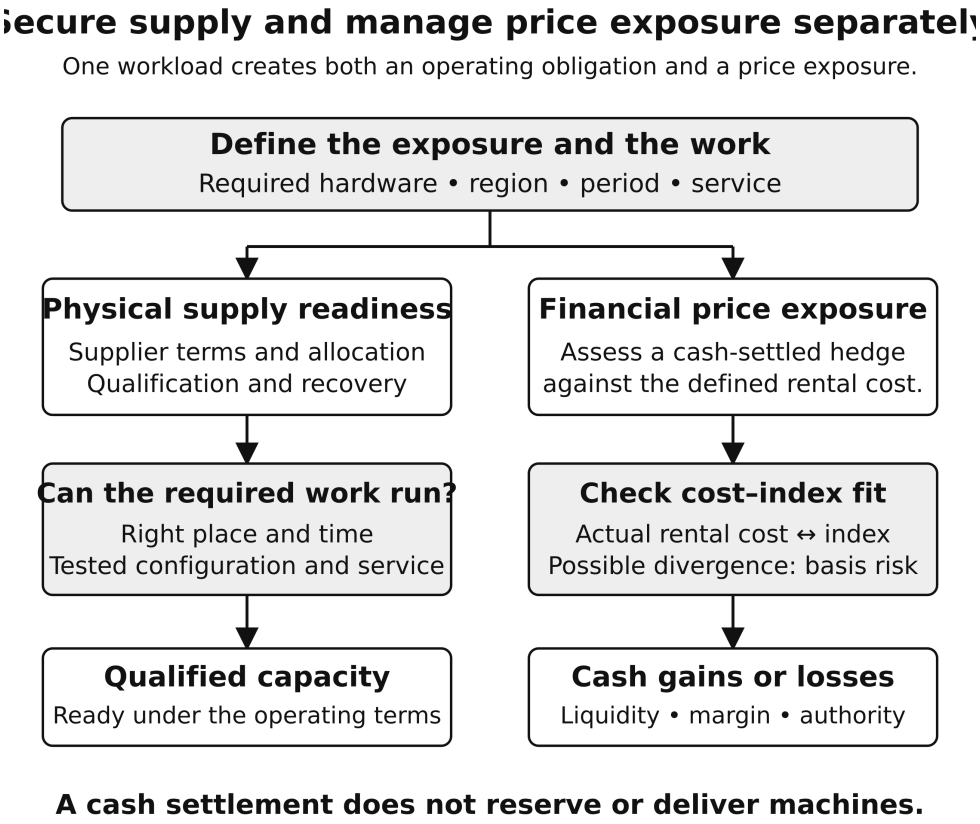


Diagram 24.1. The operating plan secures and qualifies physical supply; the financial plan evaluates price exposure and cost–index fit. A cash-settled hedge can leave basis risk and does not reserve or deliver machines.

Explore this section in the full chapter →

Why exchange trading has been difficult

The first obstacle to exchange trading is heterogeneity. An H100 and a B200 are different resources, and the same chip in two regions or on two networks may support different work. A token buyer is buying a service, not a GPU-hour. A standard contract has to say which differences it includes, which it adjusts for, and which stay with the buyer; Architect's proposed delivery network, for example, would publish basis tables by SKU, memory configuration, and location.[16]

Explore this section in the full chapter →

[16]Architect Financial Technologies, "Architect Financial Technologies Partners with Compute Desk to Launch Exchange-for-Physical Market for GPU Compute", July 8, 2026: the American Innovation Exchange is "forthcoming" and "pending regulatory review"; ComputeConnect would publish "standard basis tables for different SKUs, memory configurations, and locations." Company announcement, checked October 4, 2026.

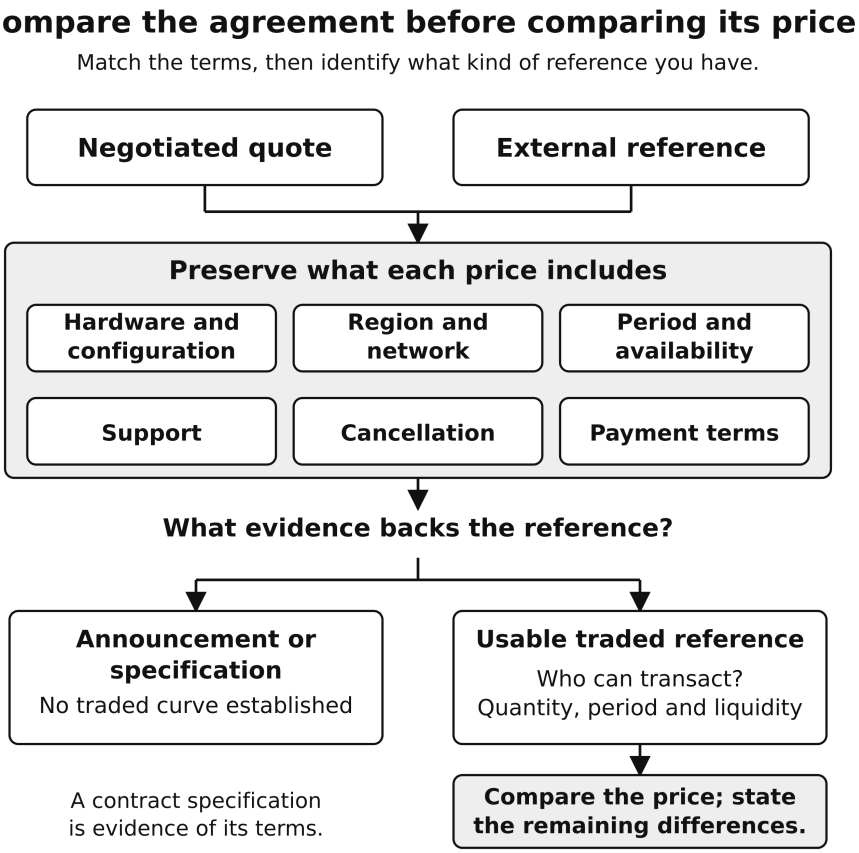


Diagram 24.2. A quote and an outside reference become comparable only after their hardware, configuration, location, period and commercial terms are stated. An announced specification describes terms; a usable traded reference also needs evidence of quantity, period and liquidity.

Explore this section in the full chapter →

What is available and what is announced

Status as of October 4, 2026:

Instrument	What is bought	Price set by	Example (status)
Provider spot	Interruptible capacity now	Provider, without bidding	EC2 Spot (operating)[17]
Prepaid capacity block	Named instances for a future window, prepaid	Provider, from supply and demand at purchase	AWS Capacity Blocks (operating)[18]
Commitment	Lower rates for a spending or usage obligation	Published or negotiated terms	Chapter 20 (operating)
Procurement marketplace	Competing quotes, then a direct supplier contract	Supplier offers	Compute Exchange (operating)[19]
Physical forward	Delivery in a future period at a price fixed now	The two parties	Compute Exchange (offered); OneChronos–Auctionomics (announced)[20]
Resale rights	Permission to sublease contracted time	Other buyers	SF Compute (operating)[21]
Cash-settled index future	Exposure to a monthly average rental index	Exchange trading	CME, ICE, Nodal, Architect (none trading)[22]
Perpetual future	H100 price exposure, no expiry	Crypto-venue trading	Seven venues (trading, thin)[23]

Explore this section in the full chapter →

[17]AWS, "Amazon EC2 Update — Streamlined Access to Spot Capacity, Smooth Price Changes, Instance Hibernation", November 28, 2017, on the change from bidding to smoother provider-set prices.

[18]AWS, Capacity Blocks and pricing and billing documentation, checked October 3, 2026, for the discount and billing rules described.

[19]Compute Exchange, marketplace site, checked October 3, 2026: quotes through a "live auction across verified providers," a services agreement signed directly with the chosen provider, and forward contracts on bilateral terms. Vendor descriptions.

[20]Auctionomics and OCX Group, "Auctionomics and OneChronos Partner on First Tradable Financial Market for GPU Compute", Business Wire, July 29, 2025: price discovery "through bi-lateral forwards"; "first" is the companies' claim. OneChronos, "Compute Markets", checked October 4, 2026: "Coming soon."

[21]San Francisco Compute Company, "About", checked October 3, 2026: clusters leased "on contracts that let our customers sublease." Vendor description.

[22]CFTC, product record 62544, "Approval Pending (90)," checked October 4, 2026; Division of Market Oversight, letter to CME Group, September 21, 2026, extending review of NYMEX Submission No. 26-370 to November 9, 2026, for "novel or complex issues." An extension is not a disapproval.

[23]Loris Tools, "H100 Perpetual Futures", checked October 4, 2026, 21:41 UTC: seven venues, $1.51 million of 24-hour volume, $80,800 of open interest, a $2.647 reference price, and 84.0% of volume on one venue. A third-party, unaudited aggregation of crypto-venue data; an order of magnitude only.

If suitable contracts start trading, I would take four questions to Finance and Treasury first. How closely has the index tracked our actual rate, month by month? How much trades in the months we would hedge? Who has authority to trade? Who funds the margin when the position moves against us? A hedged shortage is still a shortage.[24]

Explore this section in the full chapter →

[24] CME Group, "Learn about Basis: Grains", checked October 3, 2026, and *Self-Study Guide to Hedging with Grain and Oilseed Futures and Options*, on basis (cash minus futures), long hedges, and margin. Applying grain mechanics to compute still requires checking the actual contract.

If we all want the same GPUs

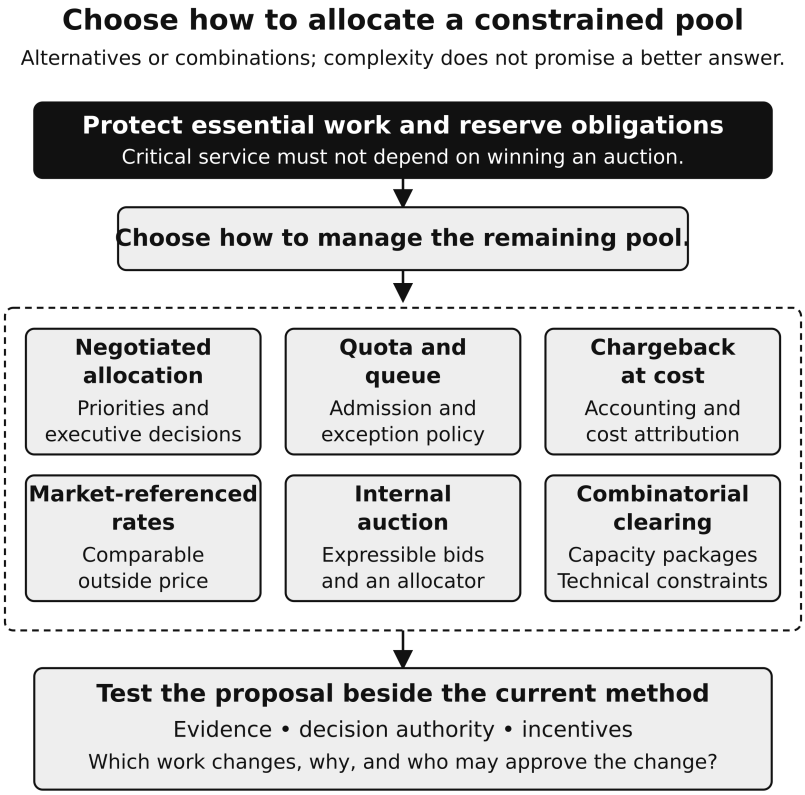


Diagram 24.3. Protect essential work, then choose among allocation and pricing mechanisms for the remaining constrained pool. These are alternatives or combinations; test the proposal against the current method with evidence, decision authority and incentives visible.

Explore this section in the full chapter →

The question the curve answers

None of this has to wait for an exchange. I would take the next GPU commitment due for renewal and write it as one row: hardware and configuration, region, term, price per GPU-hour, minimum spend, cancellation and transfer rights, and payment timing. Beside it go the nearest outside prices for the same period (the on-demand rate, a Capacity Block quote, a published index, a marketplace quote), each with what differs and roughly what the difference is worth. That page goes to procurement and Finance before the negotiation, with one question for Treasury: if a usable contract existed, could we hedge this exposure, and who would fund the margin? After signing, I would record our actual rate beside the index every month. That record is the basis estimate a future hedge will need, and it strengthens the next negotiation whether or not these exchanges ever become deep markets.

Explore this section in the full chapter →