CHAPTER 4 / 3 MINUTE BRIEF

The Simplest Practice

Read the full chapter → / Read this brief online →

The main idea

If you're starting a capacity and cost practice, I would resist the temptation to begin by designing the process for the entire organization. You'll need to understand that organization eventually, and it will need more than one approach. But there's a more useful place to begin: find a decision somebody needs to make and stay involved long enough to see what it takes to make it well.

You might start with a service that's about to grow. The engineering team can explain what it runs today, somebody in the business has a plan that will change the demand, and Finance wants to understand the cost of supporting that plan. Getting those people into the same discussion sounds straightforward until you discover that the growth estimate assumes a launch date, the launch date assumes available capacity, and the capacity depends on a purchasing decision that hasn't been made. Now you have something worth working on.

Explore this section in the full chapter →

A portfolio small enough to follow

For the first decision, we'll concentrate on the web service and the measurement Chapter 2 gave it. Its operators have tested a particular request mix at the configurations we're considering: one worker can complete 250 requests per second while holding ninety-fifth-percentile (p95) latency to 200 milliseconds and errors to no more than 0.1 percent of offered requests. That evidence exists because we've assumed it. In a real discussion, I'd want the operators to explain how they qualified the rate and what change in request mix, configuration, or downstream behavior would make it unreliable.

The service has 64 workers arranged in four independent failure groups of 16, providing a tested 16,000 requests per second when everything is available. The business has

also agreed that the service must continue after losing any one group. That requirement changes the capacity we can promise:

$$(64 - 16) \times 250 = 12{,}000 \text{ requests/second.}$$

Twelve thousand is the limit I'd use for this decision; as Chapter 2 showed, sixteen thousand answers a question about normal operation. Even the lower number covers only the loss of one independent group; we haven't established what happens when an entire region fails.

Explore this section in the full chapter →

Borrowed capacity comes with a return date

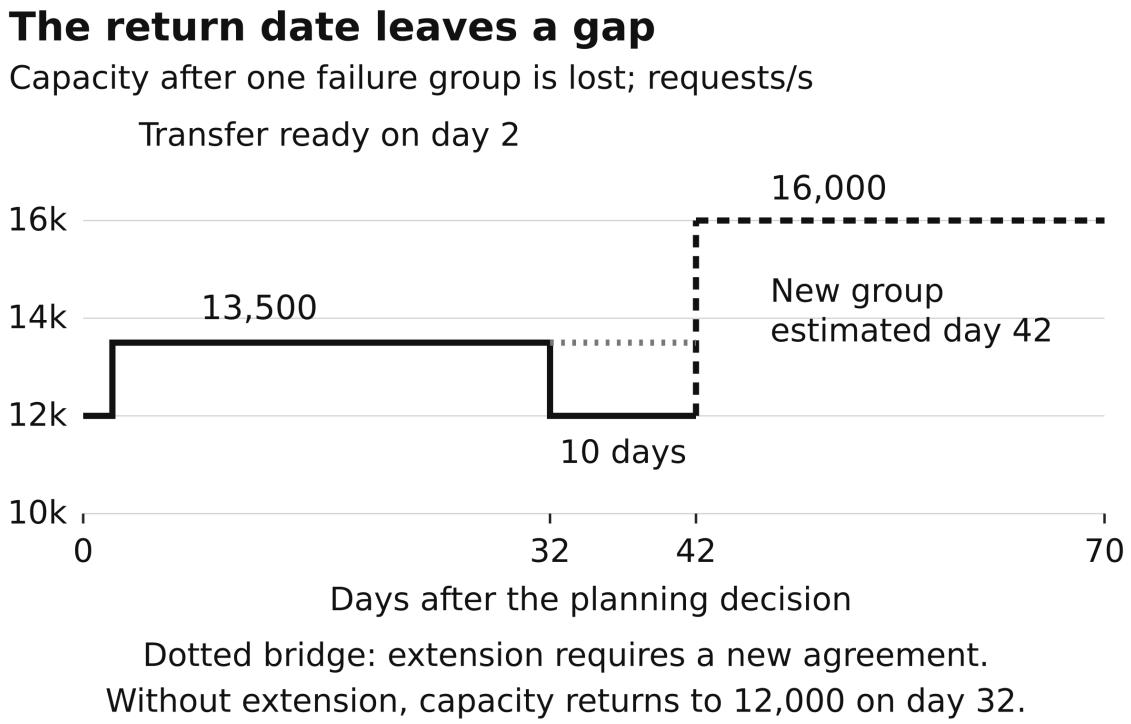


Diagram 4.1. The temporary transfer expires on day 32; the new group's qualified readiness is only estimated for day 42. The dotted bridge requires a further batch agreement.

Explore this section in the full chapter →

Make the decision legible

For this proposal to proceed, five of the people in Chapter 3's authority map have to commit to something:

Who	Commitment	Needed before
Product lead	Stage the rollout at the estimated 2,000 additional requests/s, and keep the ability to stop expansion	the rollout
Batch owner	Accept a bounded delay with a catch-up plan	the transfer
Web operations lead	Qualify the transferred workers, including their behavior when a group fails	the rollout
Budget owner	Authorize the $640-a-day purchase within a delegated limit, contingent on supply and the actual terms	the purchase
Responsible business executive	Accept the remaining exposure, or change the launch scope or date	the rollout

Several may be settled in one conversation. They're still five commitments, and the decision record should make a missing one obvious. If batch refuses the delay, we no longer have this option. If the business refuses staging, 14,000 still exceeds 13,500, and changing the failure requirement to make the arithmetic pass would mean offering a different service. We may need to discuss that possibility, but with the person entitled to change the obligation and with its consequence understood.

The agreement also needs a way to reopen the discussion before the next scheduled review. Demand running above the staged estimate, a changed request mix, a larger launch, an infeasible catch-up plan, or delayed delivery can each remove a condition we relied on. Waiting for the monthly meeting would only be sensible if we could afford to wait. Whoever learns about the change should know whom to contact and what decision it may affect.

Explore this section in the full chapter →