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first and briefly, starting with the capacity team's own, so that nobody argues a decision on a number we have already withdrawn.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 →