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.[1] 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

[1] 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 →