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 →