CHAPTER 2 / 3 MINUTE BRIEF

Capacity Is a Decision

Read the full chapter → / Read this brief online →

The main idea

Anytime I hear “yeah, we’ve got enough capacity,” I want to ask what they mean. Enough for which operating mode? A complex system runs under steady load, through a seasonal swing, and at the peak that arrives once a year. When someone who runs one breezily tells me everything is fine, no problems, I start wondering whether they have ever experienced a failure, because the honest answer to “do you have enough capacity” is a complex one. Something like: for the work in front of us right now, yes; here is the evidence for that; here is the traffic we expect over the next 30 to 90 days; and here is what we are watching so we can prepare before it becomes a problem. No measurement can give that answer by itself, because it depends on what the organization has decided to promise, who is entitled to rely on the promise, and who answers for keeping it.

Explore this section in the full chapter →

Headroom has a purpose

Headroom is often described as the difference between a measured limit and expected demand, but that leaves out the question of which limit we mean. In our original configuration, normal capacity of 16,000 minus expected demand of 10,000 leaves 6,000 requests per second. Under one-group loss, the calculation is 12,000 minus 10,000, leaving 2,000. These describe different states of the same service, not two reserves we can add together.

Some of that room is carrying more than one responsibility in the plan: growth while everything works, and service preserved when something fails. The healthy service's 6,000 looks like room for the rollout's 4,000 and also for the 4,000 that a lost group takes with it, but it cannot be both, because together they need 8,000. Test the combined requirement in the condition you intend to promise. Counting each responsibility separately and then celebrating the apparent total can make us feel considerably better prepared than we are.

Explore this section in the full chapter →

Autoscaling still needs someone to supply the capacity

I like automation that responds to changing demand. What worries me is the point at which the word autoscaling starts doing work that belongs in a supply plan. An autoscaler can ask for resources when a signal changes; it cannot, by that request alone, establish that the right resources will be available, ready, and connected to their dependencies in time. Azure's quota guidance makes the distinction explicit: quota is a subscription's permission to deploy, capacity is the compute actually available in a region or zone, and a deployment with enough quota can still fail for lack of capacity.[1] That leaves two questions to settle before the autoscaler needs an answer — will we be allowed, and will the supply be there — with options such as qualifying another size or region, or securing an eligible capacity reservation where the workload requires it.

Explore this section in the full chapter →

Make the promise understandable

The person approving a launch should be able to explain what the customer will receive and what has to remain true for us to deliver it. If that explanation requires reconstructing a month of technical discussions, we have more work to do. I would give them a short account of the useful work, quality, deadline, operating condition, and usable supply, with named owners and the underlying evidence available to whoever needs to challenge it.

[1] Microsoft Learn, "Check vCPU quotas", checked October 3, 2026. Defines quota as a subscription's permission to deploy and capacity as the compute available in a region or zone, says a deployment with sufficient quota can still fail for lack of capacity, and suggests a different size, zone, or region, or an on-demand capacity reservation.

For our web service, the account might run like this. At the tested request mix, the service carries the expected 10,000-request peak with 2,000 to spare after losing one of its four sixteen-worker groups. The full rollout, at 14,000, would leave it 2,000 short in that condition. A staged rollout reaches 12,000 and uses all of the margin, so it also needs eight batch workers borrowed for thirty days, if the batch owner accepts the delay and the workers pass qualification. Sixteen new workers would carry the full rollout once funded, but they are estimated to be six weeks from qualified readiness, and the borrowed workers are due back before then.

Explore this section in the full chapter →