Cutting Cloud Spend Without Cutting Reliability

A quarter of right-sizing, reserved capacity, and saying no

by

Dana Whitfield

August 27, 2026

Contents

1. The Bill

   Where the money actually went

2. What We Changed

   Four changes, in order of impact

3. Results

   What it cost us and what we saved

The Bill

Where the money went

- 62% of spend was compute sitting idle outside business hours
- 21% was over-provisioned instance sizes, chosen once and never revisited
- 17% was genuinely needed capacity

What We Changed

Right-sizing

- Ran two weeks of utilisation data through a sizing recommender
- Downsized 140 instances with no measurable latency regression
- Automated a weekly re-check so sizes do not drift back up

Scheduled scale-down

- Non-production environments now scale to zero outside 7am–8pm local
- Saved 62% on the non-production bill with zero engineering complaints

Reserved capacity

- Bought 1-year reservations for the stable 70% baseline
- Kept the volatile 30% on-demand for genuine elasticity

Results

Nine months later

- Monthly cloud spend down 34%
- p99 latency unchanged across every service we touched
- Zero incidents attributable to the changes

Backup: methodology

- Utilisation sampled at 1-minute resolution for 14 days per service
- Recommender required 20% headroom over p99 observed usage