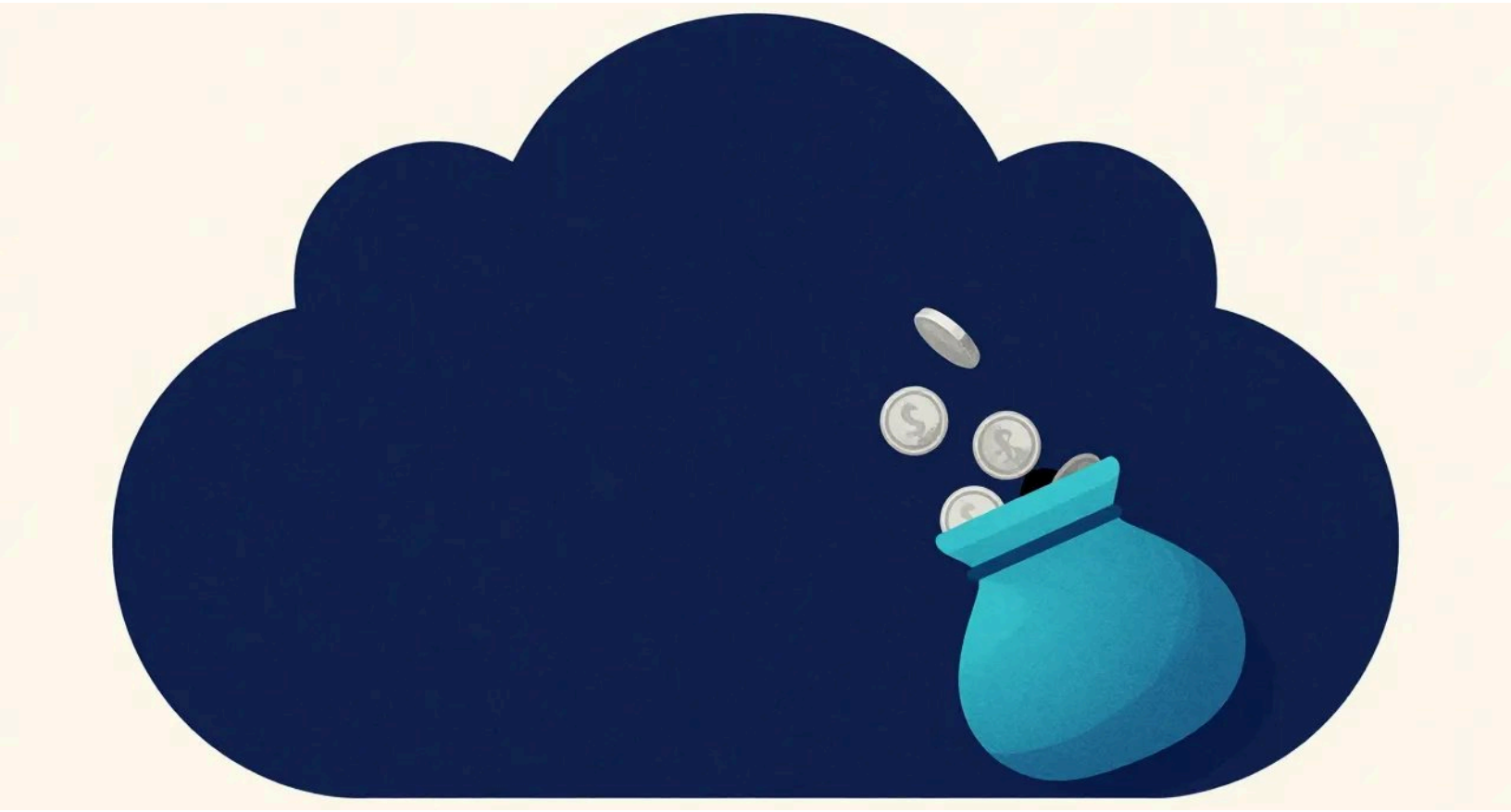


EXECUTIVE PAPER · INFRASTRUCTURE

# Your Cloud Bill Is an Engineering Problem

Engineers write the cloud bill one design choice at a time. CFOs and CTOs who want the number to move have to manage those choices.

AUTHOR

Chris Watkins

SERVICE

Infrastructure

FOR

CFO, CTO · 12 min read

AT A GLANCE

**The problem.** Cloud spending keeps climbing, and the share respondents call wasted rose in 2026 for the first time in five years. Most of that money is committed long before an invoice arrives, in how systems were designed and in what nobody turned off.

**Why it persists.** The bill lands in finance while the decisions that create it sit with engineering. Engineers rarely see what their choices cost, and finance cannot tell an expensive design from a necessary one. Cost reviews chase individual line items and leave the architecture underneath untouched.

**The recommendation.** Measure before you cut: allocate every dollar to a workload and every workload to a named person. Rank fixes by dollar impact and work down the list, from switching off idle resources through commitments to redesign, and only then consider moving workloads out of the public cloud. Review cost per unit of business output monthly, with the CFO and CTO in the same room.

## The Decision in Front of You

Most companies treat the cloud bill as a procurement problem. Finance negotiates a discount, sets a budget, and asks engineering to stay under it. When the number overshoots, someone launches a cleanup sprint, the bill dips for a quarter, and then it climbs again.

The decision is whether to keep managing the invoice or start managing the engineering decisions that produce it. Managing the decisions costs leadership time and some engineering capacity, and it takes a year to show its full effect.

The stakes are large. IDC expects worldwide public cloud spending to pass $1 trillion in 2026, growing more than 21% a year [4]. Flexera's 2026 survey of 753 cloud decision-makers found 85% name managing cloud spend as a top challenge, and 17% exceeded their public cloud budget in the past year [1][2]. Respondents estimated 29% of their infrastructure and platform spend was wasted [2].

## What's Happening

### Waste is rising again after five years of decline

Flexera's 2026 report marks the first increase in estimated waste in five years, even though 63% of respondents now have a dedicated FinOps team [1][2]. FinOps, short for cloud financial operations, is the practice of bringing finance, engineering, and business owners together to manage variable technology spend.

The FinOps Foundation's 2026 State of FinOps survey, with 1,192 respondents representing more than $83 billion in annual cloud spend, explains part of the pattern [3]. Workload optimization and waste reduction remain practitioners' top priority, yet many report diminishing returns because the obvious waste is already gone. One respondent put it as having already hit the "big rocks" [3]. What remains is embedded in system design, where a central FinOps team has little authority.

## Engineers cannot see what they spend

A 2025 survey of 700 developers and engineering leaders at large US and UK companies, sponsored by the software vendor Harness, estimated that 21% of enterprise cloud infrastructure spend goes to underused resources, about $44.5 billion in 2025 [5]. Fewer than half of developers had real-time data on idle resources. Teams took an average of 31 days to find and remove idle or orphaned resources and about 25 days to rightsize overprovisioned ones [5]. More than half said their commitment purchases were based on guesswork. The sponsor sells cost tooling, so read the dollar figure as an upper bound. The operational findings line up with the independent surveys.

Ownership is the gap. In the same study, 62% of developers said they wanted more control over and responsibility for cloud costs [5]. The FinOps Foundation's own framework lists "everyone takes ownership for their technology usage" as a core principle [7]. AWS's Well-Architected guidance tells customers to attribute costs transparently to workload owners so those owners can measure return and act [6].

## The reporting line works against engineering ownership

The State of FinOps 2026 data shows 78% of FinOps practices now report to the CTO or CIO, and only 8% report to the CFO [3]. That shift moves cost management closer to the people who make design decisions. It also means the CFO can lose sight of the program unless the review is shared. The same survey found practices with executive engagement at VP level or above had two to four times more influence over technology selection than those engaged only at director level [3]. Practitioners also ranked pre-deployment architecture costing as a top tool request, a sign that the useful moment to control cost is before a system ships [3].

## Three recent cases

**37signals, 2022 to 2025.** The company behind Basecamp and HEY was spending about $3.2 million a year on AWS when it began leaving in 2022 [12]. It bought roughly $700,000 of Dell servers, recovered that outlay during 2023, and cut its cloud bill by almost $2 million a year with the same operations team [10]. In 2025 it moved about six petabytes off Amazon S3, which had cost just under $1.5 million a year, onto $1.5 million of purchased storage [11][12]. Its chief technology officer projected total savings of more than $10 million over five years [10]. AWS waived about $250,000 in data transfer fees for the exit [12]. The lesson is narrower than it looks. 37signals runs steady, predictable workloads with an experienced operations team, which is the profile where owning hardware pays.

**GEICO, 2013 to 2025.** The insurer began moving more than 600 applications to the cloud in 2013. A decade later its vice president of platform and infrastructure engineering said the migration had been mostly lift-and-shift and that cloud bills had risen about 2.5 times [13]. She described running legacy applications unchanged in the cloud as prohibitively expensive. GEICO is now moving storage and other workloads to a private cloud built on OpenStack and Kubernetes [13]. The 2.5 times figure is her account and has not been independently verified. The lesson is that moving an application without redesigning it carries its old inefficiencies into a metered environment.

**AWS billing estimates, July 2026.** A configuration change in AWS's billing system on July 16, 2026 produced estimated charges in the billions and even trillions of dollars for ordinary accounts [14]. Actual invoices were unaffected. AWS's own alarms detected the anomaly but did not stop the pipeline, and customers raised the alarm about four and a half hours later. AWS then paused estimated billing and disabled budget and cost anomaly alerts for all customers as a precaution [14]. At least one customer removed workloads in response [14]. A company whose only cost control is a provider budget alert lost that control for a day. Cost data needs its own owner and its own sanity checks, the same as any other financial feed.

SIDEBAR: BUDGET ALERTS ARE NOT SPENDING LIMITS

Provider budget alerts send a notification when spending crosses a threshold. They do not stop the spending. The July 2026 AWS incident showed that the alerts themselves can be switched off during a provider's outage [14]. A spending control that matters needs an internal owner, a check against your own usage data, and pre-approved authority to shut a workload down.

## Why the Usual Approach Fails

The common response to a high cloud bill is a cost-cutting sprint. A team hunts for idle servers, buys a round of commitments, and declares victory. Each of those steps is useful. The pattern fails for structural reasons.

Cleanup treats symptoms. Idle resources come back because nothing in the delivery process stops them from being created and forgotten. The 31-day average to find them means a month of waste for every new one [5].

Commitments get bought blind. AWS Savings Plans offer up to 72% off on-demand prices in exchange for a usage commitment [9]. Without reliable usage data by workload, a company either buys too little and leaves savings on the table or buys too much and locks in payment for capacity it will later shrink. Harness found more than half of developers described commitment purchases as guesswork [5].

Architecture costs are invisible in the invoice. Data transfer between regions, chatty services, oversized storage tiers, and logs retained forever show up as large aggregate line items with no owner. GEICO's experience shows how design choices made during migration can set the cost base for a decade [13].

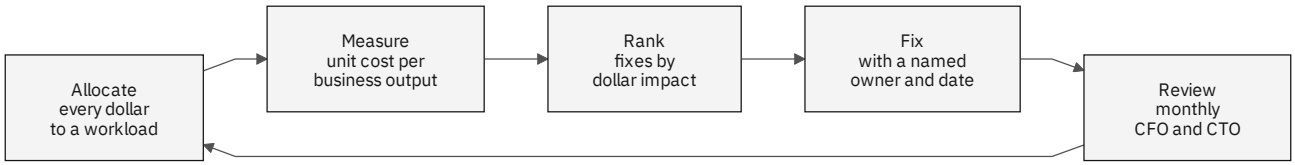
The unit cost has no owner. Finance holds the total and engineering runs the systems, yet the cost per order or per customer usually sits with no one, even though it is the number that tells you whether spending growth is healthy. The FinOps Framework treats unit economics as a core capability for exactly this reason [7].

## The Framework: Measure, Rank, Own

The model is simple. Measure spend by workload and by unit of business output. Rank fixes by dollar impact. Give each cost line and each fix a named owner who answers for it at a monthly review.

---

**Exhibit 1. The cost control loop: every dollar maps to a workload, and every workload maps to a person**




Source: author's framework, drawing on the FinOps Framework 2025 [7] and the AWS Well-Architected cost optimization principles [6].

**Allocate.** Tag or account-separate every resource so it rolls up to a workload, and assign each workload to one accountable person. The open FOCUS billing specification, version 1.2 ratified in May 2025, gives a common format for cost data across AWS, Microsoft, Google, Oracle, and some software vendors, which makes this practical in multi-cloud estates [8]. Target at least 90% of spend allocated before making large decisions.

**Measure.** Divide each workload's cost by the business output it supports: orders processed, active customers, invoices issued. AWS's guidance calls this measuring overall efficiency, tracking business output alongside the cost to deliver it [6]. A rising bill with a falling unit cost is growth. A rising unit cost is a problem regardless of the total.

**Rank.** List every candidate fix with its estimated annual savings, the effort it takes, and the risk it carries. Work down the list in order of dollar impact. Exhibit 2 shows the usual levers and where each tends to fall.

**Exhibit 2. Six levers, from quickest to deepest: rank them by dollar impact in your own estate**

| Lever | What it involves | Evidence on impact | Typical owner |
|---|---|---|---|
| Shut off idle and orphaned resources | Delete unattached storage, idle instances, forgotten environments | 21% of infrastructure spend estimated as underused (2025) [5] | Workload owner |
| Schedule non-production | Turn off dev and test environments outside working hours | Up to 75% saving on those environments [6] | Engineering lead |
| Rightsize | Match instance and database sizes to measured load | Rightsizing took teams about 25 days on average (2025) [5]; savings vary by estate | Workload owner |
| Commit | Buy savings plans or reservations against measured baseline usage | Up to 72% off on-demand pricing [9] | FinOps lead with CFO approval |
| Redesign | Cut data transfer, re-tier storage, refactor lift-and-shift applications | GEICO bills rose about 2.5 times after lift-and-shift (2024 account) [13] | CTO and architecture lead |
| Relocate | Move steady, predictable workloads to owned or leased hardware | 37signals cut about $2 million a year from a $3.2 million bill [10][12] | CTO and CFO jointly |

Source: author's analysis; impact figures from the cited sources. Savings in any one estate depend on its starting point.

Commitments sit in the middle of the list on purpose. Buy them after cleanup and rightsizing, so the commitment covers the smaller, measured baseline.

**Own.** Every line on the ranked list gets a person and a date. Each workload owner reports a monthly unit cost and explains it when it moves. Executives have to set this expectation, because a FinOps team without authority can only recommend.

# Costs, Risks, and Tradeoffs

**Exhibit 3. What this approach costs, and when it is the wrong call**

| Area | Cost or risk | Mitigation |
|---|---|---|
| Engineering capacity | Tagging, refactoring, and reviews take time from feature work | Fund cost work as a standing share of engineering capacity, sized to the savings it returns |
| Tooling | Allocation and reporting tools carry their own license cost | Start with provider-native tools and FOCUS-format exports [8] before buying a platform |
| Commitment lock-in | Over-committing turns a variable cost into a fixed one | Commit only against measured baselines after cleanup; ladder commitments across terms |
| Relocation risk | Owning hardware brings capacity planning, hardware failure, and staffing risk | Relocate only steady workloads, with a tested exit plan and staff who have run hardware before |
| Culture | Engineers may read cost ownership as blame | Review unit cost trends; do not tie individual pay to the raw bill |
| When this is wrong | A small company with a modest bill and few engineers gets little from per-workload allocation and monthly reviews | Use provider budgets, a quarterly cleanup, and a single named owner for the whole bill |

Source: author's analysis.

Relocation deserves its own caution. 37signals is a well-run company with a steady workload and a team that had operated hardware before [10][12]. A business with spiky demand, a small operations team, or heavy use of managed services faces a different calculation.

# How to Act

**Exhibit 4. A one-year roadmap**

| Window | Actions | Owner | Milestone |
|---|---|---|---|
| First 30 days | Export billing data in FOCUS format. Map every account and resource group to a workload. Name one owner per workload. | CTO with FinOps lead | At least 80% of spend allocated to a named owner |
| First 30 days | Pick one unit metric per major workload, such as cost per order or per active customer. | CFO with business unit leaders | Unit metrics agreed for the top five workloads by spend |
| First 90 days | Delete idle and orphaned resources. Schedule non-production environments. Rightsize the top 20 workloads by spend. Add cost checks to the deployment pipeline. | Engineering leads | Idle resources found within 7 days instead of 31; at least 90% of spend allocated |
| First 90 days | Buy commitments against the post-cleanup baseline. Set an internal anomaly check on your own usage data alongside provider alerts. | FinOps lead with CFO approval | Commitment coverage policy approved; anomaly alerts reach a named owner |
| First 365 days | Review architecture of the five most expensive workloads, starting with data transfer, storage tiers, and lift-and-shift systems. Evaluate relocation only for steady workloads with stable demand. | CTO and architecture lead | Unit cost down for the top five workloads; written decision on each relocation candidate |
| Ongoing | Monthly review of unit cost and the ranked fix list. | CFO and CTO jointly | Review held every month with owners present |

Source: author's recommendation.

# Questions to Ask Your Team

1. What share of last month's cloud bill can we assign to a specific workload and a specific person?
2. What is our cost per order, per customer, or per transaction, and which way has it moved over six months?
3. Which five workloads cost the most, and who owns each one?
4. How long does an idle resource survive in our environment before someone deletes it?
5. Were our current commitments sized from measured usage, and how much of them are we using?
6. Which of our systems were moved to the cloud without being redesigned?
7. If our provider's budget alerts went dark for a day, what would tell us that spending had jumped?

## Closing

The cloud invoice is the last step in a chain of engineering decisions. Finance can negotiate the rate. The volume comes from design choices and habits that engineering controls, and executives are the ones who can hold engineers accountable for them. Measure spend by workload and by unit of output, fix the largest items first, and give every cost a named owner who explains it at a monthly review shared by the CFO and CTO.

# Sources

[1] Flexera. "Flexera Finds Cloud Value is Rising While AI Waste Grows." Press release, 2026 State of the Cloud Report. 2026. https://www.flexera.com/about-us/press-center/flexera-finds-cloud-value-is-rising-while-ai-waste-grows.

[2] CFOtech UK. Coverage of Flexera 2026 State of the Cloud Report. 2026. https://cfotech.co.uk/story/flexera-warns-ai-cloud-costs-strain-technology-budgets.

[3] FinOps Foundation. "State of FinOps 2026." 2026. https://data.finops.org/.

[4] BizTechReports. Coverage of IDC Worldwide Software and Public Cloud Services Spending Guide. 2026. https://www.biztechreports.com/news-archive/2026/3/20/global-public-cloud-spending-to-surpass-1-trillion-in-2026-driven-by-paas-and-ai-platform-adoption-idc-march-23-2026.

[5] Harness. "FinOps in Focus Report 2025." Press release; survey of 700 developers and engineering leaders by Coleman Parkes. 2025. https://harness.io/press-and-news/finops-in-focus-report-2025.

[6] Amazon Web Services. "Design principles." AWS Well-Architected Framework, Cost Optimization Pillar. https://docs.aws.amazon.com/wellarchitected/latest/framework/cost-dp.html.

[7] FinOps Foundation. "FinOps Framework 2025." 2025. https://www.finops.org/wp-content/uploads/2025/05/English-FinOps-Framework-2025.pdf.

[8] DBTA. "The FinOps Foundation Unifies Format for Cloud Billing Data with Version 1.2 of FOCUS." 2025. https://www.dbta.com/Editorial/News-Flashes/The-FinOps-Foundation-Unifies-Format-for-Cloud-Billing-Data-with-Version-12-of-FOCUS-169986.aspx.

[9] Amazon Web Services. "Savings Plans." https://aws.amazon.com/savingsplans/.

[10] Hansson, David Heinemeier. "Our cloud exit savings will now top ten million over five years." 2024. https://world.hey.com/dhh/our-cloud-exit-savings-will-now-top-ten-million-over-five-years-c7d9b5bd.

[11] Hansson, David Heinemeier. "It's five grand a day to miss our S3 exit." 2025. https://world.hey.com/dhh/it-s-five-grand-a-day-to-miss-our-s3-exit-b8293563.

[12] The Register. "37signals is completing its on-prem move, deleting its AWS account to save millions." 2025. https://www.theregister.com/2025/05/09/37signals_cloud_repatriation_storage_savings.

[13] The Stack. Interview with GEICO's vice president of platform and infrastructure engineering on cloud repatriation. 2024, updated 2025. https://thestack.technology/warren-buffetts-geico-repatriates-work-from-the-cloud-continues-ambitious-infrastructure-overhaul.

[14] InfoQ. Report on the July 2026 AWS billing estimates incident. 2026. https://www.infoq.com/news/2026/07/aws-billing-estimates-incident/.

---

The views in this paper are the author's own and do not represent any employer. Scenarios are hypothetical. This paper is not legal, financial, or compliance advice.