

EXECUTIVE PAPER · INFRASTRUCTURE

Some Workloads Belong Back On-Premises

Cloud-first and cloud-exit are both policies that ignore the workload. CFOs and CTOs should place each system where its demand, data and risk make it cheapest to run well, and check the answer every year.

AUTHOR

Chris Watkins

SERVICE

Infrastructure

FOR

CEO, CFO, CTO · 11 min read

AT A GLANCE

The problem. Most companies decided where to run their systems once, as a policy, and have not revisited it. Cloud bills for steady workloads keep rising, while the cost of leaving a cloud provider has dropped since 2024. Surveys show most large companies intend to move some workloads back, yet fewer than one in ten plan a full exit.

Why it persists. Placement is argued as a matter of belief. Cloud advocates compare a discounted invoice to list-price hardware; exit advocates compare hardware to the invoice and leave out staff, power and migration. Neither side models the specific workload.

The recommendation. Score every major workload on demand shape, data volume, rate of change, operating skill and risk. Put steady, predictable, data-heavy systems on owned or colocated hardware when a full cost model favors it. Keep spiky, fast-changing and short-lived systems in the cloud. Repeat the review every year, because prices, exit fees and hardware markets now move fast enough to flip the answer.

The Decision in Front of You

For a decade the default answer was simple: new systems go to the public cloud, and old ones follow when budget allows. That default is now being questioned in the boardroom. In the 2024 Barclays CIO Survey, 83% of respondents said they planned to move some workloads back to private cloud or on-premises infrastructure [1]. IDC found in 2024 that only 8 to 9% of companies planned to repatriate everything [2].

Those two figures describe the real decision. Almost nobody is leaving the cloud. Many companies have found that certain workloads cost more there than they should, and are moving those specific workloads. Flexera's 2026 survey of more than 750 cloud decision-makers found 73% of organizations already run hybrid environments [3], and coverage of the same report put the share of workloads moved out of public cloud at 23%, up from 21% a year earlier [4].

The choice for executives is between a blanket policy and a placement decision made per workload with numbers. A blanket policy is easy to explain and expensive to run. On a steady, data-heavy workload, a wrong placement can cost a mid-size company several hundred thousand dollars a year. On a spiky one, the error runs the other way.

What's Happening

Companies move pieces, rarely everything

A 2025 Node4 survey of 601 mid-market executives found 49% planned to move a few applications out of public cloud in the following year, 43% a significant number, and 5% all of them [5]. The main reason given was performance: applications that ran slower in the cloud than users would accept [5].

IDC's research shows what moves. Companies most often bring back production data, backup and disaster recovery, and compute capacity, and larger organizations repatriate more than smaller ones [2] [6]. Cost is the common thread. In IDC's late 2023 survey, about half of cloud buyers had spent more than expected that year, and 59% expected overruns again in 2024 [2].

Leaving got cheaper

Data transfer fees, often called egress fees, were long the main penalty for leaving a cloud provider. Google Cloud stopped charging them for customers switching providers in January 2024 [7]. AWS followed in March 2024. Its terms have tightened and clarified since: as of October 2026, the free transfer covers moving all data, or all data for one service, off AWS, and customers outside the EU have 180 days to finish [8].

Regulation is pushing the same way. The EU Data Act limits switching charges to the provider's direct costs during a transition period, and from January 12, 2027 bans them entirely [9]. In the UK, the Competition and Markets Authority's July 2025 cloud market decision recommended considering special status designations for AWS and Microsoft [10]. Lower exit costs make it cheaper to test placement decisions and to reverse them.

Owning hardware got harder to price

The other side of the comparison is moving too. CBRE reported in September 2026 that vacancy in the eight largest North American data center markets was 1.4%, and that asking rents rose in every deployment size during the first half of the year, by as much as 8.3% [11]. In Northern Virginia, the largest market, vacancy was 0.2% [11].

Component prices have been volatile. AI demand absorbed memory and flash supply through 2025 and 2026, and TrendForce forecasts summarized in one 2026 analysis called for NAND contract prices to rise 55 to 60% in a single quarter [12]. A storage array bought in 2024 may cost far more to replace in 2027. Any placement model that assumes last year's hardware prices is already out of date.

AI workloads sharpen the question

GPU capacity is the most expensive compute most companies have ever rented. Deloitte's Tech Trends 2026 report advises evaluating on-premises investment once cloud costs for an AI workload reach 60 to 70% of the cost of equivalent hardware, and suggests keeping variable training work in the cloud while

running steady production inference on owned infrastructure [13]. The logic matches every other workload: steady demand favors ownership, variable demand favors renting.

The argument is older than AI. Andreessen Horowitz estimated in 2021 that committed cloud spend averaged 50% of cost of revenue at public software companies that disclosed it, and that repatriating could cost a third to a half as much as the cloud equivalent [14]. Those figures are five years old and came from interviews, so treat them as background for the argument.

Two recent cases

37signals, 2022 to 2025. The company behind Basecamp and HEY was paying AWS about $3.2 million a year. After moving compute to roughly $700,000 of Dell servers, its 2024 bill fell to about $1.3 million, all of it for Amazon S3 storage, with the same operations team [15]. In 2025 it moved almost six petabytes off S3, which cost just under $1.5 million a year, onto about $1.5 million of purchased Pure Storage hardware with under $1 million of support over five years [16]. AWS waived roughly $250,000 in transfer fees [17]. The company projects savings above $10 million over five years [15]. These are the company's own figures and projections. A separate paper in this series covers its cost-control lessons; the placement lesson is that 37signals runs steady workloads with an experienced team, the profile where ownership pays.

Figma, 2025. Figma's July 2025 IPO filing reported $749 million of 2024 revenue, up 48% in a year, and disclosed that on May 31, 2025 it renewed its cloud hosting agreement, committing to spend at least $545 million over five years [18]. Its platform runs on AWS and its software is built around that provider's services [18]. That is about $109 million a year, or roughly 15% of 2024 revenue. A company growing 48% a year cannot size owned hardware with much confidence, and rebuilding around another platform would slow product work, so staying is a defensible placement. The commitment is also a placement decision with numbers attached, made for a fixed term and due for review at renewal.

GEICO offers a third pattern: after a decade of mostly unchanged applications in the cloud, it is moving storage-heavy work to a private cloud while keeping other work in public cloud [19].

Why the Usual Approach Fails

Placement is usually decided by policy. A "cloud-first" mandate sends every new system to a provider regardless of its shape. A "cloud exit" campaign, often launched after a painful invoice, pulls everything back regardless of fit. Both skip the per-workload arithmetic.

The comparisons are rigged by omission. A cloud business case uses negotiated discounts and leaves out data transfer, support plans and the engineering time spent managing cost. A repatriation case uses the hardware quote and leaves out the people, power, space, spares, refresh cycle and the months of migration. 37signals could run its servers with the same team [15]; a company without that skill would need to hire it.

Idle capacity is ignored. Owned hardware is sized for peak demand and paid for whether it is busy or not. A system that runs flat around the clock uses what it buys, while one that peaks for six weeks a year carries capacity that sits idle for the other forty-six.

The decision is never revisited. Prices, discounts, exit fees, colocation rents and component costs all changed materially between 2024 and 2026 [8][9][11][12]. A placement that was right three years ago may be wrong now in either direction.

The Framework: The Placement Scorecard

Score each major workload on five tests. Workloads that score toward the left column are candidates for owned or colocated hardware. Workloads that score toward the right belong in the cloud. Mixed scores call for a full cost model.

Exhibit 1. The Placement Scorecard

Test	Favors owned or colocated	Favors public cloud
Demand shape	Flat; peak under about 1.5 times average	Spiky, seasonal or unknown; peak several times average
Data	Large, stable data that stays put or is read heavily	Small data, or data that must sit next to cloud-only services
Rate of change	Mature system, stable architecture for three or more years	New product, fast experimentation, frequent redesign
Operating skill	Team already runs hardware, networks and storage	No in-house infrastructure team, or a team that is hard to hire
Risk and lead time	Can wait weeks for hardware; data residency or control matters	Needs capacity in hours; needs many regions; uses managed services hard to replace

Source: author's framework, consistent with IDC and Deloitte findings on which workloads move [2][13].

For each candidate, build a five-year model that counts everything on both sides. The following illustrative estimate shows why demand shape matters more than any other test.

Exhibit 2. Illustrative annual cost of one workload, cloud versus owned

Cost line	Steady workload, owned	Spiky workload, owned (sized for 3x peak)
Hardware, $900,000 or $2.7 million amortized over 5 years	$180,000	$540,000
Colocation power and space, 60 kW or 180 kW at $250 per kW-month	$180,000	$540,000
Added staff, 1.5 people at $200,000 loaded	$300,000	$300,000
Hardware support at 10% of purchase price per year	$90,000	$270,000
Network and transit	$36,000	$36,000
Migration, $250,000 one-time spread over 5 years	$50,000	$50,000
Total owned per year	$836,000	$1,736,000
Cloud, $90,000 per month after discounts	$1,080,000	$1,080,000
Result	Owned saves $244,000 a year (23%)	Cloud saves $656,000 a year

Source: author's illustrative estimate. Assumes the cloud bill tracks average use in both cases through autoscaling, while owned hardware must be sized for peak. Colocation rate is an assumption for a small deployment; CBRE reports rents rising in 2026 [11]. Staff, support and migration figures are assumptions; substitute your own.

The steady case saves about $1.2 million over five years, and the margin is thin. If the workload needs 2.7 added people instead of 1.5, the savings disappear. That is why operating skill sits on the scorecard.

Costs, Risks, and Tradeoffs

Exhibit 3. Tradeoffs of per-workload placement

Area	Cost or risk	Mitigation
Migration	Moves take months and can fail in the middle	Move one workload at a time with a rollback path; time it to contract renewals
Skills	Hardware, storage and network skills must be hired or rebuilt	Use colocation remote-hands services; start with storage and backup, which need fewer skills
Capacity lead time	Hardware and power can take months to secure in tight markets	Reserve colocation space early; keep cloud as overflow for growth
Hardware price swings	Memory and flash prices can jump within a quarter	Price refreshes at purchase time; keep the five-year model current
Lost cloud discounts	Moving spend away can lower committed-use discounts on what remains	Model the discount tiers before moving; renegotiate at renewal
Two operating models	Running cloud and owned infrastructure doubles some tooling and process	Use the same deployment and monitoring tools on both
When this is wrong	A company with spiky or unpredictable demand, fast product change, heavy reliance on managed cloud services, or no infrastructure team will usually pay more to own hardware	Stay in the cloud and spend the effort on cost engineering and commitment management

Source: author's analysis.

How to Act

Exhibit 4. A one-year roadmap

Window	Actions	Owner	Milestone
First 30 days	List the 20 workloads with the largest cloud or data center cost. Record monthly cost, peak-to-average demand, data volume and contract dates for each.	CTO with finance	Inventory covering at least 80% of infrastructure spend
First 90 days	Score each workload on the Placement Scorecard. Build full five-year cost models for the top five candidates in each direction, including staff, power, migration and exit fees.	Infrastructure lead with FP&A	Five costed placement cases presented to CFO and CTO
First 90 days	Confirm current exit terms with each provider, including the AWS and Google transfer waivers and, for EU operations, Data Act switching rights [7][8][9].	Procurement with legal	Written exit terms on file for every major provider
First 365 days	Move one or two workloads where the model shows at least 20% savings with a skills plan in place. Measure actual cost and performance against the model monthly.	CTO	First moved workload running at or below modeled cost for three months
First 365 days	Hold the first annual placement review. Re-score every workload with current prices and decide moves for the next year.	CFO and CTO jointly	Review completed and placement decisions recorded in the budget

Source: author's recommendation.

Questions to Ask Your Team

1. Which ten workloads cost us the most to run, and where does each one run today?
2. For each, what is the ratio of peak demand to average demand over the past year?
3. What would it cost to run our largest steady workload on owned or colocated hardware, counting staff, power and migration?
4. Which workloads would we struggle to move because they depend on services only one cloud provider offers?
5. When do our major cloud contracts and commitments renew, and what are our exit terms?
6. Do we have the people to run hardware, or would we need to hire them?
7. When did we last check whether a workload is in the right place?

Closing

Most companies will run a mix of cloud and owned infrastructure, and the money is in getting each workload into the right part of that mix. Steady, data-heavy systems with an experienced team behind them often cost less on owned or colocated hardware. Spiky, fast-changing systems usually cost less in the cloud. Score each workload, model the full cost, move only where the numbers hold, and repeat the exercise every year as prices and exit terms change.

Sources

[1] Pulsant. "Cloud repatriation." E-book citing the Barclays CIO Survey 2024. 2025. https://www.pulsant.com/knowledge-hub/ebook/cloud-repatriation.

[2] IDC. "Storm Clouds Ahead: Missed Expectations in Cloud Computing." Blog citing IDC Server and Storage Workloads Survey and Cloud Pulse 4Q 2023. October 28, 2024. https://idc.com/resource-center/blog/storm-clouds-ahead-missed-expectations-in-cloud-computing.

[3] Flexera. "Flexera Finds Cloud Value is Rising While AI Waste Grows." Press release, 2026 State of the Cloud Report. March 18, 2026. https://www.flexera.com/about-us/press-center/flexera-finds-cloud-value-is-rising-while-ai-waste-grows.

[4] TecnoAndroid. "Cloud repatriation: il 23% dei workload lascia il cloud." Coverage of Flexera 2026 State of the Cloud Report. September 22, 2026. https://www.tecnoandroid.it/news/cloud-repatriation-il-23-dei-workload-lascia-il-cloud-1974900/.

[5] ITPro. "Enterprises are keen on cloud repatriation, but not for all workloads." Coverage of Node4 survey of 601 mid-market executives. June 4, 2025. https://www.itpro.com/cloud/cloud-computing/enterprises-are-keen-on-cloud-repatriation-but-not-for-all-workloads.

[6] The Register. Coverage of IDC research on cloud repatriation of specific workloads. October 30, 2024. https://www.theregister.com/2024/10/30/cloud_repatriation_about_specific_workloads/.

[7] Light Reading. "AWS follows Google Cloud in ending cloud switching costs." March 6, 2024. https://www.lightreading.com/cloud/aws-follows-google-cloud-in-ending-cloud-switching-costs.

[8] Amazon Web Services. "Free data transfer out to internet when moving out of AWS." AWS News Blog. March 5, 2024, updated September 30, 2025 and October 8, 2026. https://aws.amazon.com/blogs/aws/free-data-transfer-out-to-internet-when-moving-out-of-aws/.

[9] European Union. Regulation (EU) 2023/2854 (Data Act), Article 29, "Gradual withdrawal of switching charges." 2023. https://www.springlex.eu/en/packages/data-act/data-act-regulation/article-29/.

[10] UK Competition and Markets Authority. "Cloud services market investigation." Final decision. July 31, 2025. https://www.gov.uk/cma-cases/cloud-services-market-investigation.

[11] CBRE. "North American Data Center Demand Continues to Outpace Supply Despite Record Construction." Press release, North America Data Center Trends H1 2026. September 8, 2026. https://www.cbre.com/press-releases/north-american-data-center-demand-continues-to-outpace-supply-despite-record-construction.

[12] Feng, Ruohang. "Cloud Exit 2026." Analysis summarizing TrendForce memory and flash price forecasts. August 15, 2026. https://blog.vonng.com/en/cloud/cloud-exit-2026/.

[13] Digital CxO. "Deloitte Tech Trends 2026: Moving from AI Experimentation to Enterprise Impact." Coverage of Deloitte Tech Trends 2026. December 15, 2025. https://digitalcxo.com/?p=23761.

[14] Wang, Sarah, and Casado, Martin, Andreessen Horowitz. "The Cost of Cloud, a Trillion Dollar Paradox." May 27, 2021. https://a16z.com/the-cost-of-cloud-a-trillion-dollar-paradox/.

[15] Hansson, David Heinemeier. "Our cloud exit savings will now top ten million over five years." October 17, 2024. https://world.hey.com/dhh/our-cloud-exit-savings-will-now-top-ten-million-over-five-years-c7d9b5bd.

[16] Hansson, David Heinemeier. "It's five grand a day to miss our S3 exit." March 26, 2025. https://world.hey.com/dhh/it-s-five-grand-a-day-to-miss-our-s3-exit-b8293563.

[17] The Stack. Coverage of 37signals' S3 exit and AWS egress waiver. May 8, 2025. https://www.thestack.technology/dhh-aws-egress-s3-pure/.

[18] Figma, Inc. Form S-1 Registration Statement. U.S. Securities and Exchange Commission. July 1, 2025. https://www.sec.gov/Archives/edgar/data/1579878/000162828025033742/figma-sx1.htm.

[19] The Stack. Interview with GEICO's vice president of platform and infrastructure engineering on cloud repatriation. October 17, 2024, updated October 20, 2025. https://thestack.technology/warren-buffetts-geico-repatriates-work-from-the-cloud-continues-ambitious-infrastructure-overhaul.

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.