// WHITE PAPER · PUBLISHED 16 SEPTEMBER 2026

Placement Economics: The Right Answer, Applied Everywhere

Cloud, colocation, and the case for deciding one workload at a time

Part 5 of 10 in *The Money Language Series*. Each piece stands on its own. The others: Part 1 · Part 2 · Part 3 · Part 4 · Part 6 · Part 7 · Part 8 · Part 9 · Part 10.

Jason Newell · MAX Research Collective · Report WPR-BIZ-006 · Published 16 September 2026
Licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/
Original: https://www.maxresearchcollective.com/articles/placement-economics.html
Cite as: Newell, J. (2026). Placement Economics: The Right Answer, Applied Everywhere. MAX Research Collective, WPR-BIZ-006. https://www.maxresearchcollective.com/articles/placement-economics.html

The most expensive infrastructure decisions I have reviewed were right about the technology. What went wrong was the scope: a conclusion drawn from one workload got applied to the whole portfolio, and it stayed applied for years after it stopped being true of anything but the workload it came from.

”Cloud-first” was a policy. ”Repatriate everything” is becoming one. The workload is the unit of decision, and a company running fifty workloads owes itself fifty answers; most of those answers agree with each other, and the few that disagree are where the money is.

The three placements

	PUBLIC CLOUD	COLOCATION	OWNED FACILITY
You rent	Compute, storage, services	Space, power, cooling, network	Nothing
You own	Nothing	The hardware	Everything
Capital required	None	Moderate, the hardware	High, hardware and building
Cost structure	Operating expense, metered	Capital hardware and operating facility	Heavy capital and operating
Scaling down	Minutes	Not really	No
Scaling up	Minutes, at a premium	Weeks to months, on procurement	Months
Staff needed	Cloud and platform engineers	Platform, hardware lifecycle, remote hands	All of that plus facilities
Realistic fit	Almost everyone, for something	Steady workloads at scale	Hyperscalers and regulated giants

Owned facilities sit outside this piece. The live question for most technology leaders is cloud against colocation, per workload, with a hybrid answer at the end.

The repatriation evidence, and what it rests on

The best-documented case is 37signals. They were spending roughly $3.2 million a year on cloud. They bought about $700,000 of Dell hardware into colocation racks they already had and reported close to $2 million a year in compute savings, which is about two-thirds of the original bill. They then spent around $1.5 million on 18 petabytes of Pure Storage to replace S3; that storage costs under $200,000 a year

to run. Their projection is cumulative savings above $10 million over five years, on faster machines with more capacity, which is the sort of number that settles an argument in a board meeting and starts a different one in the room where somebody has to keep the hardware alive.

Dropbox did the same thing earlier and larger, moving roughly 90 percent of storage off AWS into three colocation facilities, investing over $53 million and estimating about $75 million saved across two years. Pointing the other way, GEICO's platform engineering leadership described a decade-long cloud journey in which they still had not moved everything, the bills had risen around 2.5 times, and reliability had got worse rather than better.

An Andreessen Horowitz survey found that among companies spending over $2 million a year on cloud, a large majority had repatriated at least one workload, and cost was the dominant reason given. Read the sentence again. A majority moved at least one workload, which is a very different finding from a majority leaving, and it is the finding that matches what the rubric below produces: a split rather than a direction.

Now the conditions, because the numbers above travel without them. 37signals is a mature and profitable company with steady traffic, an exceptionally capable operations team, and colocation relationships already in place. Growth uncertainty, which is what usually kills an owning case, is close to zero there. That combination is rare. They were also running expensive managed services, Aurora and OpenSearch among them, and a disproportionate share of the savings came from leaving those behind (their CTO has said so publicly). The published figures also do not carry the full cost of future hardware refreshes or the operations roles needed to sustain the estate.

None of that makes the case invalid. It makes the case conditional, and the conditions are the whole point.

A scoring rubric

The rubric below scores one workload at a time. Higher totals point toward owning and lower totals point toward cloud. It structures the conversation rather than

settling it, and the first question it forces is the useful one: which of the fifty workloads is this decision actually about?

FACTOR	0 POINTS, CLOUD	1 POINT	2 POINTS, OWN
Demand variance	Spiky, seasonal, unpredictable	Moderate swings	Flat and predictable
Utilization	Under 25 percent average	25 to 50 percent	Over 60 percent sustained
Data volume	Small	Moderate	Multi-petabyte
Egress	Minimal	Some	Heavy and continuous
Workload age	New, direction uncertain	Established	Commodity, years of stability
Managed service dependency	Deep: Aurora, BigQuery, Lambda	Some	Mostly open source and portable
Team capability	No infrastructure bench	Some platform skill	Real SRE and hardware lifecycle experience
Capital access	Constrained	Available	Cheap and available
Compliance	Needs inherited certifications	Mixed	Already runs its own controls
Growth horizon	Could grow tenfold or die	Steady growth	Known and bounded

A total of 0 to 6 is a cloud workload and there is not much to argue about. A total of 7 to 13 is hybrid territory, where the answer sits in the specific line items rather than in the score. A total of 14 to 20 is a colocation case worth building properly.

Here is the pattern that shows up most often. A company scores 3 on its customer-facing tier and 16 on its storage and analytics tier — one instrument, two answers —

and then makes a single decision covering both. The savings live almost entirely in the second group, and so does the operational burden. Splitting the two is usually right and is almost never what gets proposed first, because a split is more work than a policy.

Five-year cost, including the parts people leave out

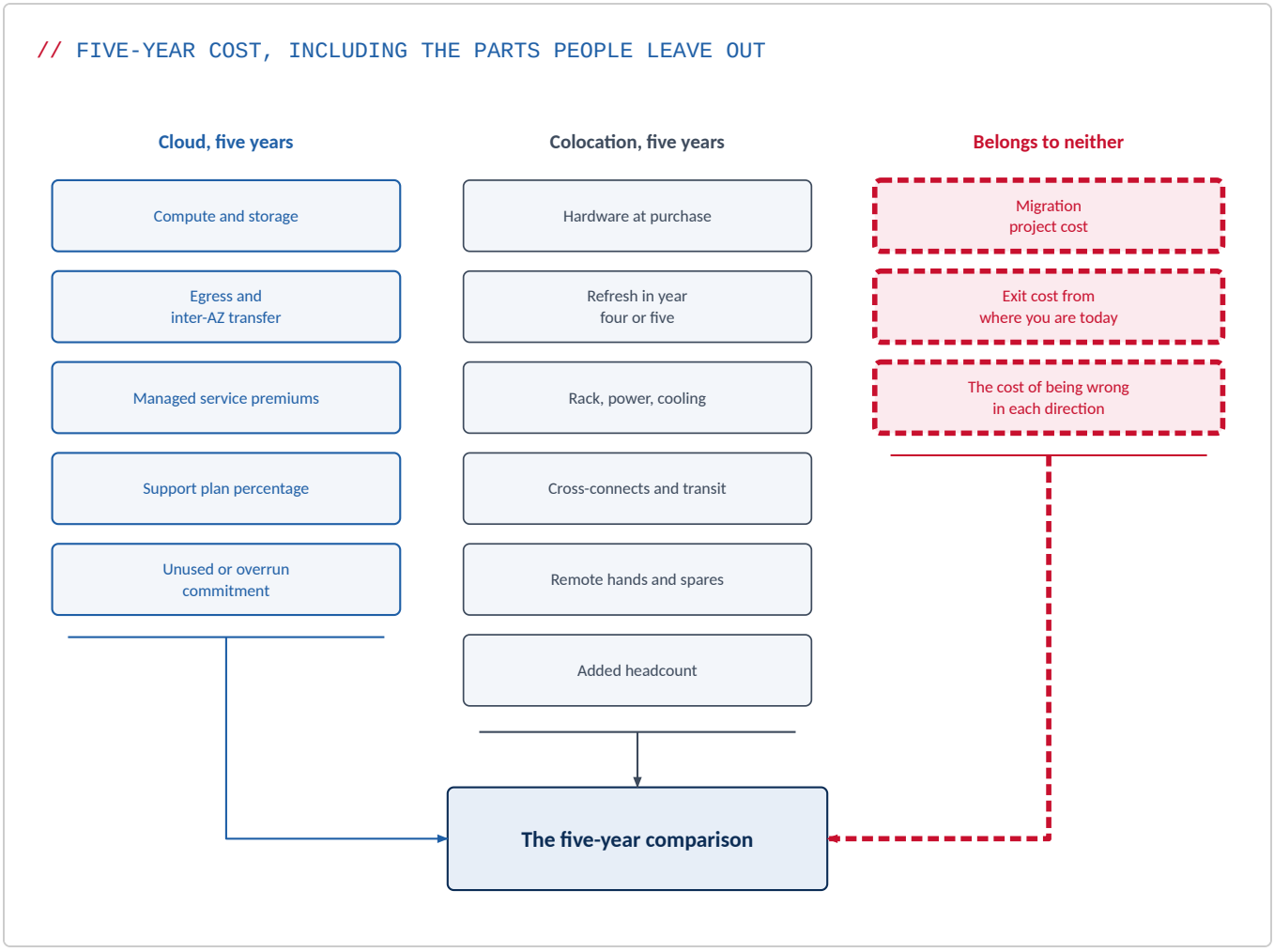


Figure 1. Two stacks and the one that belongs to neither. The third column is what decides most comparisons and it is the column most often missing from them.

The omissions, in rough order of how often they turn up missing:

1. Exit cost hides in four places: egress charges, data gravity, dual running, and rewriting whatever was built against a proprietary service. No migration case I

have reviewed has ever overstated it.

2. Hardware refresh belongs in any five-year model that sits on a four-year refresh policy, because that model needs a second purchase in it. Part 4 is about why those two numbers have to agree.
3. The headcount delta is real, since colocation has people in it. Hardware lifecycle and capacity planning are roles with salaries attached.
4. Support plan percentages scale with the thing they support. Cloud enterprise support is often priced as a share of spend, so it grows exactly when the bill does.
5. Commitment mismatch cuts both ways. Reserved capacity you never use is waste; usage above the commitment is priced at a premium. A model has to carry both tails.
6. The cost of being wrong is the line nobody writes down. If demand triples, what does each option cost, and what does it cost if demand halves instead?

Commitments are liabilities in a costume

Reserved Instances, Savings Plans and enterprise agreements all land on the profit and loss statement as operating expense. In substance they are multi-year contractual obligations, often with take-or-pay terms, which means you pay the committed amount whether or not you consume it.

A three-year commit signed against an optimistic forecast is a debt-shaped obligation that nobody on the technology side is thinking about as debt. When growth comes in under plan, the organization pays for capacity it never used, and the FinOps dashboard reports that as waste rather than as a commercial decision somebody made in a planning cycle two years earlier.

Flexera's 2026 report put self-estimated wasted infrastructure and platform spend at 29 percent, the first increase in five years, and attributed it largely to AI workloads making forecasting and rightsizing harder. That figure comes from a population where 63 percent run formal FinOps teams and 71 percent run a cloud centre of excellence. The discipline matured and the waste still went up. The workloads got harder to predict faster than the practice got better at predicting them.

Hybrid is the destination

Most organizations that run this analysis honestly land in the same shape: cloud for the variable and globally distributed workloads, colocation for the steady and data-heavy ones, and a deliberate policy about what is allowed to depend on a proprietary managed service. The shape is stable across industries because it follows the shape of the workloads rather than the shape of anybody's strategy deck.

That last clause is the one to settle early. Portability is a design decision, and it gets made long before anyone needs it. A managed database is a good trade in year one and a lock-in cost in year five; taking that trade is fine, and taking it without pricing it is the mistake.

WORKLOAD TYPE	TYPICAL PLACEMENT	WHY
Customer-facing product tier	Cloud	Variable, global, still evolving
Development, test, CI	Cloud, aggressively scheduled	Bursty, and it should be off at nights and weekends
Bulk object storage	Colocation above a petabyte or two	Storage and egress dominate, so the savings are large
Batch, ETL, analytics compute	Either, colocation at high utilization	Predictable, schedulable, dense
GPU training	Decided by utilization	Above roughly 60 to 70 percent sustained, owning wins
GPU inference	Cloud first, colocation at scale	The shape of demand decides
Disaster recovery	Cloud	Paying for idle capacity is what cloud is for

What I actually do

1. Refuse to make one decision for the whole portfolio. The deliverable is a scored list of workloads and a proposed split rather than a direction.
2. Build the five-year model with exit cost in it, in both directions, and present the break-even year rather than the total. A total invites a debate about assumptions; a break-even year invites a decision.
3. Put a portability policy into the architecture standards, naming which managed services are approved and what the escape plan is for each one (an escape plan nobody has costed is a wish).
4. Treat every multi-year commitment as debt inside the planning process. Who signed it, against what forecast, and what the downside costs.
5. Tag everything from the first day. None of this analysis is possible without knowing what runs where and for whom, and the same tags are what make the cost-of-revenue split in Part 2 possible.
6. Ask a provider for migration funding before committing to a direction, then ignore that money when choosing the architecture. Part 7 covers why those two have to stay separate.
7. Put the whole question back on the table once a year. The right call in 2023 is not automatically the right call now, in either direction.

Glossary

TERM	DEFINITION
Colocation	Renting space, power and cooling in someone else's data centre for hardware you own
Cross-connect	A physical network link to another tenant or a carrier inside the same facility
Data gravity	The tendency of a large data set to make everything around it expensive to move
Egress	What a cloud provider charges to move data out
FinOps	Managing cloud spend as a shared engineering and finance discipline
Portability	How cheaply a workload can be moved somewhere else
Remote hands	Data centre staff who physically service your equipment
Repatriation	Moving a workload from public cloud back to owned or colocated infrastructure
Reserved Instance, Savings Plan, committed use discount	Discounted capacity bought with a multi-year commitment
Take-or-pay	You pay the committed amount whether you use it or not
Cloud centre of excellence	A central governance function for cloud practice and spend

References

- DataCenterDynamics, coverage of the 37signals cloud repatriation, 2023 to 2025

- The New Stack, "Why Companies Are Ditching the Cloud: The Rise of Cloud Repatriation", carrying the GEICO commentary
- Andreessen Horowitz, cloud cost research and repatriation survey data
- Flexera, 2026 State of the Cloud Report
- Dropbox infrastructure migration disclosures and engineering blog

Best next read

Part 6, Lift-and-Shift against Re-Architect. You have decided where a workload should live. Part 6 is about how it moves, and why the cheap migration is so often the expensive one.

The full series

PART	TITLE
1	CapEx, OpEx, and EBITDA: The Same Dollar, Two Different Stories
2	Reading the Financials: The Primer for Everyone Who Never Had To Care
3	Capitalize or Expense: Your Jira Board Is Now Audit Evidence
4	Depreciation and Useful Life: The Chip Outlives Its Economics
5	Placement Economics: The Right Answer, Applied Everywhere, which is this one
6	Lift-and-Shift vs Re-Architect
7	Vendor Economics
8	AI Spend: Three Budgets Pretending to Be One
9	AI and Headcount
10	Outsourcing After Section 174A

This series is an introduction for technology leaders. It is accounting and tax context rather than accounting or tax advice. Treatment is fact-specific. Model it with your CFO and auditor before you commit.

CAREER & LEADERSHIP WPR-BIZ-006

// ABOUT THE AUTHOR

Jason Newell leads organizations and people in technology, and is a researcher, deep technologist and builder — from a CRAY to agentic AI. He founded MAX Research Collective to build AI for regulated industries that runs where its owner has control, with healthcare as its strongest focus. About Jason → Connect on LinkedIn →

Newell, J. (2026). MAX Research Collective, WPR-BIZ-006. CC BY 4.0