emma

INSIGHT BRIEF · CLOUD COST & RESILIENCE

Why You Can't Cut the Cloud Bill in Half

A banking and fintech leader's playbook for turning cloud cost, multi-cloud sprawl, and sovereignty from three separate fire drills into one control plane — under DORA

CTOS, CIOS, AND HEADS OF CLOUD

PLATFORM

INFRASTRUCTURE AT BANKS, FINTECHS, AND FINANCIAL INSTITUTIONS

Why You Can't Cut the Cloud Bill in Half — and the Number That Actually Moves

You've probably been handed a target to cut the cloud bill — ideally roughly in half.

You almost certainly can't hit it the way you've been asked to. And the three reflexes that promise to get you there each quietly make your real problem worse.

Public cloud spending is on track to pass roughly $1.48 trillion by 2029 (Gartner), and in financial services the line is growing faster than the market average — the US sector alone will spend close to half a trillion dollars on technology this year (Forrester).

For most banks and fintechs, cloud has become one of the largest controllable operating lines on the books. Which is exactly why it lands on your desk with a number attached.

But the number attached is usually the wrong one.

J.R. Storment, who runs the FinOps Foundation, has a line for it: the dirty little secret of cloud spend is that the bill never really goes down. The data bears him out — in Azul cloud trends survey of 300 CIOs, **83% spent more on cloud than they had forecast, by an average of 30%; only 2% came in under.**

And in financial services specifically, the squeeze is sharper: in Capgemini's survey of 600 financial-services executives across 13 markets, 84% of banks and insurers said they adopt cloud primarily to cut operating costs — yet fewer than 40% are satisfied with the outcomes, with inefficient cost management among the named reasons they fail to realise the value they paid for.

Managing cloud spend has been enterprises' top cloud challenge for years — 85% still name it in 2026 — and spend still rises every year.

Usage grows, AI lands on top of compute, and absolute spend climbs even when you're getting more efficient. Chase "half" as an absolute and you'll either fail — or cut something you were legally required to keep running.

From the field

Across 400+ recent cloud buying conversations we analyzed, cost came up in roughly **three out of four** (73%) and security or compliance in more **than four out of five** (83%).

But the most common cost pain wasn't "it's too expensive." It was a version of: "we only find out what we spent when the bill arrives."

The anxiety is rarely the number itself — it's not being able to see it, predict it, or steer it. And in a regulated institution, that same blind spot is now a supervisory problem, not just a finance one.

Three mandates, three reflexes — and why each one backfires

And in a regulated institution, that cost target never arrives alone. The same year that brings "cut the bill" from your CFO brings "address provider concentration risk" from your board reading DORA, and "reduce exposure to US-controlled infrastructure" from your compliance and sovereignty review. Three mandates, three owners, three deadlines — landing on the same desk, competing for the same platform team.

Each mandate triggers an obvious move. Each move is intuitive. And each, done the obvious way, makes the other two mandates harder — which is the real reason the fire drills never end.

Reflex 1 — The cost mandate: "Cut the bill in half."

The instinct is to attack the total. The trap is twofold.

First, you can't cut what you can't attribute. In most estates 15–30% of spend sits unallocated — shared services, multi-tenant platforms, Kubernetes clusters that serve many products at once. So the cuts land on the visible half while the growing half hides.

Second, the realistic prize is smaller and different than "half." Industry data puts cloud waste at about 29% in 2026 (Flexera's first uptick in five years, driven by AI). A disciplined first pass — eliminating idle and zombie resources, right-sizing, scheduling non-production, and buying commitments — typically takes 25–30% out of the bill. Real money.

But then the easy wins are gone. And the reflex to keep chasing an absolute target turns into pressure to defer the redundancy, the second region, the tested failover — the exact things DORA now obliges you to maintain.

You can cut cost or you can cut resilience. Sometimes, only one of those is yours to give away.

lustrative. Eliminating idle/zombie resources, right-sizing, commitments and non-prod scheduling are real, one-time wins; the absolute bill then tends to climb again with usage and AI. The durable target is the ratio, not the headline number.

Reflex 2 — The resilience mandate: "Go multi-cloud to de-risk."

And in a regulated institution, that cost target never arrives alone. The same year that brings "cut the bill" from your CFO brings "address provider concentration risk" from your board reading DORA, and "reduce exposure to US-controlled infrastructure" from your compliance and sovereignty review. Three mandates, three owners, three deadlines — landing on the same desk, competing for the same platform team.

Each mandate triggers an obvious move. Each move is intuitive. And each, done the obvious way, makes the other two mandates harder — which is the real reason the fire drills never end.

Reflex 3 — The sovereignty mandate: "Get off US suppliers/ go sovereign."

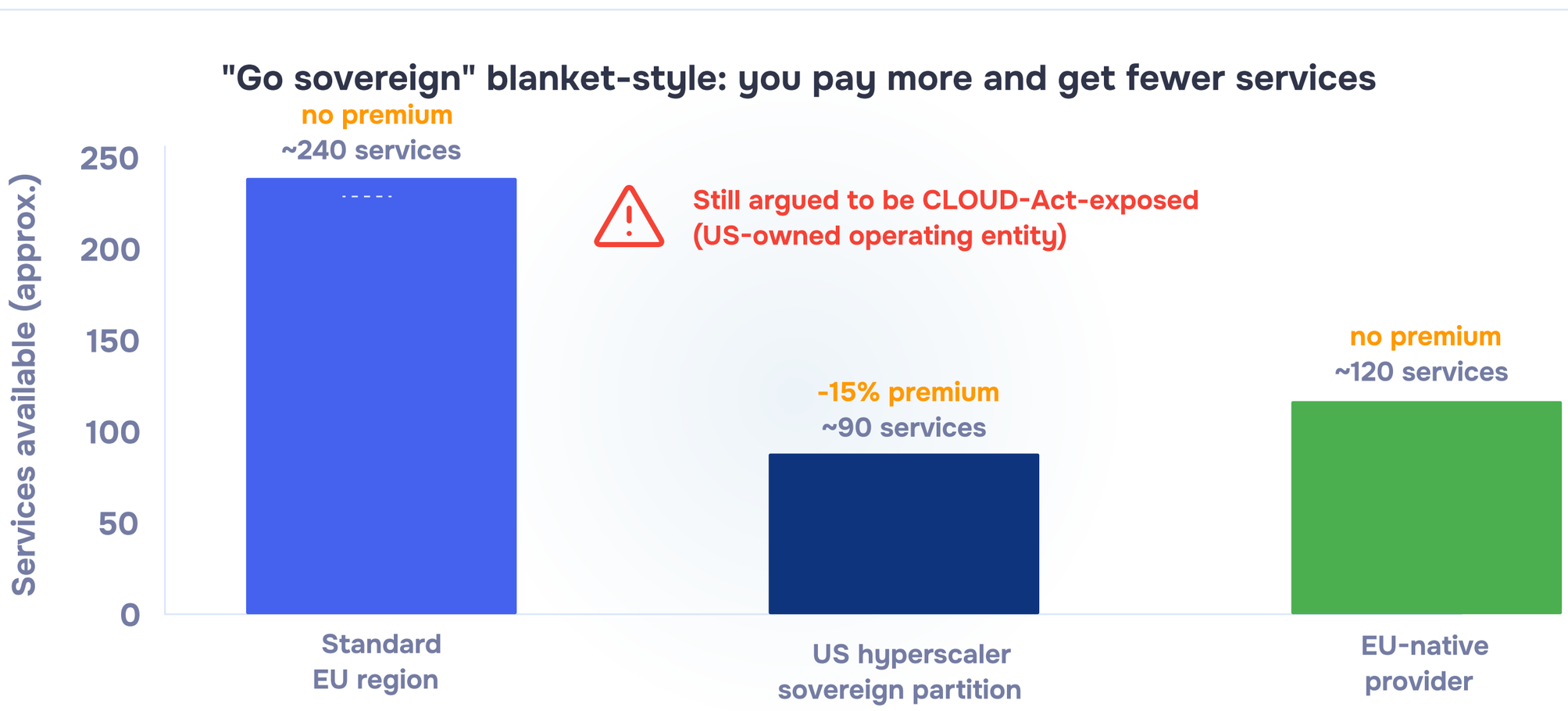
The pressure to move data off US-controlled infrastructure is real and rising; European sovereign-cloud spending is forecast to grow about 83% in 2026. But the obvious move — lift workloads onto a hyperscaler's new "sovereign" offering — carries hidden costs.

AWS's European Sovereign Cloud went live in January 2026 in Brandenburg, Germany, with independent EU-resident operations and its own identity, billing and DNS. Independent analysis puts it at roughly a 15% premium over standard EU regions, with about 90 services versus 240+ in the commercial regions, two availability zones at launch, and key services (including GPU instances) absent.

And the harder point: because the operating entity is still ultimately US-owned, critics — and a notable moment in front of the French Senate — argue it remains exposed to the US CLOUD Act.

So you can pay a premium, lose services, and deepen lock-in (exiting a sovereign partition is harder than a normal migration), and still not close the legal question you started with.

"Sovereign" is not a badge a vendor applies on your behalf — and neither is "compliant."



Illustrative, drawn from launch pricing/service-count analysis. Decide per workload by what its data actually requires-most estates find only a minority of workloads truly need the sovereignty premium

The one-sentence version:

The one-sentence version: your bill tells you what you spent — not who spent it, whether you can move it, or whether you could survive losing it. Answer those three questions and all three mandates can be addressed reliably.

The reframe: you don't have three problems. You have three unanswered questions.

Every reflex above broke in the same place.

You couldn't cut the bill because you couldn't say who spends it. You couldn't de-risk with a second provider because you couldn't actually move anything to it. You couldn't go sovereign because a migration you can't reverse — or survive — is just lock-in with a new flag on it.

Three mandates, one missing foundation. Underneath the finance project, the platform project, and the compliance project sit the same three questions:

01

Who spends it?

Every euro mapped to a product, a team, and a legal entity — including the shared clusters that defeat tagging. This is what lets you cut cost without cutting blind, and it's the same mapping your DORA register needs.

02

Can you move it?

A tested, timed answer — not a clause in a contract. This is what makes a second provider real, an exit credible, and sovereignty a placement choice instead of a one-way migration.

03

Can you survive losing it?

Known, rehearsed failure behavior for every critical function — including the provider's global control plane your "redundant" architecture secretly depends on. This is what concentration risk actually means to your supervisor.

Right now, finance, platform, and compliance are each building a partial version of these answers — their own tools, their own timelines, their own mandate. And because no team can see the other two mandates in its numbers, every win is someone else's setback: the cost cut lands on the redundancy your resilience plan requires; the second provider inflates the bill you were told to halve; the sovereignty migration deepens the lock-in your exit plan was supposed to dissolve

Answer the three questions once, on one operating model, and all three mandates draw from the same well: the attribution that cuts your bill is the register your supervisor audits; the exit you test for DORA is the portability that kills your lock-in premium and prices your sovereignty decision.

Two numbers prove you've done it

in a software company, cloud is cost of goods sold, and one number covers it: unit economics. In a bank, cloud is a supervised dependency, so the wall gets two.

Cost-to-serve per product or legal entity

the ratio, trended, not the absolute bill. This is "who spends it," made measurable. It answers: was it worth it?

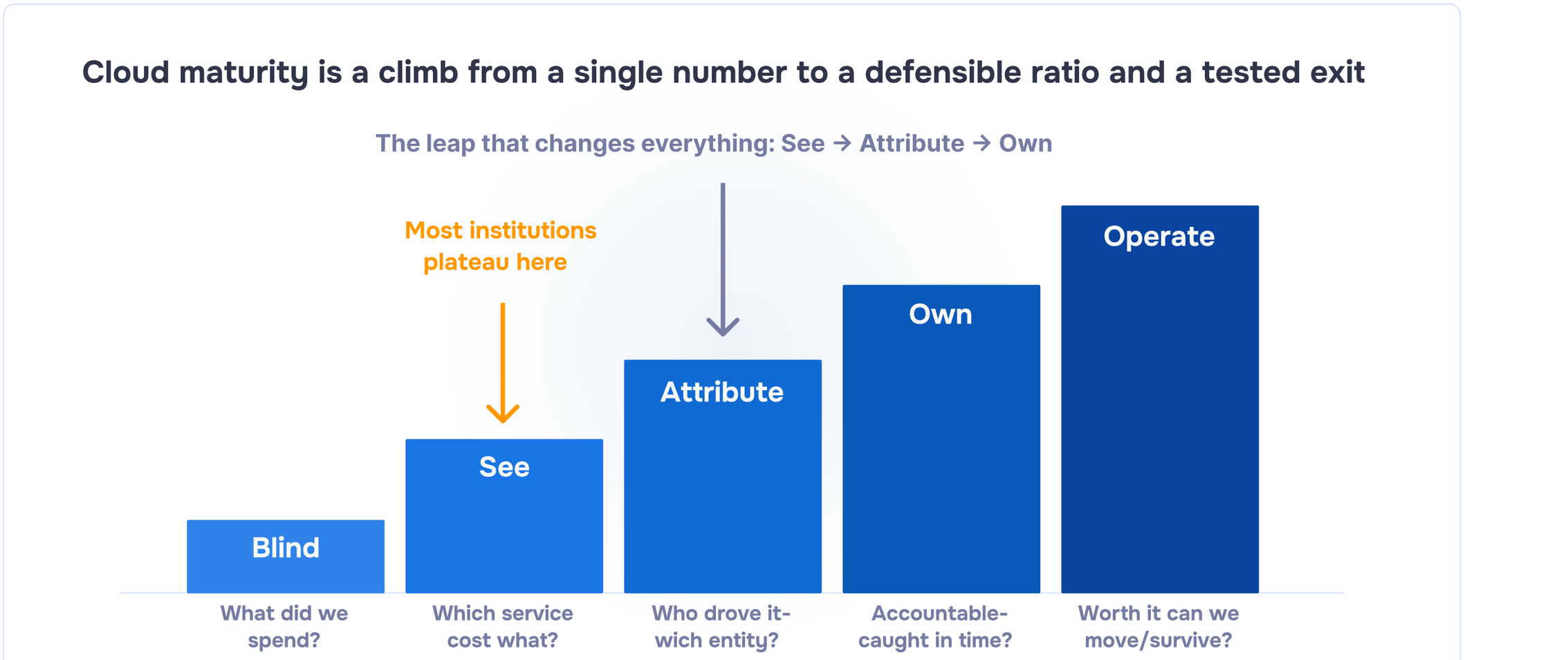
Time-to-exit per critical workload

hours or days, tested, not estimated. This is "can you move it" and "can you survive losing it," made measurable. It answers: could we leave — a question your Lead Overseer can now ask you directly.

Where are you on the ladder?

Cloud maturity in a regulated institution is a climb from a single number to a defensible ratio and a tested exit.

Most teams are stuck on the lower rungs — they bought dashboards and reached "See," but never made cost an owned signal or exit a tested fact. Visibility nobody acts on is just decoration; an exit plan nobody has tested is just a document.



Every reflex above broke in the same place.

You couldn't cut the bill because you couldn't say who spends it. You couldn't de-risk with a second provider because you couldn't actually move anything to it. You couldn't go sovereign because a migration you can't reverse — or survive — is just lock-in with a new flag on it.

Three mandates, one missing foundation. Underneath the finance project, the platform project, and the compliance project sit the same three questions:

STAGE	WHAT IT LOOKS LIKE	WHERE THE COST HIDES
Blind	One big monthly number per provider. Surprises arrive by invoice.	"What did we spend?"
See	Spend broken down by service and account in native consoles, one cloud at a time.	"Which service cost what?"
Attribute	Spend mapped across every cloud to product, team and legal entity — including shared and untagged resources.	"Who and what drove it — and which entity carries it?"
Own	Every cost has a named owner; showback in the engineer's workflow; anomalies caught in real time; the register reconciles to the bill.	"Who's accountable — and did we catch it in time?"
Operate	Cost-to-serve tied to product; time-to-exit tested for critical workloads; placement and resilience are guardrails in the design review, not a cleanup job.	"Was it worth it — and could we move or survive losing it?"

Most institutions plateau at **See**. The leap that changes everything is **See → Attribute → Own** — the point where cost stops being finance's monthly mystery and becomes an engineering signal and a compliance artefact at the same time.

A 12-question diagnostic for your next architecture and risk review

Read these aloud with your platform, finance and risk/compliance leads in the room together, because the answers are interlocked.

Score it honestly: each question is a yes only if someone can answer it now, from a system, without a week of digging. Fewer than 8 confident yeses means the gap is visibility and portability, not budget — and no amount of cost-cutting fixes a visibility gap.

Cost & attribution

01 What share of our bill is unallocated — not tied to any product, team or legal entity?

02 What did last night's deploy cost, and which product owns it?

03 What are our top five cost drivers this month, and who owns each one?

04 What does it cost to serve one customer or one product — and which way is that ratio trending against revenue?

05 What did our AI / ML features cost last month, broken out from general compute?

Portability & exit (DORA)

06 For each critical or important function, how long — tested, not estimated — to stand it up at another provider?

07 Does our register of what runs where reconcile to the actual bill, this month?

08 Which workloads are pinned to a single region or a single provider's proprietary service we couldn't replicate?

09 When did we last test an exit, and did the board see the result?

Resilience & sovereignty

10 Which "multi-region" workloads secretly depend on one provider's global control plane (identity, DNS, routing)?

11 Which workloads carry a hard data-residency or jurisdiction requirement — and which are paying a sovereignty premium they don't need?

12 If our largest provider had a 15-hour regional outage tomorrow, what stops working — and do we know before it happens?

From insight to action: six moves that actually move the number

Skip the table stakes ("tag everything," "buy a dashboard," "right-size your instances"). These are the moves that separate institutions that control cloud from institutions that just watch it — and each one pays off in cost, resilience and compliance at the same time.

Move 1. Attribute to the legal entity, not just the team

Generic FinOps stops at "allocate cost to teams and products." In a regulated group that's half the job: your spend also has to map to the legal entity that carries it, because that's the unit your register, your intragroup recharges, and your supervisor think in.

Do this: for shared and untagged resources — the clusters and platforms that defeat tagging — split the cost by a measured consumption key (CPU-hours, requests, or storage per namespace), not an even spread, and roll it up to entity level. Reconcile the result against the actual invoice to the cent; the gap you can't explain is your unallocated number, and closing it is the first KPI.

You get two artefacts from one exercise — a cost-to-serve ranking finance can act on, and a register that reconciles to the bill.

Start with one unit metric, but make it one tied to a decision you're about to make: a product's pricing, a renewal, a build-vs-buy.

Move 2. Treat the control plane as your real concentration risk, not the logo

The October 2025 AWS outage is the lesson written in neon. A DNS failure in a single region (us-east-1) cascaded for roughly 15 hours and locked customers out of major UK banks — including institutions that had paid for multi-region and multi-availability-zone redundancy.

Why did the redundancy not save them? Because "global" services — identity, DNS, routing, secrets — route through that one region regardless of where your workloads run. Analysts noted it was the fourth major us-east-1 incident in five years.

Do this: stop mapping concentration risk by provider name and start mapping it by hidden dependency. For each critical function, write down the failure that takes it offline even when your "redundant" architecture is healthy — the global control plane it silently calls, the single-region datastore, the proprietary service with no equivalent elsewhere. A filled-in example:

PRIMARY REGION	CRITICAL FUNCTION	HIDDEN SINGLE POINT OF FAILURE	STATED RTO	REAL RTO IF THAT DEPENDENCY FAILS
Customer auth / login	eu-central-1	Global IAM control plane in us-east-1	15 min	Unknown — untested
Payments ledger	eu-west-1	Primary datastore single-region; replica read-only	30 min	~4 hrs (manual promotion)
Statement generation	eu-west-1	Provider-proprietary queue, no equivalent elsewhere	"best effort"	No exit path

The right-hand column is your real resilience posture — and the gap between columns four and five is what you take to the board.

Move 3. Commit to the trough, but size the bet to your exit horizon

Reserved instances and savings plans are the largest single cost lever — commitments run 40–72% cheaper than on-demand. The non-obvious discipline is how much to commit.

Cover only your usage floor — the trough you never drop below — at roughly 70–80%, not 100%, and layer flexible savings plans over reserved instances.

But in a regulated institution there's a second constraint civilians miss: **never commit past your tested exit horizon.** A three-year reservation on a workload you've told your regulator you can exit in twelve months is a contradiction you'll have to explain.

Size every commitment against both your baseload certainty and the exit window in your register:

WORKLOAD	USAGE FLOOR (TROUGH)	COMMITMENT TARGET	INSTRUMENT	TESTED EXIT WINDOW	MAX COMMITMENT TERM
Core ledger compute	~80% of peak	75%	Compute Savings Plan (flexible)	12 months	**1 yr (not 3)**
Batch / reporting	~40% of peak	40%, spot for the rest	RI + spot	6 months	1 yr

Move 4. Treat the control plane as your real concentration risk, not the logo

Make "time-to-exit" a measured number, not a contract clause

DORA (Article 28) requires documented exit strategies for material ICT providers, and for the designated critical providers — AWS, Microsoft Azure and Google Cloud were all named on the first official list in November 2025 — those exits must be tested sufficiently and reviewed periodically.

Most institutions treat this as paperwork. Turn it into a metric instead: measure, per critical workload, how long it actually takes to stand it up at an alternative, and trend that number down.

What "tested" actually means — not a tabletop discussion, but a real rehearsal: deploy the workload to a second provider or region from infrastructure-as-code, restore the data from backup, repoint a slice of traffic, and time it end to end. Start once a year on your single most critical function, then widen coverage and shorten the clock each cycle. The first run will be embarrassing; that's the point — it converts an assumption into a number.

This single number does triple duty — it is your DORA exit evidence, it is what kills vendor lock-in (a workload you can move in days has no lock-in premium), and it is what makes sovereignty a placement decision rather than a migration crisis.

The EU Data Act, in force since September 2025, now obliges providers to support switching and curbs egress charges, so the technical barrier is falling; the discipline barrier is yours to remove.

Move 5. Commit to the trough, but size the bet to your exit horizon

Place workloads by their actual data-residency requirement, not by a blanket sovereignty migration

Don't answer "are we sovereign?" with a single yes/no for the whole estate — that's how you end up paying a 15–30% premium and losing services on workloads that never needed it.

Classify each workload by what the law actually requires of its data, and place it accordingly. A worked placement record:

WORKLOAD	DATA CLASS	RESIDENCY/ JURISDICTION REQUIREMENT	PLACEMENT DECISION	PREMIUM ACCEPTED	TIME-TO-EXIT
KYC / customer PII for EU clients	Regulated, sensitive	EU jurisdiction; minimise CLOUD-Act exposure	EU-owned sovereign or EU-native provider	Yes (~15%)	Target < 30 days
Marketing site / public content	Public	None	Standard EU region, cheapest	No	Days
Internal analytics (anonymised)	Internal	EU residency, no sovereignty	Standard EU region	No	Weeks

Most estates discover that only a minority of workloads truly require the sovereignty premium — and that the right answer for the sensitive minority is often an EU-owned provider rather than a US hyperscaler's sovereign partition, precisely because of the ownership question.

Decide it per workload, with the data classification in front of you.

Move 6. Give each cost a regulated owner, and put the number in the design review before the code exists

"Get engineers to care about cost" has failed for years because caring isn't a mechanism. What works: give each team a cost number they own the way they own latency and uptime — surfaced in the pull request and dashboard they already live in — and make the central FinOps function enablement, not police.

Then add one line to your architecture/design-review template, next to availability and data classification: a new service must declare **its target unit cost, its resilience tier, and its time-to-exit before it ships.**

Forecast the unit cost at 10× volume: if cost-to-serve rises as you scale instead of falling, you've found an architecture problem while it's still cheap to fix — not after it's baked into the cost base and the register.

Governance as prevention, not monthly cleanup.

Seven anti-patterns to avoid

Optimising before you can attribute

you right-size the visible half of the bill and miss the shared-infra half that's actually growing.

Chasing an absolute "half"

the bill rises with usage; the durable win is the ratio (cost-to-serve), not a headline number you'll never hold.

Counting logos as resilience

a second cloud you can't fail over to is cost and complexity with none of the risk reduction the regulator means.

Mistaking multi-region for safety

if your "redundant" workload depends on one provider's global control plane, you have one region's resilience and three regions' bill.

Treating the exit plan as paperwork

an untested exit is not an exit; under DORA, for critical providers, it must be tested annually.

Blanket "go sovereign"

you pay a premium and lose services on workloads that never needed it, and may not close the legal question anyway.

Committing past your exit horizon

a multi-year reservation on a workload you've promised to be able to exit in months is a contradiction you'll have to defend.

Who needs to be in the room

Cloud cost in a bank is not a finance project handed down for cleanup. The spend is created by architecture decisions, it shows up as both operating cost and supervised dependency, and its risks are now a board-level obligation. Solving it is a four-legged stool:

Engineering & Platform

own the architecture choices that create cost and the dependencies that create concentration risk — and need both signals in their workflow to act.

The technical executive (CTO / CIO)

owns the trade-off between velocity, resilience and cost — and owns the ICT risk picture to the management body, which under DORA can no longer be delegated away.

Finance & Product

own the cost-to-serve ratio, intragroup recharges, and pricing that the unit metric directly informs.

Risk & Compliance

own the register, the exit testing, and the concentration-risk assessment — which are the same artefacts the other three legs produce, viewed through a supervisory lens.

What aligns them is a single shared pair of numbers everyone trusts: **cost-to-serve per product, and tested time-to-exit per critical workload, both trended over time.**

When platform, finance, product and risk argue from the same numbers, the conversation shifts from blame to trade-offs — which is where good decisions get made.

Who needs to be in the room

Don't launch a programme. Build one artefact and take it to the table you already have scheduled.

Pick your three most critical functions and fill in two tables for each: the control-plane dependency map and the placement record shown above.

You will almost certainly surface at least one "redundant" workload that isn't, one workload paying a sovereignty premium it doesn't need, and one exit that has never been tested.

That is your agenda — concrete, defensible, and visible to finance and your supervisor in the same breath.

The one number to put on the wall

Stop staring at the total. Put two numbers where the team can see them: **cost-to-serve per product**, **trended against revenue**, and **tested time-to-exit for each critical workload, trended down.**

The first turns "are we overpaying?" from a recurring anxiety into a question you can answer and defend. The second turns "can we leave, switch, or survive losing this?" from a regulatory headache into a number you improve quarter over quarter.

Together they convert cloud cost from three separate fire drills — finance, resilience, sovereignty — into one engineering habit. Which is the only version of cloud control that lasts.

This brief was put together by the team at emma. If a second set of eyes on your own estate would help — a quick read on your top cost drivers, your unallocated spend, your hidden control-plane dependencies, and where your tested time-to-exit really stands across clouds — that's a conversation we're glad to have. No deck required.

Sources

- Gartner, Forecast: Public Cloud Services, Worldwide, 2023–2029, 3Q25 Update (public cloud growth of 21.3% in 2026, market reaching $1.48T by 2029).
- Forrester, US Tech Forecast 2026 (US financial services tech spend ~$495B, 17.1% of US tech spend, growing faster than market).
- Flexera, 2026 State of the Cloud Report (cloud waste ~29%, first uptick in five years, AI-driven; managing spend the top cloud challenge; 15–30% of spend unallocated under tag-dependent approaches).
- FinOps Foundation, State of FinOps 2026 (multi-cloud ~76% of enterprises; unified cross-cloud visibility the top challenge; AI cost management now near-universal; "the bill never really goes down"); industry analyses of optimization levers (commitments 40–72% vs on-demand; right-sizing 15–25%; non-production scheduling 10–20%; mature FinOps reduces waste from ~32–40% to ~15–20%; typical 25–30% reduction).
- EU Digital Operational Resilience Act (Regulation (EU) 2022/2554), in application since 17 January 2025: Article 28 (documented exit strategies; tested at least annually for critical providers), Article 30 (contractual provisions), board/management-body accountability. ESAs (EBA, EIOPA, ESMA), first list of 19 designated Critical ICT Third-Party Providers, 18 November 2025 (including AWS, Microsoft Azure, Google Cloud); EBA as Lead Overseer for banking. ECB Guide on outsourcing cloud services, finalised 15 July 2025.
- AWS us-east-1 outage, 20 October 2025 (DNS / DynamoDB control-plane failure, ~15 hours; UK banking and trading services disrupted despite multi-region architectures; Analysts noting the fourth major outage in that region in five years). CrowdStrike outage, July 2024.
- AWS European Sovereign Cloud (live 15 January 2026, Brandenburg, Germany; ~€7.8B investment; independent EU-resident operations); independent pricing analysis (~15% premium; ~90 services vs 240+; two AZs at launch); CLOUD Act exposure debate (US-owned operating entity; July 2025 French Senate testimony). EU Data Act (in application since September 2025; switching support and limits on egress charges). Gartner, sovereign cloud spending forecasts (~$80B in 2026; European spending +83% year over year). Alternative EU-owned and partner providers: S3NS (Thales/Google), Bleu (Orange/Capgemini), STACKIT, OVHcloud, Scaleway.
- emma proprietary analysis of 400+ recorded cloud surveys (theme frequency: pricing in 73%, security/compliance in 83%; recurring customer language).