

The Cloud, Demystified

Reading time: ~10 min · Prereqs: Lessons 01–02 · *Snapshot 2026 — cloud products and prices evolve quarterly*; see [refresh-checklist.md](#).

Why this matters to you

"The cloud" is no longer a buzzword or a strategic question — it's the default. Whether you join a three-person startup or a Fortune 500 bank, your team is almost certainly running on AWS, Microsoft Azure, or Google Cloud, and quite likely on two of them at once. Cloud cost is now a top-three line item at many companies, which means conversations about *where things run* and *what they cost* spill out of the engineering room into product, finance, and the boardroom. Knowing the three layers of the cloud cake and the basics of how the bill is computed lets you sit at that table and follow along — even if you never personally launch a server.

1. What "the cloud" actually is

Strip away the marketing and "the cloud" is one idea: **renting computing infrastructure from someone else's data center, billed by usage**. Instead of buying servers, racking them, cooling them, and replacing them every four years, you swipe a credit card and get a virtual machine in two minutes. Instead of capacity-planning for the peak, you pay only for what you actually use.

This was a fringe idea in 2006, when Amazon launched S3 (cheap object storage) and EC2 (rentable virtual machines) and almost nobody believed a real company would put production workloads on someone else's hardware. By the mid-2010s the migration was in full swing. By 2026 the defaults have flipped: new companies start in the cloud; on-premise is the deliberate, defended choice. The shift is from **capex** (big up-front purchases of hardware) to **opex** (monthly operating expense), plus instant scale, plus no more "rack and stack" weekends.

2. The three layers: IaaS, PaaS, SaaS

Cloud offerings stack into three layers. The further up you go, the more the provider does for you — and the less control (and flexibility) you have.

- **IaaS — Infrastructure as a Service.** Raw virtual machines, raw block storage, raw networking. You manage the operating system, the runtime, the application, the patches. Examples: AWS `EC2`, Google `Compute Engine`, Azure `Virtual Machines`. Most "lift and shift" migrations land here — it's the layer that looks most like the old data centre.
- **PaaS — Platform as a Service.** The provider runs the database, the queue, the runtime, the container scheduler; you bring code and config. Examples: AWS `RDS` (managed Postgres/MySQL), `Lambda` (functions), `Fargate` (managed containers); GCP `Cloud Run`, `BigQuery`; Azure `Functions`, `Azure SQL`. Less ops work, more vendor lock-in.
- **SaaS — Software as a Service.** The provider runs the whole application; you use it via a browser or an API. Examples: Gmail, Notion, Figma, Slack, Stripe, Salesforce. Yes — Stripe and Slack are "cloud services" too. From the user's seat there's no server to think about at all.

Different teams live at different layers. Platform engineers spend their days in IaaS and PaaS. Frontend engineers might only ever touch PaaS and SaaS. Finance just sees the consolidated invoice.

3. The big three (and the rest)

Three providers dominate global market share. The numbers below are the 2025 figures; treat them as a snapshot, not a constant.

- **AWS — Amazon Web Services.** The biggest, the broadest catalog (200+ services), often the most expensive in any given category, and the one most teams hire for. Roughly **31% market share**. Started the modern cloud era.
- **Microsoft Azure.** Second-largest, deeply integrated with Microsoft 365, Active Directory, and the Windows server stack. Default choice in enterprises that are already Microsoft shops. Roughly **25%**.
- **Google Cloud (GCP).** Third by share but strong in data, analytics, AI/ML (Vertex AI), and Kubernetes ergonomics — GCP gave Kubernetes to the world. Roughly **12%**.

- **Others. Cloudflare** (edge-first compute and storage), **Oracle Cloud** (Oracle workloads and aggressive pricing), **IBM Cloud**, **Alibaba** (dominant in China). Niche but real.

Most large companies use **two** clouds — either deliberately, for redundancy and bargaining leverage, or accidentally, because an acquired company brought its own. Most startups use **one**.

4. The pricing surprise

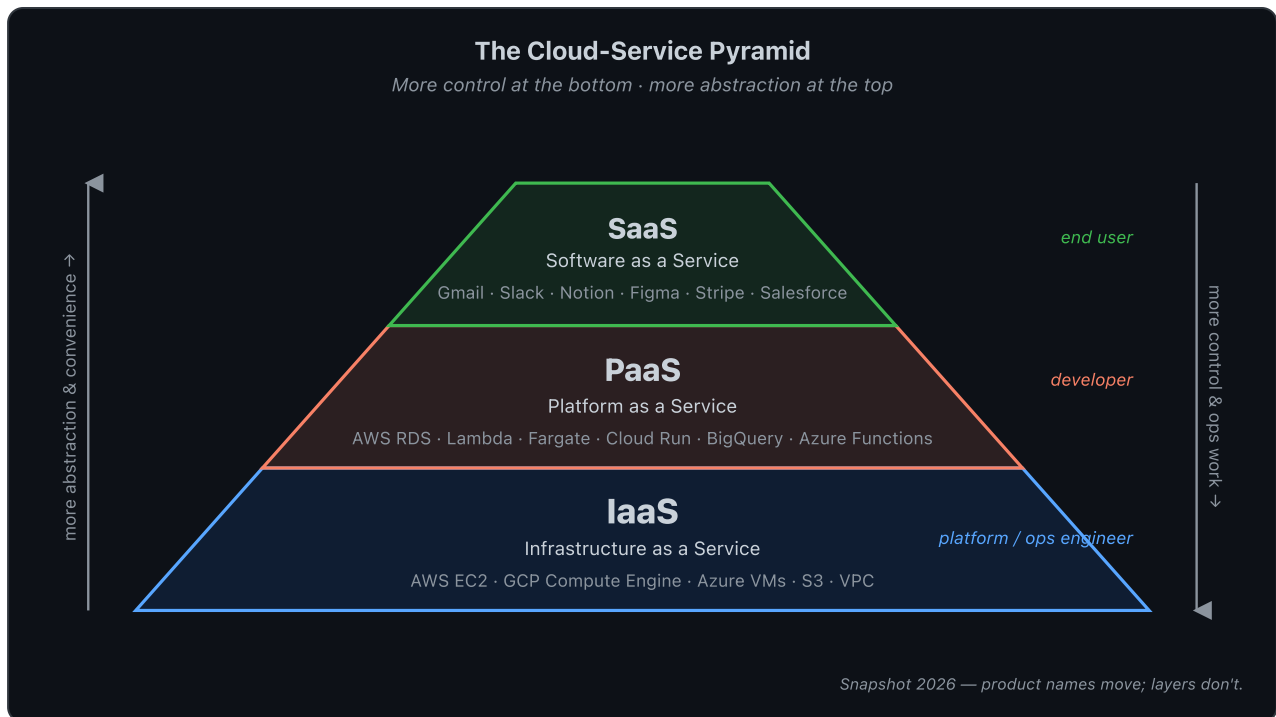
Cloud bills are measured in tiny units — seconds of compute (or milliseconds, for serverless functions), GB-hours of stored data, GB transferred *out*, per-API-call. The surface looks simple. Reality contains four classic surprises:

- **Egress.** Moving data *out* of the cloud (to your users, or to another cloud) is expensive. Moving data *in* is free. This single asymmetry shapes architecture decisions across the industry.
- **Idle resources.** A forgotten EC2 instance keeps billing 24/7 until somebody kills it. A single mid-sized VM left running by mistake can quietly burn \$200–\$500 in a month.
- **Cross-region traffic.** Even inside one provider, moving data between regions (say, US East to EU Frankfurt) costs money. Architectures that span regions for redundancy pay for that redundancy.
- **Premium support.** Production accounts typically pay a multi-thousand-dollar-per-month support tier on top of the resource bill, just to get a human on the phone when things break.

This is why **FinOps** (Cloud Financial Operations) is now a recognised profession, complete with certifications and a foundation. Lesson 10 returns to deployment and operational economics.

5. The picture: the cloud-service pyramid

The three layers fit together like a pyramid — broad and infrastructural at the bottom, narrow and user-facing at the top. Example products on each layer:



6. "Cloud-native" vs "lifted into the cloud"

Two very different patterns hide under the word "cloud":

- **Cloud-native.** The application was designed around managed services from day one. It assumes auto-scaling fleets, ephemeral storage, managed message queues, identity-as-a-service, and serverless functions. Restarts and node failures are normal, not emergencies. New companies almost always build this way.
- **Lift-and-shift.** An existing on-premise application was packaged up and moved into VMs inside the cloud, where it behaves exactly the way it did on the company's old hardware — same architecture, same scaling assumptions, same fragility. Most large enterprises are still mid-migration, with a mix of both patterns living side by side.

Lift-and-shift pays the cloud bill but rarely gets the cloud's elasticity benefits. Cloud-native gets both — at the cost of a redesign and an irreversibly deeper dependence on the provider's catalog.

What people argue about

Three live debates in the cloud world right now:

- **Single-cloud vs multi-cloud.** Multi-cloud is fashionable for redundancy and negotiating leverage with providers. Critics counter that the operational tax — two sets of identity systems, two consoles, two networking models, twice the training — outweighs the benefits for most teams. There is no industry consensus.
- **Cloud repatriation.** A small but high-profile cohort of companies (37signals being the loudest) are moving workloads *back* from the cloud onto self-owned hardware, arguing they now have predictable steady-state load and the cloud bill is no longer worth it. Most companies have not followed. The jury is still out on whether this is a trickle or a trend.
- **Hyperscaler lock-in.** The deeper you go into a provider's managed services (Lambda, DynamoDB, BigQuery, Cosmos DB), the harder it is to leave. Some teams deliberately stay close to "boring" portable primitives (just VMs, just Postgres, just S3-compatible storage). Others embrace the managed services for the velocity gains and accept the lock-in. The right answer depends on the company — and the bill.

Reading on

- **AWS Free Tier** — aws.amazon.com/free. Sign up, poke around the Console, click through the launch wizards without launching anything. No dollars required for the first 12 months of small resources, plus an always-free set.
- **The Cloud Resume Challenge** — cloudresumechallenge.dev. A popular hands-on entry path into cloud careers: build and host a simple resume site using real cloud primitives. Free, community-maintained.
- **FinOps Foundation** — finops.org. Community, certifications, and reading material if cloud cost is your angle into the field. Especially useful for finance and product folk who own a cloud P&L line.

LAST-REVIEWED: 2026-05-22