CLOUD COVER

Price Swings, Sovereignty Demands, and Wasted Resources

July 2025

By Filippo Scognamiglio, Sukand Ramachandran, Michael Engelhardt, Prasanna Santhanam, Abhinav Gupta, Pankaj Sherawat, JR Quinones, Andrew Dunbar, Basem Abdelrahman, and Daniel Øllgaard

The **Nimbus Pricing Index** (NPI) for the first half of 2025 reveals the latest average list price for a basic enterprise-cloud package from the three main providers. **(See Appendix 1.)**

Amazon Web Services (AWS) kept its pricing largely stable globally, with occasional strategic price reductions in highly competitive markets such as São Paulo. Microsoft reduced Azure's prices aggressively in competitive Asia-Pacific countries—specifically, India, Australia, Malaysia, and Thailand. And Google consistently but moderately increased its rates worldwide for Google Cloud to reposition it in premium regions or to offset operational costs. Even so, Google Cloud remains competitive in major hubs, including Tokyo, Singapore, and London. **(See the sidebar "About the Series.")**

Across regions, there are notable pricing differences to use AI-specific graphics processing unit (GPU) instances, with AWS and Azure offering a clear cost benefit in the Eastern US over Google Cloud. **(See Exhibit 1.)** AWS and Azure also frequently have very close pricing in the Western US and in European hubs. Regionally, the cheapest AI-specific GPU instances are still in North America and the Nordic countries, while most of those in Europe and Asia-Pacific range from $5,000 to $6,500. **(See Appendix 2.)** This range generally applies to Africa and the Middle East as well, and prices peak above $7,000 in South America. AWS and Azure shadow each other almost everywhere, diverging mainly in South America and the Middle East, whereas Google Cloud is slightly cheaper in Asia-Pacific and Europe where it has T4 capacity.

About the Series

BCG is pleased to present the latest installment of our Cloud Cover series. The goal of this series is to share the latest data, insights, and news on the evolving cloud industry, with a particular focus on three major cloud service providers: Amazon Web Services, Microsoft Azure, and Google Cloud. In addition to reviewing price movements in the cloud industry, this update focuses on sovereignty demands and rising cloud waste in 2025.

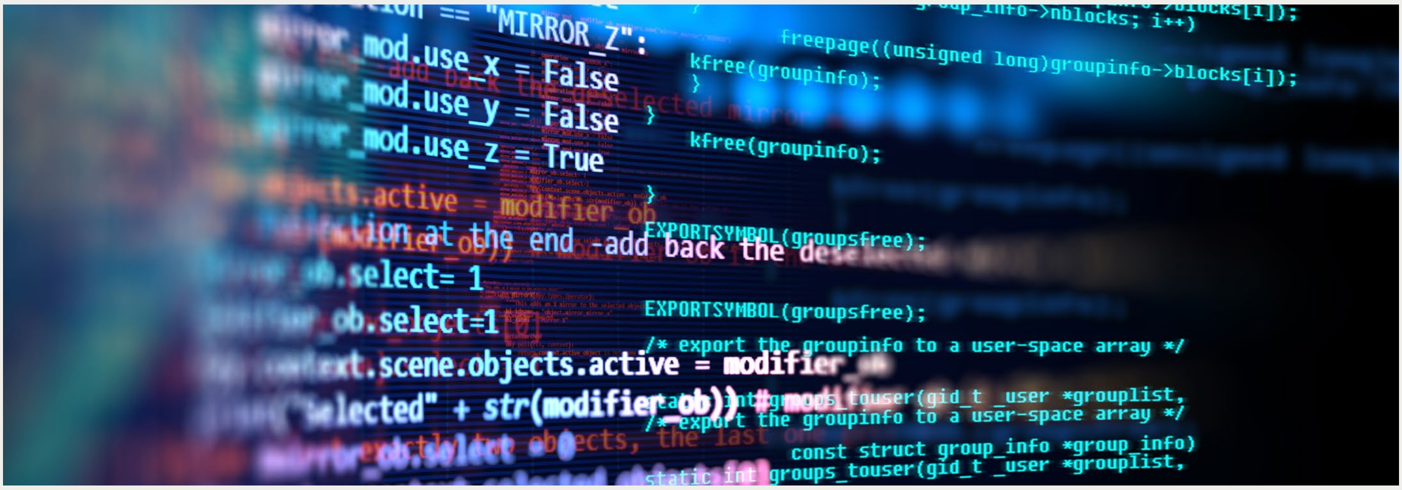
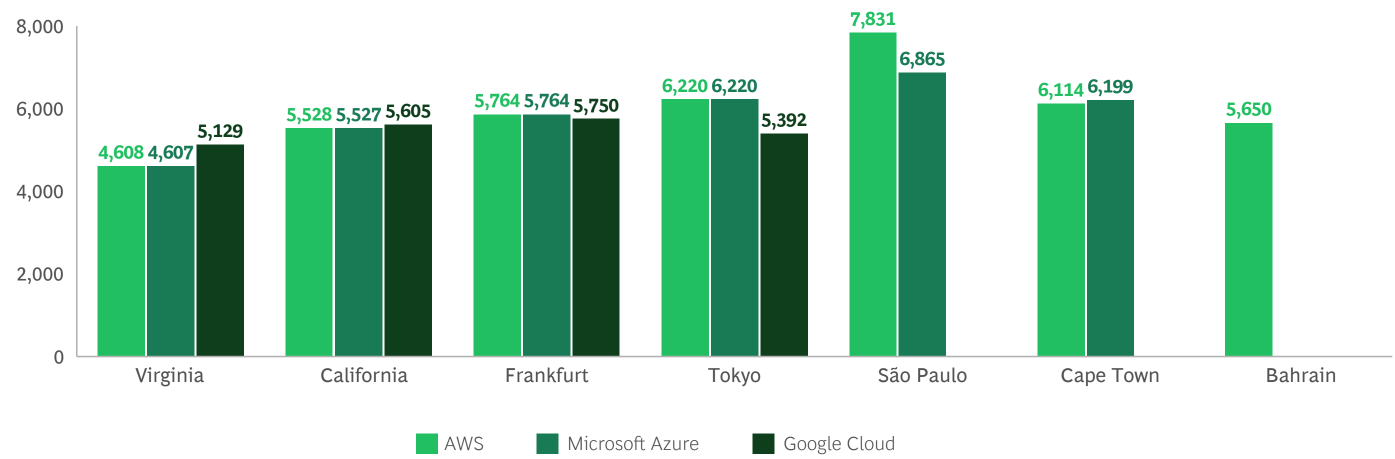


EXHIBIT 1

The Pricing of AI-Specific GPU Instances Varies Across Leading Providers in Key Global Markets

Annual price of AI-specific GPUs ($)



Source: BCG's NPI for AI-specific GPU instances, H1 2025.
Note: GPU = graphics processing unit; H1 = first half. AWS and Microsoft Azure are the only providers in São Paulo and Cape Town, and AWS is the only **provider** in Bahrain.

The Sovereign Cloud Redefines National Security

A sovereign cloud is a nationally controlled zone of public-cloud technology. It delivers the same scalability and on-demand services that are expected from AWS, Microsoft, or Google. The sovereign cloud has emerged as a strategic imperative for governments and local enterprises to ensure control over national data, retain operational independence, and maintain economic benefits.

- **Regulation and Privacy.** The European Union's General Data Protection Regulation, France's SecNumCloud rules, and India's Digital Personal Data Protection Act all insist that certain data remain locally governed. With a sovereign cloud, enterprises can comply with local regulations while continuing to access cloud-native capabilities securely.

- **Geopolitics.** The Clarifying Lawful Overseas Use of Data Act in the US allows US authorities to subpoena data from any US-based provider even if that data sits in Europe or Asia. A country can use a sovereign cloud to build a jurisdictional firewall.

- **Cyber Resilience.** Conflicts such as the Russia-Ukraine war demonstrate how easily cross-border cloud access can be disrupted or weaponized. A sovereign cloud guards against nefarious foreign influence or attacks.

- **Economic Strategy.** By implementing a sovereign cloud, a country can nurture a domestic cloud ecosystem and keep jobs, tax revenue, and intellectual property value at home.

Analysts expect sovereign-cloud infrastructure as a service (IaaS) spending to leap from $37 billion in 2023 to $169 billion by 2028—a compound annual growth rate of 36%, versus about 24% for general IaaS spending.

As a result of these powerful drivers, by 2028, 65% of nations are expected to implement a digital sovereignty plan with three pillars:

- **Data Sovereignty.** All customer data—including backups and metadata—stays inside national borders, encrypted with keys that the customer (or a designated national body) controls.

- **Operational Sovereignty.** The day-to-day administration is carried out only by people who are legally subject to the local jurisdiction. Foreign authorities cannot compel those staff members to hand over data.

- **Technological Sovereignty.** The cloud stack can be inspected, escrowed, or replicated, ensuring that the nation isn't reliant on a single foreign vendor during a crisis.

While the reasons to implement a sovereign cloud are similar across countries, as are the stakeholder groups within the sovereign ecosystems, each country decides the level of sovereignty that it wants to achieve.

Vendor Models That Address Sovereign Cloud Demand

Vendors offer two primary sovereign-cloud models that address governments' digital sovereignty needs:

- **Hyperscaler Cloud with Sovereignty Features.** Vendors make this model available to the general public, but customers can add security features for an extra charge. Customers also manage data encryption through their network.

- **Sovereign Cloud with Hyperscaler Software.** Vendors provide the physical infrastructure and configure the technical setup, and a local partner (for example, OVH US, Thales, or T-Systems) conducts the physical operations.

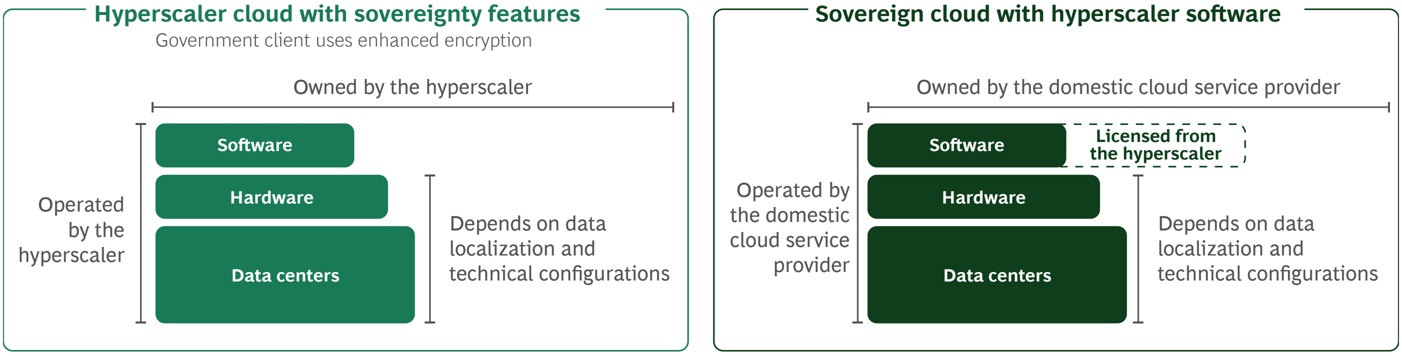
It's important to weigh the benefits and risks in each model. **(See Exhibit 2.)**

In our work with clients, we have identified five key factors when choosing between the two models. **(See Exhibit 3.)**

Ultimately, each model varies in sovereignty level, cost, and scalability in ways that appeal to different countries. There are multiple use cases globally for both models. **(See Exhibit 4 and Exhibit 5.)**

EXHIBIT 2

Governments Should Weigh the Key Benefits and Risks of both Sovereign Cloud Models



Hyperscaler cloud with sovereignty features		Sovereign cloud with hyperscaler software	
Benefit	Leveraging the investments and capabilities of hyperscalers for a secure cloud	Benefit	Separating ownership from foreign laws with a limited stake from the hyperscaler
Risk	Ensuring the security of the cloud environment through emerging encryption technology	Benefit	Accelerating knowledge transfer through a joint venture model
Risk	Retaining the technical expertise in government to oversee emerging encryption technology	Risk	Ensuring that the service quality of the cloud is equal to a global cloud offering

Sources: Expert interviews; BCG analysis.

EXHIBIT 3

Key Factors to Consider When Choosing a Model

Key factor	Hyperscaler cloud with sovereignty features	Sovereign cloud with hyperscaler software
Data sovereignty safeguard	A country's data may still fall under a foreign jurisdiction if the cloud infrastructure is owned by the hyperscaler	All national data is governed by local law (local ownership)
Technology control	A nation has little or no control over the underlying stack and is dependent on a foreign vendor	A country has full control of the infrastructure and services
Operational resilience	Service availability is exposed to the hyperscaler's home country's policy and global outages	The cloud is architected to operate even if it's disconnected from the global cloud backbone
Capex responsibility	Capex is shouldered by the hyperscaler, but it still needs tax breaks or incentives	Capex is borne by local or joint venture entities, but the implementation requires larger public–private funding
Time to market	The cloud can be operational in two to three years, which is faster thanks to the hyperscaler's expertise	The cloud may be operational in two to three years, but it also may take longer while local talent and governance mature

Source: BCG analysis.

EXHIBIT 4

The Hyperscaler Cloud with Sovereignty Features Has Many Use Cases Globally

Netherlands	Sweden	UAE
The Netherlands's National Cyber Security Center is piloting Microsoft Cloud for Sovereignty with dedicated landing zones and policy initiatives to align with the country's BIO regulation for sovereign cloud usage.[1]	Atea supports Sweden's government agencies in enhancing citizens' digital experiences and facilitating advanced predictive health care by using Microsoft Cloud for Sovereignty to analyze highly sensitive data for sovereign cloud usage.	G42 empowers the UAE's public sector and regulated industries to use new platform capabilities for data security by accessing the latest Microsoft Azure cloud and AI features while ensuring compliance with local privacy and regulatory needs.
Italy	**Singapore**	**Germany**
Leonardo aims to provide Italian public administrations, including ministries and local governments, with secure and compliant cloud services using Microsoft Cloud for Sovereignty.	The Smart Nation and Digital Government Group partnered with AWS to create dedicated local zones, enhancing data security for sensitive workloads in Singapore's cloud environment.	AWS will simplify its service adoption for the German and European public sectors and data-sensitive organizations through the AWS European Sovereign Cloud.

Source: BCG analysis.
Note: UAE = United Arab Emirates. Examples of use cases.
[1]The Dutch government's Information Security Baseline (BIO) is a set of guidelines and measures designed to ensure information security across all levels of the Dutch government. It acts as a unified framework, replacing previous individual baselines used by different government bodies, promoting consistency and efficiency.

EXHIBIT 5

The Sovereign Cloud with Hyperscaler Software Also Has Many Use Cases Globally

France	Germany	China
Capgemini and Orange have partnered to create Bleu. Bleu provides hyperscaler trusted cloud services from locally owned and operated data centers in France. This platform ensures compliance with French government regulations and critical infrastructure data sovereignty requirements.	SAP, in collaboration with Arvato and Microsoft, has established the subsidiary Delos Cloud to deliver a sovereign cloud platform for Germany's public sector. The infrastructure is locally owned, and operations are managed by Arvato, ensuring all data is processed and stored in Germany, adhering to local data sovereignty laws.	The company 21Vianet is partnering with a hyperscaler to offer a physically separated instance of cloud services, ensuring compliance with Chinese regulations. The infrastructure is locally owned and operated, supporting public- and private-sector needs from data centers in China.

Sources: Company websites; BCG analysis.
Note: Examples of use cases.

Specific Vendor Offerings

Several major cloud providers have developed sovereign cloud solutions that are tailored to the different regulatory and sovereignty needs worldwide. At a minimum, these solutions address data sovereignty, operational sovereignty, and technological sovereignty. **(See Exhibit 6.)**

The Economics of a Sovereign Cloud

Dedicated sovereign-cloud offerings focused on specific regions require a screened staff, have fully isolated infrastructures, and are compliance heavy. The result is a price premium over public clouds. For example, Google Sovereign Cloud is priced 10% to 20% over the public cloud, while Oracle EU Sovereign Cloud charges a 15% to 30% price premium. Meanwhile, ultraisolated government-cloud offerings are air-gapped with strict residency and security controls. For example, AWS GovCloud is offered at a 20% to 30% price premium over the public cloud, and Microsoft Azure Government carries a 15% to 25% price premium.

EXHIBIT 6

Cloud Solutions by Vendor

Vendor	Solution	Key features	Availability
AWS	AWS European Sovereign Cloud	Separate and independent EU cloud, encrypted operations, and EU personnel only	EU (Germany first)
Microsoft	Microsoft Cloud for Sovereignty	Regional compliance controls, AI-driven security, and localized cloud zones	EU and UAE
Google Cloud	Google Distributed Cloud	Fully air-gapped, no public cloud dependency, and local governance	EU
Oracle	Oracle EU Sovereign Cloud	Physically separated, EU-based key management, and GDPR compliant	EU
SAP and Arvato	SAP's sovereign cloud	Joint venture model, full BSI compliance, and data hosted locally	Germany
Alibaba Cloud	China's sovereign cloud	Fully localized cloud infrastructure and strict government compliance	China
G42 and Microsoft	UAE's sovereign cloud	Partnership model, AI and security features, and built on Microsoft Azure	UAE

Source: BCG analysis.
Note: EU = European Union; GDPR = General Data Protection Regulation; UAE = United Arab Emirates. BSI is Germany's Federal Office for Information Security.

While the list prices for sovereign cloud regions and ultraisolated government-cloud zones are higher than standard public-cloud regions, cloud providers increasingly offer discounts to anchor tenants—especially early, credential-building customers. Deep incentives (such as subsidized bandwidth and free credits) can narrow the real-world premium well below the list price range.

The Outlook for the Sovereign Cloud

We have identified three key trends that will shape sovereign-cloud adoption from 2025 through 2028.

Sovereign AI Stacks. When a country insists that all data (training sets, model checkpoints, and telemetry) stays on its soil, it also needs local AI horsepower. That requirement is driving the implementation of a new layer of sovereign AI infrastructure. Hyperscalers are already partnering with national operators to stand up in-country GPU "gigafactories," while new rules such as the EU Artificial Intelligence Act require local custody of model weights and audit logs. As chip subsidies and anchor-tenant discounts kick in, companies will be able to fine-tune large language models inside protected regions, making sovereign AI a practical, policy-driven extension of today's sovereign data clouds. **(See Exhibit 7.)**

Hybrid Sovereign Landing Zones. Few organizations can afford to keep every workload in a premium sovereign-cloud region. The emerging pattern is a hybrid landing zone that stitches together two resource pools:

- **Global public-cloud pods,** which are cheap, elastic, and ideal for developing, testing, and anonymized analytics
- **Sovereign partitions,** which are dedicated subtenants or completely separate regions that are holding regulated data

For example, Microsoft's open-source Sovereign Landing Zone adds extra policy packs and management groups on top of the standard Azure Landing Zone so workloads can step up from public-cloud pods to sovereign partitions with the same DevOps tooling:

- **Data Classification Tags.** These let administrators label every resource public, restricted, or secret, enabling automated routing to the right side of the hybrid fence.

EXHIBIT 7

The Building Blocks of Sovereign AI Stacks

Building block	Why it matters	Examples
In-country GPU superclusters (10 to 100 petaflops)	Keeps raw training data and fine-tuned models within the jurisdiction	Oracle and Nvidia's sovereign-AI-anywhere service lets governments deploy H100 clusters inside national data centers G42 and iGenius are planning Europe's largest sovereign AI cluster in Puglia, Italy, and powering it by Nvidia's Blackwell GPUs
Confidential-computing enclaves	Prevents cloud administrators—even local ones—from seeing model weights	Microsoft Azure and Google Cloud now offer EU-hosted servers that keep data encrypted even when in use so that no individual cloud staff member can see it
National AI model registries	Certifies all large language models used in public services	Bleu will host certified GPT-France models for health care compliance
Air-gapped, machine learning operations pipelines	Runs full CI/CD inside the sovereign perimeter	Oracle Alloy's sovereign AI models and compute services are physically hosted and controlled in Japan

Source: BCG analysis.
Note: GPU = graphics processing unit; EU = European Union; CI/CD = continuous integration and continuous delivery; GPT = generative pretrained transformer.

- **Encrypted Cloud-to-Cloud Links.** By using TLS 1.3 (the latest version of the Transport Layer Security protocol) with a company's own keys—and sending the traffic straight across without detouring through any cloud provider's relay servers—the customer can move nonpersonal data cheaply while keeping the sensitive bits locked down.
- **Shared ID and Policy Plane.** One tenant for Azure role-based access control and Azure active directory can span both landing zones, avoiding the two-cloud, two-teams overhead.

Subsidies and Tax Breaks. Sovereign cloud regions still carry a 15% to 30% price premium over standard public-cloud regions, but governments are offsetting much of that extra cost with incentives to entice their private sector to move workloads to the sovereign cloud. **(See Exhibit 8.)**

The next wave of sovereign clouds will blend local-first AI, cost-smart hybrids, and government incentives. This will make compliant, in-country clouds the new normal rather than a premium exception.

Rising Cloud Waste and How to Optimize Spending

Cloud computing has revolutionized the way enterprises operate, offering unparalleled scalability, flexibility, and innovation. However, with the rise in cloud adoption comes a significant and growing challenge: cloud waste.

Studies indicate that up to 30% of cloud spending is wasted due to inefficient usage and a lack of cost control. Given that cloud costs now account for as much as 17% of IT budgets and about 80% of companies expect cloud spending to grow further, addressing this inefficiency is an imperative. **(See Exhibit 9.)**

We have identified five key reasons for cloud waste:

- **Decentralized Cloud Procurement and Governance.** The self-service nature of cloud platforms empowers teams to procure resources independently. While this accelerates innovation, it also can result in uncontrolled provisioning of virtual machines, a lack of visibility into actual usage, and ineffective cost accountability.
- **Overprovisioning and Poor Resource Utilization.** Many organizations provision cloud resources on the basis of peak demand rather than actual usage. Instances and storage are often oversized, and unused resources continue running, leading to unnecessary costs.
- **Ineffective Pricing and Discount Strategies.** Organizations often fail to leverage available pricing models effectively. Many still operate with on-demand pricing, neglecting cost-saving options such as reserved instances, savings plans, and spot instances. Poor contract negotiations with cloud providers can further compound the issue.

EXHIBIT 8

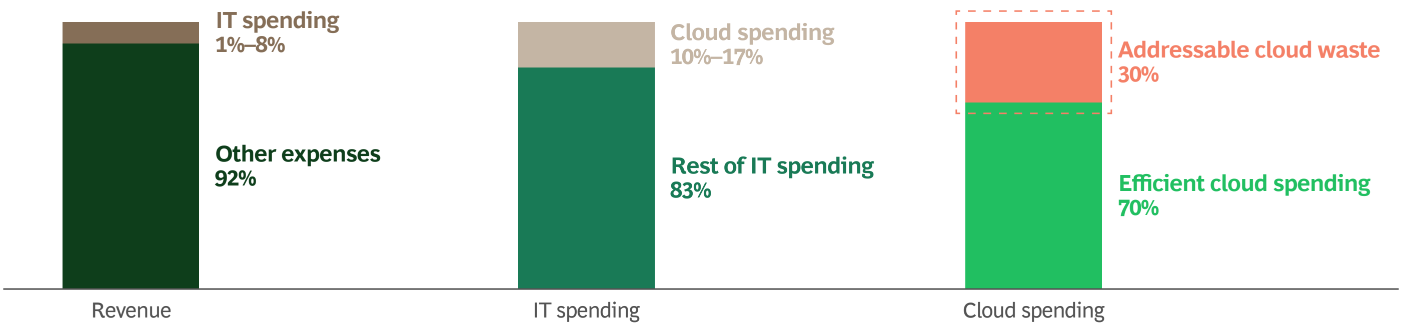
Sovereign Cloud Government Incentives

Jurisdiction	Incentive package	Impact
EU (IPCEI CIS)	Up to €1.2 billion in state aid was approved in December 2023; that triggered €1.4 billion in private capex for next-generation cloud and edge computing	Lowers capex for operators of EU sovereign-cloud regions; first grants were expected in 2024 to 2025
Netherlands	€71 million national subsidy window for firms joining the IPCEI cloud project	Encourages private investment after early sovereign-cloud pilot
Middle East free zones (UAE and Saudi Arabia)	0% corporate tax for 50 years plus power tariff rebates for data center builds	Helped to incentivize Microsoft and Oracle to launch sovereign clouds for GCC countries
Regional tax credits (Canada and Australia)	Accelerated depreciation and green energy credits for compliant data centers	Operators pass savings on to customers, trimming the 30% premium to about 20%

Sources: Government websites; BCG analysis.
Note: EU = European Union; IPCEI = important projects of common European interest; CIS = cloud infrastructure and services; UAE = United Arab Emirates; GCC = Gulf Cooperation Council.

EXHIBIT 9

About 30% of Enterprise Cloud Spending Is Addressable Waste



Sources: *Flexera 2024 State of the Cloud Report*; Gartner.

- **A Lack of FinOps Practices.** Many enterprises lack a structured framework for cloud financial operations (FinOps), leading to poor alignment between technical teams and financial goals. Without continuous monitoring and optimization, cloud costs can spiral out of control.

- **Data Storage Inefficiencies.** Cloud storage costs often escalate due to excessive data retention, a lack of tiered storage strategies, and redundant data copies. Organizations often fail to implement archival and life cycle management policies, resulting in bloated storage expenses.

Fortunately, there are several powerful levers to mitigate cloud waste. **(See Exhibit 10.)**

DEMAND OPTIMIZATION:

- **Rightsizing and Auto Scaling.** Ensure that compute instances match actual workload demands.

- **Auto Shutdown Policies.** Shut down during idle time outside of business hours.

- **Serverless and Containerization.** Transition workloads to serverless or containerized architectures for cost efficiency.

EXHIBIT 10

The Nature of Cloud Waste Varies by Company, and There Are Five Levers to Reduce This Waste

Demand optimization

Provide fit-for-purpose cloud capabilities to prevent waste and reduce costly resource usage

Example: Size resources to match baseload demand rather than provisioning for peaks; leverage analytics to identify underutilized or rarely utilized capacity bands

Architecture optimization

Leverage well-architected cloud technologies to avoid driving up costs and generate a better ROI

Example: Challenge policies and business drivers; rearchitect for cloud-native tech-like containers or serverless architectures instead of lift and shift

Technical optimization

Focus on enhancing performance, scalability, and cost efficiency in a cloud environment

Example: Optimize workloads by automating with IaC and CI/CD, leveraging managed and spot resources, tuning storage, and enhancing observability

Commercial optimization

Adopt appropriate pricing models and contract terms to substantially reduce cloud costs

Example: Analyze the characteristics of all resource pricing models in use and select the most cost-effective options (such as, spot instances, reserved instances, or saving plans)

Operating model and governance

Create accountability on cloud spending and provide a sustainable way to operate, ensuring that resources (such as time, budget, and cloud spending) align with business goals

Example: Define FinOps roles and responsibilities between functions sustained by right processes and tools; implement show-back and charge-back models to allocate to product teams, fostering accountability

Source: BCG analysis.
Note: ROI = return on investment; IaC = infrastructure as code; CI/CD = continuous integration and continuous delivery.

ARCHITECTURE OPTIMIZATION:

- **Refactoring for Cloud-Native Efficiency.** Adopt microservices and serverless architectures to optimize cost-performance balance.
- **Efficient Data Management.** Implement data compression, tiered storage, and life cycle policies to minimize storage costs.

TECHNICAL OPTIMIZATION:

- **Improved Utilization.** Automate provisioning and scale intelligently to lift utilization. Codify infrastructure and deployment using infrastructure as code and continuous integration and continuous delivery, leveraging spot and ephemeral capacity for bursts.
- **Continuous Optimization.** Tune compute and storage classes, adopt a managed platform as a service, and embed observability to drive continuous optimization.

COMMERCIAL OPTIMIZATION:

- **Reserved Instances and Savings Plans.** Move workloads from on-demand to reserved capacity, reducing costs by up to 65%.
- **Multiyear Commitment Discounts.** Engage in long-term negotiations with cloud providers for additional discounts.
- **Competitive Benchmarking and Vendor Management.** Continuously compare pricing across providers to secure a better deal.

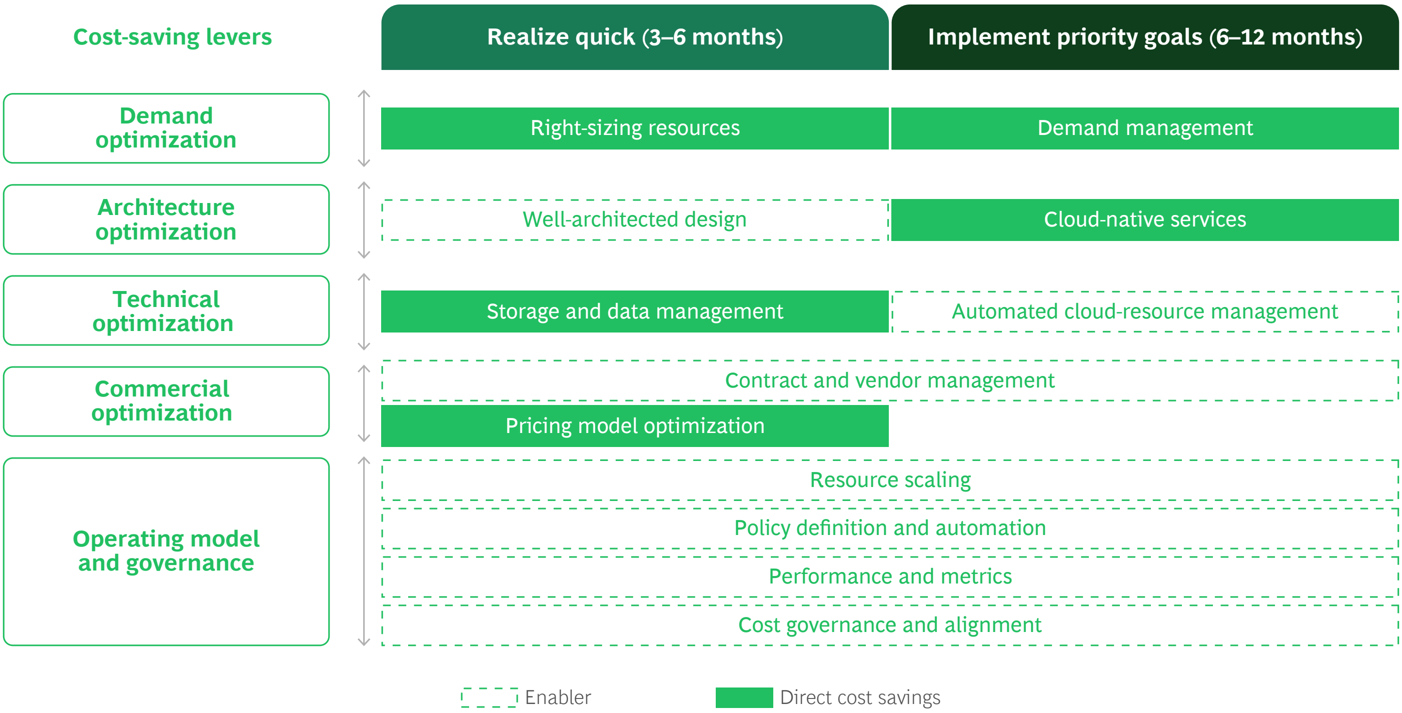
OPERATING MODEL AND GOVERNANCE:

- **Charge-Back and Show-Back Models.** Foster cost accountability by tracking expenses at the team or project level.
- **Automated Cost Monitoring and Alerts.** Use FinOps dashboards and anomaly-detection tools to enforce financial discipline.
- **Regular Audits and Cost Reviews.** Conduct periodic cost audits to identify and eliminate inefficiencies.

To put these guiding principles into practice and target the 30% of addressable cloud waste, companies should identify and prioritize quick wins, which can reduce addressable waste by 6% to 14% by swiftly diminishing cloud excess and securing a rapid return on investment through immediate cost reductions. **(See Exhibit 11.)**

EXHIBIT 11

From Quick Wins to Priority Goals, Cost-Saving Levers Can Unlock Value Within 12 Months



Source: BCG analysis.

Companies should also invest in priority goals where targeted efforts can drive substantial, quantifiable returns and significantly reshape cost structures, reducing addressable waste by 8% to 20%. Meanwhile, companies should probably minimize efforts focused on long-tail events, which can only reduce addressable waste by 3% to 6%.

Cloud computing remains essential for business innovation, but without strong cost controls, inefficiencies will continue to drive waste. Organizations need to move beyond basic cost-cutting measures and adopt a structured FinOps approach to ensure cloud spending aligns with business value. By leveraging best practices in usage, pricing, architecture, governance, and vendor negotiations, companies can significantly reduce cloud waste and unlock sustainable cloud cost savings.

APPENDIX 1

NPI for General-Purpose GPU Instances, H1 2025

	Average annual price ($)		
Location	**AWS**	**Microsoft Azure**	**Google Cloud**
North America			
US East			
Virginia	1,461	1,325	1,646
South Carolina			1,474
Ohio	1,461		1,474
US East-Central			
Iowa		1,420	1,474
Illinois		1,325	
US Central			
Texas		1,491	1,667
Wyoming		1,491	
US West-Central			
Arizona		1,288	
Utah			1,726
Nevada			1,646
US West			
California	1,680	1,490	1,726
Oregon	1,461		1,474
Washington		1,288	
Mexico	1,549		
Canada			
Toronto	1,609	1,426	1,619
Calgary	1,627		
Quebec City		1,426	
Montreal			1,619
Asia-Pacific			
Hong Kong	1,963	1,769	1,938
India			
Hyderabad	1,595		
Mumbai	1,593	1,580	1,726
Pune		1,400	
Chennai		1,804	
Delhi			1,726
Singapore	1,824	1,596	1,724
Indonesia			
Jakarta	1,837		1,880
South Korea			
Seoul	1,783	1,562	1,815
Busan		1,505	
Malaysia	1,571		
Thailand	1,571		
Australia			
Sydney	1,825	1,596	1,960
Melbourne		1,650	1,731
Canberra		1,596	
Japan			
Tokyo	1,869	1,637	1,814
Osaka	1,798	1,770	1,815
Taiwan			1,643

	Average annual price ($)		
Location	**AWS**	**Microsoft Azure**	**Google Cloud**
Northern Europe			
UK			
London	1,645	1,455	1,820
Cardiff		1,512	
Belgium			1,581
Germany			
Frankfurt	1,700	1,504	1,820
Berlin		1,864	2,114
Ireland	1,589	1,450	
Netherlands		1,512	1,582
Nordic region			
Finland			1,582
Norway			
Oslo		1,656	
Sweden			
Stockholm	1,531		
Gavle		1,358	
Central Europe			
France			
Paris	1,655	1,512	1,682
Marseille		1,894	
Poland			
Warsaw			1,820
Switzerland			
Zurich	1,863	1,655	1,963
Geneva		2,131	
Southern Europe			
Italy			
Milan	1,659	1,472	1,682
Turin			1,821
Spain			
Madrid		1,406	1,704
Middle East			
Bahrain	1,974		
Israel			
Tel Aviv		1,518	1,593
Saudi Arabia			
Dammam			2,240
UAE			
Dubai		1,577	
Abu Dhabi		2,044	
Qatar			
Doha		1,573	1,741
South America			
Brazil			
São Paulo		2,050	2,288
Rio de Janeiro		2,745	
Chile			
Santiago			2,058
Africa			
South Africa			
Cape Town	1,919	2,188	
Johannesburg		1,691	1,865

Source: BCG analysis.
Note: GPU = graphics processing unit; H1 = first half; UAE = United Arab Emirates. Prices are not shown if a povider does not offer cloud services in an area.

APPENDIX 2

NPI for AI-Specific GPU Instances, H1 2025

	Annual price ($)		
Location	**AWS**	**Microsoft Azure**	**Google Cloud**
North America			
US East			
Virginia	4,608	4,607	5,129
South Carolina			4,742
Ohio	4,608		
Alabama			4,742
US East-Central			
Iowa			4,742
Illinois		4,607	
US Central			
Texas		5,527	
US West-Central			
Arizona		4,607	
Utah			5,605
Nevada			5,129
US West			
California	5,528	5,527	5,605
Oregon	4,608		4,742
Washington		4,607	
Canada			
Toronto	5,116	5,116	
Montreal			5,218
Asia-Pacific			
Hong Kong	7,096		6,196
India			
Mumbai	5,072		5,605
Pune		5,072	
Chennai		6,500	
Delhi			5,693
Singapore	6,447	6,447	5,308
Indonesia			
Jakarta			5,787
South Korea			
Seoul		5,668	5,392
Australia			
Sydney	5,992		6,232
Melbourne			6,732
Canberra		5,992	
Japan			
Tokyo	6,220	6,220	5,392
Osaka	6,220		5,392
Taiwan			5,005

	Annual price ($)		
Location	**AWS**	**Microsoft Azure**	**Google Cloud**
Northern Europe			
UK			
London	5,387	5,387	5,750
Belgium			4,909
Germany			
Frankfurt	5,764	5,764	5,750
Berlin			7,303
Ireland	5,142		
Netherlands		5,764	4,911
Nordic region			
Sweden			
Stockholm	4,888		
Gavle		4,888	
Central Europe			
France			
Paris	5,387	5,387	5,501
Poland			
Warsaw		5,747	5,752
Switzerland			
Zurich			6,286
Southern Europe			
Italy			
Milan	5,396	5,387	5,501
Turin			6,117
Spain			
Madrid			5,596
Middle East			
Bahrain	5,650		
Israel			
Tel Aviv		6,342	5,217
Saudi Arabia			
Dammam			7,588
Qatar			
Doha			5,762
South America			
Brazil			
São Paulo		7,831	6,865
Chile			
Santiago			6,782
Africa			
South Africa			
Cape Town	6,114		
Johannesburg			6,199

Source: BCG analysis.
Note: GPU = graphics processing unit; H1 = first half; UAE = United Arab Emirates. Prices are not shown if a povider does not offer cloud services in an area.

About the Authors

Filippo Scognamiglio is a managing director and partner in the New York office of Boston Consulting Group. You may contact him by email at **scognamiglio.filippo@bcg.com**.

Sukand Ramachandran is a managing director and senior partner in BCG's London office. You may contact him by email at **ramachandran.sukand@bcg.com**.

Michael Engelhardt is a managing director in the Munich office of BCG Platinion. You may contact him by email at **engelhardt.michael@bcgplatinion.com**.

Prasanna Santhanam is a partner and director in BCG's Singapore office. You may contact him by email at **santhanam.prasanna@bcg.com**.

Abhinav Gupta is a knowledge expert in BCG's Bengaluru office. You may contact him by email at **gupta.abhinavIKA@bcg.com**.

Pankaj Sherawat is a senior knowledge analyst in BCG's Bengaluru office. You may contact him by email at **sherawat.pankaj@bcg.com**.

JR Quinones is a senior analyst in BCG's Boston office. You may contact him by email at **quinones.jr@bcg.com**.

Andrew Dunbar is a lead analyst in the London office of BCG Expand. You may contact him by email at **andrew.dunbar@expandresearch.com**.

Basem Abdelrahman is a lead analyst in the Toronto office of BCG Expand. You may contact him by email at **basem.abdelrahman@expandresearch.com**.

Daniel Øllgaard is a senior principal in the New York office of BCG Expand. You may contact him by email at **daniel.ollgaard@expandresearch.com**.

For Further Contact

If you would like to discuss this article, please contact one of the authors.

Boston Consulting Group partners with leaders in business and society to tackle their most important challenges and capture their greatest opportunities. BCG was the pioneer in business strategy when it was founded in 1963. Today, we work closely with clients to embrace a transformational approach aimed at benefiting all stakeholders—empowering organizations to grow, build sustainable competitive advantage, and drive positive societal impact.

Our diverse, global teams bring deep industry and functional expertise and a range of perspectives that question the status quo and spark change. BCG delivers solutions through leading-edge management consulting, technology and design, and corporate and digital ventures. We work in a uniquely collaborative model across the firm and throughout all levels of the client organization, fueled by the goal of helping our clients thrive and enabling them to make the world a better place.