Enterprise Spotlight: New Thinking about Cloud Computing

Credit: puhhha/Shutterstock

Cloud computing is highly established in its many forms, and IT's view has shifted from understanding the concept to making the best use of it to solve both technology and business challenges. The time is ripe to step back and revisit your thinking about the cloud.

CONTENTS

From the editors of Foundry's enterprise IT sites:

CIO CSO COMPUTERWORLD InfoWorld NETWORKWORLD

The cloud architecture renaissance of 2025

BY DAVID LINTHICUM, INFOWORLD.COM

The perfect storm is coming that will force enterprises to rethink their cloud strategy. Cloud architecture will take center stage during 2025. This isn't just another hype cycle.

First, we need to talk about the elephant in the room: generative AI. The computational demands of running generative AI models make traditional cloud deployments look like a kid's lemonade stand. According to Gartner's projections, enterprise AI workloads will consume more than 30% of total cloud infrastructure capacity by 2025. Considering the elevation of AI-driven cloud spending, that transition is underway right now.

Here's the kicker — and I've been shouting this from the rooftops to anyone who will listen — public cloud costs are becoming the board room's newest headache. The "lift-and-shift" parties of the past decade have created massive technical debt. CFOs are choking on their morning coffee when they see the bills. We're talking about companies spending two or three times what they initially budgeted for cloud services, and that's before adding AI workloads into the mix.

The most common consulting request I have these days is to figure out why IT is spending so much on public cloud resources. The pleas come from boards of directors, CEOs, and CFOs — people who had little interest in IT infrastructure only a decade ago.

Here's where it gets interesting. Smart money is focused on reducing cloud costs by better optimizing resource utilization. By smart money, I'm talking about institutional investors pushing for a more sophisticated approach to cloud architecture. They're no longer buying the "all-in on public cloud" story. Instead, they pose challenging questions regarding hybrid cloud architectures, integration of edge computing, cloud-native optimization patterns, and the possibility of returning workloads to on-prem environments.

Think about it this way: If you're running large language models and need to process sensitive data, do you want to pay premium rates for public cloud GPU instances? This is why we're seeing a renaissance in private cloud architectures but with a twist. Now they're being designed from the ground up to support AI workloads while maintaining data sovereignty.

What about the vendors? They're scrambling to catch up. Traditional cloud providers are racing to offer better hybrid solutions, while enterprise tech companies are finally getting their acts together with usable private cloud platforms. The consulting firms are changing their messages from "Use your cloud partners," to "Let's rethink what we've been doing for the past 10 years."

The bottom line is that 2025 won't be just another year of cloud evolutions. Rather, it's shaping up to be the year we see a fundamental shift in how we architect our systems. Innovative enterprises are already preparing for this shift, as are many who might find themselves on the wrong side of the technology curve.

GAME PLAN FOR 2025

Enterprises must take several steps to prepare for the coming cloud architecture renaissance. The good news? Enterprise goals can be met by adopting and applying new concepts and processes to existing and proven technologies. There's nothing magical about this approach.

First, get your house in order. The next three to six months should be spent deep-diving into current cloud spending and utilization patterns. I'm talking about actual numbers, not the sanitized versions you show executives. Map out your AI and machine learning (ML) workload projections because, trust me, they will explode beyond your current estimates. While you're at it, identify which workloads in your public cloud deployments are bleeding money — you'll be shocked at what you find.

Next, develop a workload placement strategy that makes sense. Consider data gravity, performance requirements, and regulatory constraints. This isn't about following the latest trend; it's about making decisions that align with business realities. Create explicit ROI models for your hybrid and private cloud investments.

Now, let's talk about the technical architecture. Your focus must be on optimizing data pipelines, integrating edge computing, and meeting AI/ML infrastructure requirements. Multicloud connectivity isn't optional anymore — it's a requirement for survival. But here's the catch: You must also maintain ironclad security and compliance frameworks.

The organizational piece is critical, and most enterprises get it wrong. Establish a Cloud Economics Office that combines infrastructure specialists, data scientists, financial analysts, and security experts. This is not just another IT team; it is a business function that must drive real value. Investment priorities need to shift, too. Focus on automated orchestration tools, cloud management platforms, and data fabric solutions.

DOLLARS MATTER

Financial management is crucial. Implement proper chargeback mechanisms and develop explicit total-cost-of-ownership models. Make people accountable for cloud spending. You'll be amazed how behavior changes when departments see the actual costs of their cloud usage. Watch out for FinOps. Although there is value in FinOps, the way some "FinOps consultants" are explaining and implementing it leads to false metrics — just saying.

This transformation should span 12 to 24 months, starting with assessment and planning, moving through pilot projects, and ending with full-scale implementation. But remember, this isn't just an IT project. It's a business transformation initiative that needs buy-in from all stakeholders.

The winners in 2025 won't be the enterprises that spend the most on cloud services. It will be the organizations that build intelligent, flexible cloud architectures that align with their business goals. Now is the time to start, before the market forces your hand and you're left playing catch-up. You've been warned, and you know I'm not above saying "I told you so." ■

CIOs are rethinking how they use public cloud services. Here's why

BY ROBERT MITCHELL, CIO.COM

Over the past few years, enterprises have strived to move as much as possible as quickly as possible to the public cloud to minimize CapEx and save money. Increasingly, however, CIOs are reviewing and rationalizing those investments. Are they truly enhancing productivity and reducing costs?

"In the rush to the public cloud, a lot of people didn't think about pricing," argues Tracy Woo, a principal analyst at Forrester. And for some organizations, annual cloud spend has increased dramatically. "Cloud spending is going up and budgets are tightening, so they're asking what's going on and how do we right this ship."

In 2025, the plan, according to Ron Hollowell, SVP and CTO at Reinsurance Group of America (RGA), is to focus on right-sizing their public cloud footprint by maturing processes around work intake, distribution criteria, and implementation practices across private and public clouds. "Expense optimization and clearly defined workload selection criteria will determine which go to the public cloud and which to private cloud," he says.

As VP of cloud capabilities at software company Endava, Radu Vunvulea consults with many CIOs in large enterprises. "This will be a year when we talk more about hybrid cloud, multi cloud, and repatriation to on-premises," he says. The reasons include higher than expected costs, but also performance and latency issues; security, data privacy, and compliance concerns; and regional digital sovereignty regulations that affect where data can be located, transported, and processed.

"The primary driver for leveraging private cloud over public cloud is cost," Hollowell says. He sees public cloud as the most cost efficient for seasonal or bursty, on-demand workloads. "For workloads with more consistent capacity demands, the economics can be more attractive for private cloud and fixed-capacity solutions."

For many other CIOs, the primary motivator is cost as well, argues Vunvulea. While up to 80% of the enterprise-scale systems Endava works on use the public cloud partially or fully, about 60% of those companies are migrating back at least one system. “We see this more as a trend,” he says.

Where are those workloads going? “There’s a renewed focus on on-premises — on-premises private cloud or hosted private cloud — versus public cloud, especially as data-heavy workloads such as generative AI have started to push cloud spend up astronomically,” adds Woo. “By moving applications back on premises, or using on-premises or hosted private cloud services, CIOs can avoid multitenancy while ensuring data privacy.” That’s one reason why Forrester predicts four out of five so called cloud leaders will increase their investments in private cloud by 20% this year.

That said, 2025 is not just about repatriation. “Private cloud investment is increasing due to GenAI, costs, sovereignty issues, and performance requirements, but public cloud investment is also increasing because of more adoption, generative AI services, lower infrastructure footprint, access to new infrastructure, and so on,” Woo says.

HIDDEN COSTS OF PUBLIC CLOUD

For St. Jude’s Research Hospital, the public cloud is a good way to get knowledge into the hands of researchers who aren’t part of their ecosystem today, says SVP and CIO Keith Perry. The hospital uses on-premises supercomputers to generate much of its research data, and the movement of that data into and out of the public cloud can become expensive. “The academic community expects data to be close to its high-performance compute resources, so they struggle with these egress fees pretty regularly,” he reveals.

But data-heavy workloads can be expensive, especially if constant, high-compute is required. “Another driver is data movement, not only in terms of dollars but in performance,” Hollowell says. “So we carefully manage our data life cycle to minimize transfers between clouds.”

Woo adds that public cloud is costly for workloads that are data-heavy because organizations are charged both for data stored and data transferred between availability zones (AZ), regions, and clouds. Vendors also charge egress fees for data leaving as well as data entering a given AZ. “So for transfers between AZs, you essentially get charged twice, and those hidden transfer fees can really rack up,” she says. And Vunvulea says the cost of data transfer, especially in terms of petabytes, is high, and data transfer and synchronization can be complex. “We’ve seen AI projects where around 45% of cloud costs are generated by moving data from the public cloud to another location,” he says. “And if you put the full systems in place with everything you need around the service, you can have a solution that costs three or four times more than the initial estimation.”

For example, organizations that build an AI solution using OpenAI need to consider more than the AI service. Adding vaults is needed to secure secrets. Security appliances and policies also need to be defined and configured to ensure that access is allowed only to qualified people and services. Secure storage, together with data transformation, monitoring, auditing, and a compliance layer, increase the complexity of the system. Around the AI service, you need to build a solution with an additional 10 to 12 different cloud services that fulfill the needs of an enterprise system.

Jeff Wysocki, CIO at mining firm The Mosaic Company, acknowledges those budget-busting concerns, but he says CIOs may be able to work with their public cloud provider to get those costs under control. For example, Mosaic recently created a data-heavy Mosaic GPT safety model for mining operations on Microsoft’s Bing platform, and is about to roll that out in a pilot. It contains years of safety information that Mosaic built into the model, so contractors working at a mining site can enter questions around safety and see how to handle a given situation.

“We made changes to our architecture to get around the cost issues,” he says. How Mosaic’s team built the models, as well as how Microsoft architected the solution, helped to keep the project within budget. “We made some changes with Microsoft to get the cost down to something we can consider a reasonable return.”

Mosaic’s ERP system initially resided in a private cloud but now runs in an SAP private cloud, says Wysocki. But, he adds, some servers will always be on premises, and that’s unlikely to change, although there may be edge server solutions with cloud synchronization. “I don’t see that evolving too much beyond where we are today.” Between 80 and 85% of the company’s IT operations are in the cloud, and he expects it to stay that way.

AI PROJECTS CAN BREAK BUDGETS

Because AI and machine learning are data intensive, these projects can greatly increase cloud costs. Organizations don’t have much choice when it comes to using the larger foundation models such as ChatGPT 3.5 and 4.0 because the scale of compute power required would be too costly to reproduce in-house, says Sid Nag, VP, cloud, edge, and AI infrastructure services and technologies at Gartner.

By 2027, however, more than 50% of the GenAI LLMs enterprises use will be industry-specific, Gartner predicts. These will be a much smaller carve-out of the very large-scale general-purpose foundation models, and could be run elsewhere. Even after organizations use tools such as RedHat's InstructLab to augment those industry-specific models with company-specific data, they're still small by comparison. "Industry-specific models…require fewer resources to train, and so could conceivably run on on-premises, in a private cloud, or in a hosted private cloud infrastructure," argues Nag.

But, says Vunvulea, the computation power and infrastructure needed to train or optimize the model isn't easy to find or buy on prem. "Computation needs are one of the most important factors," he says. Fortunately, cloud vendors also offer off-the-shelf AI platforms that enterprises can use to train their models against their own data. "So you don't need to configure the on-premises system, even if you decide to run it there."

But should you? "I'd be cautious about going down the path of private cloud hosting or on premises," warns Nag. "Decision-makers with fiduciary responsibility are going to balk at the idea of going back to the days of CapEx unless there are compelling reasons to do so."

Cloud vendors continue to provide more AI and ML services as part of their platform-as-a-service offerings, Vunvulea says. You start with a pretrained model, bring your own data, and just use the service without any problems. "We're getting close to the point when the models available from public cloud vendors are mature enough to cover up to 90% of the standard needs of most companies," he says. The question as to whether to use those services or not will come down to cost: Do the numbers make sense for your business model?

INEXPENSIVE BUT UNDERPERFORMING

At first, says Woo, CIOs focused on reducing cost, but that doesn't always align with performance considerations or end goals. Even when the public cloud is the less costly option, it may not be the best fit if potential latency or other performance issues are factored in. That's particularly true for industries that can't tolerate latency, such as in payment processing and financial services, says Vunvulea.

"The latency between the instrument producing the data and the compute power that processes it is an important variable in determining data location," says St. Jude's Perry. In some cases, that instrument needs an almost instantaneous connection to high-performance compute resources. “Due to the latency between research instrumentation and our high-performance computers on premises and the public cloud, using the public cloud to perform real-time checks doesn’t make sense.” And as more public cloud hyperscalers build large-scale GPU clusters that can handle high-performance computing, you also have to factor in the cost, he says.

Genomic sequencing is one area where offloading some processing from local supercomputers to the public cloud might make sense — if the price is right. Some of the workflows associated with genomic sequencing become somewhat standardized over time, Perry says. In those cases it may make more sense to optimize the pipelines for scale and run them in the cloud, depending on the cost. “We’ve worked on moving some of our genomic sequencing pipelines into the cloud to free up cycles on our on-premises high-performance compute,” he says.

Performance is certainly important, but not the deciding factor when choosing whether to host an application in the public cloud — with the exception of some that run on edge servers at Mosaic’s mining operations sites, says Wysocki. “For us there’ll always be a need for edge computing that needs to be on the device or near it to be effective.”

A QUESTION OF LOCATION

Security, privacy, and cost are the three main factors for us,” adds Wysocki. But so far, security and privacy haven’t been major issues with public cloud services.

Hollowell says RGA is satisfied with the security of its public cloud service. “We’re utilizing foundation models from Anthropic, Mitral, and others through AWS’s Bedrock service, which provides data isolation and security,” he says, enabling the company to provide ChatGPT-like functionality in a secure environment.

But digital sovereignty issues are a different matter, says Woo. In countries with strict localization rules, public cloud may be a non-starter. “You can opt for on-premises private cloud or hosted private cloud where you manage it or someone else does,” she explains. “Either way you have control over where your data resides.”

But the regulatory landscape isn’t the only factor, Hollowell says. “In some geographies, data localization and privacy requirements are embedded directly into customer contracts,” he says. In such cases, a private cloud may offer a more flexible solution. So a hybrid approach between on prem and cloud is the best choice for large organizations running in multiple countries, says Vunvulea. And with respect to regional regulations, the choice of public cloud provider matters. “For example, Oracle cloud is one of the best options if you want to run workloads from inside a specific location in the Middle East,” he says, where each country has its own regulations for handling data. No single cloud provider has a presence in all those countries, but Oracle has a big footprint there, so you can run on-premises workloads with Oracle and other cloud vendors.

But there’s a downside to hybrid cloud, warns Hollowell. “Managing interoperability and performance for large data sets across public and hybrid cloud environments remains a key challenge to address, he says.

MAINTAIN YOUR FLEXIBILITY — AND BE READY TO ADJUST

Going forward, reveals Hollowell, “our strategic intent is to evaluate hosting decisions through the lens of evolving business requirements for new features, combined with natural application life cycle management practices, rather than simply moving everything to the public cloud.” Applications with consistent capacity requirements that can be satisfied with traditional converged infrastructure will run in a private cloud, while those that don’t consistently require high compute will remain candidates for public cloud.

For Perry, constructing the right IT infrastructure for his organization’s applications is all about using the right building materials. “Public cloud is just one of the materials we need to build an architectural solution,” he says, and you have to strike the right balance.

Unfortunately, optimizing the mix of on-premises, private cloud, and public cloud services is a moving target. “I can’t say that everything is in the right place because the technology is evolving constantly,” Perry says. Cloud technologies are always changing, so be ready and able to change with the times, he advises. Making sure you have the right tools to do that is extremely important since the tools you have today might not be the ones you’ll need tomorrow.

That need to change things up as the technology advances is also a reason why you should avoid vendor lock-in, says Vunvulea. That’s a conundrum because to run cloud workloads in the most optimized way, you may need to use the vendors’ most advanced, proprietary features.

But in the end, he says, you want to avoid lock-in to have the flexibility to move more easily between on-premises, public cloud, and private cloud. ■

Cloud trends: Repatriation and sustainability make their marks

BY BRIAN ADLER, INFOWORLD.COM

What are the top priorities and challenges related to the use of cloud computing? The Flexera 2025 State of the Cloud Report draws on the insights of 759 cloud decision-makers and users globally who took part in a survey in late 2024. The results illustrate the evolution of ongoing trends in past years, while simultaneously spotlighting the emergence of new forces driving cloud usage.

WORKLOADS MOVE BACK TO DATA CENTERS

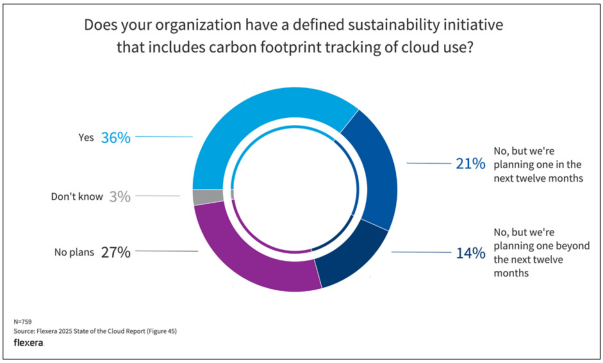
A noteworthy shift of applications and data back from cloud to data centers — known as repatriation — is happening. Slightly more than one-fifth (21%) of workloads and data have been repatriated. However, ongoing migration to cloud and net-new cloud workloads outstrip these cloud exits, resulting in continued cloud growth.

Analysts and experts have, for some years now, indicated that organizations are moving cloud workloads back to their own data centers, often due to the inefficiencies and expenses that result from failing to refactor applications for cloud. Although net-new cloud workloads are still increasing, the frequency of repatriation is notable.

SUSTAINABILITY GAINS GROUND

Cloud sustainability initiatives are becoming top of mind for many respondents. More than half (57%) of respondents either have or plan to have a defined sustainability initiative that includes carbon footprint tracking of cloud use within the next 12 months. With more than a third (36%) of all respondents already tracking their cloud carbon footprint, the need to do so has clearly been gaining traction.

Among European respondents, the number tracking their cloud carbon footprint rises to 43%. The gap between European respondents and respondents overall is closing; as an increasing number of global



N=759
Source: Flexera 2025 State of the Cloud Report (Figure 45)

Credit: Flexera

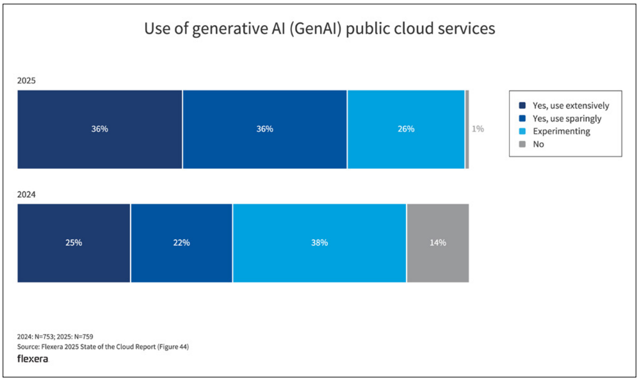
More than a third (36%) of all respondents are already tracking their cloud carbon footprint.

organizations adopt and adhere to important sustainability standards, this gap is expected to shrink even further.

GENERATIVE AI IS BECOMING MAINSTREAM

Not a surprise: Adoption of AI-related public cloud services is exploding. Almost half of respondents indicate that their organizations already use artificial intelligence/machine learning (AI/ML) platform-as-a-service (PaaS) services. This year's survey also shows a surge in the use of data warehouse services, which are often used to feed AI models.

Generative AI use is also booming. Nearly three-quarters (72%) of organizations already use GenAI either sparingly or extensively; another 26% are currently experimenting with GenAI. Not only is GenAI here to stay, but it's becoming mainstream.



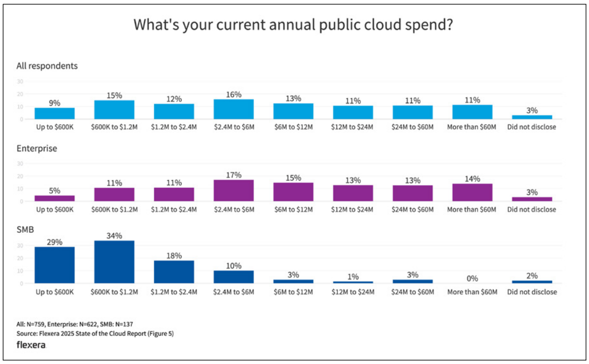
2024: N=753; 2025: N=759
Source: Flexera 2025 State of the Cloud Report (Figure 44)

Credit: Flexera

Almost three-quarters (72%) of organizations already use GenAI either sparingly or extensively.

CLOUD SPEND AND SECURITY ARE THE TOP CHALLENGES

Managing cloud spend is the top cloud challenge for organizations of all sizes, reported by 84% of respondents. As additional workloads find their way into the cloud, the need to manage and optimize the associated spend becomes paramount. Nearly nine out of 10 (87%) identify "cost efficiency/savings" as their top metric for assessing progress



All: N=759, Enterprise: N=622, SMB: N=137
Source: Flexera 2025 State of the Cloud Report (Figure 5)

Credit: Flexera

A third (33%) of respondents are spending more than $12 million every year, up from 29% last year.

against cloud costs, making it the leading metric in this category, jumping from 65% a year ago. Similarly, “cost avoidance,” which can be achieved with proper license management, rose from 28% to 64% during the same period. As software-as-a-service (SaaS) usage increases, the focus on SaaS licensing is gaining increased attention, given the significant impact that SaaS expenses have on driving up cloud bills.

Following cloud spend as the top cloud challenge is security. Reported by 77%, security — always a top concern in the digital age — is the second-largest challenge for cloud initiatives. Among the tools used for managing multicloud initiatives, security tools take the number one spot, with 55% of all respondents using them.

PUBLIC CLOUD ADOPTION CONTINUES TO ACCELERATE

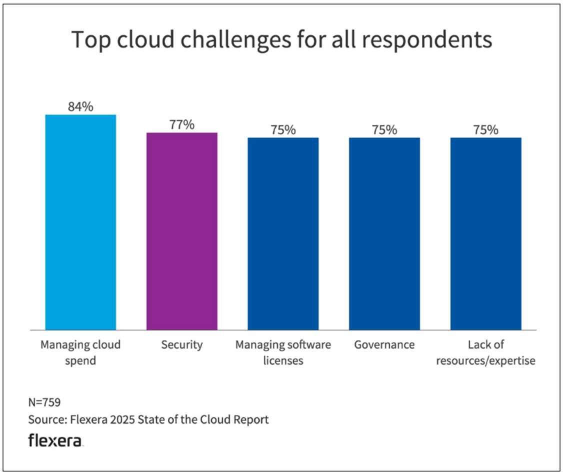
Public cloud spend continues to increase, with a third (33%) spending more than $12 million a year, up from 29% of respondents last year. Among enterprises (with more than 1,000 employees), the number spending this amount goes up to 40%. Even as cloud costs rise, more workloads are finding a home in the cloud. SaaS expenses remained fairly consistent year over year.

An area of hesitance is around sensitive data. Organizations remain cautious about moving sensitive data to the cloud, although more than a third indicate that all non-sensitive data will move to the cloud.

CENTRALIZED INITIATIVES GROW

The approach to governing and optimizing cloud and SaaS costs is shifting from vendor management teams towards cloud centers of excellence (CCoEs) and FinOps teams, representing a centralized approach to cloud. Today 69% of respondents have a CCoE or central cloud team.

Additionally, cloud cost optimization strategies are increasingly being handled



N=759
Source: Flexera 2025 State of the Cloud Report

Credit: Flexera

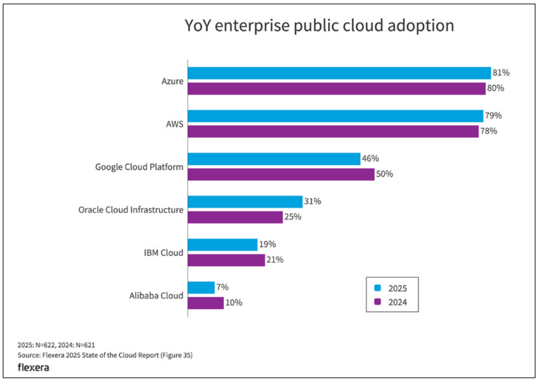
Respondents said managing cloud spend (84%) and security (77%) are the top cloud challenges.

by FinOps teams. Nearly three-fifths (59%) of respondents now indicate they have a FinOps team for some or all of their cloud cost optimization strategies, up from 51% a year ago. As FinOps gains additional traction within the cloud community, particularly with SaaS and data centers now part of the FinOps scopes, reliance on FinOps teams across organizations is anticipated to rise.

AWS AND AZURE COMPETE FOR DOMINANCE

Year over year, this ongoing research shows that there has been little change among the leaders, with many organizations seemingly having found their steady state regarding the cloud — or mix of clouds — they're using. Among all respondents, it boils down to a race that continues between Amazon Web Services (AWS) and Microsoft Azure as leading public cloud providers. A close contest in recent years, the two providers trade leads, based on the number of workloads running.

Historically, enterprises are more likely to use Azure than are small- to medium-size businesses (those with fewer than 1,000 employees). Today, among enterprises, AWS holds a slight lead (53%) over Azure (50%) among organizations that run "significant workloads," while Azure (81%) has the lead over AWS (79%) when also including "some workloads."



2025: N=622, 2024: N=621
Source: Flexera 2025 State of the Cloud Report (Figure 35)

Credit: Flexera

Amazon Web Services (AWS) and Microsoft Azure are the leading public cloud providers.

As part of cloud strategy, organizations continue to embrace multicloud: 70% of respondents embrace hybrid cloud strategies, using at least one public and one private cloud, while the remaining 30% use only public clouds or private clouds. Large enterprises (with more than 10,000 employees) make use of multicloud tools more than smaller organizations, regardless of the tool type.

LOOKING AHEAD

Growing cloud usage, initiatives to optimize costs, competition between the top cloud providers, and the ongoing use of AI all promise to be hallmarks of cloud programs in 2025. The new emphases on repatriation and sustainability will modulate how cloud initiatives are managed.

Brian Adler is senior director of cloud market strategy at Flexera. He was previously a senior director analyst at Gartner and a member of the FinOps Foundation governing board. This report appeared in InfoWorld's New Tech Forum, which provides a venue for technology leaders — including vendors and other outside contributors — to explore and discuss emerging enterprise technology in unprecedented depth and breadth. The selection is subjective, based on our pick of the technologies we believe to be important and of greatest interest to InfoWorld readers. InfoWorld does not accept marketing collateral for publication and reserves the right to edit all contributed content. ■