INFO~TECH
RESEARCH GROUP

Growth *Is* Crashing for the Public Cloud Hyperscalers

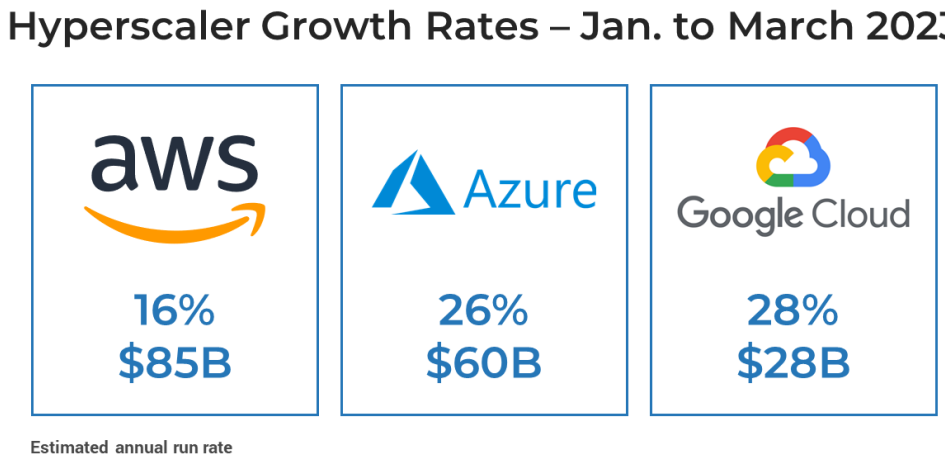
What Comes Next, and How Can You Benefit?

By Scott Bickley and Josh Mori

Cloud Hyperscaler Performance Overview

The continuing theme from Q4 of 2022 into Q1 of 2023 across the Big Three public cloud hyperscalers can be summed up in the phrase "customer workload optimization," which is the process of cloud customers performing deep-dive exercises to drive costs and inefficiencies out of their recurring cloud spend. Cloud workload optimization has a material effect on the hyperscalers' earning growth, driving growth rates down and countering the positive effects of new cloud growth projects (migration and digital transformation). None of the Big Three were spared rolling out of Q1 of 2023.

How long will this growth slump last? The hyperscalers only provide forward guidance one quarter at a time, and Q2 is projected to be a weak quarter. Microsoft has stated that cloud optimization efforts will end eventually, because once the cost rationalization efforts are realized, cloud optimization is typically a one-time event that is then maintained over time. At that point, revenue growth can re-accelerate, but likely at a new and lower baseline rate. Cloud migrations and digital transformation projects are not stopping; they are experiencing elongated decision-making cycles and an enhanced focus on making more calculated bets in the current challenging macroeconomic environment.



Source: Info-Tech Research Group

These growth rates are down materially over fiscal year 2022 (FY22) numbers:

AWS exited FY22 at a growth rate of 21%; it ended Q1 23 at 16% but exited the month of April at a meager 11% growth rate. Google and Microsoft maintained much more stable growth rates, although down moderately from 2022 levels.

There is a silver lining in this dismal growth story, at least for Microsoft initially. In fact, Microsoft realized a 1% tailwind to Azure revenue growth in its most recent earnings report, and it just launched commercially as it entered its Q3! This 1% of Azure growth translates to 4% of the total revenue increase between 26% and 27%. The various new AI-based services (e.g. ChatGPT and other large language models) delivered over the internet require an immense level of compute, storage, unstructured data management, managed databases, security, and health monitoring resources. Much of the compute power required is derived from specialized GPUs from Nvidia or, increasingly, from proprietary chipsets designed by the cloud providers.

Theme of the Year: Cloud Optimization

Customers have been in various phases of optimizing their cloud workloads for the past two years. The hyperscalers offer similar consumption models for their infrastructure-as-a-service (IaaS) technology offerings but with notable and subtle differences. Yet all these environments can be right sized to provide the requisite services at an ideal cost point. Without going into laborious detail in this note, it is important to understand some of the key mechanisms enabling cloud cost rationalization.

The past few years leading into the pandemic, as well as the initial exit, were a free-for-all where enterprises entered a frantic race against the competition to transform and digitalize at scale as the primary consumption mechanism shifted to cloud and e-commerce wholesale. As companies ramped up these efforts, little regard was given to cloud costs, which were deprioritized against delivery of the critical capabilities required for these organizations to thrive, or at least survive. All of this has changed now.

As organizations dig deep into their cloud bills and usage, the waste detected can be staggering, often reaching up to 30% of overall cloud spend. Whether using third-party services, creating a FinOps function, or using good old-fashioned grit and determination, numerous methods are available to achieve the end goal of identifying and reducing waste.

Actions such as these exemplify the path organizations are taking today:

- Lean provisioning and right-sizing
- Use of spot instance pricing for applicable workloads

- Autoscaling
- Savings plans/reservations
- Automated provisioning with guardrails
- Reducing storage buffer
- Tier storage based on requirements
- Custom storage tiers where capacity is not tied to IOPS (licensed separately)
- Network optimization – reduce inter-region traffic/use cost-effective network services (e.g. DirectConnect)
- Constrained virtual machines (lower virtual machine consumption while preserving IOPS, memory, storage capacities of each virtual machine)
- Optimized log ingestion options

The blessing in disguise for the big cloud players is that these optimization efforts are generally done once, with the changes yielding the highest benefits being prioritized first where technically feasible. Microsoft and Amazon have both implied that optimization efforts will last throughout FY23 and that they will assist their customers in realizing these benefits. Info-Tech still sees hyperscalers pushing hard to expand cloud spend, so take this "assistance" with a grain of salt, as your mileage may vary from the public-facing statements. In the end, ongoing optimization efforts should become embedded in the daily management practices of most organizations, with FinOps and IT asset management functions joining forces to institutionalize these behaviors.

The team over at siliconANGLE aptly note that the ability to dial spend up or down in an IaaS world should be viewed as a feature, not a bug. There's a dual headwind here as cloud optimization efforts dovetail with an overall slowing in client consumption demand due to the deteriorating macroeconomic environment. As companies both optimize and adjust their consumption patterns downward, cloud growth patterns may be more muted, should they turn positive later in 2023.

AI to Accelerate Cloud IaaS Growth

AI as a key technology trend promising the next wave of growth for the enterprise is not a new concept, and many software vendors have already incorporated machine learning (ML) and AI functionality into their products for years now. The commercialization and widespread availability of generative AI large language models (LLMs) have changed the game, seemingly. Organizations should exercise caution, as every time there is a new "breakthrough" in the technology space, it is touted as heralding the next paradigm shift that will revolutionize all aspects of a business. It may just be true this time around.

The innovation of the generative AI user interface has evolved from basic search, click, and review to that of a natural language query model where instructions can be spoken or typed into the system and then iterated with ease until the desired output manifests. The natural language processing (NLP) model, coupled with the incredible processing capacity and super-charged AI models, will enable deeper insights than could ever be derived from legacy research processes. This is not arguable and represents a major value creator over a short period of time. It's safe to say this technology has already unleashed the next product race in the tech world.

Big Tech has the advantage because they are the only ones with pockets deep enough to scale the infrastructure required to run these LLMs at scale. For example, SemiAnalysis states that it costs $700,000 per day to run ChatGPT. Microsoft structured its initial deal with OpenAI with the underlying deal structure taking the form of Azure credits in exchange for a share of future profits from the commercialization of the AI LLM technology. Microsoft recently went all-in with OpenAI to the tune of a $10 billion investment in exchange for additional profit-sharing until it recoups its investment and then some. It will take time for this technology to proliferate through the Global 2000 and beyond, likely a few years at most. During this time, expect the size, scale, and competitive advantage gaps to dramatically widen to the advantage of those organizations who can afford to dive into this new world and start reaping the benefits of LLM-based AI productivity improvements.

Customer Impacts of Generative AI

Shock and awe. That's the attitude of just about everyone, marveling at the wonders of generative AI and dreaming of the possibilities of an AI-infused future, replacing the mass skepticism of recent years. But this marvelous new toolset is not free; it comes with a large price tag, as noted above. The large cloud players are already investing billions of dollars per quarter to modernize their legacy infrastructure while investing in as many new GPU chips as they can get their hands on.

Buyer beware. When declining growth rates of IaaS consumption are coupled with a massive CapEx infusion to fund the generative AI tsunami, somebody must pay those bills. That somebody will be the customer, both through direct fees to access the technology and through higher prices (or, at a minimum, a slowing of price declines) being passed through to the customer.

Let's look at what Big Tech is already saying about how generative AI will possibly impact their workforces. A disturbing trend is already emerging out of this quarter's earnings reports as many executives in Big Tech are planning to pause hiring for jobs that *may* be able to be replicated with the use of AI.

According to Bloomberg, IBM CEO Arvind Krishna stated that IBM would now start pausing hiring for back-office roles that could be replaced by AI to the tune of up to 30% of those 26,000 jobs or 7,800 if realized over the next five years. Other executives like Meta's CFO, Susan Li, were more circumspect in their commentary but noted that they anticipate positive workforce benefits over time. Dropbox lands in the middle but has announced plans to lay off 500 employees or 16% of their workforce to focus on AI-related growth initiatives, likely oriented toward reconfiguring the workforce to enable AI feature development. It is readily apparent that AI will soon be impacting the workforce from multiple angles, with the net result being pressure applied to overall headcount growth in the years ahead, should AI-driven productivity improvements manifest as currently projected – which is a big "if."

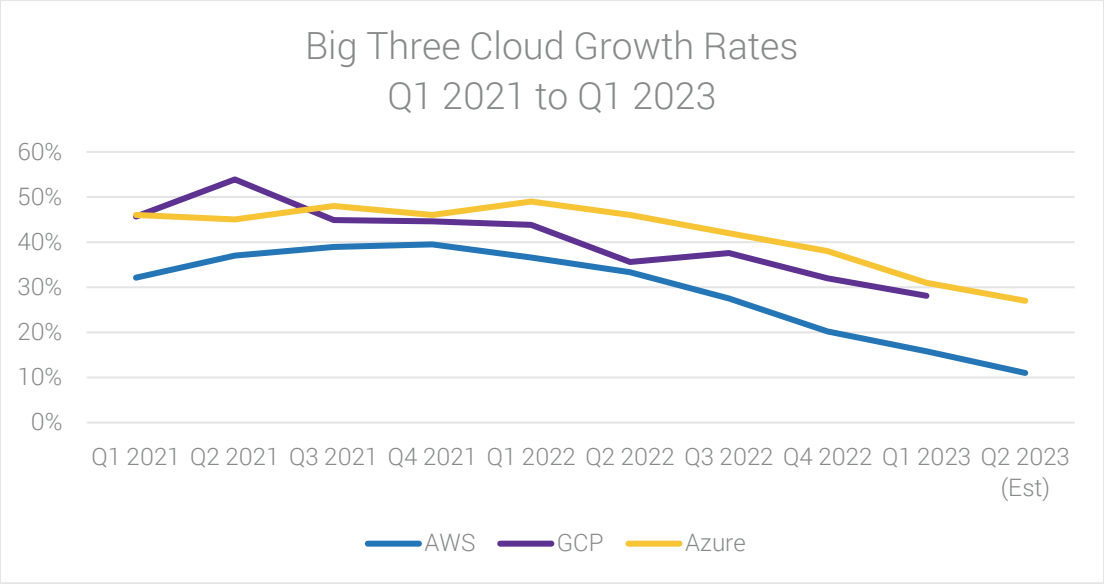
So how are customers going to pay for this non-trivial expense? The up-front investment costs in generative AI labor and infrastructure are massive. Despite the hype parade, true and scalable monetization of this technology will take years to refine and mature, and the "haves" of the Big Three will continue to create space between themselves and the "have nots" in terms of feature, function, and scale.

Developer Productivity

Another area that merits a cost-benefit analysis is the area of software development. With the launch of AI coding assistants (e.g. Microsoft's GitHub Copilot), initial reports are promising as many developers are estimating a 1.55X increase in productivity. Over time, efficiencies will continue to increase as mundane and time-consuming developer tasks are automated. These tasks include suggestions for code, test script creation (unit, regression, smoke, load), code review against defined standards, integration checks (e.g. API discovery), and more. This enhanced productivity will yield more application production and possibly fewer developers to tow the workload. This is an example of what the future might look like as budgeted spend gets split between headcount and AI services to enhance productivity. Don't get caught flat-footed! Currently, GitHub Copilot is in preview mode and is free of charge. This *will* change. Once all developers are dependent on or used to Copilot, it will be difficult to deny them a budget for these tools, no matter the cost.

Next Steps to Drive Value and Savings

The headwinds to cloud growth are now clear, but here is another look:



Note annual run rates: AWS – $85B; GCP – $28B; Azure – $60B+ Estimated
Source: Info-Tech Research Group (derived from earnings statements)

From a long-term perspective, AI will become the next catalyst powering growth in the cloud infrastructure space that will span far and wide, like the mobile app expansion of the last decade but with more concentration in the IaaS space. However, the road between today and tomorrow is an expensive tollway to navigate.

It is also necessary to put these declining cloud growth rates in perspective with the broader growth of overall IT spend, which is ~3%; when adjusted for inflation, it is shrinking! Over the long term, AI should be a deflationary force, perhaps on par with the internet or the emergence of the mobile phone. Until then, somebody has to pay the bills as the era of cheap money seems to be over for the foreseeable future.

Options:

1. Double down on cloud optimization efforts now.
2. If scale justifies the effort, create a FinOps team to ensure cloud lifecycle optimization is built into the process.
3. Budget for cloud pricing to remain at current levels or to rise slightly as the hyperscalers are now being judged by profit margin as much as revenue growth.
4. Consider hybrid cloud models that return a positive ROI for on-premises workloads.

5. Repatriate workloads to on-premises if the use case does not benefit from native cloud capabilities and the cost is higher in the cloud.
6. Negotiate flexible terms in your cloud agreement that allow for a rollover of unused capacity to the next term.
7. Be prudent, yet aggressive, in adopting generative AI and other AI/ML capabilities to automate workloads with the aim of reducing future headcount levels.
8. Seek out new instance offerings and lower pricing/performance on instances running the cloud providers' proprietary silicon.

OEM-Specific Commercial Strategies

Discounts with hyperscalers are not linear and require a mix with manifold dev activities to optimize costs:

Microsoft Azure

1. Choose procurement method (Server Cloud Enrollment, Microsoft Azure Consumption Commitment, Cloud Solution Provider, etc.).
    a) Server Cloud Enrollment (SCE) – Customers can take advantage of standardizing key server technologies and enjoy the best discounts available for their server licenses.
    b) Microsoft Azure Consumption Commitment (MACC) – The largest benefit of a MACC agreement is the 50% growth allowance, which offsets shortfall and allows customers to scale as needed with a lower up-front commitment.
    c) Cloud Solution Provider (CSP) – Deeper cross-service discounting and partner value-added services can be realized under this model. However, billing fluctuates monthly and is determined by Microsoft billing adjustments for Azure Cloud Services.
2. No cost for instances that you configure but don't run.
3. Upgrade clusters (Azure Kubernetes Service) and take advantage of the hybrid benefit, which allows customers to reuse licenses already owned for certain Microsoft products.

Google Cloud Platform

1. Service category discounts (i.e. Big Query)
2. Product weighting and contract merging (Anthos, Maps, Workspace, etc.)
    a. Negotiate retroactive invoice credits for and any delta between discounts across the different contracts
3. Commit flexibility and mid-term contract reviews – Customer ability to adjust commit or term within a defined range

AWS

1. AWS Migration Acceleration Program
    a) Discounts up to 25% for migration-specific workloads
2. Savings plans for heavy Compute, EC2, and SageMaker AI usage
3. Multiyear commitments are still the biggest driver for higher discounting.
    a) Promos available for ERP-specific and other containerized workloads