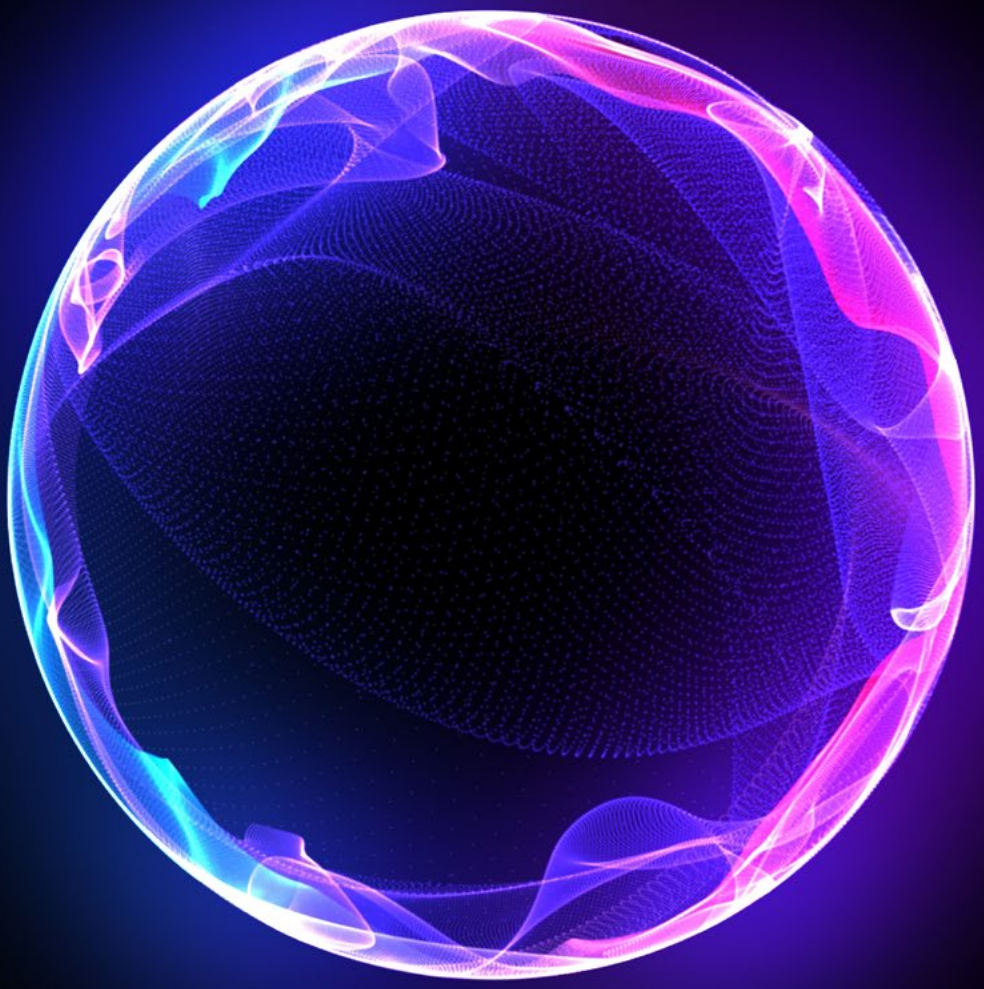


GlobalData.

NOVEMBER 2025

Solving the AI Infrastructure Dilemma: Choosing Cloud, On-Prem, or Hybrid for Strategic, Secure & Scalable Deployment

SPONSORED BY

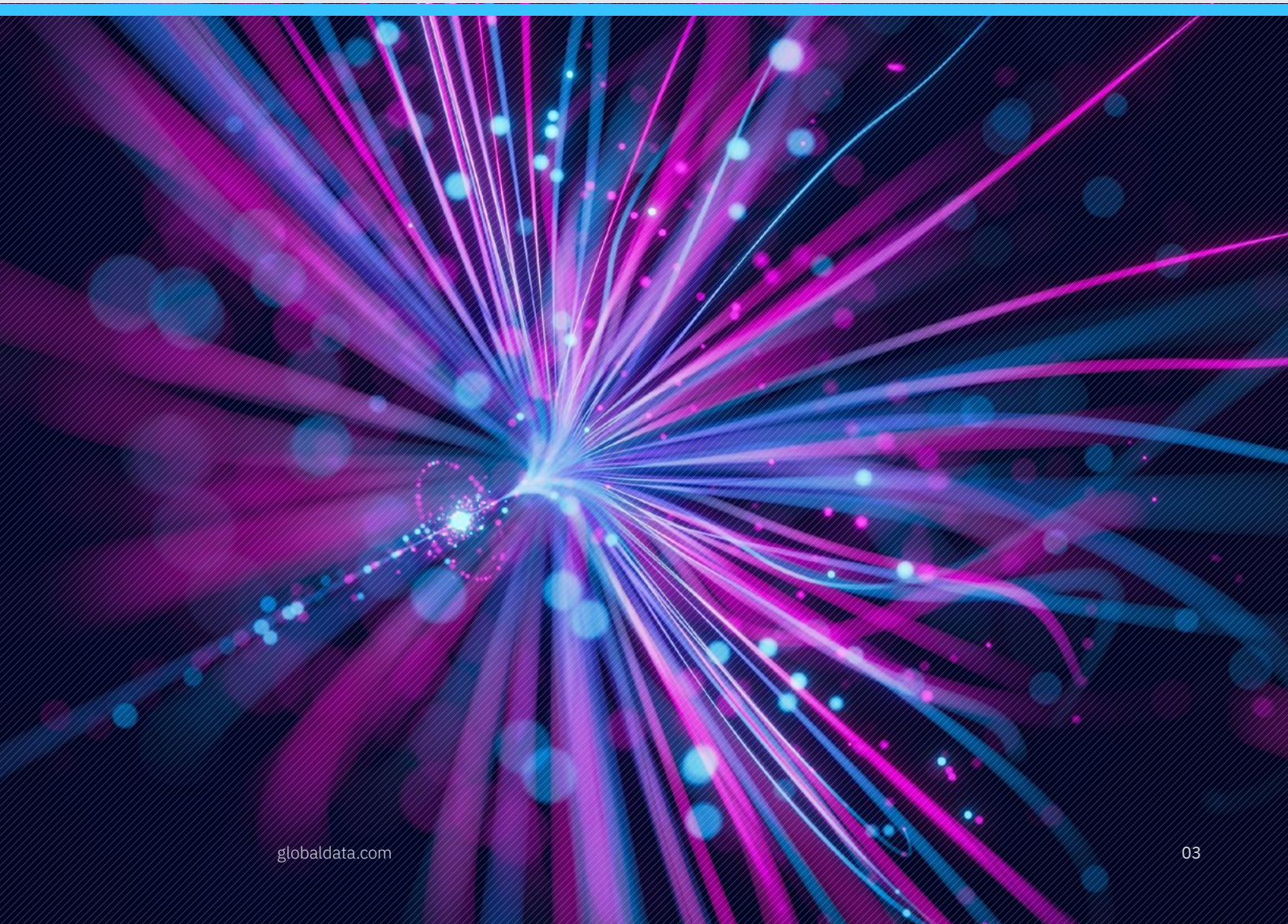
cisco

Contents

Introduction

"Past results are not indicative of future performance."

This standard phrase of legal boilerplate is familiar pabulum but contains wisdom that transcends its origins in the financial sector. Enterprises, institutions, and governments are all heavily investing in AI right now and if using the past performance standard, it will lead to the belief that AI workloads properly belong in the cloud. Years of cloud-first policy in IT has reduced the discussion of on-premises workloads to nearly nothing, mostly legacy and OT applications. Further, business objectives have evolved and achieving macro business goals requires more than a single approach to implementation of AI. The reality of general IT strategy today is centering on the hybrid approach, where the workload is placed where it makes the most sense, according to business needs, risk management, and overall costs. Understanding why that is and what has changed is key when considering where to place production AI workloads.



What once was and what it is today

When the cloud era began, enterprise workloads were in the on-premises data center, often virtualized with VMware or another machine virtualization solution. Cloud offered advanced functionality, most of the hardware abstracted away, automated, and deployable with only a few clicks of the mouse. Resources could be elastically expanded; there was no capital outlay. Compared to the standard machine virtualization of the day, it was a revelation. From there the strategic direction for most IT workloads was cloud first.

But the cloud model isn't perfect, and things have changed. Technology purchasing decisions are driven today by broader business needs and ambitions from business leaders, than by IT fiat like it was in the past. Also, as clouds have matured, cloud providers expanded their offerings – a good thing, but it led to add-on style billing. The resulting cloud bills are arcane at best and sprouted an entire trend in financial operations to identify exactly where the money was being spent and which workloads were attributed to it. The demand for cloud since its inception has continued to be strong – prices for base cloud services such as compute and general storage have trended down, mostly due to technological advancements. Further, competition has driven prices down, especially in government and very large enterprise deals. But hidden fees such as data egress from the cloud, faster/specialized storage, advanced networking, and of course GPUs have been trending up – these technologies are not commoditized. This has led to the widespread adoption of hybrid cloud, with workload being put in the cloud or on-premises, based on need. This trend will continue to grow and there is already some repatriation of cloud workloads back to an on-premises or colocation data center.

Technological advancement has been occurring in on-premises systems as well. The level of abstraction and self-service once exclusive to the cloud is now available on-premises through unified management platforms and validated full-stack solutions. The operational burden of on-premises workloads vs. cloud workloads has narrowed considerably, to the point where it isn't an overriding factor. Even subscriptions and scaling and use based consumption are available on premises by some vendors.

There are other factors as well. Security for on-premises environments is easier to achieve and maintain – there is no shared infrastructure and fully validated and integrated security is much more mature for on-premises deployments. From a pure performance standpoint, on-premises deployments can be created to fit the needs of the workload with higher speeds than cloud typically offers.

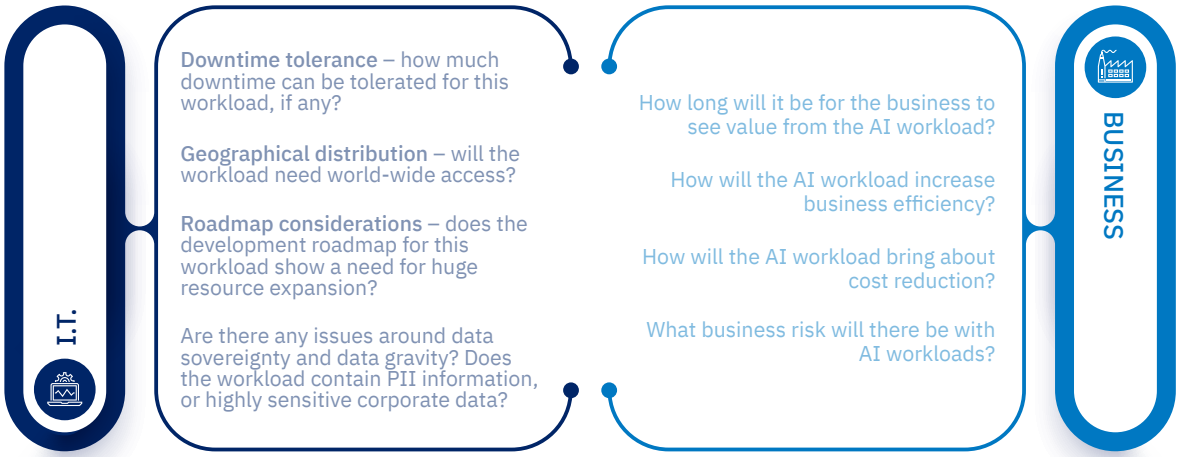
This by no means invalidates the value of cloud computing – but it does mean that production AI workloads can be run in on-premises environments or in the cloud. Hybrid clouds have obsoleted the cloud as the default choice and on-premises workloads can offer advantages enterprises should consider, especially in the AI era.

It’s not one or the other

In choosing on-premises or cloud for production AI workloads, enterprises should not view this decision to be binding across all AI workloads or across any given AI workload’s lifecycle. The use cases for AI are nearly infinite – AI workloads will have significantly different business needs.

While AI technologies are relatively new and have a steep learning curve, evaluation of AI workloads still encompasses a lot of standard, familiar factors to assess risk/reward from an IT perspective and from the standpoint of maximum business benefit.

Figure 1



Source:
GlobalData

These questions need to be answered for AI workloads. Even if the primary strategy for AI workloads is on-premises, there still may be later workloads that make more sense in the cloud. And vice versa. Not every on-premises AI installation needs to be huge – some AI production workloads may need to be near the edge and where the actual work is being done, for safety, latency, or legal reasons. Being flexible and putting the workload where it belongs is the heart of hybrid cloud philosophy.

Further, there are practical reasons to be flexible when it comes to AI workloads. A new, proposed AI workload can be prototyped and possibly go through proof of concept using a cloud instance. If the concept doesn’t work out as planned, there will be no need to expand on-premises hardware or risk disruption of existing AI workloads. However, if the new proposed AI workload is large, prototyping and proof of concept can be done with existing capacity on-premises without an excessive cloud bill and data exfiltration costs. The decision then can be made to bring it into a production on-premises or in the cloud. Development of highly sensitive AI workloads can also be done in an on-premises data center, mitigating security, data, regulatory, and privacy issues that would stem from a cloud environment. Flexibility should be maintained – the right environment for the right workload, based on usage projections, costs, and risk evaluations.

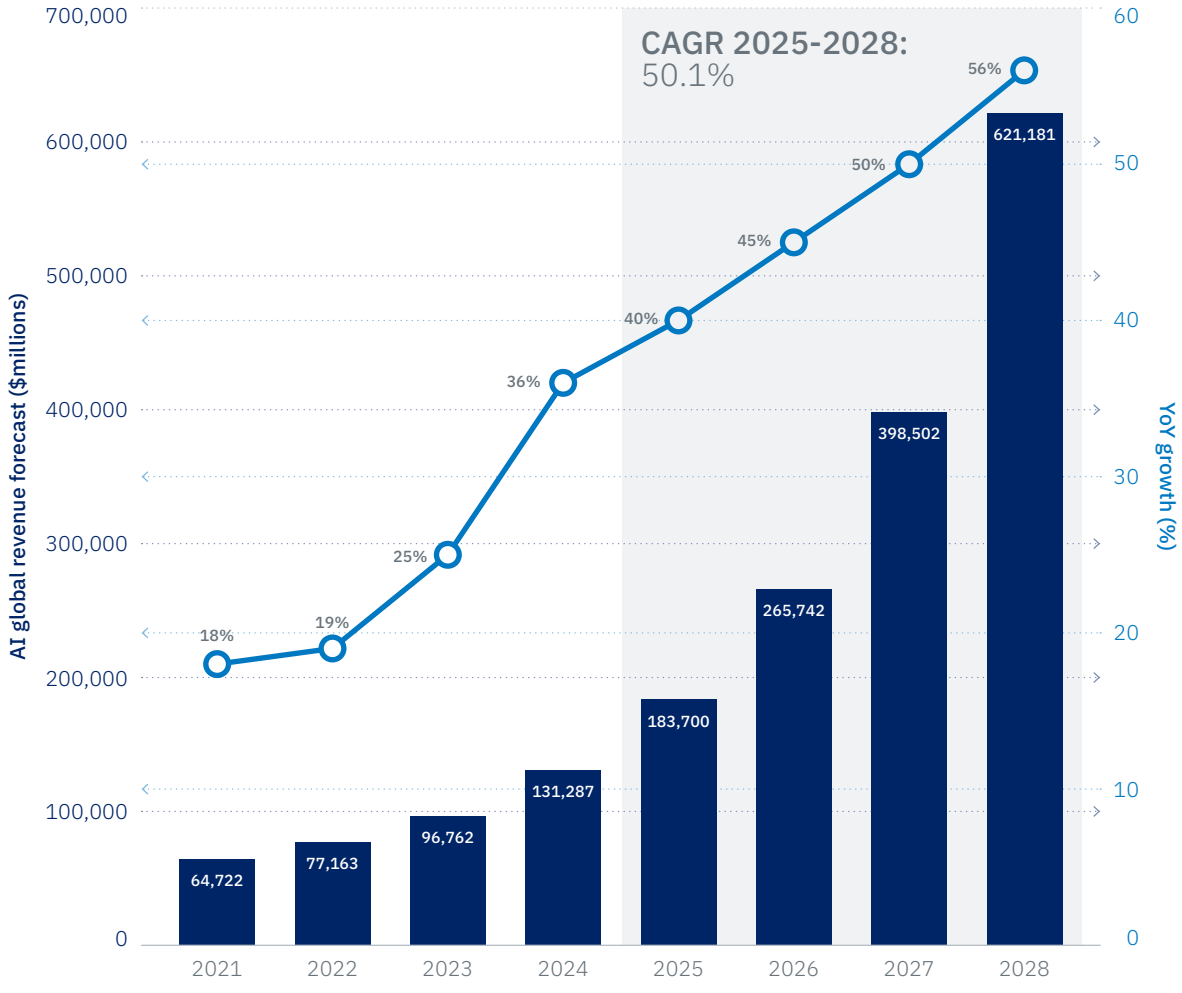
While cloud computing is a popular solution for many AI projects, building a storage infrastructure on-premises, if properly done can often be less expensive at scale for data that has to be kept and reused over time. The cost of renting GPU resources for a long time in the cloud can become expensive and unpredictable. For this reason, organizations that invest heavily in ML/AI will probably find they are more in control of their budget with an on-premises architecture.

GlobalData forecasts enterprise AI revenue (including AI consulting and support services; AI hardware; AI platforms; and specialized AI applications) to grow consistently in the mid double-digits every year during the period from 2025 to 2028.

As AI matures, companies across a wider section of vertical industries will increasingly deploy production applications. Late adopters will be those in more "conservative" sectors like finance, healthcare and other industries with strict compliance, sovereignty and privacy requirements - which will likely lead to greater use of on-premises environments over time. Estimating the TCO of an on-premises environment compared to a cloud environment involves careful consideration of each organization's specific AI workloads, compliance and sovereignty requirements, and strategic priorities. While digital transformation has seen enterprises migrating a significant part of their workloads to the cloud, many enterprises still prefer to host their critical workloads and applications on-premises. In the last few years, many organizations have moderated their "cloud first" strategy. This has led to repatriating certain workloads back on-premises.

Figure 2

AI global revenue forecast ($millions); YoY growth (%)



Source: GlobalData

Cloud production AI – the upside

Consider the overall positives for running production AI in the cloud, the first and most obvious for production AI workloads in the cloud is scalability, sometimes referred to as elasticity. Resources can be scaled up as needed – both base cloud workload resources such as CPU and storage, but also GPU resources which are necessary for AI. In the event of an AI workload that experiences 'lumpy' or uneven resource consumption, the ability to scale up is an important feature. The elasticity to scale back down as AI workload demand lowers can save considerable money on costs.

Clouds also have very fast lead times – there is rarely a wait, whereas in on-premises there must be the cycle of ordering, receiving, installation, and configuring equipment. Cloud providers have what is needed ready to be tapped and configured in many use cases. Nor with cloud do enterprises have to muster a large amount of capital at the beginning of the project. Faster initial setup means a faster return on investment (ROI) and faster time to workload productivity.

From a managed services standpoint, cloud providers have strong options, such as ready-to-use models, API services, and other AI-related ecosystem services. Cloud providers often have pre-built frameworks, that when combined with API services can speed development time. Cloud providers also have strong characteristics when it comes to operations – cloud providers have long experience in operating complex data center environments and the budget to keep those critical operational teams fully staffed.

Clouds also offer a way to keep pace with technology, upgrading often and early. For companies that exist on the bleeding edge, that is a boon. The upgrade cycles are much longer in on-premises AI data centers. But considering most innovation today is around the software itself; this is less of an issue for an ongoing production environment.

How an enterprise handles the costs around its production AI installation matters as well. Cloud offers the convenience of pay-as-you-go services, accounted for as expenses. For many companies, this is a model they are used to and can manage, as opposed to capital depreciation.

Cloud production AI – the downside

Locating enterprise AI production workloads in the cloud has some considerable downside as well, especially considering issues, such as regulations around privacy, legal/ professional compliance, data sovereignty, and geopolitical considerations.

The first issue is legal/regulatory. Different countries have different rules when it comes to data sovereignty, that is where the data must reside. Privacy laws, both protecting information and the right to have personal data deleted also vary a great deal between national jurisdictions. The situation around data sovereignty and privacy laws is fluid – geopolitical tensions and changes in international trade are causing a surge in these kinds of regulations and previous international agreements are inadequate. Even within nations, there are legal issues – municipalities and provinces/departments/ states, can all have variations on data residency and data privacy, especially when overarching federal government bodies fail to standardize regulations. Patchwork legal requirements for far-flung cloud resources complicate running production AI workloads.

In truth, cloud computing remains complex, and increasingly complicated offerings are leading to a situation where enterprises are losing control and grasp of the intricacies inherent to the technology.

For example, pay-per-use or subscription-based billing models in cloud services can make it challenging to predict and manage total costs. The dynamic nature of resource usage and the need for ongoing management and optimization contribute to the uncertainty in total cost of ownership.

In a GlobalData survey of 357 technology professionals undertaken in the second quarter of 2025, only 42% of respondents said they fully understood cloud computing. This represented a significant decrease of eight percentage points from 50% in the same survey the previous quarter:

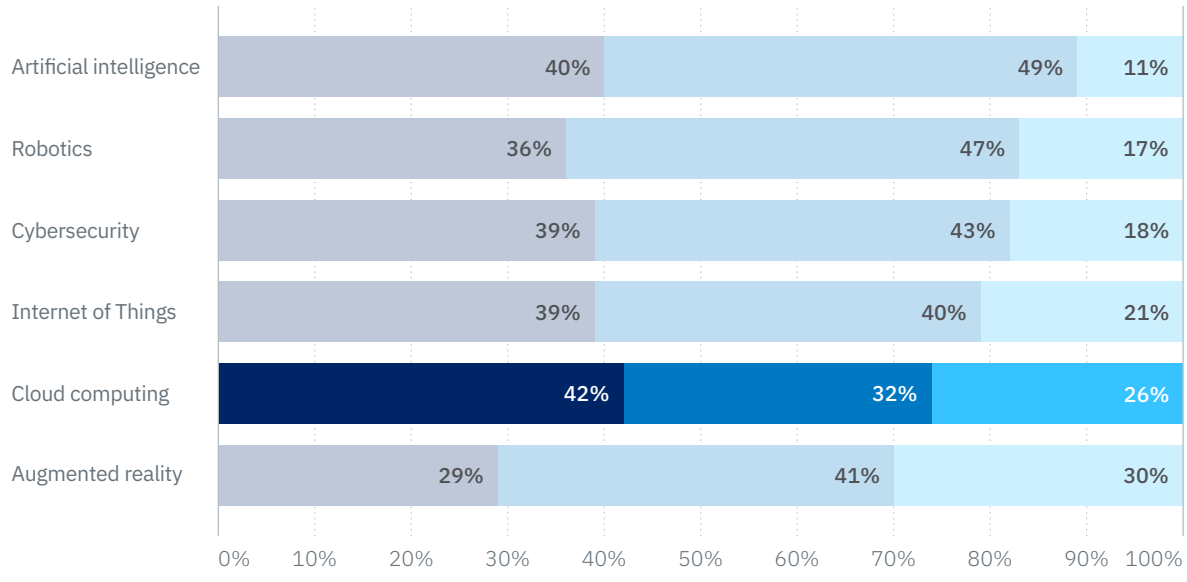
Figure 3

Technology understanding survey, Q2 2025

Key:
- I fully understand this technology
- I partially understand this technology
- I don't know anything about this technology

Question 4: Which statement best describes your understanding of each of the technologies listed?

Source:
GlobalData
N=357 *(poll participants are professionals from more than 30 industries)*



Industry requirements

Specific industries have their own regulations and requirements. Healthcare, finance, and the government sector all have laws, regulations, and standard practices that must be complied with, usually with regular audits and reports. In more common applications, there are mechanisms to ensure compliance in the cloud, but with production AI being as new as it is, these same mechanisms are often not available or immature.

Data gravity

Then comes the matter of data and the invisible pull it puts on application location, often called data gravity. The concept of data gravity means that in cases where the data is large, cumbersome, and expensive to move, applications are more likely to move to where the data is gathered, rather than data moving to where the application is. This principle has a profound effect when it comes to deciding to go with a cloud or on-premises production AI system. The amount of data that AI workloads use, particularly for training, is huge, often many petabytes. If an AI's data store does not exist in the cloud, moving it up to the cloud would be not only expensive, but may simply be impractical to do.

In cases where the data is created and collected in the field, moving a constant stream of data from the field locations to the cloud also creates difficulties. Edge AI data centers at or near the data collection point serve that purpose. The principle of data gravity, combined with the costs and difficulty moving that data means it is very difficult for an enterprise to move production locations.

In more complex use cases, the location of multiple corporate data lakes needs to be proximate to the AI frameworks and GPU hardware to maintain a reasonable pipeline for production and ongoing development.

Lock-in

Many cloud providers use proprietary, in-house APIs and storage formats. This is a long-understood problem that makes it difficult for an enterprise to move to another cloud environment or to an on-premises data center for workloads, and it applies to production AI workloads as well. The dangers and sunk costs into the current cloud provider can make it prohibitively expensive and/or risky to change providers or move on-premises. Further, egress fees for data, software re-architecture, and of course existing contractual agreements all factor in.

Similarly, if the data is in the cloud, the gravity of that data makes it difficult to switch away from a particular cloud provider. In on-premises production AI facilities, the enterprise can have greater flexibility and minimize the possibility of long-term functional lock-in.

Utilization rates in on-premises environments are an essential component of TCO estimates. The higher the utilization rate, the more organizations are "sweating their assets" and the better ROI and TCO of on-premises environments compared to cloud. The breakeven point marks the time after which on-premises infrastructure typically becomes more cost-effective than cloud, a point after the initial investment, past which the ROI will increasingly grow after a certain period. With consistent high utilization, organizations should be able to achieve much lower costs over time. Of course, the specific needs of the AI workload(s) have a huge impact on how quickly that breakeven point is reached. Meanwhile, the complexity associated with cloud environments means there is a high degree of unrealized value. Strategic consultancy McKinsey estimates that with cloud environments, breakeven generally comes at around 50% of cloud adoption, a factor which, together with unrealized use cases and redundant cloud foundations can erase the benefits cloud architectures provide or even eliminate value altogether.[1]

1 McKinsey & Company, In search of cloud value: Can generative AI transform cloud ROI?, by Chhavi Arora, Will Forrest, Mark Gu, and James Kaplan, November 15, 2023.

Complexity, security and the cloud

One of the bigger issues with cloud-based infrastructure is the fact that it is largely shared. Cloud providers have done a good job at keeping customer data separate, but they are also one of the biggest targets for exploitation. For companies that have highly sensitive data and intellectual property, the risk of unauthorized data exfiltration is higher in the cloud. Several cloud-specific risk factors apply. First, the threat of insiders at the cloud provider, enterprises have no visibility or control over who works on the underlying infrastructure or software systems at the cloud provider. Next is the risk of misconfiguration on the part of the cloud provider – not a common issue, but it can happen.

The model of security for cloud providers is also problematic. Not in the sense that cloud providers are careless or lacking security knowledge. It's the split of responsibilities between the cloud provider and the enterprise customer. The cloud provider secures the infrastructure itself, but the enterprise customer is the one tasked with the responsibility of securing their own data, access controls, and applications. This can leave gaps that are not obvious, despite well-documented best practices created by the cloud provider.

Lastly, when it comes to long term ROI, cloud-based production AI systems cost more in the long run and are less flexible. Further, the nature of cloud billing and usage-based billing brings the inability to accurately predict expenses. This complicates budgeting and future planning. Cloud also prevents customers from choosing to squeeze more out of their investments by delaying upgrades and maximizing existing infrastructure investments.

Disaster recovery and business continuity

Establishing effective business continuity and disaster recovery strategies is essential when deploying live AI workloads. But when it comes to cloud computing environments, there can be hidden costs involved. Disaster recovery, whether on-premises or in the cloud, entails having a "back up" zone, a replicated environment that can be accessed when disaster strikes, ensuring that operations can continue without disruption, and critical systems and data are not lost. But the nature of cloud environments means that when disaster strikes, the recovery process entails transferring data across cloud regions. The complexity of multi-region synchronization means that disaster recovery and business continuity operations can become a cumbersome affair from a financial standpoint. Organizations often underestimate costs.

When it comes to live AI workloads in production environments, enterprises should bear in mind that in all likelihood there will be unexpected expenses beyond the initial setup, in the form of compute (overprovisioned resources, idle compute), inefficiencies in cross-region data transfer, storage and management (data replication), and all manner of operational complexities. These can sometimes make up a large part of the spending.

Production environments and the behavior of AI workloads at scale can be onerous to measure and there is a high degree of unpredictability. Moreover, when it comes to cloud computing, among the large cloud computing providers, each presents different pricing structures that make it even more complicated to predict and control spending patterns.

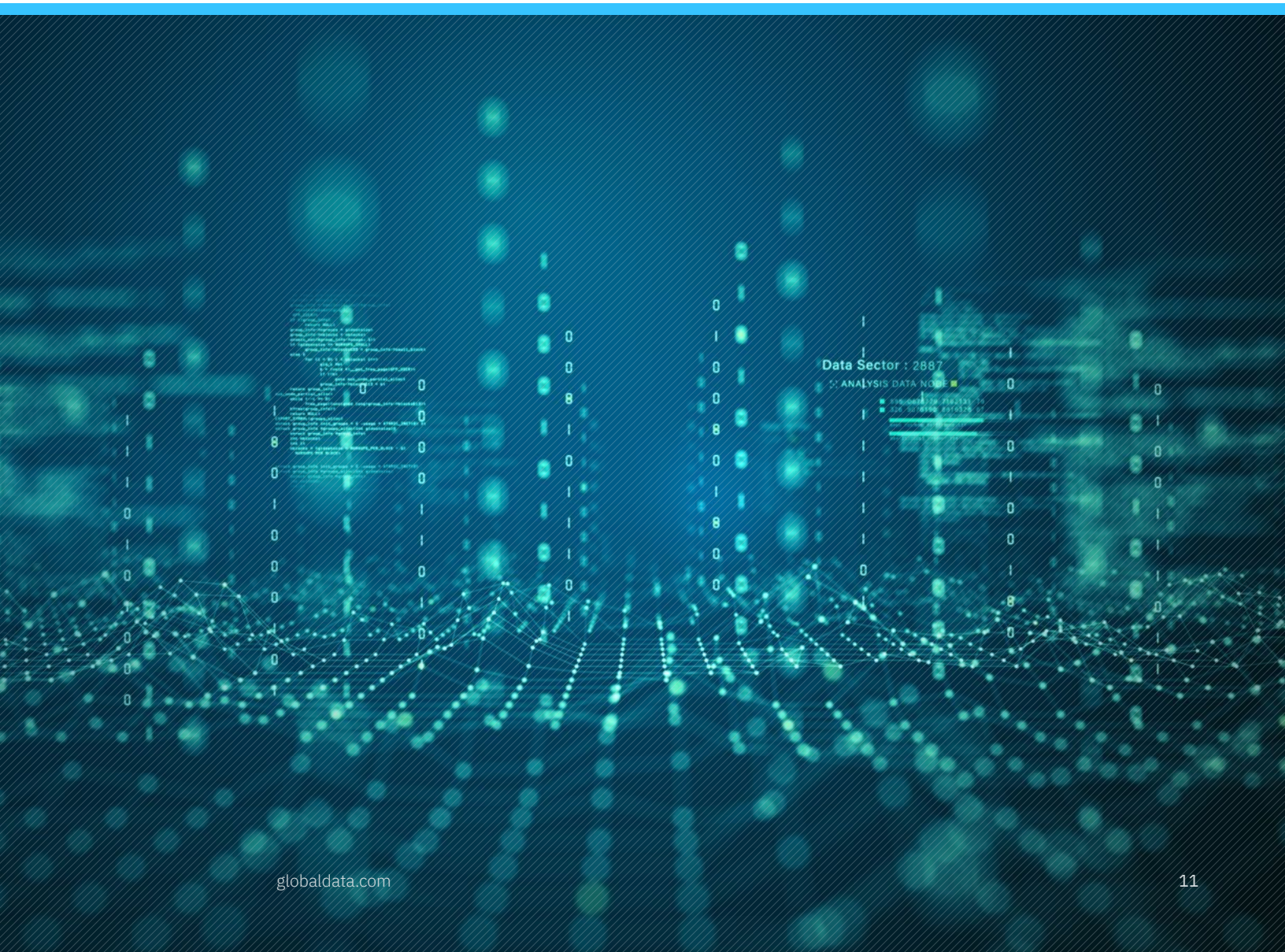
In the cloud, the existence of different availability zones throws a big wrench in the works. And AI often presents unique challenges. Data replication is one of the key hurdles. AI workloads, especially when using large AI models, involve continuous data replication across regions or availability zones to ensure resilience. But pricing can be confusing in this area; some providers may have more explicit

transfer fees, which makes them appear as the more expensive option, while others may present intra-region data replication as free, but make customers pay in other ways such as premium pricing for zone-redundant SKUs.

Moreover, cloud providers charge for both the storage of replicated datasets and the network traffic required to keep secondary sites in sync. The high volumes of AI training data, which can reach petabytes in size, can incur significant monthly transfer fees.

Accelerator usage also poses challenges. AI requires the utilization of specialized semiconductors called GPUs (graphic processing units), to process the massive amounts of compute required to run high-end AI applications. GPU standby overhead means that oftentimes idle GPUs still incur full compute cost. Overprovisioning for GPU availability can incur exaggerated costs.

Alongside the traditional Hyperscalers, so-called Neoclouds are cloud computing companies specializing in providing GPUs as a service to enterprises that require powerful accelerators for their high-end workloads. These companies, which are somehow a novelty in the market, present a much simpler pricing structure than their hyperscale competitors, because they offer “bare-bones” compute/ GPUs, without the myriad tools, marketplaces, libraries and repositories available in hyperscale platforms. Competing on price, they can appear as an attractive alternative for enterprises eager to find the coveted hardware necessary to process AI workloads. However, data movement and storage still represent the kind of expenses enterprises don’t budget for.



AI workloads on-premises

There are a lot of reasons to consider an on-premises solution for production AI workloads. With years of operational improvements to private cloud environments, the barrier of operational complexity has faded. On-premises data center environments, even heavy, high-performance environments like those needed for AI now have cloud-style interfaces, with orchestration and monitoring capabilities.

But the overriding reason to run AI workloads in an on-premises environment is both simple and profound. That reason is control. Pre-validated, modular infrastructure, available from most major vendors, make it much simpler to deploy. This kind of AI infrastructure can be utilized either in their own data center or in a co-location facility. Enterprises gain the control to make their own choices, rather than having choices made for them by a cloud or specialty-service provider Let's examine other factors.

Sovereignty, privacy, and data security

One of the biggest issues for enterprises today is dealing with the data itself – regulations on privacy, data security, and especially sovereignty is constantly evolving. On-premises hosting of AI workloads prevents accidental access by 3rd parties on shared infrastructure. This is particularly important for a number of verticals, including government, healthcare, finance, and manufacturing. The first and foremost of the regulations that effect all verticals is the General Data Protection Regulation (GDPR) that is designed to protect the personal data and privacy of individuals by giving them greater control of their own data and data about them. Despite this regulation being law in the EU since 2018, this regulation still represents an ongoing challenge. Regulations similar to GDPR are on the rise across the world. In the United States, there is not a country-wide privacy regulation which has led to individual states passing their own privacy laws, adding to the patchwork of regulations enterprises must follow. At the state level, the California Privacy Rights Act (CPRA), also known as Proposition 24, recently expanded the California Consumer Privacy Act (CCPA). The provisions of the CPRA went into effect in 2023 to enhance privacy rights for California residents and introduce new obligations for businesses that handle their personal information.

Data privacy can be a stumbling block for the deployment of AI workloads in production environments. Data is the blood of AI, but sensitive data can be most at risk when transferred to the public cloud. This is stopping many enterprises from going live with AI applications because they are fearful, particularly in industries with strict compliance requirements, of infringements on data privacy. In the EU, healthcare companies particularly in sectors like pharma must follow regulations to secure private data, and moving data to the public cloud becomes an almost insurmountable challenge with very long implementation cycles. Adherence to GxP guidelines is crucial for pharmaceutical companies, as it is often a legal requirement and is seen as a fundamental aspect of patient safety.

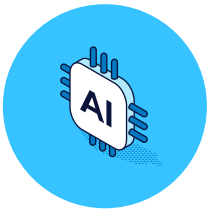
With AI, enterprises need to be more careful than ever how privacy data is protected, where it is stored, and how it is processed in conjunction with AI. There is already further regulation specific to AI when it comes to personal data, with the AI Act (2024) in the EU as well as more pending legislation in the U.S., Brazil, India, China and the U.K. On-premises AI workloads are easier to regulate and audit – the cloud ecosystem does not come into it. Security

solutions are also critical, and it is critical to ensure that a production AI data center has the latest, fully integrated and full-stack solution for AI. With many security platform vendors, this security solution can be shared across the general IT data center and the production AI facility, for operational and cost savings.

Many of the privacy laws also have aspects of data sovereignty in them as well, with restrictions on how and where data must be stored, and if it can be passed across national borders. The legal landscape around data sovereignty is already complicated and increases in geopolitical tensions, and nationalistic movements promise to add to the chaos. Data sovereignty is an issue that must be dealt with – on-premises data centers for AI workloads help companies keep in compliance – they have all control over data distribution.

Lastly, when it comes to data protection, the on-premises option is also superior to cloud. The specifics of data protection at rest, data protection in flight, and of course the scope of an AI's ability to access data can be specified and tailored to the exact needs of the vertical market or specialized enterprise need.

Additionally, while it hasn't reached full volume as of yet, the problem of post-quantum computing is waiting for enterprises. When quantum computers reach a certain level of computing power, these systems can be used to break today's encryption. Post-quantum encryption needs to be implemented, and that means every system and every device. Enterprises who control their own IT systems will be much better off when it comes to managing the risk. While there are no quantum computers yet able to break today's encryption, the danger is still there – the practice of capture and store data in flight for later decryption is happening now. Enterprises can mitigate this by ensuring that encrypted data stays on their own networks in their own data centers, which is much more difficult in the cloud environment.



Getting it just right – tailored environments

Enterprises use off the shelf style solutions all the time. These solutions can be characterized as mass produced and generally one-size-fits-all solutions that are designed for the widest appeal and broadest use case. That's fine for many things, but something as critical to the future of an enterprise such as its AI implementation, it's important to have a solution that is designed for the exact need. The time and expense of creating a comprehensive AI solution for the enterprise can be maximized by ensuring that the infrastructure it runs on is tuned for the task. With cloud-based installations, options are limited to what is offered in the menu with no guarantee those same options will continue to be offered. In contrast, modern on-premises AI infrastructure is often built on modular 'scale units' that allow organizations to start with a specific GPU density and expand incrementally. This building-block approach, based on validated designs, ensures predictable performance and allows the infrastructure to be tailored to the exact needs of a workload—whether for training, fine-tuning, or high-throughput inferencing.

A good example of this would be around latency. AI requires considerable network bandwidth and the less latency introduced into the equation, the fast the AI task. On-premises AI provide that with more certainty than other more cloudy methods. There is also the consideration of access to storage, which also needs to be low latency. On-site storage brings the data closer to the point of work, prevents multiple copies from being created, and makes it easier secure along with the rest of the AI production infrastructure.

This brings up the savings that can be had by not duplicating infrastructure with on-premises production AI data centers. Cloud-based systems often have duplicative qualities due to their nature. Often cloud-based production AI creates duplicative data – instances of data for training, instances for inference, instances for backup, for example. This can increase egress costs as well as security risks. Workloads in the cloud are often specific, each requiring their own VM or container. Security such as firewalls proliferate per region

or cloud account. Centralized monitoring is more difficult – cloud-specific monitoring tools tend to specialize in a single cloud service provider. Often this is completely different software than the standard enterprise centralized monitoring/ observability solution.

While it's hard to quantify without a specific use case, on-premises solutions for AI workloads can save money and operational complexity by sharing and standardization. This includes typical enterprise solutions such as the enterprise security stack, the enterprise monitoring/ observability solution, network policy, and more efficient use of VMs and containers.

People and price

The use of on-premises data centers for production AI workloads also comes with new opportunities for both the enterprise and the IT infrastructure staff. There is a misconception that the industry-wide shortage of trained, experienced AI engineers, architects, and developers means that there is a similar lack of people to run production AI infrastructure. The reality is that in a practical sense, the infrastructure used for AI is operationally similar to existing systems. Of course, there are some exceptions such as specialized GPU clusters, and the specialized back-end linkages such as NV-Link by NVIDIA. But those systems are designed and sold in optimized configurations – changes on that end of the AI hardware chain are generally handed by AI architects and developers.

The opportunity is to provide a training path and career path for existing IT infrastructure staff, helping with retention and allowing enterprises to hire more junior infrastructure staff for existing IT infrastructure. This doesn't mean that there will be no training requirements – of course a new system to run production AI will require training – but that training would be necessary in the cloud context as well.

The real question here is this – can the enterprise really save money – hard dollars – by putting their production AI in an on-premises data center? The answer is yes. But it's important to understand exactly how and where that money is saved. With the on-premises approach for production AI workloads, the hard savings occur farther down the line, as the relentless cloud fees pile up. Ongoing operational costs for the on-premises data center are less than that of a similar production environment in the cloud. For many use cases, after the first two-three years the on-premises data center can offer savings of anywhere between 10% to 30% over cloud. That's not considering any of the soft dollar and risk management benefits of on-premises installations.

Ground downside

There are a few downsides to an on-premises production AI facility. Of course, the first one would be in up-front costs and longer lead times. On-premises AI solutions require a large investment up front – solutions like that are very high performance. For some enterprises, the initial costs may be a barrier to entry, but most vendors selling these solutions have in-house financing options. Further, there are financial firms that can also handle the difficulties of the up-front costs. Financing does mean a longer ROI, but on-premises still represent considerable cost savings vs cloud.

Planning for an on-premises solution also has longer lead times. In cloud solutions, only available resources are shown, and cloud providers are constantly increasing capacity to serve more customers. Most major AI infrastructure vendors offer pre-validated or turnkey solutions that take a lot of the effort out of planning on-premises hardware. Plus, these pre-validated solutions help to prevent over-provisions. Cloud is the original environment of 'move fast and break things' and it's very easy to over-provision, provision too soon, or provision resources and simply not use them. Parsing cloud bills is difficult at best.

With an on-premises environment, on-demand spin ups for new AI applications are generally not available, as it is in the cloud environment. For enterprises that want that level of convenience, several vendors offer the ability to over-provision

an environment and simply pay for capacity when its needed via their lifecycle application. However, with the more stringent planning requirements for on-premises environments, the need to do emergency spin-ups should be considerably less.

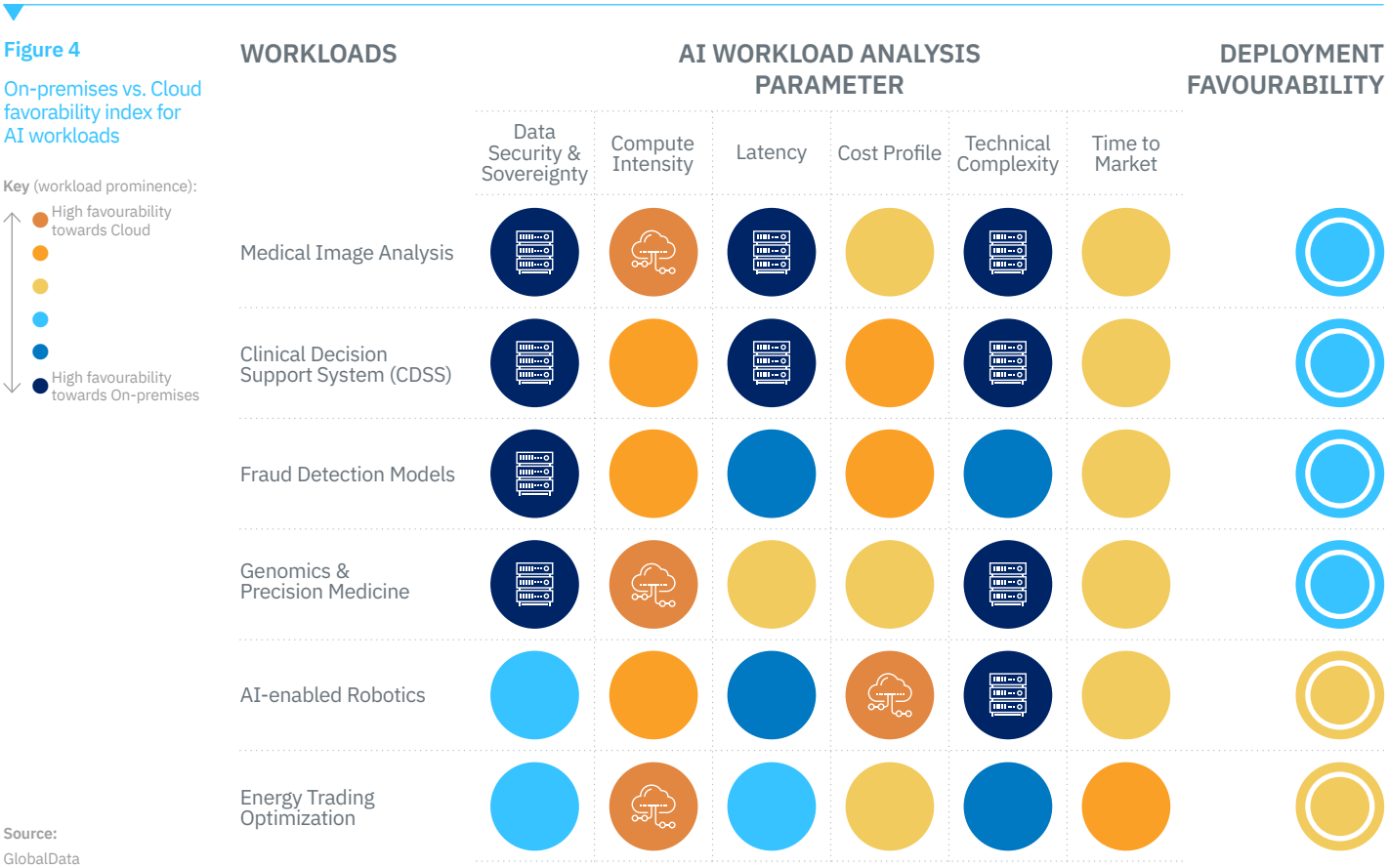
While this varies considerably by country, region, and vertical, there are some enterprises that prefer to claim IT costs as expenses rather than as capital. For companies like that, again financing is a good option – vendors and financial firms have offerings to allow enterprises to lease their IT equipment, converting the costs to an expense basis. There are also some vendors that offer on-premises 'subscriptions' that are accounted for as expenses.

Lastly, one of the biggest stumbling blocks to using on-premises for AI workloads is the impression that companies must build data centers. It is true that most existing data center space is inadequate for AI production workloads. The heat and power density of AI systems is high and climbing. Even now 30kw per rack is common, 60kw per rack will be the norm soon, and in a few years, per-rack power expenses could possibly grow past 200kw per rack. That level of power consumption and needed heat dissipation is simply not available in a data center that wasn't built for those specifications. Liquid cooling is becoming increasingly common.

However, building a data center isn't the only option. Co-location providers can give the benefits of an on-premises data center without needing to build. There are also programs in places such as the EU to supplement data center construction so that enterprises to co-locate with other regional enterprises. There is also the option to partner with other enterprises for a joint venture to ease the costs of construction and create a facility to be used by both enterprises or more. Often there are governmental programs to help with costs, from local, regional, or national governments. The point is that there are options – breaking ground is just one of them.

Figure 4

On-premises vs. Cloud favorability index for AI workloads



Source:
GlobalData

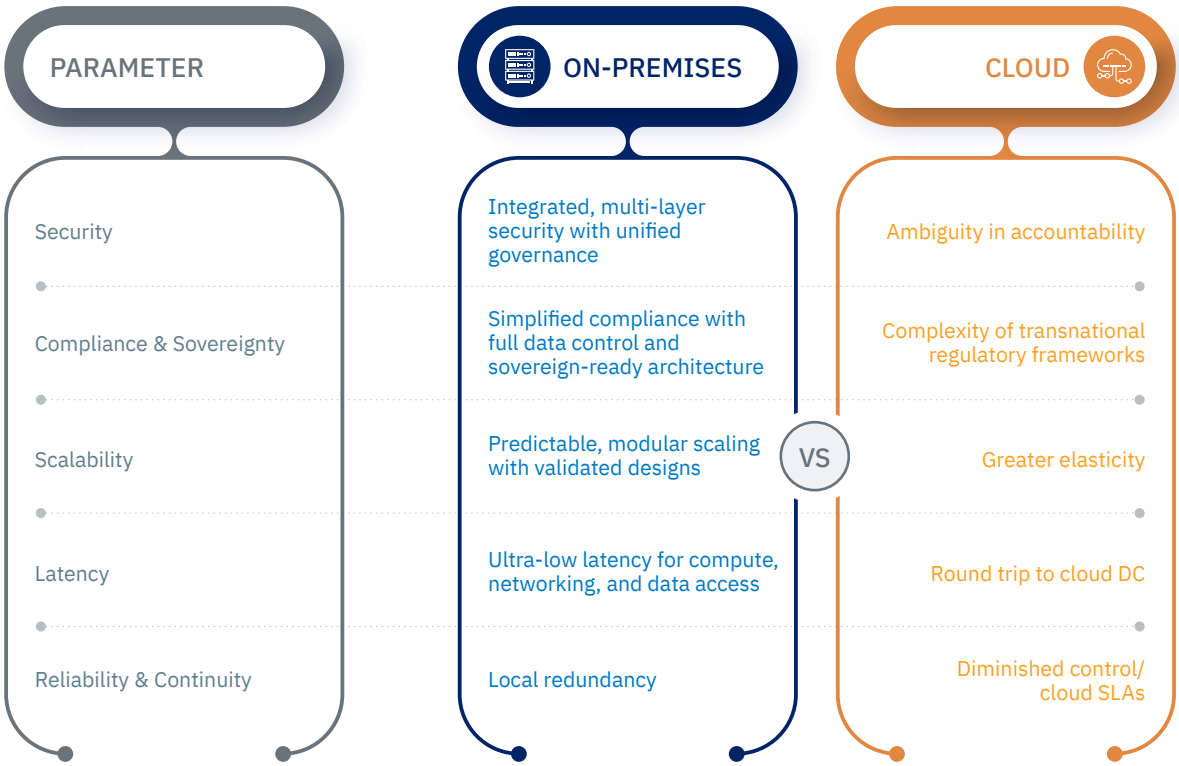
The bottom line

On-premises data centers for production AI loads are not a backwards-looking dream, but a real, viable option for many AI use cases. The reflex to assume that AI = Cloud is common, but there are real reasons. The primary reason is control – the ability to control the environment, the data in the environment, and keep a clear eye on both performance and costs while achieving cost synergies with existing IT systems for things such as monitoring/observability and security. Real control means better control and understanding of the risks. Risk can now be measured in known conditions, rather than simply the parameters handed out by the cloud provider. This also extends to personnel, enterprises can use their own exacting standards to hire, rather than just leaving that aspect to the cloud provider. It does mean taking more responsibility, both by IT and by corporate management – but the flexibility to make decisions that are optimal for the enterprise can lead to significant long-term savings.

Both the cloud and the on-premises data center are valid options. However, the emergence of turnkey, full-stack on-premises AI platforms means enterprises no longer have to choose between the control of on-prem and the agility of the cloud. Enterprises should remain flexible, choosing the right environment for the right workload, with the confidence that enterprise-grade AI can be deployed securely and efficiently wherever it is needed.

Figure 5

On-premises vs. Cloud – key parameters for AI workloads



Source:
GlobalData

Cisco secure AI factory with NVIDIA

Organizations need to show real business value from their AI investments but face many challenges. Setting up AI systems is complex because it involves combining AI development and delivery software, a Kubernetes platform, GPU accelerated compute, high performance networking, High performance storage, security, and observability tooling into one system. Making this process easier and more efficient helps reduce costs and risks, keeping organizations competitive in AI. Security is also essential. As AI grows more advanced, it's important to protect against risks like prompt injection, data leaks, hacking, and unauthorized access. Strong security builds trust in AI and allows teams to innovate with confidence.

However, to enable enterprises' AI success requires a modern, fully validated architecture – one that embeds security at all layers of the AI stack, including software and hardware, and automatically expands and adapts as the underlying infrastructure changes.

Cisco Secure AI Factory with NVIDIA meets these challenges by providing an integrated, high-performance, and secure AI infrastructure solution from Cisco, NVIDIA, and strategic partners to help accelerate the adoption of AI for enterprises. This solution provides a comprehensive, secure, and flexible AI infrastructure that helps enterprises build and secure data centers to develop and run AI workloads, dramatically simplifying the process for these businesses, while safeguarding critical data and workloads throughout the AI ecosystem.

This integrated solution combines proven AI infrastructure, inherent security features, and advanced AI software tools. By leveraging enterprise data and AI models as foundational inputs, the Cisco Secure AI Factory with NVIDIA empowers AI teams to operationalize AI and GenAI pipelines with agility and purpose. This acceleration in delivering AI tokens and applications fosters remarkable advancements in prediction, generation, and reasoning, enabling enterprises to unlock business value and realize their aspirations through the power of AI.

- Unlike other AI solutions, Cisco Secure AI Factory with NVIDIA embeds security at every layer of the stack to help securely develop and deliver AI tokens and applications.
- Features high-performance enterprise-proven networking, accelerated compute and storage, along with AI software helps accelerate the various phases of the AI/GenAI pipelines resulting in faster development and delivery of AI tokens and applications.
- Includes pre-validated deployment options help operationalize the solution. Enterprises have the flexibility to choose between a predefined and vertically integrated stack, or a more modular and flexible system with greater design choices.

Learn more here
https://www.cisco.com/site/us/en/solutions/artificial-intelligence/secure-ai-factory/index.html

GlobalData.

We are the trusted, gold standard intelligence provider to the world's largest industries

We have a proven track record in helping thousands of companies, government organizations, and industry professionals profit from faster, more informed decisions.

Our unique data-driven, human-led, and technology-powered approach creates the trusted, actionable, and forward-looking intelligence you need to predict the future and avoid blind-spots.

Leveraging our unique data, expert analysis, and innovative solutions, we give you access to unrivaled capabilities through one platform.

LONDON

John Carpenter House
7 Carmelite Street
London
EC4Y 0AN
UK

Tel: +44 20 7936 6400

NEW YORK

441 Lexington Avenue
2nd Floor
New York
NY 10017
USA

Tel: +1 646 395 5460

SYDNEY

56 Clarence Street
Level 11
New South Wales
Sydney 2000
Australia

Tel: +61 2 8073 7642

MEXICO CITY

Mote Pelvoux 111-2 Piso
Lomas de Chapultepec
Mexico D.F, 11000
Mexico

Tel: +52 55 5284 2945

GlobalDataPlc
GlobalDataPlc
GlobalData.com