Computacenter

Apple devices in an edge strategy

Cloud isn’t the only option for AI

The adoption of AI is surging because generative models have made the technology easy to use and widely accessible.

AI is now a strategic capability for organisations because it can help them transform productivity, customer engagement and the employee experience.

The public cloud underpins this shift by providing the scalable compute and data platforms needed to train and build, deploy and govern modern AI at large scale.

Running AI in the cloud isn’t the only option though. Bringing AI closer to the user through far edge and user edge deployments unlocks speed, resilience, privacy and cost efficiency that public cloud architectures can’t match.

Far edge and user edge deployments are better suited to many use cases, delivering faster, more secure performance that enables employees to do more work in less time, wherever they are.

Organisations need a clear understanding of the advantages and disadvantages of different AI deployments in order to build an AI strategy with well-defined use cases that match the needs of different domains of work.

Explore

Premium
Business Partner

The hierarchy of AI

AI inference functions can be run in a variety of places, each offering different performance considerations for users.

Public cloud AI

Offers access to massive compute, cutting edge models and integrated data that allows organisations to innovate and scale quickly. However, cost uncertainty, variable latency and unreliable connectivity pose issues for real-time workloads. In addition, there are regulatory and data sovereignty constraints as well as the risk of an organisation becoming dependent on a single cloud provider.

Near edge AI

Runs in a private data centre or on a private cloud managed by the organisation and accessible to users at different sites or remotely. Organisations gain a higher level of control over data and compliance than is possible with the public cloud, as well as predictable performance and the ability to tailor infrastructure to specific workloads. Capital investment, specialised skills and ongoing operational effort are all required.

Far edge AI

Runs on site and is accessible only to in-house users. This model offers fast, reliable, low latency processing and keeps sensitive information on site for stronger privacy and compliance. It also avoids dependence on external connectivity for mission critical operations. Dedicated hardware and skilled local operations are needed and a private cloud will lack the scalability of the public cloud.

User edge AI

Runs on the user's device, delivering ultra low latency inferencing for real-time needs, enabling work even without connectivity, protecting privacy by keeping data on the device and reducing cloud costs.

Run AI where it makes most sense

Any decision about where and how to deploy AI inferencing needs to reflect different workplace requirements. These fall into three domains:

- **HQ users** – knowledge workers who need tools that support focused work, collaboration and flexibility.
- **Frontline users** – a large group often underserved by technology; enabling them directly improves operational consistency and customer experience.
- **The intelligent edge** – places where people, devices and systems interact.

Five parameters for an AI strategy

AI deployment models differ in where intelligence runs, how much compute is available and how tightly they integrate with the cloud. An effective AI strategy must consider a variety of workplace needs based on use cases and their demands for latency, security, data sovereignty, resiliency and cost.

Far edge and user edge offer the fastest and most consistent latency for use cases demanding real-time performance. Public cloud and near edge AI typically suffer more variable latency better suited to non-real-time workloads with higher processing needs.

While far edge and user edge AI minimise external data exposure, they expand the attack surface because every device or site must be secured and updated individually. However, with modern management approaches, securing distributed devices is no more complex than managing end-user endpoints, and organisations can maintain consistent policy enforcement across their estate. Public cloud and near edge models offer stronger centralised controls, yet risk data exposure through insecure connections as data traverse external networks.

Data sovereignty depends on how the AI model handles data location, movement and jurisdiction. Far edge, user edge and near edge AI deployments keep data within a physical boundary to meet privacy and regulatory requirements by avoiding cross-border transfers.

Public cloud AI may require data to move across regions, increasing the need for precise governance.

Resiliency is affected by the way each AI model handles failure, redundancy and dependency. Far edge and user edge deployments avoid internet dependency and keep working through WAN outages, but every device or site becomes a point of failure, requiring strong IT support and possibly on site redundancy. Near edge and public cloud AI benefit from centralised infrastructure redundancy but rely on network connectivity.

Costs encompass capital expenditure, operational expenditure and utilisation efficiency. Far edge and user edge AI eliminate public cloud operational costs but require distributed hardware that must be managed and maintained. Near edge AI incurs significant investment and ongoing maintenance costs but avoids spiralling public cloud fees. Public cloud AI avoids Capex but requires strict controls to avoid runaway fees being incurred.

Today, AI systems contribute 2.5–3.7% of worldwide greenhouse gas emissions, already surpassing the aviation industry's 2% share and growing at

[1] AI Environment Statistics 2026: How AI Consumes 2% of Global Power and 17B Gallons of Water, AllAboutAI.com, 4 December 2025

While public cloud AI runs on energy efficient high utilisation infrastructure, near edge and far edge AI reduce network energy and use long lifecycle hardware.

User edge AI minimises data movement to reduce operational emissions and lowers energy use by relying on low power devices.

Tackle rising token and data transfer costs

One of the most important benefits of shifting to far edge and user edge AI deployments is the elimination of public cloud AI operational costs in the shape of token and data-transfer fees. These fees can create budgeting and cost control difficulties for many organisations because the costs are hard to predict.

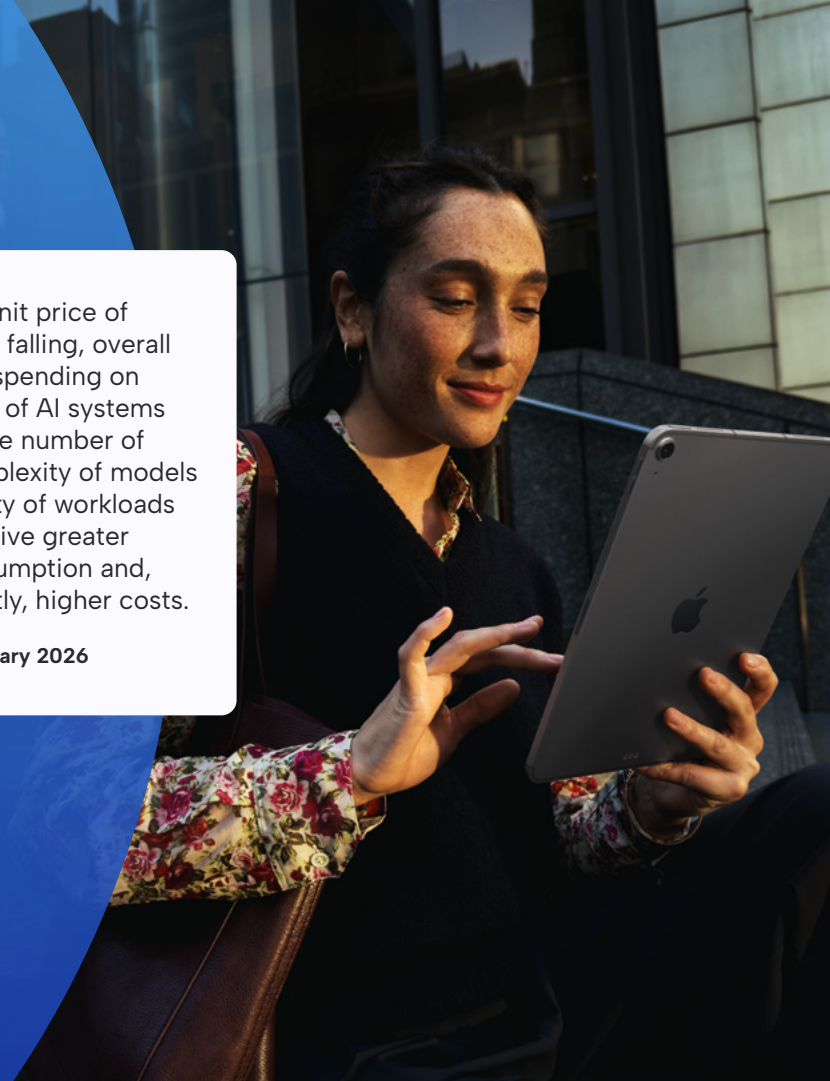
Token fees and data transfer fees are separate costs. Tokens are text building blocks that AI creates from a user's input and then uses to process queries and generate language-based responses. Public cloud AI models charge for these tokens. Data transfer fees are charges incurred when data leaves a cloud provider's network.

While token prices have **declined by up to 90% in recent years**[2], an organisations overall spend may rise because models are bigger, users become more comfortable embedding AI into their personal workflows, agentic workflows are more complex and multimodal inputs generate more tokens. If a product suddenly becomes popular, token consumption can explode and so can the bill.

Likewise, data transfer fees are rising mainly because AI systems move more data than traditional applications. Even if per GB pricing stays flat, the volume of data crossing cloud boundaries is increasing, so total data transfer can keep climbing.

[2] Agentic AI's Token Paradox: When Cheaper Means More Expensive, Forbes, 3 November 2025

> While the unit price of AI tokens is falling, overall enterprise spending on and scaling of AI systems is rising. The number of users, complexity of models and intensity of workloads will likely drive greater token consumption and, consequently, higher costs.
>
> **Deloitte, January 2026**



The pillars of an effective AI strategy

Identifying use cases is central to AI deployment strategy because it shows which models best meet the organisation's operational needs without costs spiralling out of control.

Public cloud AI is suited to use cases that are compute intensive or create predictable workloads, while near edge AI is best for regulated, unpredictable enterprise workloads, or applications that rely on internal systems and data stores. Far edge and user edge AI are increasingly able to offer distinct performance and security advantages for many use cases.

Far edge AI typically meets the needs of high volume, low latency workloads that exceed device capacity and is especially useful to support safety-critical applications in factory or healthcare settings unable to tolerate connectivity failures.

User edge AI offers the lowest latency for real-time, unpredictable workloads and meets the needs of use cases where data must stay local, or for field operations with unreliable connectivity. User edge AI also provides personalised experiences on PCs, laptops and mobiles.

Use cases best served by far edge and user edge AI

Far edge use cases

High capacity, real time, site level intelligence for:

- Factory automation and quality inspection using cameras and sensors for millisecond level decisions
- Retail analytics such as footfall tracking, loss prevention vision models and real time shelf monitoring
- Smart building and smart campus control including HVAC optimisation, energy management and security systems
- Healthcare environments where imaging, diagnostics, or patient monitoring data must be secured on site
- Logistics warehouses using robotic coordination, pallet scanning, routing optimisation, and vision based tracking

User edge use cases

Instant responses, strong privacy and the ability to run without a network for:

- Real time vision and audio processing such as on device transcription, barcode recognition, defect spotting
- Field service, frontline workers and emergency responders in remote or low connectivity environments
- Security and authentication including on device biometrics, anomaly detection, and continuous identity verification
- Live translation for better interactions in customer service centres, healthcare settings and field service situations
- Confidential data handling for legal, financial or medical professionals who cannot send documents to the cloud

Apple devices underpin a modern AI strategy

Apple devices have become an integral part of many organisations' IT estates, offering ease of use, stylish design, robust security and high performance that enhances employee productivity, creativity and connectivity.

Apple devices can also play a pivotal role in an organisation's intelligent edge strategy to deploy and scale up far edge and user edge AI. Devices from the iPhone to the iPad, as well as Mac mini and Mac Studio, provide powerful local compute, predictable cost and low latency performance without relying on cloud inference.

Mac mini is well suited for far edge and user edge AI, offering good local performance from a compact, quiet and cost effective package. The device efficiently runs individual or department level AI tasks, helping organisations avoid token fees and reduce data-transfer costs by keeping processing local.

Mac Studio provides a step up in capability, delivering higher throughput and memory for heavier, sustained workloads such as larger models, distributed inference, or media intensive pipelines. It functions well as a site level AI node that can support multiple users or continuous workloads without relying on cloud GPUs.

Distributed inference with Mac

Distributed inference involves running AI models across devices, edge nodes and cloud resources so workloads are processed according to latency, cost and security demands. By shifting away from using only public cloud AI, organisations can reduce data transfer and token spending, improve responsiveness and keep sensitive information local.

Mac mini and Mac Studio play an important role as compact, reliable edge compute nodes: the Mac mini handling lightweight, local inference for small teams, while the Mac Studio supports heavier, shared workloads across a site. This enables organisations to run routine inference close to the user while using cloud resources only when scale or model size requires it, benefiting from the most cost-effective AI deployments.

Computacenter lifecycle support

The challenge facing many organisations is how to ensure their AI strategy will efficiently and securely deliver its expected business value as edge devices, networks and local data storage proliferate.

Computacenter's portfolio of services and specialists focus on public cloud, private cloud, workplace IT and intelligent applications that address the challenges. We help organisations to define and deploy a modern AI strategy that meets their needs, from building a near edge AI data centre to deploying and managing Apple devices for individual users.

With Computacenter, organisations can scale up and manage their Mac technology as cost-effectively and efficiently as possible.

As an Apple Premium Business Partner and Apple Authorised Services Provider, Computacenter offers a full set of lifecycle services that span the selection, implementation, maintenance, repair, replacement and retirement of Apple devices.

Get in touch

To discover how Computacenter can help you achieve the most effective modern AI deployment, please contact your Computacenter Account Manager, call **01707 631000** or email **enquiries@computacenter.com**

Computacenter

Computacenter is a leading independent technology and services provider, trusted by large corporate and public sector organisations. We are a responsible business that believes in sustainable long-term value creation. We help our customers to source, transform and manage their technology infrastructure to deliver digital transformation, enabling people and their business. Computacenter plc is a public company quoted on the London Stock Exchange (CCC.L) and a member of the FTSE 250. Computacenter employs over 21,000 people worldwide.
www.computacenter.com