FT TECH FOR GROWTH FORUM

Why it is vital you understand the infrastructure behind AI

The FT Tech for Growth Forum is supported by

AVEVA

HCLTech

Glossary

- **Training**
  Teaching a model how to perform a given task.
- **The inference phase**
  the process by which an AI model can draw conclusions from new data based on the information used in its training.
- **Latency**
  The time delay between an AI model receiving an input and generating an output.
- **Edge computing**
  Processing on a local device. This reduces latency so is essential for systems that require a real-time response, such as autonomous cars, but it cannot deal with high-volume data processing.
- **Hyperscalers**
  Providers of huge data centre capacity such as Amazon's AWS, Microsoft's Azure, Google Cloud and Oracle Cloud. They offer off-site cloud services with everything from compute power and pre-built AI models through to storage and networking, either all together or on a modular basis.
- **AI compute**
  The hardware resources that run AI applications, algorithms and workloads, typically involving servers, CPUs, GPUs or other specialised chips.
- **Co-location**
  The use of data centres which rent space where businesses can keep their servers.
- **Data residency**
  The location where data is physically stored on a server.
- **Data sovereignty**
  The concept that data is subject to the laws and regulations of the land where it was gathered. Many countries have rules about how data is gathered, controlled, stored and accessed. Where the data resides is increasingly a factor if a country feels that its security or use might be at risk.

Foreword

Visit the editorial hub at
ft.com/tech-for-growth-forum

Businesses are putting a lot of time into deciding which artificial intelligence models to adopt — and they are spending a lot of money with Microsoft, OpenAI, Google, Anthropic and others.

Executives should aim to be as familiar as possible with the vast computing infrastructure behind AI as this will help them choose where to invest.

While some businesses will opt for the cheapest cloud offerings that support AI, others will need to think deeply about which computing infrastructure best supports their operations, especially if they are in highly regulated or sensitive sectors.

Concerns are growing about energy use in the vast data centres that handle huge AI workloads. Choosing a green solution will be paramount for companies with an ESG commitment. Others will want to ensure that they can access the top-of-the-range AI solutions that most suit them.

These decisions are likely to prove crucial to companies' future performances. This report is designed to act as a guide for those who must navigate the coming AI age.

We are always keen to hear more. You can reach us at forums@ft.com.

Murad Ahmed
Tech for Growth Forum Editor and Technology News Editor
Financial Times

Why it is vital you understand the infrastructure behind AI

Adoption of artificial intelligence is booming, but without the right infrastructure — from chips to cooling — most companies risk falling behind, writes *Lucy Colback*

As demand increases for AI solutions, the competition around the huge infrastructure required to run AI models is becoming ever more fierce. This affects the entire AI chain, from computing and storage capacity in data centres, through processing power in chips, to consideration of the energy needed to run and cool equipment.

When implementing an AI strategy, companies have to look at all these aspects to find the best fit for their needs. This is harder than it sounds. A business's decision on how to deploy AI is very different to choosing a static technology stack to be rolled out across an entire organisation in an identical way.

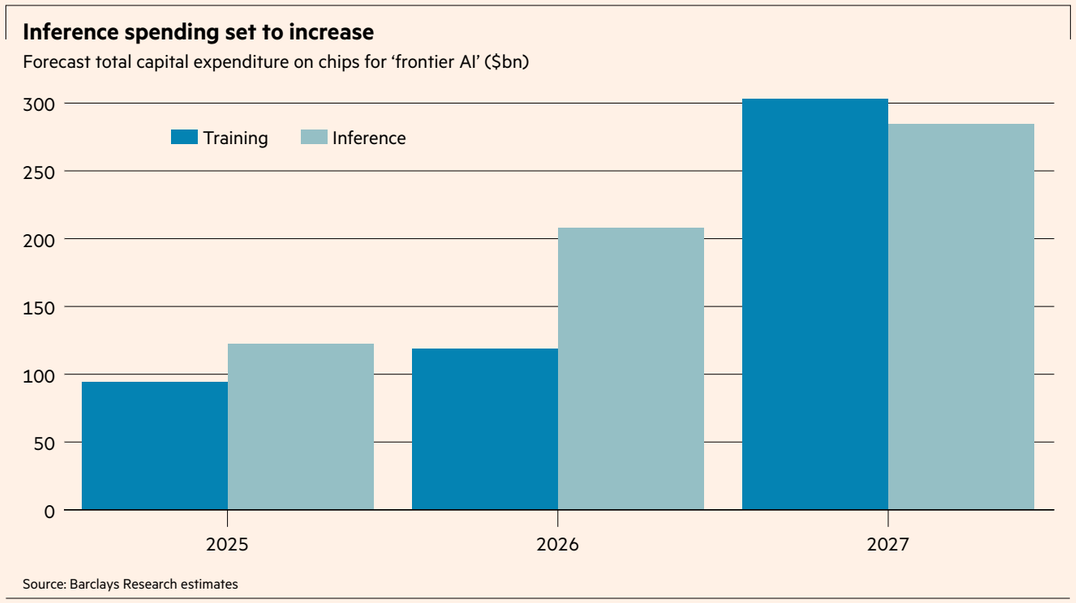
Businesses have yet to understand that a successful AI strategy is "no longer a tech decision made in a tech department about hardware", says Mackenzie Howe, co-founder of Atheni, an AI strategy consultant. As a result, she says, nearly three-quarters of AI rollouts do not give any return on investment.

Department heads unaccustomed to making tech decisions will have to learn to understand technology. "They are used to being told 'Here's your stack'," Howe says, but leaders now have to be more involved. They must know enough to make informed decisions.

While most businesses still formulate their strategies centrally, decisions on the specifics of AI have to be devolved as each department will have different needs and priorities. For instance legal teams will emphasise security and compliance but this may not be the main consideration for the marketing department.

"If they want to leverage AI properly — which means going after best-in-class tools and much more tailored approaches — best in class for one function looks like a different best in class for a different function," Howe says. Not only will the choice of AI application differ between departments and teams, but so might the hardware solution.

One phrase you might hear as you delve into artificial intelligence is "AI compute". This is a term for all the computational resources required for an AI system to perform its tasks. The AI compute required in a particular setting will depend on the complexity of the system and the amount of data being handled.



The decision flow: what are you trying to solve?

Although this report will focus on AI hardware decisions, companies should bear in mind the first rule of investing in a technology: identify the problem you need to solve first. Avoiding AI is no longer an option but simply adopting it because it is there will not transform a business.

Matt Dietz, the AI and security leader at Cisco, says his first question to clients is: what process and challenge are you trying to solve? "Instead of trying to implement AI for the sake of implementing AI…is there something that you are trying to drive efficiency in by using AI?," he says.

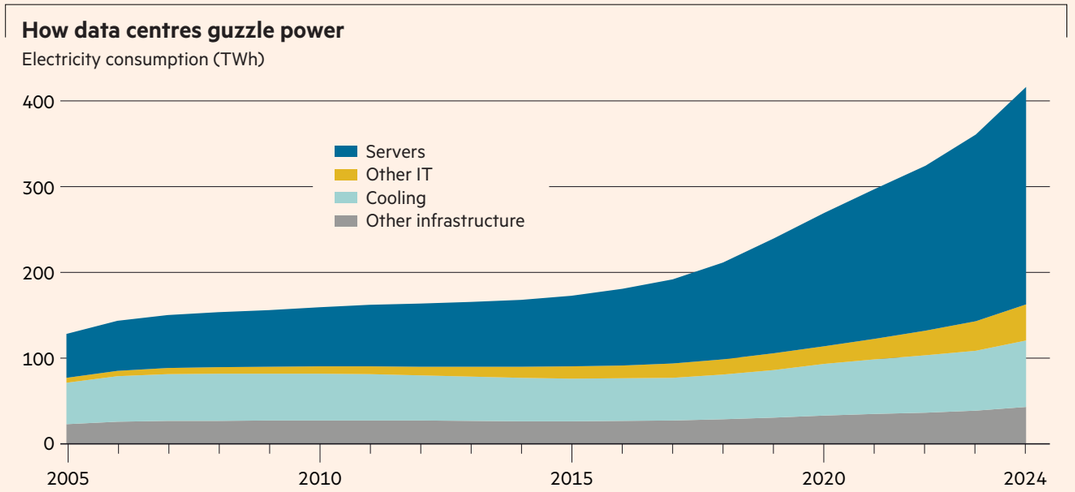
Companies must understand where AI will add the most value, Dietz says, whether that is enhancing customer interactions or making these feasible 24/7. Is the purpose to give staff access to AI co-pilots to simplify their jobs or is it to ensure consistent adherence to rules on compliance?

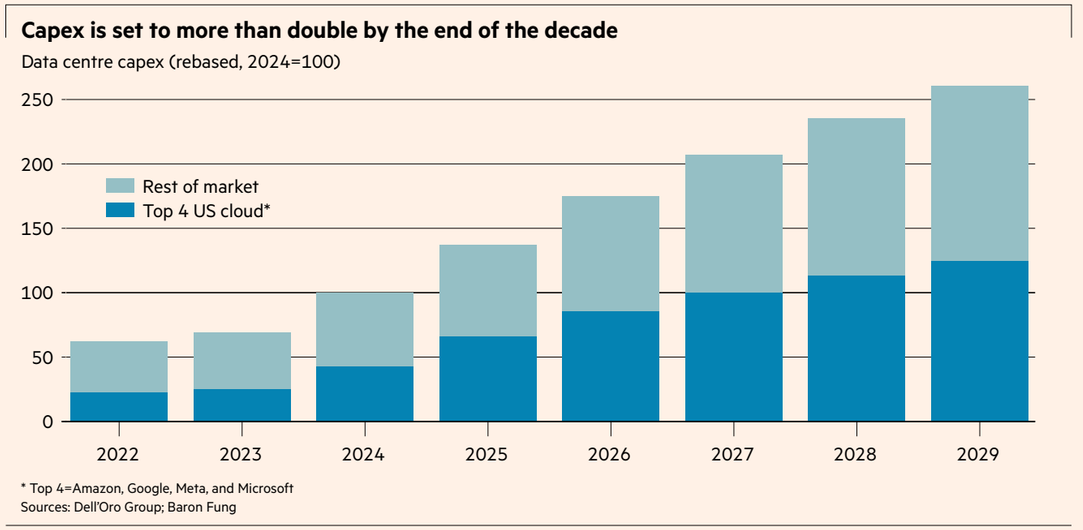
"When you identify an operational challenge you are trying to solve, it is easier to attach a return on investment to implementing AI," Dietz says. This is particularly important if you are trying to bring leadership on board and the initial investment seems high.

Companies must address further considerations. Understanding how much "AI compute" is required — in the initial phases as well as how demand might grow — will help with decisions on how and where to invest. "An individual leveraging a chatbot doesn't have much of a network performance effect. An entire department leveraging the chatbot actually does," Dietz says.

Infrastructure is therefore key: specifically having the right infrastructure for the problem you are trying to solve. "You can have an unbelievably intelligent AI model that does some really amazing things, but if the hardware and the infrastructure is not set up to support that then you are setting yourself up for failure," Dietz says.

He stresses that flexibility around providers, fungible hardware and capacity is important. Companies should "scale as the need grows" once the model and its efficiencies are proven.





The data server dilemma: which path to take?

When it comes to data servers and their locations, companies can choose between owning infrastructure on site, or leasing or owning it off site. Scale, flexibility and security are all considerations.

While on-premises data centres are more secure they can be costly both to set up and run, and not all data centres are optimised for AI. The technology must be scalable, with high-speed storage and low latency networking. The energy to run and cool the hardware should be as inexpensive as possible and ideally sourced from renewables, given the huge demand.

Space-constrained enterprises with distinct requirements tend to lease capacity from a co-location provider, whose data centre hosts servers belonging to different users. Customers either install their own servers or lease a "bare metal", a type of (dedicated) server, from the co-location centre. This option gives a company more control over performance and security and it is ideal for businesses that need custom AI hardware, for instance clusters of high-density graphics processing units (GPUs) as used in model training, deep learning or simulations.

Another possibility is to use prefabricated and pre-engineered modules, or modular data centres. These suit companies with remote facilities that need data stored close at hand or that otherwise do not have access to the resources for mainstream connection. This route can reduce latency and reliance on costly data transfers to centralised locations.

Given factors such as scalability and speed of deployment as well as the ability to equip new modules with the latest technology, modular data centres are increasingly relied upon by the cloud hyperscalers, such as Microsoft, Google and Amazon, to enable faster expansion. The modular market was valued at $30bn in 2024 and its value is expected to reach $81bn by 2031, according to a 2025 report by The Insight Partners.

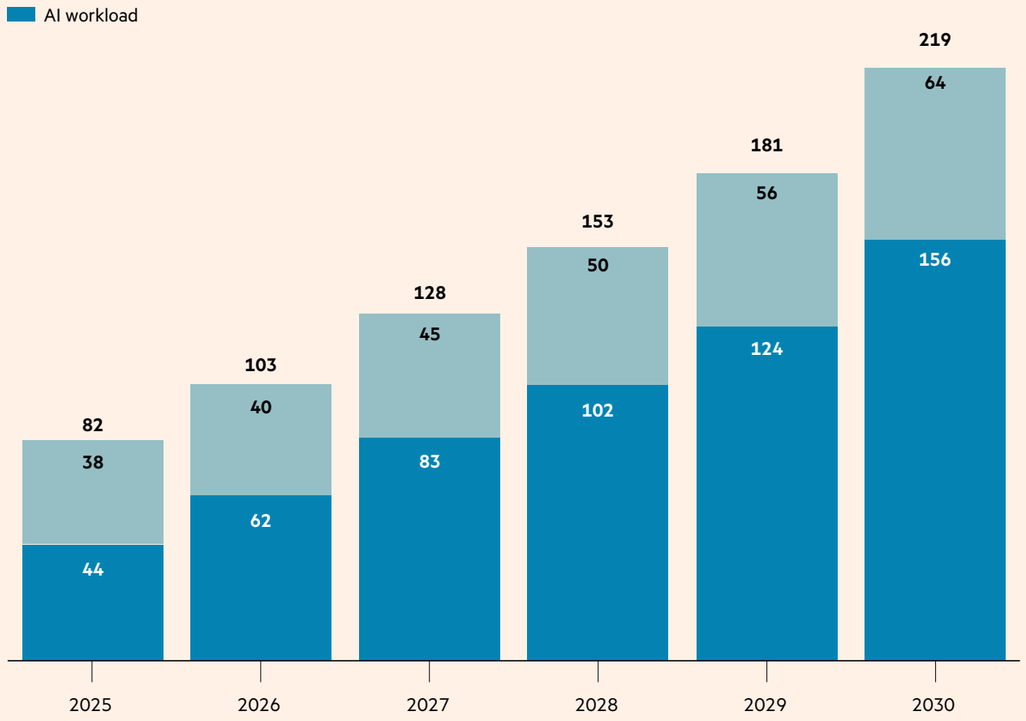
Modular data centres are only a segment of the larger market. Estimates for the value of data centres worldwide in 2025 range from $270bn to $386bn, with projections for compound annual growth rates of 10 per cent into the early 2030s when the market is projected to be worth more than $1tn.

Much of the demand is driven by the growth of AI and its higher resource requirements. McKinsey predicts that the demand for data centre capacity could more than triple by 2030, with AI accounting 70 per cent of that.

While the US has the most data centres, other countries are fast building their own. Cooler climates and plentiful renewable energy, as in Canada and northern Europe, can confer an advantage, but countries in the Middle East and south-east Asia increasingly see having data centres close by as a geopolitical necessity. Access to funding and research can also be a factor. Scotland is the latest emerging European data centre hub.

AI 'workloads' are likely to drive more need for computing infrastructure

Estimated global data centre capacity demand, 'continued momentum' scenario (gigawatts)



Figures may not sum to totals because of rounding
Sources: McKinsey Data Center Demand Model; Gartner reports; IDC reports; Nvidia capital market reports

Choose the cloud . . .

Companies that cannot afford or do not wish to invest in their own hardware can opt to use cloud services, which can be scaled more easily. These provide access to any part or all of the components necessary to deploy AI, from GPU clusters that execute vast numbers of calculations simultaneously, through to storage and networking.

While the hyperscalers grab the headlines because of their investments and size — they have some 40 per cent of the market — they are not the only option. Niche cloud operators can provide tailored solutions for AI workloads: CoreWeave and Lambda, for instance, specialise in AI and GPU cloud computing.

Companies may prefer smaller providers for a first foray into AI, not least because they can be easier to navigate while offering room to grow. Digital Ocean boasts of its simplicity while being optimised for developers; Kamatera offers cloud services run out of its own data centres in the US, Emea and Asia, with proximity to customers minimising latency; OVHcloud is strong in Europe, offering cloud and co-location services with an option for customers to be hosted exclusively in the EU.

Many of the smaller cloud companies do not have their own data centres and lease the infrastructure from larger groups. In effect this means that a customer is leasing from a leaser, which is worth bearing in mind in a world fighting for capacity. That said, such businesses may also be able to switch to newer data centre facilities. These could have the advantage of being built primarily for AI and designed to accommodate the technology's greater compute load and energy requirements.

. . . or plump for a hybrid solution

Another solution is to have a blend of proprietary equipment with cloud or virtual off-site services. These can be hosted by the same data centre provider, many of which offer ready-made hybrid services with hyperscalers or the option to mix and match different network and cloud providers.

For instance Equinix supports Amazon Web Services with a connection between on-premises networks and cloud services through AWS Direct Connect; the Equinix Fabric ecosystem provides a choice between cloud, networking, infrastructure and application providers; Digital Realty can connect clients to 500 cloud service providers, meaning its customers are not limited to using large players.

There are different approaches that apply to the hybrid route, too. Each has its advantages:

- Co-location with cloud hybrid. This can offer better connectivity between proprietary and third-party facilities with direct access to some larger cloud operators.
- On premises with cloud hybrid. This solution gives the owner more control with increased security, customisation options and compliance. If a company already has on-premises equipment it may be easier to integrate cloud services over time. Drawbacks can include latency problems or compatibility and network constraints when integrating cloud services. There is also the prohibitive cost of running a data centre in house.
- Off-site servers with cloud hybrid. This is a simple option for those who seek customisation and scale. With servers managed by the data centre provider, it requires less customer input but this comes with less control, including over security.

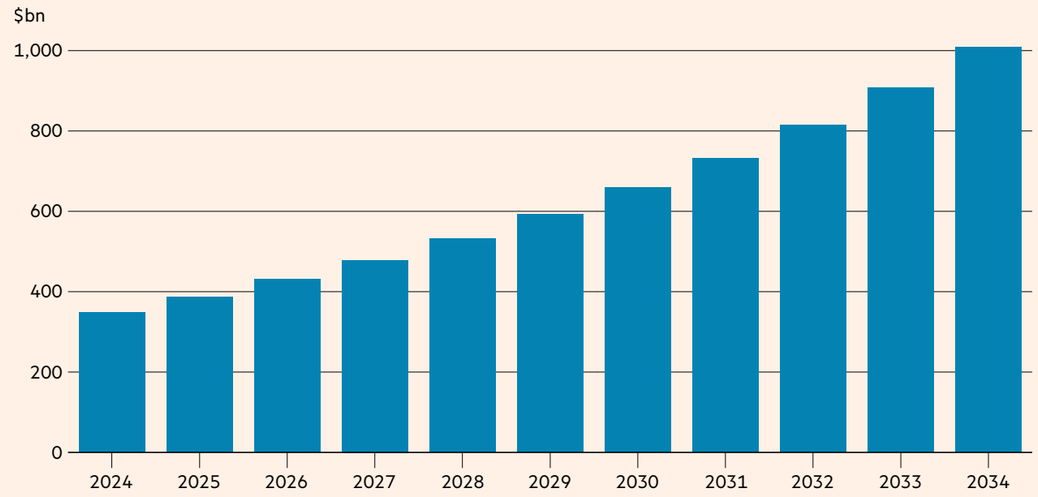
In all cases whenever a customer relies on a third party to handle some server needs, it gives them the advantage of being able to access innovations in data centre operations without a huge investment.

Arti Garg, the chief technologist at Aveva, points to the huge innovation happening in data centres. "It's significant and it is everything from power to cooling to early fault detection [and] error handling," she says.

Garg adds that a hybrid approach is especially helpful for facilities with limited compute capacity that rely on AI for critical operations, such as power generation. "They need to think how AI might be leveraged in fault detection [so] that if they lose connectivity to the cloud they can still continue with operations," she says.

Using modular data centres is one way to achieve this. Aggregating data in the cloud also gives operators a "fleet-level view" of operations across sites or to provide backup.

Predicted size of the global data centre market



Source: Precedence Research

Chips with everything, CPUs, GPUs, TPUs An explainer

Chips for AI applications are developing rapidly. The examples below give a flavour of those being deployed, from training to operation. Different chips excel in different parts of the chain although the lines are blurring as companies offer more efficient options tailored to specific tasks.

GPUs, or graphics processing units, offer the parallel processing power required for AI model training, best applied to complex computations of the sort required for deep learning.

Nvidia, whose chips are designed for gaming graphics, is the market leader but others have invested heavily to try to catch up. Dietz of Cisco says: "The market is rapidly evolving. We are seeing growing diversity among GPU providers contributing to the AI ecosystem — and that's a good thing. Competition always breeds innovation."

AWS uses high-performance GPU clusters based on chips from Nvidia and AMD but it also runs its own AI-specific accelerators. Trainium, optimised for model training, and Inferentia, used by trained models to make predictions, have been designed by AWS subsidiary Annapurna. Microsoft Azure has also developed corresponding chips, including the Azure Maia 100 for training and an Arm-based CPU for cloud operations.

CPUs, or central processing units, are the chips once used more commonly in personal computers. In the AI context, they do lighter or localised execution tasks such as operations in edge devices or in the inference phase of the AI process.

Nvidia, AWS and Intel all have custom CPUs designed for networking and all major tech players have produced some form of chip to compete in edge devices. Google's Edge TPU, Nvidia's Jetson and Intel's Movidius all boost AI model performance in compact devices. CPUs such as Azure's Cobalt CPU can also be optimised for cloud-based AI workloads with faster processing, lower latency and better scalability.

Many CPUs use design elements from Arm, the British chip designer bought by SoftBank in 2016, on whose designs nearly all mobile devices rely. Arm says its compute platform "delivers unmatched performance, scalability, and efficiency".

TPUs, or tensor processing units, are a further specification. Designed by Google in 2015 to accelerate the inference phase, these chips are optimised for high-speed parallel processing, making them more efficient for large-scale workloads than GPUs. While not necessarily the same architecture, competing AI-dedicated designs include AI accelerators such as AWS's Trainium. Breakthroughs are constantly occurring as researchers try to improve efficiency and speed and reduce energy usage. Neuromorphic chips, which mimic brain-like computations, can run operations in edge devices with lower power requirements. Stanford University in California, as well as companies including Intel, IBM and Innatera, have developed versions each with different advantages. Researchers at Princeton University in New Jersey are also working on a low-power AI chip based on a different approach to computation.

In an uncertain world, sovereignty is important

Another consideration when assessing data centre options is the need to comply with a home country's rules on data. "Data sovereignty" can dictate the jurisdiction in which data is stored as well as how it is accessed and secured. Companies might be bound to use facilities located only in countries that comply with those laws, a condition sometimes referred to as data residency compliance.

Having data centre servers closer to users is increasingly important. With technology borders springing up between China and the US, many industries must look at where their servers are based for regulatory, security and geopolitical reasons.

In addition to sovereignty, Garg of Aveva says: "There is also the question of tenancy of the data. Does it reside in a tenant that a customer controls [or] do we host data for the customer?" With AI and the regulations surrounding it changing so rapidly such questions are common.

Edge computing can bring extra resilience

One way to get around this is by computing "at the edge". This places computing centres closer to the data source, so improving processing speeds.

Edge computing not only reduces bandwidth-heavy data transmission, it also cuts latency, allowing for faster responses and real-time decision-making. This is essential for autonomous vehicles, industrial automation and AI-powered surveillance. Decentralisation spreads computing over many points, which will help in the event of an outage.

As with modular data centres, edge computing is useful for operators who need greater resilience, for instance those with remote facilities in adverse conditions such as oil rigs. Garg says: "More advanced AI techniques have the ability to support people in these jobs…if the operation only has a cell or a tablet and we want to ensure that any solution is resilient to loss of connectivity…what is the solution that can run in power and compute-constrained environments?"

Some of the resilience of edge computing comes from exploring smaller or more efficient models and using technologies deployed in the mobile phones sector.

While such operations might demand edge computing out of necessity, it is a complementary approach to cloud computing rather than a replacement. Cloud is better suited for larger AI compute burdens such as model training, deep learning and big data analytics. It provides high computational power, scalability and centralised data storage.

Given the limitations of edge in terms of capacity — but its advantages in speed and access — most companies will probably find that a hybrid approach works best for them.

High-bandwidth memory helps but it is not a perfect solution

Memory capacity plays a critical role in AI operation and is struggling to keep up with the broader infrastructure, giving rise to the so-called memory wall problem. According to techedgeai.com, in the past two years AI compute power has grown by 750 per cent and speeds have increased threefold, while dynamic random-access memory (Dram) bandwidth has grown by only 1.6 times.

AI systems require massive memory resources, ranging from hundreds of gigabytes to terabytes and above. Memory is particularly significant in the training phase for large models, which demand high-capacity memory to process and store data sets while simultaneously adjusting parameters and running computations. Local memory efficiency is also crucial for AI inference, where rapid access to data is necessary for real-time decision-making.

High bandwidth memory is helping to alleviate this bottleneck. While built on evolved Dram technology, high bandwidth memory introduces architectural advances. It can be packaged into the same chipset as the core GPU to provide lower latency and it is stacked more densely than Dram, reducing data travel time and improving latency. It is not a perfect solution, however, as stacking can create more heat, among other constraints.

Everyone needs to consider compatibility and flexibility

Although models continue to develop and proliferate, the good news is that "the ability to interchange between models is pretty simple as long as you have the GPU power — and some don't even require GPUs, they can run off CPUs," Dietz says.

Hardware compatibility does not commit users to any given model. Having said that, change can be harder for companies tied to chips developed by service providers. Keeping your options open can minimise the risk of being "locked in".

This can be a problem with the more dominant players. The UK regulator Ofcom referred the UK cloud market to the Competition and Markets Authority because of the dominance of three of the hyperscalers and the difficulty of switching providers. Ofcom's objections included high fees for transferring data out, technical barriers to portability and committed spend discounts, which reduced costs but tied users to one cloud provider.

Placing business with various suppliers offsets the risk of any one supplier having technical or capacity constraints but this can create side-effects. Problems may include incompatibility between providers, latency when transferring and synchronising data, security risk and costs. Companies need to consider these and mitigate the risks. Whichever route is taken, any company planning to use AI should make portability of data and service a primary consideration in planning.

Flexibility is critical internally, too, given how quickly AI tools and services are evolving. Howe of Atheni says: "A lot of what we're seeing is that companies' internal processes aren't designed for this kind of pace of change. Their budgeting, their governance, their risk management…it's all built for that very much more stable, predictable kind of technology investment, not rapidly evolving AI capabilities."

This presents a particular problem for companies with complex or glacial procurement procedures: months-long approval processes hamper the ability to utilise the latest technology.

Garg says: "The agility needs to be in the openness to AI developments, keeping abreast of what's happening and then at the same time making informed — as best you can — decisions around when to adopt something, when to be a little bit more mindful, when to seek advice and who to seek advice from."

Industry challenges: trying to keep pace with demand

While individual companies might have modest demands, one issue for industry as a whole is that the current demand for AI compute and the corresponding infrastructure is huge. Off-site data centres will require massive investment to keep pace with demand. If this falls behind, companies without their own capacity could be left fighting for access.

McKinsey says that, by 2030, data centres will need $6.7tn more capital to keep pace with demand, with those equipped to provide AI processing needing $5.2tn, although this assumes no further breakthroughs and no tail-off in demand.

The seemingly insatiable demand for capacity has led to an arms race between the major players. This has further increased their dominance and given the impression that only the hyperscalers have the capital to provide flexibility on scale.

Sustainability: how to get the most from the power supply

Power is a serious problem for AI operations. In April 2025 the International Energy Agency released a report dedicated to the sector. The IEA believes that grid constraints could delay one-fifth of the data centre capacity planned to be built by 2030. Amazon and Microsoft cited power infrastructure or inflated lease prices as the cause for recent withdrawals from planned expansion. They refuted reports of overcapacity.

Not only do data centres require considerable energy for computation, they draw a huge amount of energy to run and cool equipment. The power requirements of AI data centres are 10 times those of a standard technology rack, according to Soben, the global construction consultancy that is now part of Accenture.

This demand is pushing data centre operators to come up with their own solutions for power while they wait for the infrastructure to catch up. In the short term some operators are looking at "power skids" to increase the voltage drawn off a local network. Others are planning long-term and considering installing their own small modular reactors, as used in nuclear submarines and aircraft carriers.

Another approach is to reduce demand by making cooling systems more efficient. Newer centres have turned to liquid cooling: not only do liquids have better thermal conductivity than air, the systems can be enhanced with more efficient fluids. Algorithms preemptively adjust the circulation of liquid through cold plates attached to processors (direct-to-chip cooling). Reuse of waste water makes such solutions seem green, although data centres continue to face objections in locations such as Virginia as they compete for scarce water resources.

The DeepSeek effect: smaller might be better for some

While companies continue to throw large amounts of money at capacity, the development of DeepSeek in China has raised questions such as "do we need as much compute if DeepSeek can achieve it with so much less?".

The Chinese model is cheaper to develop and run for businesses. It was developed despite import restrictions on top-end chips from the US to China. DeepSeek is free to use and open source — and it is also able to verify its own thinking, which makes it far more powerful as a "reasoning model" than assistants that pump out unverified answers.

Now that DeepSeek has shown the power and efficiency of smaller models, this should add to the impetus to a rethink around capacity. Not all operations need the largest model available to achieve their goals: smaller models less greedy for compute and power can be more efficient at a given job.

Dietz says: "A lot of businesses were really cautious about adopting AI because…before [DeepSeek] came out, the perception was that AI was for those that had the financial means and infrastructure means."

DeepSeek showed that users could leverage different capabilities and fine-tune models and still get "the same, if not better, results", making it far more accessible to those without access to vast amounts of energy and compute.

Partners

The FT Tech for Growth Forum is supported by *AVEVA* and *HCLTech*, our partners, who help to fund the reports.

Our partners share their business perspective on the forum advisory board. They discuss topics that the forum should cover but the final decision rests with the editorial director. The reports are written by a Financial Times journalist and are editorially independent.

Partners' views stand alone. They are separate from the FT and the FT Tech for Growth Forum.

Achieving the perfect balance: how infrastructure and insight drive AI success

Arti Garg, chief technologist, AVEVA

Nearly three-quarters of AI rollouts fail to deliver any return on investment, which is one of the key findings of this report. It is an outcome that is more easily prevented than many people realise.

Avoiding this pitfall does however require a shift in thinking. Successful adoption depends not only on the AI models and systems but on pairing these tools with expertise and insight.

At AVEVA we see first-hand that AI can deliver transformational value when businesses strike the right balance between solutions, infrastructure and expert partners.

For most organisations, several considerations come into play:

Data management — a connected, integrated approach to data management helps to unlock the potential of AI. When departments and sites share data securely and seamlessly, this can uncover unexpected insights that can compound the benefits across the operation.

Deployment approach — hybrid architectures spanning edge to cloud computing help enterprises to realise value from AI. In the industrial sector, for example, high-value applications of AI may occur in remote locations such as factories, plants or mines. These applications might require near real-time decision-making, and these locations often have intermittent internet connectivity and regulatory requirements that limit data transfer. Deploying AI solutions at the edge can ensure always-available capabilities and faster results.

Model choice — the largest, most general AI models are not always the best solution. Industrial sites are often constrained by compute capacity and the availability of electricity. In these scenarios, choosing smaller and fit-for-purpose models can reduce energy, cooling and computing requirements. In addition these models often produce more accurate and reliable results for specific problems, accelerating "time to value" from the AI solution.

At AVEVA, our "industrial intelligence" approach to AI brings together integrated data management, constraint-aware deployment and intentional model selection, driving meaningful business outcomes and allowing our customers to accelerate ROI.

For example, one US power producer has saved more than $250mn through AI-powered predictive analytics; a food and beverage brand saw 10 per cent material savings, reduced waste and improved product yield; one utility cut its energy usage by 4.6 per cent.

These examples highlight our belief that, at its core, AI should help our customers achieve their goals around sustainability and responsible resource usage. That is why we play our part in developing standards to measure the environmental effect of AI, including the IEEE P7100 Environmental Impacts of AI standards working group (of which I am chair) and the Coalition for Sustainable AI.

A belief in empowering the industrial user underpins AVEVA's approach to AI. We aim to reduce the burden for operators so that they can do their jobs more easily and safely and foster radical collaboration across the value chain. Doing this takes a vision that fuses innovation with operational insight. With decades of experience, AVEVA delivers trusted and secure solutions that manage complexity, boost efficiency and support sustainable growth.

Organisations that embrace AI while keeping in mind agility, purpose and sustainability will lead the next wave of industrial innovation. AVEVA is proud to be helping turn that promise into real-world results.

About the FT Tech for Growth Forum

The Tech for Growth Forum offers a unique perspective on how technology can drive growth for organisations, consumers and society as a whole.

Through reports, intimate events and expert insights, it empowers business and political leaders with the knowledge to harness technology for meaningful change.

Find out how to take part in the FT Tech for Growth Forum by emailing **forums@ft.com**