The rise of AI is reshaping data-centre infrastructure and site deployment strategy

May 2025

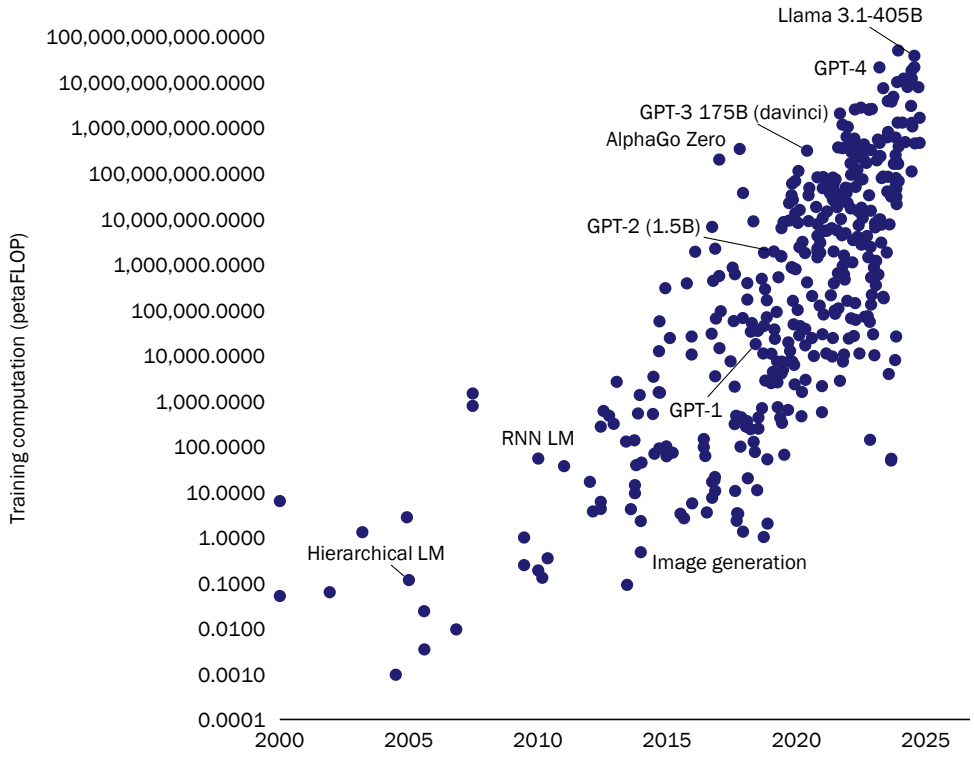
Sylvain Loizeau

Investors in digital infrastructure are well versed in the importance of keeping up with technological developments. The rapid advancement of artificial intelligence (AI) and the resulting exponential growth in computational requirements are beginning to cause major disruption in the architecture and location of data centres. Rapid adaptation is the key to avoiding stranded assets.

AI workloads require significantly more power than traditional cloud workloads

The rise of AI is dramatically reshaping the technological landscape, and nowhere is this more tangible than in the data-centre space. As large internet players, such as hyperscalers or AI specialists, race to develop and deploy frontier AI models, the infrastructure supporting these innovations is being pushed to its limits. The computational requirements of large-scale AI training and inference are unlike anything data centres have previously managed, as illustrated in Figure 1 below.

Figure 1: Increase in computational requirements for the training of frontier models, 2025



Source: Analysys Mason (based on public data)

This leap in demand for data-centre capacity is propelled by the evolution of the processing units used: moving from traditional central processing units (CPUs), to graphics processing units (GPUs) or tensor processing units (TPUs), brings much greater processing power. But GPU-/TPU-intensive workloads result in significantly greater power consumption, increased heat output, and architectural pressures that challenge conventional deployment strategies.

Inside the data centre: new pressures on power, cooling and physical design

AI accelerators, such as NVIDIA's H100 GPU, consume up to 1700W per unit (when overheads such as other electronic components, networking, power distribution, cooling, etc. are included). This is around eight times the power consumption of traditional CPUs. When deployed in dense clusters for training workloads, racks can easily exceed 40kW and are trending toward 85kW, with projections of 200–250kW per rack by 2030. Google has even started envisioning 1MW racks in the longer term. These power and heat levels will overwhelm standard data-centre cooling and structural configurations.

One major structural consequence is the inadequacy of raised floors. A traditional set-up is designed for lightweight, air-cooled racks, but the weight of new AI racks can exceed 2000kg. Reinforced floors and taller, deeper rack designs are becoming necessary, both to accommodate power distribution units (PDUs) and to handle increased cabling and airflow demands.

Cooling requirements have also escalated. Air cooling is effective for densities of up to 30kW per rack, but can no longer keep up. Liquid cooling – whether via rear-door heat exchangers, direct-to-chip systems or immersion tanks – is emerging as the only viable option. Retrofitting for liquid cooling involves significant capital investment and operational overhaul, but it has the potential to bring major gains in energy efficiency and sustainability. In some cases, facility power use can drop by 20%, and total data-centre power by over 10%.[1]

These infrastructure changes are not optional – they are foundational. Supporting growing demand for AI workflows will require a rethinking of the data centre from the floor and rack, to the cooling and electrical distribution systems.

Centralised versus distributed: strategic trade-offs in workload deployment

Beyond the rack and floor, the architecture of AI deployment also raises pressing questions. Should AI workloads be centralised within a single, high-capacity campus or distributed across several sites?

Centralisation allows for high-efficiency operations and easier synchronisation during training. It allows economies of scale for power and cooling infrastructure. However, the sheer concentration of power demand can overwhelm local grids. A single training run using 16 000 H100 GPUs can require 27MW of sustained power – comparable to the annual usage of 23 000 households.[2] Major AI players are responding by investing in

1 American Society of Mechanical Engineers (2022), *Power usage effectiveness analysis of a high-density air-liquid hybrid cooled data center.*

2 Based on the number of GPUs announced by Meta required to train the LLaMa 3.1 405B model, and considering the average annual electricity consumption per American household (10 000kWh according to the U.S. Energy Information Administration, 2024).

on-site power generation, such as solar farms, gas power plants or nuclear generation, to guarantee reliable supply and reduce grid dependency.

Distributed architectures reduce pressure on individual sites and introduce redundancy, but raise issues around synchronisation, latency and network bandwidth. Distributed training requires GPUs across multiple campuses to share updates frequently and with low delay. Even minor packet loss or millisecond-scale latency can degrade the overall training performance of the distributed system. High-throughput, low-latency, lossless inter-site networking infrastructure is becoming an absolute requirement. As an example, Analysys Mason estimates that if Meta had trained the LLaMa 3.1 405B model using a distributed architecture, network limitations would have led to an additional 1–5% in total AI training time and therefore costs. As workloads increase and data-centre interconnections expand, this issue will become an increasing point of focus for infrastructure leaders.

Training versus inference: how workload type shapes location strategy

AI workloads can be split into training (developing a new model) and inference (using an existing model to infer new data). Each comes with different sets of requirements, both in terms of data-centre infrastructure and localisation strategies.

Training typically requires a vast amount of power to process massive datasets over long durations, is relatively latency insensitive (if we disregard the inter-site latency mentioned above) and can be housed in remote, centralised data centres (considering that centralised training, when local power allows, is more efficient than distributed). Such ‘training facilities’ can be located where conditions are favourable – where land and water are available, and energy is cheap and sustainable.

Inference, by contrast, must deliver real-time (or near-real-time) responses to users. It is much more sensitive to latency and requires high availability. This pushes inference workloads closer to end users – through regional facilities or edge deployments. Inference workloads also scale differently: whereas training runs are episodic and planned, inference is constant, and its volume grows directly with product adoption. Combined inference workloads should, in the next few years, overtake training workloads, in terms of installed capacity, as AI workloads gain traction.

In effect, AI infrastructure is fragmenting. Some organisations may continue to separate training and inference across distinct facilities. Others may adopt hybrid approaches, clustering workloads based on latency, power and regulatory constraints. What is clear is that a ‘one size fits all’ approach cannot be systematically applied to data-centre strategies. New data-centre hubs are emerging as a response to the growing need for AI capacity.

Building for the future of compute

As AI workloads redefine what data centres need to deliver, the infrastructure landscape is evolving, with the emergence of new technologies, innovative data-centre designs and alternative deployment geographies. There are new opportunities for infrastructure investors, but also a significant need to stay abreast of developments to minimise the risk of obsolete assets.

Analysys Mason has extensive experience of helping infrastructure investors to optimise their positioning in a changing market. We also support data-centre operators in rethinking their data-centre design, operations and siting strategies for the AI era. Whether quantifying the evolving demand for data-centre capacity, advising on

retrofitting legacy infrastructure, building a new hyperscale campus or planning a global network of inference nodes, we bring the expertise needed to futureproof data-centre operations and investments.

If you would like to discuss these issues, please reach out to Sylvain Loizeau. You can also find Sylvain at Datacloud Global Congress 2025 on 4 June where he will be speaking on “Data centre design: Flexibility, scalability, and reliability”.