How AI can affect the colocation market

July 2023

Stéphane Piot and Sylvain Loizeau

Recent advancements in artificial intelligence (AI) could have a significant impact on various industries. As organisations adopt AI-driven applications and processes, the demand for robust cloud services that can support the related training and inference workloads will continue to increase. This article explores the main impacts that these developments will have on data-centre infrastructure.

AI workloads will require new server hardware

Faster interconnections

AI workloads frequently involve multi-node computing, where distributed systems collaborate to handle complex calculations. These distributed systems require high-bandwidth low-latency interconnections, to limit, as much as possible, the bottleneck effects that are linked to the communication between compute nodes. For a given central/graphic/tensor processing unit system architecture, the friction can be reduced in two ways:

- by increasing the density of compute and memory chipsets on server boards (intra-board and intra-rack)
- by deploying networks based on standards such as InfiniBand that feature dedicated high-speed fibre-optic-based interconnections (inter-rack).

AI-dedicated processors

Total compute requirements for training large AI models are high. As an example, recent large language models have around 100 billion parameters and take around 1000 petaflop/s.days to train (that is, a notional compute power of 30 petaflop/s could train the model in around 33 days).

Training and inference workloads for AI models involve complex matrix computations, but the time and resources required for such workloads can be reduced by using specialised processors that are designed for the task. Traditional central processing units (CPUs) are relatively inefficient at performing matrix and tensor calculations and more-specialised processors are replacing them for AI workloads. Graphics processing units (GPUs) that were originally created for rendering 3D graphics but are capable of 'general-purpose' calculations were the first to be used because they can speed up AI workflows. To further increase speeds, AI-dedicated application specific integrated circuits (ASICs), such as tensor processing units (TPUs), have been developed. All these specialised processors excel at performing AI-related computation. However, they are expensive, are in short supply and require a lot of electrical power. Indeed, data-centre specialist Danseb Consulting considers that relationships with chip makers will be important, with a co-location provider recently interviewed noting that "to support AI, you need to have a partnership with NVIDIA".

Higher density of compute power

The combination of a high density of processors within server boards and racks with the use of high-power GPUs and AI-specific processors is driving desired rack power density to new heights. Indeed, a rack of GPUs can draw up to 50kW,[1] far above the current average of ~10kW per rack.[2] Although ASICs such as TPUs are designed to be more power-efficient than GPUs, data-centre operators should expect an increase in power density, which has practical implications. Danseb Consulting considers that "the compute required for AI creates a significant opportunity for the data-centre industry", as evidenced by the recent increase in demand for racks delivering above 30kW for AI applications.

This architectural shift will have multiple impacts on data-centre infrastructure

Upgrade of power distribution

The increase in rack power density will lead data-centre facilities to require more power overall (to avoid running out of power while the data halls are half full …). Therefore, data-centre operators will need to:

- upgrade power distribution systems including power conditioners and transformers, back-up generators and uninterruptible power supply (UPS) systems
- discuss the increased demands on the electrical power grid with the relevant utility companies.

Upgrade of cooling technology

The increase in rack power density will also increase the amount of heat that will have to be dissipated. Currently, data centres mostly use traditional air-cooling methods, which can support an energy density of up to 20kW per rack.[3] As rack density increases beyond this point, data centres will probably need to enhance their cooling systems – for example, by upgrading their cold sources, such as chillers or cooling towers. Advanced heat exchangers can also improve the efficiency of cooling systems by optimising the transfer of heat between hot and cold sources.

In addition, data-centre operators may also consider emerging technologies such as:

- direct-to-chip liquid cooling, which involves running liquid coolant directly through microchannels that are integrated within the processor, removing heat at its source
- immersion cooling, which involves submerging racks in dielectric fluid, removing the need for air-conditioning infrastructure.

Liquid-cooled systems are believed to support an energy density of up to 100kW per rack5 and Alibaba estimates that immersion cooling could reduce power consumption by 36% from an air-cooled facility with 1.5 power usage effectiveness (PUE) and has deployed this technology in its Hangzhou data centre.[4]

1 phoenixNAP (USA, 2021), Why Density per Rack is Going Up. Available at: https://phoenixnap.com/blog/rack-density-increasing.

2 Uptime Institute (USA, 2020), Rack Density is Rising. Available at: https://journal.uptimeinstitute.com/rack-density-is-rising.

3 Schneider Electric – Data Science Center, Five Reasons to Adopt Liquid Cooling (Schneider Electric, 2019; White paper 279).

4 Alibaba Group (China, 2018), OCP Summit – Immersion Cooling for Green Computing. Available at: www.opencompute.org/files/Immersion-Cooling-for-Green-Computing-V1.0.pdf.

The implications for the colocation market are profound

These infrastructural and architectural shifts have significant implications for the colocation market.

- Upgrading existing data-centre facilities to meet the demands of AI workloads can be complex and capex-intensive, especially in areas/facilities with limited space (to accommodate extra cooling/power conditioning) and power availability (because local substations could be at capacity).

- The installation of advanced cooling technologies requires careful planning, especially for facilities that are already running, and substantial investment, particularly if such upgrades are to be performed on hyperscale facilities.

- Dark fibre interconnections within and between data centres will enable users to connect their AI compute nodes and workflows efficiently and with the protocol of their choice (for example, InfiniBand).

- Emerging use cases will require low-latency processing for AI inference and some could also benefit from the use of edge facilities.

Undoubtedly, the rapid advancement of AI could bring positive change to many industries.

At Analysys Mason, we have been at the forefront of the telecoms and technology sector for the past 35 years, helping clients to tackle complex challenges in digital infrastructure and regulation.

We are eager to hear your thoughts on the impact of AI on data-centre infrastructure. For more information and to discuss how we can help your organisation to navigate this evolving landscape, please contact Stéphane Piot or Sylvain Loizeau.