How does AI impact data centres?

AI is expected to bring huge changes to how we live both our personal and professsnal lives. However, this article details how AI will change its own home – data centres.

Matt Bamforth, Senior Consultant – Data Centre practice lead

August 2024

AI and Data Centres: Transforming infrastructure for the future

The AI hype has now spread far beyond the telecoms and technology world and into the mainstream. Consumers and businesses of all shapes and sizes globally are looking to the future and how AI will help them both to improve their day-to-day and to unlock new opportunities. Great expectation from end-customers means there is pressure on the underlying technology providers to make the AI hype a reality.

It is in data centres that AI will live, and delivering will be a challenge for operators. AI will require data centres to operate in a fundamentally different way to before. This article will outline the physical differences between data centres hosting cloud and colocation services, and data centres hosting AI workloads.

What is different about AI?

There are a few key reasons that AI workloads are different from cloud or colocation workloads, however they primarily stem from the very high compute requirements of AI applications – and the resultant need for significantly more power. The computational requirements of AI workloads are significantly higher than for colocation or cloud and existing IT infrastructure is struggling to cope, this has seen the need for new solutions to meet the ever increasing demands of AI.

The first of these is the increased usage of graphic processing units (GPUs), which were initially developed to handle computer graphics and animation as computer processing units (CPUs) were unable to handle the increase in compute power. Today GPUs are vital for handling the most compute-intensive workloads such as machine learning and AI; to execute one Chat GPT-3 inference would take a CPU about 32 hours compared to 1 second for a GPU. Every server will still contain CPUs, but with AI we are now seeing more and more servers needing to contain GPUs as well. The extra compute power of GPUs does not come for free, they are significantly more power hungry than CPUs and this has implications for data centre architecture and design.

Another technical solution that data centres are using is clustering which involves combining multiple servers into a single, unified system. Again, the goal of this is to increase the computational power of servers to cope with the increasing demands of AI. Clustering is an efficient and scalable way to combine server power.

Both clustering and the need for GPUs mean that AI workloads need significantly more power than cloud or colocation workloads, and this has implications for data centre design and operations.

How does AI impact data centres?

Power

In order to meet the increased power demands of AI applications we are seeing bigger and bigger data centres by power supply. The average rack density used to be sub-10kW but for some providers this is now getting near to the 50kW mark, and the prospect of 100kW racks is a real possibility. As racks demand more and more power, hyperscalers and colo providers are requesting permission for more grid power and many markets are unable to grant it. Dublin has placed a moratorium on new data centres and electricity grids globally are creaking under the pressure. Many data centres are moving towards more on-site generation as a result, with solar, wind and even small modular nuclear reactors an alternative to grid power.

Space

A consequence of the increased power requirements is densification. This means that the same amount of power is now enough for a much smaller footprint of servers: a 50kW workload may have previously been 10 racks using 5kW each, but this could now feasibly be a single 50kW rack. Any retrofitted facilities will end up

with a lot of empty space if they switch from traditional workloads to AI workloads, and facilities of the same area size will now need significantly more power if they are to be filled.

Cooling

Higher power density racks requires more powerful cooling systems. This will not mean just increasing air cooling but also finding more efficient cooling systems, such as liquid cooling. Once rack density starts to reach around 40kW data centre operators will begin to integrate liquid cooling solutions as well, so that there is a hybrid model with air and liquid cooling. Read door solutions are a good bridge between air and liquid cooling, but liquid-to-the-chip will be necessary once you start to reach the 40kW rack density level. Immersion cooling is yet to take-off but will presumably play a role as well. Liquid cooling (rear door, direct-to-chip and immersion) is projected to make up about 1/3 of cooling capacity by 2027.

Racks

More power infrastructure, more cooling equipment and generally more advanced servers mean that the physical racks will also need to adapt. Data centres will contain racks that are higher, wider, deeper and heavier than before.

Equipment lifecycles

GPUs are being upgraded every 12 – 18 months to keep up with the ever increasing demands of AI applications. This is an issue when you are designing data centres for much longer periods, you would hope that power infrastructure, for example, would be able to last you 15 years, so you do not want it to become obsolete. This especially makes it a challenge for colocation providers to hoost GPU-as-a-service providers looking to provide AI services to enterprises, at least until they begin to extend their product lifecycle over 5 years.

AI's lasting impact on the future of data centres

All of these factors will vary depending on the AI workload. For example, training LLMs is far more power-intensive than inference, and even within training there is a huge difference between massive public-facing applications like Chat GPT and more use case specific AI applications used by enterprises. However, AI is already fundamentally changing data centre design and operations and we can only expect this to continue.

Matt Bamforth is a Senior Consultant at STL Partners, specialising in data centres and digital infrastructure.

Get in touch with the author to learn more

matt.bamforth@stlpartners.com

Or visit STL Partners' Data Centres Hub

https://stlpartners.com/data-centres/