Why Hyperscale Is Leading the Data-Centre Revolution

By **Cary Springfield***, International Banker*

According to data uncovered in April by Synergy Research Group, the number of large data centres operated by hyperscale providers increased to 992 at the end of 2023 before surpassing the thousand mark by early 2024. Synergy's data also identified the United States as the clear leader in this space, with 51 percent of global hyperscale capacity as measured by megawatts (MW) of critical IT (information technology) load. Europe and China followed, each with around a third of the remaining 49 percent of capacity. And with the generative artificial intelligence (GenAI) revolution triggering an accelerated rise in demand for more expansive, powerful data-centre facilities, the hyperscale era is only just beginning.

Along with enterprise, colocation, edge and modular, hyperscale data centres are one of the five main types of data centres in demand today. As the name implies, hyperscale is a massive data centre that can scale to extreme magnitudes and is thus built to support the potentially substantial workload demands that are placed on the system whilst maintaining its performance (as measured by such important parameters as optimised network infrastructure, strong connectivity and minimal latency).

What sets a hyperscale data centre apart from other types of data centres it its sheer size. According to International Data Corporation (IDC), a data centre can be classified as hyperscale if it has a minimum of 5,000 servers and covers at least 10,000 square feet of physical area to accommodate all the necessary IT infrastructure. That said, some data centres even exceed a million square feet of space, with the construction costs of hyperscale facilities now starting to exceed $1 billion in some instances.

Hyperscale thus recognises the ballooning needs of enterprises for expansive off-premises data centres from specialist third-party providers (TPPs) that can be scaled up as and when required. As such, they are

aimed at supporting companies and cloud-service providers with the biggest processing, storage and network requirements, such as Amazon, Google and Microsoft, ultimately helping them seamlessly deliver key services to their global customer bases. In total, these three tech giants account for 60 percent of all hyperscale data-centre capacity.

Given this mammoth scope, then, it's not surprising that the most serious concern regarding hyperscale is the massive energy requirements of such power-hungry tasks as operating the servers and powering the cooling systems, as well as the sheer magnitude of the footprints tied to constructing and maintaining such vast infrastructures. In a similar vein to the mining of bitcoin and other "proof-of-work" cryptocurrencies, hyperscale has earned a reputation for requiring enormous amounts of electrical power, and the proliferation of these colossal data centres stands at odds with global environmental goals and sustainability targets.

Some estimates put the average hyperscale data centre's energy draw at 50-100 megawatts (MW) and around 25 kilowatts (KW) per rack. "Globally, data centres consume around 3 percent of the total energy generated worldwide, with hyperscale data centres accounting for 20 percent of the world's data centre electricity usage, growing to 50 percent by 2020," according to network solutions firm AFL Hyperscale. "One of the world's largest data centres, Microsoft's 700,000 sq. ft. data centre in Chicago, Illinois, has the capacity to consume 198MW of power. To put that into context, in the US, 1MW of electricity would be enough to power an average of 750 homes; theoretically, this data centre could pull an amount of power similar to that of 150,000 homes."

While hyperscale often represents some of the most efficiently run data centres from a power-consumption perspective, with extra resources able to be deployed without requiring additional cooling, electrical power or physical space, the sheer scale of these operations means that providers are increasingly looking for ways to lower their carbon footprints.

One way to achieve this is to build data centres in parts of the world where electricity is produced cleanly and cheaply. "Hyperscalers are looking to diversify their geographical footprints and find readily available and scalable power by expanding into different secondary and tertiary markets like Atlanta, Salt Lake City, Reno, Denver, Columbus, and Charlotte," according to real-estate firm JLL's "Data Centers 2024 Global Outlook" report. "Globally, Sweden and other Nordic countries offer data centres a secure source of green electricity as well as a colder climate that reduces the need for cooling operations. It's no surprise that Google, Meta, and Amazon have all established data centres in the Nordics over the past several years, and experts expect the region will see more growth in the coming years."

Some are also choosing to power their data centres using clean and renewable energy sources. One of the world's largest data centres at around 7.2 million square feet, the Citadel Campus in Tahoe Reno, Nevada, is powered by up to 650 megawatts from 100 percent renewable-energy sources, such as solar and wind farms.

The Climate Neutral Data Centre Pact, meanwhile, is a Europe-wide agreement among data-centre operators and trade associations to make data centres climate-neutral by 2030, with all signatories to the pact committing to power their facilities with 100 percent renewable energy by then. Along with carbon-free energy, signatories must also prioritise water consumption, reuse and repair servers, provide transparency regarding energy efficiency through measurable targets and explore options to recycle heat.

A report published on October 8 assessing the European hyperscale data-centre market outlook from 2024 to 2029 highlighted the Climate Neutral Data Centre Pact as being “instrumental” in making European data centres more sustainable. “Investors in the European hyperscale data centre market are exploring new locations, such as Spain, Portugal, Greece, and other countries with abundant renewable energy resources and reasonable land prices,” the report noted, adding that leading technology firms, such as Microsoft, Meta and Google, were at the forefront of developing environmentally sustainable data centres. “They employ initiatives such as using sustainable materials in construction, implementing green facades, and exploring alternative energy sources such as hydrogenated vegetable oil (HVO). Colocation data centre developers also follow suit.”

As of April 2024, Synergy calculated the pipeline of future hyperscale data centres at 440 facilities, which are either currently being planned, developed or fitted out, while data from August showed that hyperscale already accounted for a hefty 41 percent of the worldwide capacity of all data centres. Looking ahead, Synergy has projected a doubling of total hyperscale data-centre capacity in just the next four years. And while each year will see some 120-130 additional hyperscale data centres coming online, Synergy also observed that overall capacity growth would largely be driven by the expanding scale of those new data centres, mostly stemming from growth in GenAI demand.

“While both the number of hyperscale data centres and their average size continue to grow at an impressive pace, there is a lot of complexity and nuances behind those trends,” said John Dinsdale, chief analyst at Synergy Research Group. “Generally speaking, self-owned data centres are much bigger than leased data centres, and data centres in the home country of a hyperscale company are much bigger than its international facilities, though there are plenty of exceptions to these trends. We’re also seeing something of a bifurcation in [the] data centre scale. While the core data centres are getting ever bigger, there is also an increasing number of relatively smaller data centres being deployed in order to push infrastructure nearer to customers. Putting it all together, though, all major growth trend lines are heading sharply up and to the right.”

According to a Precedence Research report, moreover, the global hyperscale data-centre market size will grow from an estimated $102.1 billion in 2023 to reach $935.3 billion by 2032, at a mammoth compound annual growth rate (CAGR) of 27.9 percent during this period. What’s more, Precedence has predicted that the banking, financial services and insurance (BFSI) segment will become the largest in the hyperscale data-centre market.

“The major issues related to data storage, recovery, and cybersecurity that BFSI firms must manage, together with processing enormous amounts of data that must be processed, saved, and copied periodically, have led to the implementation of a number of cloud techniques,” the market-intelligence firm’s report also noted. “The constant need for data protection and scalability in hybrid cloud environments fuels the demand for hyperscale data centre[s] to automate and manage IT services across heterogeneous clouds.”

The report cited the IT and telecom (telecommunications) industries as accounting for another sizeable share of the hyperscale data-centre market during the 2023-32 forecast period. “Large, mission-critical facilities known as hyperscale data centres are frequently connected to well-known data-producing companies like Facebook, Google, Amazon, IBM, and Microsoft.”

While the report specified North America as holding the largest regional market share of the hyperscale data-centre sector during this period, it also identified the Asia-Pacific (APAC) region as being the fastest grower, with rapid industrialisation and market development underpinning the region's strong expansion.

"The basic IT infrastructure of the APAC area is evolving more swiftly, which results in a self-replicating cycle that promotes more investment and growth. As a result, IT management will require more authority, resources, and accountability," according to Precedence. "This area has qualified programmers who are at least on par with those found elsewhere in the world. It is also advantageous to use the infrastructure that China and India provide…two growing economies [that] are expected to contribute significantly to Asia Pacific's notable economic and technical growth during the predicted period."