GlobalData.

AI's Hidden Impact and Vertiv's Vision for the Data Center of the Future

Malcolm Rogers & Siow Meng Soh | March 28, 2024

Summary

Issue

The explosion in the use of generative AI (GenAI) services is transforming operations across many different verticals. GlobalData expects global enterprise spend of GenAI to continue to grow at a rapid pace, from $2.4 billion in 2023 to $33 billion by 2027 (For more, please see: GlobalData Market Opportunity Forecasts to 2027: Generative AI, July 23, 2023). However, the increased use of GenAI and other compute tasks that require high performance compute (HPC) and graphics processing unit (GPU) resources has a very real impact on the demands of existing data center infrastructure. For example, it has been calculated that energy usage at data centers will double from their present levels. The International Energy Agency (IEA) sees electricity consumption by data centers from workloads such as AI and cryptocurrency doubling by 2026, from an estimated 460 terawatt-hours (TWh) in 2022. Equally, it is estimated that global demand for water to cool data centers will be driven by AI to between 4.2-6.6 billion cubic meters by 2027 – about half the amount consumed in the UK each year. This is pushing current data center power and cooling systems beyond their limits and necessitating innovations to keep pace with the AI boom.

The industry is rapidly adjusting, and data center infrastructure solutions providers like Vertiv are developing new solutions to support these new demands. This report looks at new challenges driven by the increased use of AI and how Vertiv envisions the data center of the future meeting these challenges.

Key Takeaways

• While there is excitement across industries about the different use cases AI/GenAI can support, their increased power demands are creating challenges around sustainability and performance that industry needs to address.

• The massive amounts of power used by data centers is already leading toward regulations limiting new data center builds, highlighting the need for the industry to be more energy efficient and reduce dependency on power from the grid.

• Vertiv's vision for the future of data center power revolves around dynamic power management systems, drawing from a hybrid of energy sources, supported by increased deployment of uninterruptible power supplies (UPS) and battery energy support systems (BESS).

• In many cases high performance compute workloads now require liquid cooling systems to remove

heat, due to changes in chip architecture.

• Vertiv and other industry ecosystem players are currently evolving standards for the future of liquid cooled infrastructure, while still seeking to provide liquid cooling systems into existing brownfield data centers.

• Challenges still exist and while the ecosystem is being driven by AI hyperscale requirements, smaller enterprise colocation facilities will still need to be supported.

Perspective

Current Perspective

AI Driving Changes to Data Center Infrastructure

AI has evolved from predictive AI to GenAI and multi-modal GenAI, over a relatively short period of time. AI has already been adopted in various areas including chatbots, computer vision, personalization, and predictive maintenance. GenAI is further pushing the envelope, creating new possibilities. Various use cases have started to emerge such as disease diagnosis, application development (assisted code development), marketing content creation, and demand forecasting for retail businesses. Against this backdrop, it is no surprise to see many forecasters expecting a rapid growth of revenue related to GenAI. However, amid the excitement, there have been concerns, particularly the need to address ethical issues and responsible AI. One area that AI technology providers have not been highlighting is the infrastructure, including the physical facility required to host the hardware for AI workloads.

The processing power required for GenAI workloads has pushed the semiconductor industry to develop faster and more efficient AI-optimized solutions. Training a large language model (LLM), for example, with ChatGPT can take a long time even with multiple GPUs (years with a single GPU). The industry is working on solutions such as Google's TPU architecture and Nvidia's InfiniBand for faster networking speeds, interconnecting GPU servers. But besides achieving the processing speed, the hardware needs to be more compact and consume less power. Most of the LLM development and training of AI models are being done in large-scale data centers, and service providers are adding GPUs to meet the growing demand. This is driving demand for data center space as well as how it is designed. In particular, these high-density computing resources require more power to function and they need more efficient cooling systems to remove the heat generated.

Future of Data Center Power Systems

A 10MW data center was considered a large facility about a decade ago. According to data center infrastructure vendor, Vertiv, 10MW data centers may now be considered for edge locations as individual racks are rapidly growing from 50kW per rack up to 100kW per rack, and 200kW per rack may be the reality in the near future. This is due to chips become more powerful and liquid cooling technologies enable higher density of IT equipment per rack. Building new facilities in key data center

hubs will get more difficult due to the power requirements. The growing demand for energy will be a major concern as countries have commitments to meet carbon emission targets. For example, Singapore had imposed a four-year moratorium on the development of new data centers due to their high energy consumption that would impede the country's efforts to achieve its sustainability goals. While the moratorium was lifted in mid-2023, Singapore has been more selective in approving new data centers, based on criteria such as energy efficiency, AI/ML compute, and HPC capabilities, expansion of international connectivity, and the economic commitments to Singapore. In London, it was also reported that in some boroughs west of the city, the lack of sufficient electrical capacity - due to the demand from data centers - has impacted the approvals for new housing developments.

With challenges related to the use of energy from public resources, data center operators can no longer rely on the utilities market and broader grid to supply all of their power needs. Data centers will need to be able to dynamically manage power consumption between grid sources and self-generated power within millisecond time frames in order to meet the demanding nature of AI workloads. With highly dynamic AI workloads causing draws that go from idle capacity to over 100% within seconds, data centers cannot expect the grid to handle that sort of drastic change.

Vertiv's perspective on meeting the challenges of dynamic power management is a multipronged approach. The company advocates for the increased use of UPS and BESS systems to support use of hybrid energy sources, while maintaining power supply to critical systems during dynamic switching. For UPS, Vertiv envisions larger capacity systems, with its hyperscale rated systems offering capacity in MW range, as compared to more traditional UPS measured in kW. Vertiv sees UPS being used more frequently as energy bridges for power systems swapping between grid resources and generated power. Their view is also that UPS will be used to support not only IT related power but also power for cooling systems as well. Further, the company views increased BESS deployment as critical to support transition to more to green energy generation. Increased storage capacity can help ensure consistent supply of solar and wind-based power.

High Performance Compute and Liquid Cooling

Data centers packed with high-density servers will also require new systems to remove heat. There are limitations with existing systems, which are mostly based on air-cooling. According to Vertiv, legacy facilities are ill-equipped to support widespread implementation of the high-density computing required for AI, with many lacking the required infrastructure for liquid cooling. Companies will need to retrofit their data centers, which requires additional investment. Liquid cooled servers can be placed closer together driving performance enhancements by decreasing physical space between the silicon circuits. This is driving up the density of compute per space, but also thermic load required to cool the chips. Hyperscalers are now deploying liquid cooling, to achieve the cooling needed for high performance, GPU-based workloads. In the future, this could be driven even further, some chipmakers are exploring super high-performance chips that will require cooling that goes beyond what's possible with heat capacity of water. To compensate they chipmakers And Vertiv are exploring cooling with refrigerants that take advantage of a phase change to remove energy from the system without raising temperatures.

These changes require a different approach to how data centers are architected. Previously the power, cooling, and IT systems have all been discrete units with clear boundaries within the data center. Data center operators focused on power and cooling, while tenants were mainly concerned with the IT cabinet. However, as technology like direct chip cooling evolves, according to Vertiv, IT and cooling systems will be integrated, complicating the chain of responsibility. While not every data center deployment will be liquid cooled, the new architectures will require data center operators to think beyond simply driving power efficiency of air conditioners.

To support the transition, Vertiv is already supporting customers with liquid cooled systems and is working directly with industry ecosystem partners to develop standards, along with various architectures to support different liquid cooling scenarios. The company offers coolant distribution units (CDUs) that utilize existing heat exchange infrastructure via more traditional air exchange to pull heat from liquid cooling systems. Standards for the systems are still evolving but are being driven by the world's leading industry players (e.g., chip makers Samsung, Intel, and Nvidia; hyperscale infrastructure providers Equinix, Microsoft, and Google; and Vertiv). They are among the participants of the Open Compute Project's working group on data center cooling environment for HPC workloads, driving the demand for liquid cooling systems.

However, this is also an opportunity for companies to become more sustainable by adopting a more efficient cooling system. IEA estimates that data centers consumed about 1-1.3% of electricity globally, and this is expected to increase over time due to accelerating cloud adoption and AI development. While global enterprises and policy makers are keen to harness the power of GenAI, they need to review data center requirements to pave the way for innovation happening at speed and scale. For example, regulations on the use of refrigerants are already impacting data center design. As part of the wider Montreal Agreement, many countries around the world are exploring ways to reduce the reliance on refrigerants with high global warming potential (GWP – i.e., the warming potential of a refrigerant to an equivalent volume of carbon dioxide). The US recently announced that the use of refrigerants, including within data centers, with GWP value over 700 will be banned by 2027. While this is currently only a US standard, as the leading market for data centers in the world, it will help drive international standards. While there are many refrigerants on the market today, many of them do not meet these standards. Further those that do may have increased risk profiles in terms of flammability or require lower pressure, meaning more energy spent on compression.

Challenges

Retrofitting cooling and power systems can be very costly, and the use of high density liquid cooled racks can leave empty white space. Further, while much of the data center industry is being driven by hyperscalers and their requirements, their needs do not always reflect the needs of the enterprise colocation market. These players currently have even lower installed base of GPUs and high-performance racks lowering economies of scale for investing in upgrades. The industry still needs to strike a balance between driving the current generation of technology and investing sufficiently for the future.

Recommended Actions

Vendor Actions

• Enterprises developing an AI strategy needs to consider the ability of their infrastructure to support the new workloads. Besides having more GPUs and faster networking solutions, they also need to ensure that their data center power and cooling systems can keep up with the requirements. This should be factored into their ESG efforts and strategy as well.

• Some workloads, including AI models with smaller number of parameters can be deployed at the edge to meet latency and security requirements. Service providers and enterprises can evaluate modular, prefabricated data centers in the market for edge data centers. These data centers are faster to deploy and may come with more advanced power and cooling technologies.

• Enterprises are looking to gain business advantages with GenAI, but they need to manage the energy consumption and consider their commitments to climate change. This creates opportunities for data center providers to create differentiation based on their ability to run their data centers greener (including tapping into renewable energy sources and hydrogen for storage) and more efficiently. They should develop tools for users to monitor and visualize their energy consumption and efficiency.

• There is still a big industry counter argument that GPU is not the way to do at least now and today, everything else in IT is defaulted to CPU. If you go the way of GPU first, you need to change everything else. There is a cost and complexity trade off and CPU performance still supports many important workloads. Focusing on driving efficiency from existing air conditioning infrastructure is also a valid route for data center operators and their vendor partners.