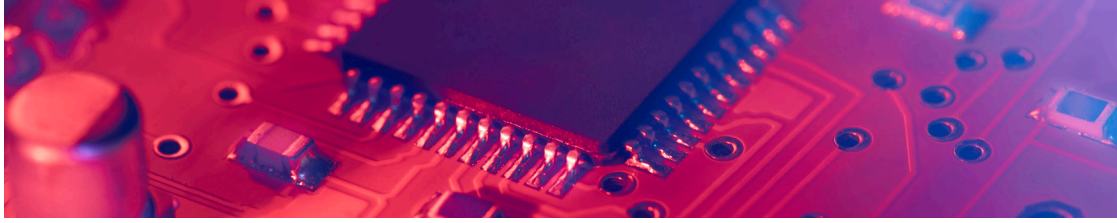
S&P Global
Market Intelligence
451 Research
Business Impact Brief



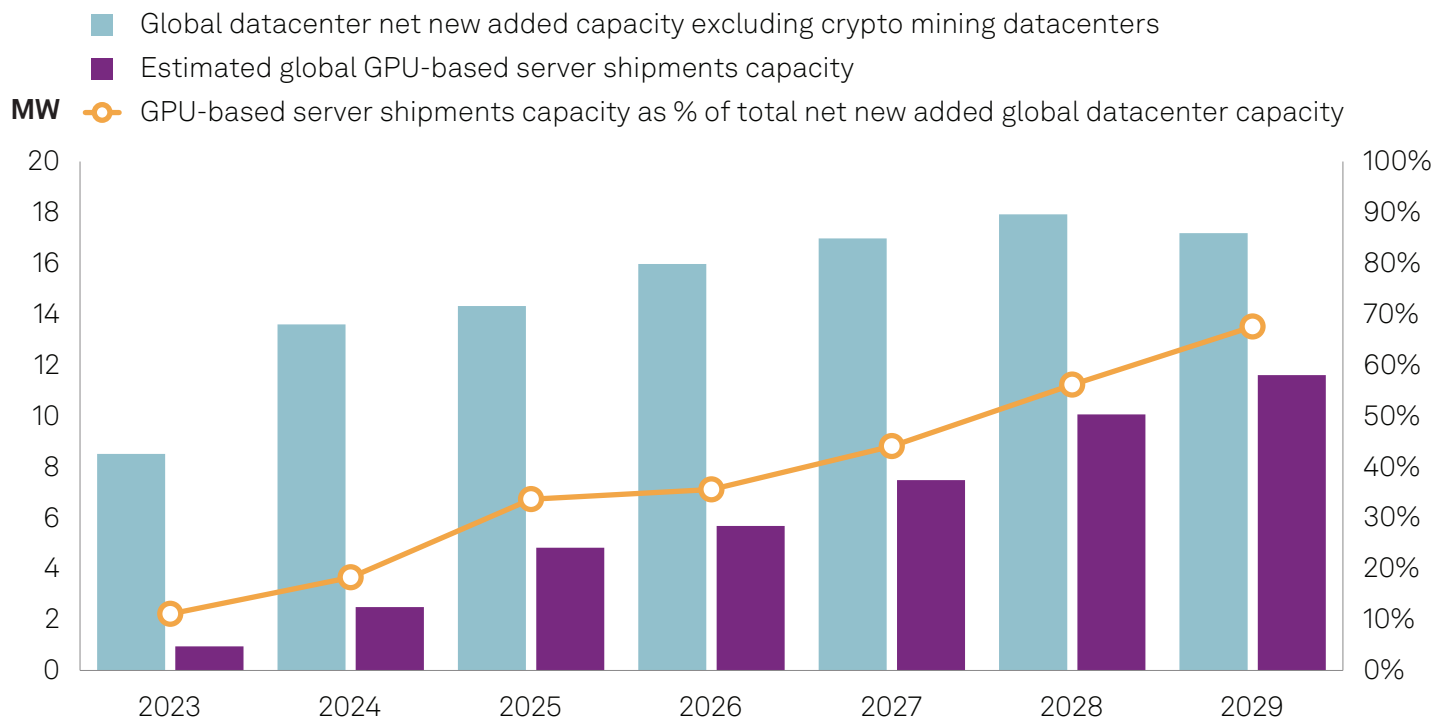
New considerations for datacenter risk exposure

The Take

In the rapidly evolving datacenter landscape, the rise in AI accelerators and graphics processing unit (GPU) rack deployments for high-performance computing and AI workloads is altering operational and financial frameworks. As demand for AI accelerators grows, the cost to deploy GPU racks can be 5-10 times the cost for traditional equipment racks, and GPU deployments often necessitate additional investments in power and cooling infrastructure. When scaled across a datacenter, this not only increases the capital expenditure required for construction or upgrades but also heightens risk exposure in the event of an outage. Whether outages are caused by weather-related disruptions, incidents such as fire, smoke or accidental activation of fire-suppression systems, or human error, the financial implications are substantial. As a result, these increased costs and potential revenue impacts are driving a reassessment of facility insurance coverage and the development of new service-level agreement frameworks.

As shown in the figure below from 451 Research's Datacenter Services & Infrastructure Market Monitor & Forecast, we project that GPU-based servers will grow to 68% of net new added global datacenter capacity by 2029. With this GPU growth comes reconsideration of datacenters' total cost of ownership since a single datacenter may house billions of dollars of equipment, entailing dramatic cost implications for even a single hour of outage.

Global GPU shipments' impact on datacenter capacity



Source: 451 Research's Datacenter Services & Infrastructure Market Monitor & Forecast, 2024.

Business impact

Due to the increased cost implications and revenue impact of outages in this new AI datacenter reality, stakeholders must recalculate the loss footprint. This will likely require reviewing the insurance premiums and coverage levels of high-density GPU deployments, which in turn will affect operational cost dynamics. Factors include not only high repair or replacement costs, but also large contractual penalties that may arise from a sustained outage, as well as the potential for delayed operational recovery due to complexity of repairs and scarcity of hardware.

Considerations also include lease agreements that may limit power usage and equipment types. In colocation facilities, multi-tenant clients may encounter shared resource constraints due to the high power and cooling requirements of GPU racks. Deploying GPU racks without proper safety considerations may lead to contractual breaches and liabilities for damages or service disruptions.

Further, the heat generated by high-density GPU racks may exceed the capacity of conventional cooling systems. Traditional air cooling becomes less efficient at such scale, prompting investment in advanced cooling technologies. However, using liquid cooling for advanced AI accelerators introduces the risk of coolant leaks, which can cause equipment damage, electrical shorts and data loss. Datacenters will need to minimize leak risk with properly architected, installed, maintained and monitored liquid cooling systems using sealed enclosures to identify and isolate coolant leaks before they can escalate to catastrophic failures.

Looking ahead

In early September 2024, xAI announced the activation of its "Colossus" cluster of 100,000 NVIDIA H100 liquid-cooled GPUs, with plans for significant expansion at its Memphis datacenter by 2025. This development signals a shift for datacenter owners, who must now confront the challenges of constructing and operating facilities with costs reaching tens of billions of dollars. The industry is entering uncharted territory, with power and cooling demands that could equal the energy consumption of a small city and capital expenditures potentially surpassing the GDP of many nations.

The transition of AI workloads from training to inference is unlocking unprecedented business efficiencies and automation, driving advancements in health services and medical treatments among many other contexts, and offering a competitive edge to companies that effectively leverage AI. However, integrating GPU racks into datacenters presents liability and infrastructural challenges. Datacenter owners, leaseholders and colocation providers must evaluate the legal and contractual implications, ensuring adherence to all regulations and agreements.

To accommodate the demands of GPU racks, power and cooling infrastructures require significant upgrades, but liquid cooling systems offer solutions while creating new risks. Strategic planning, investment in advanced technologies and clear communication among all stakeholders are crucial to successfully modernize datacenters while mitigating liabilities and operational risks.