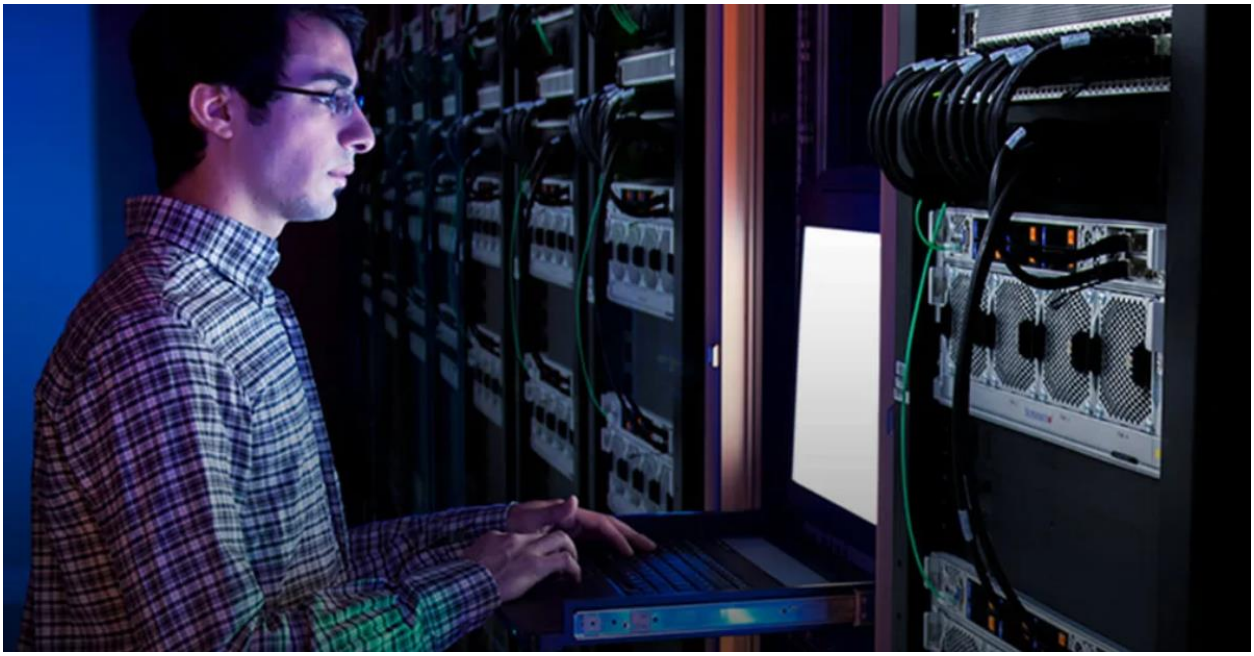
Investing in AI Infrastructure

AI architecture must scale effectively without sacrificing cost efficiency. Process and partnerships are critical.



As businesses across industries transform themselves into data-first operations, the push to find more powerful ways to leverage information, including artificial intelligence (AI) and other forms of data-intensive, high-powered computing, is clear. Revenue in the global AI market, including hardware, software and services, is expected to exceed $500 billion in 2023 and investment in AI-tailored hardware is expected to see a compound annual growth rate of 20.5% over the next five years, according to market intelligence firm IDC. Companies that underinvest in their IT capabilities today and into the future are risking their ability to compete effectively in the market.

“The danger is that you’re going to fall short of delivering on critical capabilities to the business,” says Ashish Nadkarni, IDC group vice president and general manager, infrastructure systems, platforms and technologies. “The differentiation in terms of products and services, in future-proofing the business—all of these things will become headaches and, ultimately, that is going to mean the business won’t do as well as it could.”

Building an agile, cost-effective environment that delivers on a company’s present and long-term AI strategies can be a challenge, and the impact of decisions made around that architecture will have an outsized effect on performance. “AI capabilities are probably going to be 10%-15% of the entire infrastructure,” Nadkarni says. “But the amount the business relies on that infrastructure, the dependence on it, will be much higher. If that 15% doesn’t behave in the way that is expected, the business will suffer.”

Experts like Nadkarni note that while companies can, and should, avail themselves of cloud-based options to test and ramp up AI capabilities, as workloads increase over time, the costs associated with the cloud can rise significantly as workloads scale or the enterprise expands its usage, making on-premises architecture a real consideration.

“Every time you run a job on the cloud, you’re paying for it, whereas on-premises, once you buy the infrastructure components you can run applications multiple times,” he explains. “When businesses have a solid set of use cases and know how to keep that infrastructure busy for a sustained amount of time, they’re moving in the right direction to ensure the return on investment is good.”

Good Process, Good Infrastructure

No matter the industry, to build a robust and effective AI infrastructure, companies must first accurately diagnose their AI needs. What business challenges are they trying to solve? What forms of high-performance computing power can deliver solutions? What type of training is required to deliver the right insights from data?

Context matters. Factory automation, for example, differs from video distribution, risk modeling, supply chain optimization or a host of other applications benefiting from AI capabilities. Retailers will have different needs than manufacturers, while companies with extensive edge computing demands will have different requirements for data than those without.

"It's a matter of finding the right configuration that delivers optimal performance for the workloads." — Michael McNerney, Vice President of Marketing and Network Security, Supermicro

"It's a matter of finding the right configuration that delivers optimal performance for the workloads," says Michael McNerney, vice president of marketing and network security at Supermicro, a leading provider of AI-capable, high-performance servers, management software and storage systems. "How big is your natural language processing or computer vision model, for example? Do you need a massive cluster for AI training? How critical is it to have the lowest latency possible for your AI inferencing? If the enterprise doesn't have massive models, does it move down the stack into smaller models to optimize infrastructure and cost on the AI side as well as in compute, storage and networking?"

One example of an application optimized system for AI training is the Supermicro GPU systems with AMD CPUs, including the Universal GPU system that supports both AMD Instinct MI250 OAM accelerators or NVIDIA HGX A100 or the latest NVIDIA H100. The system's modularized architecture helps standardize AI infrastructure design for scalability and power efficiency despite complex workloads and workflow requirements enterprises have, such as AI, data analytics, visualization, simulation and digital twins.

Given the volume of considerations and the scope of investment, Nadkarni believes communication across the enterprise is critical. "The first and best practice is to get all the

stakeholders together to make sure everyone is heard," he says. "Otherwise, the business will suffer because stated objectives are not aligned with infrastructure design goals."

From there, accelerated computing—the use of special processors called accelerators that work in tandem with traditional central processing units (CPUs) and enable far greater computing power without slowing the system—can help construct greater value around AI. "Data will move faster when hardware is optimally configured for specific workloads, compute needs and characteristics," says Matt Kimball, vice president and principal analyst, data center, at Moor Insights. "Accelerators shave milliseconds off AI computations, but milliseconds add up to seconds, minutes, hours and days. When trying to train big data sets, reducing that time helps get to a place of value faster, which is critical."

Partners to Scale AI

Meeting present-day demands for AI is only part of the challenge companies face in forming an effective strategy around its deployment. Businesses must take those investments and scale them appropriately, paving the way for future innovations and insights. Survey data from Omdia reveals approximately 20%-25% of enterprises are scaling AI projects across multiple business units, up from only 7% in 2020. While this still leaves tremendous space for increased adoption, the rapid growth also indicates that companies seeking to create, or even expand, competitive advantage through AI should act sooner rather than later. No matter the industry, however, businesses must navigate significant challenges when scaling workloads to optimize critical tasks.

Companies must match high-performance demands with appropriately high-performance memory, storage and networking capabilities, and answer questions around physical spaces as well. How can companies make sure hardware is adequately powered and cooled, as newer CPUs and GPUs demand an increased amount of electricity? How can they accommodate changes to physical spaces if more infrastructure is required?

“We’re in a constant state of change. The workloads we’re now deploying are themselves becoming very bespoke, with unique requirements around compute,” Kimball says. “The pace of innovation is only increasing. Investment needs to be made in platforms that have broad utility—the capabilities, functionality and underlying architecture to support everything from machine learning to cloud native platforms to big data analytics.”

Creating the right architecture requires not only constant communication between business units, but also meaningful partnerships and coordination with technology providers. “It’s not just me and my hardware vendor, me and my software vendor,” Kimball notes. “It’s every part of the value chain working together to deliver solutions.” He adds that vendors like Supermicro offer services that help tailor agile, customizable and scalable architectures to accelerate computing capabilities. The right collaboration can help companies avoid both over- and under-provisioning operating AI environments with efficiency and value.

Nadkarni agrees.

“It’s a partnership,” he says. “You can’t do this in a vacuum.”

Learn more about Supermicro at ***https://www.supermicro.com/en/products/aplus***