White Paper

The Power of Now: Gaining Deep and Timely Insights with Performance-Intensive Computing Infrastructure

Sponsored by: Supermicro and AMD

Peter Rutten Ashish Nadkarni
August 2022

IDC OPINION

Data-driven insights form the basis on which organizations can gain competitive differentiation now and in the future. Externally, competitive differentiation is delivered via new products and services or existing products and services continually enhanced via new user engagements and experiences. Internally, differentiation can be achieved via streamlining and optimizing business operations. Being data driven is compressing time to value of insights, thereby creating top- and bottom-line impact on the business.

Performance-intensive computing infrastructure (PIC-I) required for scaling artificial intelligence (AI), Big Data and analytics (BDA), modeling and simulation (M&S), and engineering applications like electronic design automation (EDA) is gaining traction at businesses, large and small, and within every organizational entity (corporations, governments, universities, nonprofits). For the infrastructure, the primary service-level objective is performance at scale. The scale that we are referring to for an organization to gain deep insights from large and/or diverse data sets requires a performant infrastructure designed from the ground up. Further, companies often deploy PIC-I separate from their corporate IT infrastructure, given that these applications should not be hindered during execution. Performance-intensive computing (PIC) applications tend to be batch oriented but are being increasingly retooled to take on a (real-time) streaming analytics element, making the service quality of the underlying infrastructure even more crucial. General-purpose infrastructure — however current or performant it may be — is not designed to host PIC applications. Rather, it is designed to provide stable business operations, delivering service quality required for business-critical and mission-critical revenue-generating applications.

PIC workloads are proven for scale, performance, and security features in many public cloud deployments, where organizations can often do initial development or testing. IT buyers are now seeking to deploy PIC-I solutions on premises or at colocation facilities. Their objective here is to focus on business outcomes with similar value, performance/watt, security, and sustainability objectives as the public cloud. Fortunately, IT vendors are rising to the occasion. For example, silicon vendors like AMD along with their solutions partners like Supermicro have developed infrastructure solutions that are specifically designed for PIC applications.

IDC has found that the leading cause of failure in many PIC initiatives (like AI, for example) is the lack of understanding of, and therefore proper investments in, fit-for-purpose infrastructure. Organizations must also understand that successful PIC initiatives must factor in the influence of data gravity and significance on deployment choices. To eliminate the barriers to broad and secure PIC deployments, businesses must invest in the right infrastructure stack as part of their IT strategy.

SITUATION OVERVIEW

We are about to begin a new chapter on the relationship between IT and the business. This new chapter will further push IT into a strategic role, one that increases its influence on business outcomes.

The previous chapter was focused on elevating the strategic importance of IT. We saw this play out during the COVID-19 pandemic with IT organizations focusing on business resiliency in the wake of unprecedented changes like work from home, shift to online customer engagement, security challenges, and resetting of global supply chain routes. The "new normal" as we now call it would not have been possible had it not been for the unwavering commitment by IT organizations — led by CIOs — ensuring the digital strategy of the business could be fortified and expanded, with many business models introduced during the pandemic now occupying a permanent berth within the organization.

IDC expects this new chapter to play out in the next four to five years; it thrusts IT into a strategic role, to implement a new infrastructure foundation for the business to gain timely and deep insights, at scale. Investments in new workloads will be more significant than those used for corporate IT and other business applications.

In the past several years, these PIC workloads have evolved at an accelerated pace, and an overwhelming majority of respondents in IDC's 2021 *IT Enterprise Infrastructure Survey* agreed that they are important or even critically important to their business. And while some organizations may already be in advanced stages with their investments today, many others are just getting started. IDC's 2021 *AI InfrastructureView Survey* found that only a third of organizations have reached maturity with AI. Many cite lack of proper infrastructure as an inhibitor to deploying these apps at scale. General-purpose infrastructure cannot get the job done.

Investments in Digital Infrastructure Focus on Business Differentiation

According to IDC's Future of Digital Infrastructure predictions, by 2025, 70% of companies will invest in alternative computing technologies to drive business differentiation by compressing time to value of insights from complex data sets. These data sets will be sourced and analyzed by artificial intelligence (AI), Big Data and analytics (BDA), modeling and simulation (M&S), and workloads used in engineering organizations such as electronic design automation (EDA) and computer-aided engineering (CAE).

Introducing Performance-Intensive Computing Infrastructure

IDC is seeing a convergence of infrastructure requirements for the aforementioned use cases and workload groups that require highly performant and scalable compute platforms, network fabrics, and storage systems. In the public cloud, organizations can, in the short term, host these workloads on separate instances. Many of these workloads tend to be batch oriented in nature, and organizations can simply switch off instances when not in use. However, on premises or in colocation facilities, the use of separate infrastructure to host all these workloads can lead to underutilization of resources as well as significant costs to procure and manage.

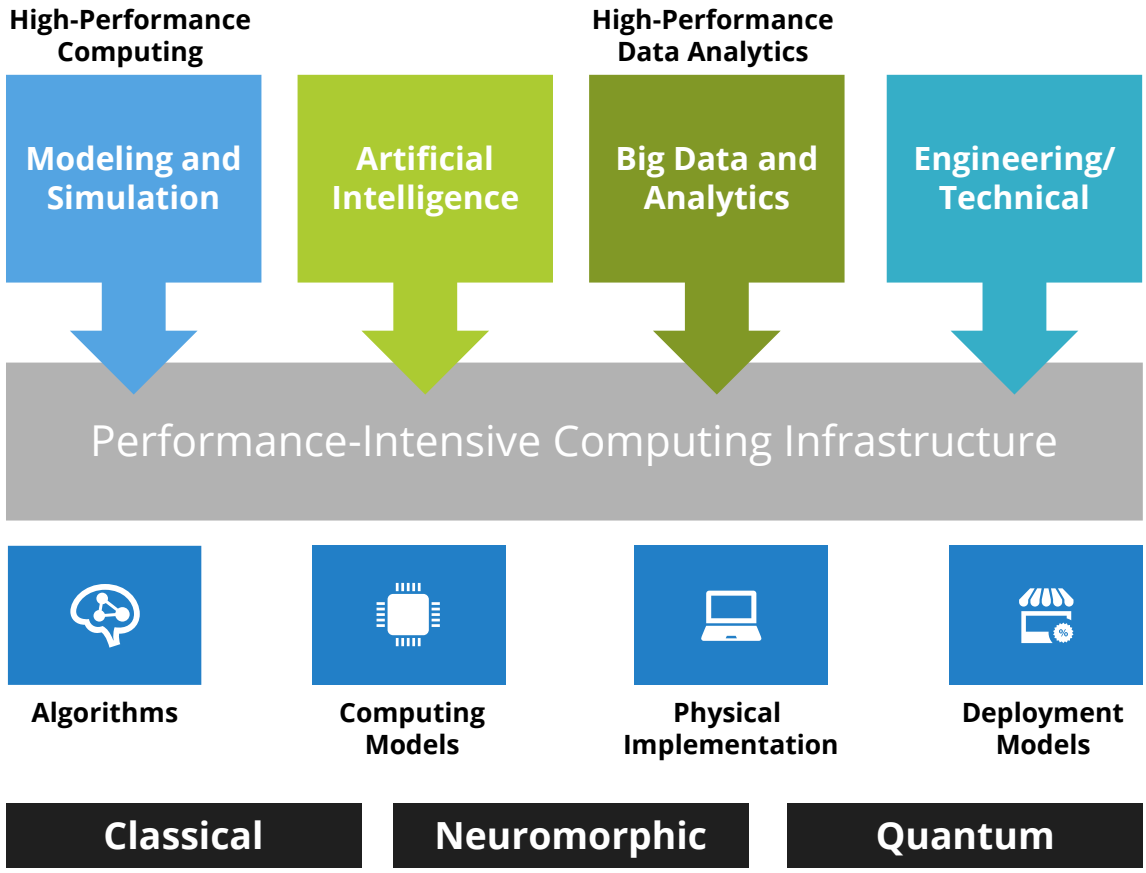
A common infrastructure strategy, therefore, calls for IT organizations to take a "converged" approach to highly performant infrastructure. IDC calls this Performance-Intensive Computing Infrastructure. IDC defines performance-intensive computing as the process of performing large-scale, mathematically intensive computations, commonly used in analytics, machine learning, and technical computing and now increasingly for artificial intelligence and Big Data and analytics in the commercial space.

PIC approaches are also used for processing large volumes of data or executing complex instruction sets in the fastest way possible, often at times with reduced precision.

Figure 1 illustrates a recent trend in IT. As organizations evolve their IT infrastructure, PIC workloads are becoming mainstream, permeating all businesses, with every organizational entity (corporations, governments, universities, nonprofits) evaluating its IT infrastructure and assessing ways to achieve optimal efficiency while scaling performance.

FIGURE 1

Convergence of Workloads for a Common Infrastructure



Source: IDC, 2022

With various sectors and industries undergoing digitalization, they are also investigating current and future data requirements with varying levels of data security, evaluating technology integration and automation while delivering better outcomes for digital transformation, security, and energy efficiency. A multitude of use cases from modeling and simulation to artificial intelligence, analytics, and engineering – where the IT infrastructure requirements are performance and scalability – are becoming increasingly important across all industries; a few examples are explained in the sections that follow.

Artificial Intelligence

Examples of AI use cases include asset/fleet logistics and management, augmented claims processing in the insurance industry, augmented customer service agents across all industries, diagnosis and treatment in healthcare, expert shopping advisors and product recommendations in retail, and fraud analysis and investigation in banking.

Key business objectives: IT automation and improved operations coupled with critical business insights, energy efficiency, and security features

Big Data and Analytics

Examples of BDA use cases (across most industries) include customer relationship analytics applications; end-user query, reporting, and analysis tools; enterprise performance management applications; production planning applications; and supply chain and product analytic applications.

Key business objectives: Improved business and market insights as well as better product development

Modeling and Simulation

Modeling and simulation use cases are diverse. Examples (across most industries) include real-time risk management in banking; asset liability matching in insurance; real-time, high-frequency trading in security and investment services; noise reduction studies in discrete manufacturing; and 3D models for drug discovery in healthcare, mechanical engineering, and fluid and mechanical dynamics simulations.

Engineering and Technical Workloads

Lately, many engineering and scientific (technical computing) applications like electronic design automation and computer-assisted design (CAD) benefit from an increase in compute-intensive infrastructure to get the job done meaningfully.

Key business requirements: Improved business operations, accelerated product development cycles, business automation and efficiency, and new and deep insights into customer behavior

Implementing Performance-Intensive Computing Infrastructure

Operationalizing many of the use cases described previously requires a three-step strategy: technology, organization (people), and process. A common challenge that organizations find when they operationalize many of the new workloads is a lack of understanding of the critical people, process, and technology capabilities required to succeed, especially dealing with challenges associated with managing the life cycle of the apps and associated data. Security and energy efficiency are also rising to the top of the IT agenda.

From a people, processes, and methodology perspective, it requires new process and organizational approaches including investing in collaboration between the business and technical teams, as well as between the development and operations teams. For example, with artificial intelligence, businesses are required to scale their AI operations, which include workflow collaboration, model prototyping and training, model deployment, model performance evaluation, and ongoing model monitoring.

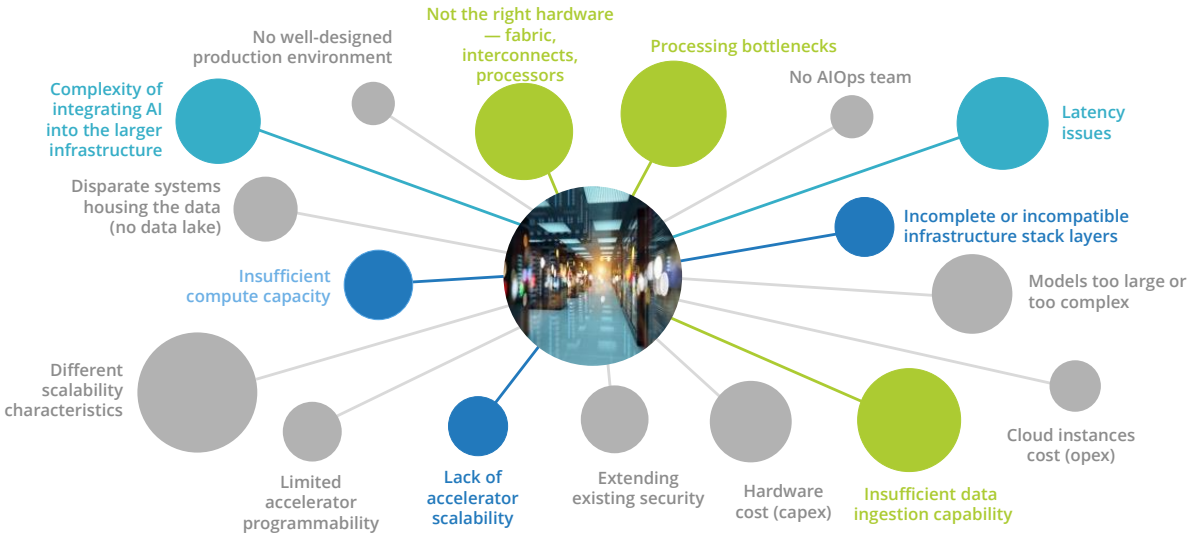
From a technology perspective, there are five areas organizations need to evaluate as part of their PIC-I strategy. They must take a full-stack approach contemplating:

- Software technologies and platforms that deliver base functionality for downstream app developments and help cross the chasm between developers, data scientists, and IT operations teams
- Purpose-built infrastructure that can scale performance to support the burgeoning compute and data persistence requirements of the apps
- Deployment locations for infrastructure to enable ubiquitous consumption and insights across the entire organization
- Security technologies built in at the silicon level, which help organizations take a modern approach to decrease risks to important assets, mitigate malware, and protect against internal and physical attacks
- Technology investments that are aligned to the company's corporate environmental, social, and governance (ESG) goals (By deploying energy-efficient architectures, IT organizations can help achieve these goals while also meeting the needs of the business.)

Infrastructure is one of the most misunderstood and underestimated parts of any PIC workload stack. As Figure 2 shows, it is also one of the main reasons PIC projects fail. For example, IDC research shows that in the early days of AI, little consensus existed among organizations as to what type of infrastructure would be most suitable for running their AI workloads. As a result, they tried everything and often ended up with unsatisfactory results. Cloud service providers and hyperscale datacenters that were implementing AI functionality have the luxury of using vast quantities of general-purpose servers for scaling their AI and HPC workloads. They also have huge IT teams to maintain the service quality of this infrastructure. Most enterprises cannot emulate such approaches, and any attempts to do so often led to their AI initiatives failing in production. The right solution here is to build a more performance-oriented infrastructure for these performance-intensive applications and workloads because running on general-purpose infrastructure will not provide the scale needed to gain meaningful insights and, therefore, competitive advantage.

FIGURE 2

Reasons for PIC Projects Failing



Source: IDC, 2022

The amount of compute that PIC workloads demand is proving insatiable. For example, during the AI development and especially during the model training stage, businesses must find performant and efficient infrastructure, more so for larger and more complex AI models that power modern enterprises while also allowing data scientists to iterate as often as needed without wasting time waiting for training runs to complete. Another example is mainstream HPC. As enterprises start to embrace algorithms like Monte Carlo simulations into their use cases, they are increasingly finding themselves as users of modeling and simulation workloads.

This is where the convergence factor comes into play. IDC is seeing enterprise infrastructure for AI model training evolve from standalone, accelerated bare metal servers to increasingly high-performance, tightly connected server clusters. The most popular AI models, like those for natural language processing, may consist of tens of billions of parameters, which means that, just as with modeling and simulation workloads, performance is key. IDC calls this implementation a massively parallel computing (MPC) architecture, an emerging computing platform and data management architecture that relies on massive parallelization for processing large volumes of data or executing complex instruction sets in the fastest way. MPC architecture is commonly deployed for modeling and simulation, so as enterprises embrace performance-intensive computing, they can leverage the same infrastructure stack. Note that MPC is one technology approach within PIC.

IT organizations and, therefore, the businesses stand to gain from investing in the right infrastructure stack that can meet the performance required today but crucially whose architectural strategy is based on a vision aimed at future performance, flexibility, and responsiveness to the changing needs of the business.

AMD Enables a Shift to Fit-for-Purpose Silicon for Performant Workloads

As a company, AMD has focused on high-performance and adaptive computing for next-generation datacenter infrastructure. It has invested in multigenerational road maps to continue delivering a portfolio of products to build and deliver performance-intensive computing, cloud scale, and hybrid IT solutions. Taking various innovation, packaging, and process approaches, AMD has delivered industry-differentiated capabilities. These features are highly sought after in performance-intensive infrastructure solutions where optimized performance is crucial. In recent years, AMD has also greatly expanded its portfolio to include server central processing units (CPUs), graphics processing unit (GPU) accelerators, SmartNICs, DPUs, FPGAs, and adaptive SoCs. The processors, accelerators, fabric interconnect, and software combined provide a unified system architecture for PIC that can extend all the way to exascale computing. AMD Infinity Architecture empowers system builders and cloud architects alike to unleash the very latest in server performance without sacrificing power, manageability, or the ability to help secure their organization's most important assets, its data.

With the latest additions to the AMD portfolio, the company has taken direct aim at technical computing applications, which tend to be among the more complex and demanding workloads in the datacenter. These applications are typically used for product design and therefore critically important for the enterprise. Some examples are Finite Element Analysis (FEA) and Structural Analysis (SA), used to simulate the design of physical systems; Computational Fluid Dynamics (CFD), used to simulate interactions between physical systems across various applications, from consumer products to aerospace; and electronic design automation tools, used to simulate and optimize chip design.

For these types of applications, large cache and high clock frequency are critical to attaining better performance. More L3 cache helps ensure that critical data is closer to the core, reducing latency in the system. AMD Infinity Fabric, a high-speed chiplet interconnect with 8 memory channels per socket providing 410GBps DRAM bandwidth at peak (for third-generation EPYC platforms), contributes to increasing performance and is the die-to-die connectivity. The demo focused on the verification process in chip design confirmed that 3D V-Cache allows businesses to finish verification much faster, hence enabling them to get to market faster. Businesses can do more testing in the same amount of time, improving the quality of the design. AMD works closely with leading ISVs in the technical computing space to enable their existing applications to benefit from this performance increase. These partners' solutions are fully certified with AMD EPYC processors today.

The Xilinx SoC portfolio, made up of Versal and Zynq series devices, integrates the software programmability of a processor with the hardware programmability of an FPGA, providing superior system performance, flexibility, and scalability. The Alveo family of datacenter accelerator cards provides optimized acceleration for workloads in financial computing, machine learning, computational storage, video streaming, and data search and analytics. The industry's first customizable, programmable, and computational storage drive (CSD), the SmartSSD CSD, dramatically accelerates data-intensive applications by pushing compute to where the data lives. Function-offload accelerator (aka DPU or SmartNIC) offerings include Alveo platforms that provide software-defined hardware acceleration for function offloads and comprehensive solutions that combine network, storage, and compute acceleration onto a single platform.

Supermicro Delivers Flexible AMD-Powered Datacenter Solutions

Supermicro is an AMD premier partner focused on manufacturing datacenter infrastructure in San Jose, California. Since its founding three decades ago, Supermicro has specialized in power-efficient datacenter infrastructure. It set out to develop datacenter infrastructure that could be highly flexible and consistently utilized while being energy efficient. Supermicro understood that achieving these two goals simultaneously meant taking a clean slate approach to system design. It also knew that it required integration at rack level and a ground-up approach that started from highly efficient system boards to provide a total IT solution specifically tailored for workload requirements, instead of providing general-purpose, cookie-cutter infrastructure.

As the computational intensity of processing units (including accelerators such as graphics processing units and high core count central processing units) and power consumption of memory and storage components increased, Supermicro's engineering team maintained a focus on power efficiency. Many of the company's systems are deployed in advanced datacenters and aid in delivering the highest power usage effectiveness while keeping energy consumption at a minimum.

The rapid proliferation of performance-intensive computing workloads such as AI training and modeling and simulation, which rely on high core count processors, high-performance accelerators, and fast memory access, is pushing the cooling abilities of traditional air-cooled platforms and systems. This presents a problem for many CIOs whose investments in innovative infrastructure could directly conflict with their stated environmental, social, and governance objectives.

Supermicro has invested in liquid cooling technologies, including direct-to-chip cooling, immersion cooling, and chiller doors, to solve the challenges of operating infrastructure for performance-intensive computing in a thermally efficient manner. These liquid cooling technologies effectively cool the systems while extending the serviceable life of the infrastructure and enabling the organization to meet its sustainability objectives. In environments where liquid cooling is not warranted, Supermicro can deliver efficient air-cooled solutions using innovative placement of components and effective thermal design.

Supermicro's AMD EPYC CPU-Based Solutions for PIC Infrastructure

Supermicro's AMD EPYC processor-powered datacenter solutions are engineered to deliver optimized performance for artificial intelligence, analytics, modeling and simulation, and engineering workloads. They are designed to be modular and serve as building blocks for a multitude of options and configurations. Design flexibility also allows Supermicro's systems to be used in a common infrastructure for a diverse set of options for precise workload optimizations. Some examples of Supermicro's servers designed for performance-intensive computing workloads are explained in the sections that follow.

Building Blocks Optimized for PIC Workloads

A cluster required for large-scale PIC is typically purpose built and not efficient with general-purpose servers. Instead of introducing a new system for each permutation, Supermicro introduced its universal GPU that is flexible, versatile, and modular, leveraging as many common building blocks as possible to meet the various needs of large-scale training clusters. The system's "future proofed" design allows to standardize on one GPU platform with multiple configurations for all datacenter needs with optimized thermal management.

The open, modular, and standards-based server is built to deliver maximum acceleration power for large-scale AI, machine learning, and HPC workloads. Powered by the latest AMD EPYC processors,

the solution supports four OAM form factor AMD Instinct MI250 GPUs with integrated AMD Infinity Fabric Link technology that provides a total of up to 600GBps of aggregate bandwidth between modules. The Universal GPU System also supports up to 10 x 2.5in. NVMe U.2 or SAS/SATA drives for fast and flexible storage configuration. An optional 1U height module kit increases airflow so the server can accommodate up to 700W per GPU. For a challenging environment with limited power budget, the Universal GPU System also supports liquid cooling. The system is designed to process massive amounts of data to train machine learning models in areas such as video detection, drug discovery, autonomous driving, and robotics.

The 4U 8 PCIe GPU system is a flexible, highly configurable GPU server, powered by dual AMD EPYC 7003/7002 Series processors and having the flexibility to support the fastest GPUs from the leading vendors, including the AMD Instinct MI210 PCIe accelerator and the NVIDIA A100 PCIe GPU. This 4U server also supports AMD Infinity Fabric Link or NVIDIA NVLink Bridge technologies for accelerated GPU-to-GPU connectivity. The system offers flexibility to host up to 8 PCIe form factor GPUs with 16 lanes of PCIe 4.0 connecting directly to each of 8 GPUs, with no latency-inducing switching or bandwidth sharing for intense GPU accelerated workloads. One additional x16 PCIe 4.0 slot powers the fastest 200Gbps InfiniBand interconnect for HPC clusters or dual 100 Gigabit Ethernet ports ideal for AI, ML, and HPC.

Supermicro offers variety of HPC-AI-optimized systems and rack integration with AMD CPUs, used by automaker, chip maker, national lab, and broad enterprise applications.

Hyperdense, Multinode, and Hot-Swappable Systems for PIC

A well-designed PIC system offers the flexibility to scale up or down based on different requirements and needs. Supermicro's SuperBlade system hosts up to 20 individual SuperBlade servers in a single 8U enclosure and features 20 hot-pluggable blade servers. Each SuperBlade node can support one AMD EPYC 7003 Series Processor and up to two single-width GPUs or one double-width GPU. SuperBlade is pushing the envelope in the HPC market by delivering petaflops of performance with the combined innovations of AMD EPYC 7003 Series Processors with AMD 3D V-Cache Technology and AMD Instinct MI210 accelerators. The EDA industry can benefit when running RTL simulations on the SuperBlade powered with AMD 3D V-Cache technology. The 8U enclosure integrates one 200G HDR InfiniBand switch and two 25G Ethernet switches per node in a highly scalable HPC environment that is suited to process distributed ML workloads without losing flexibility or performance. With multiple world record performances, the SuperBlade provides a resource-saving architecture that combines performance, density, and advanced networking tools in one compact build to power all enterprise, cloud, AI/ML, and HPC applications.

Supermicro's Twin family of products are designed for performance-intensive computing applications and high-density environments where you need many discrete servers with high-speed interconnects for networked or clustered operations. These highly customizable multinode systems can be tailored for specific workload requirements. Supermicro's BigTwin hosts up to 512 cores in the 2U four-node server with dual AMD EPYC processors per node. When building a cluster of high-density servers for PIC, the last thing it needs is for a failure to interrupt the workloads. Supermicro's no-cable architecture increases the reliability of the platform. The server midplane connects directly to the power supplies and each server in the rear and to the disk cages in the front. This design eliminates cables and sockets within the chassis, reducing the chance of failure. In the event of a disk or node failure, all components are hot pluggable in both the front and rear of the chassis. Supermicro's TwinPro is another 2U four-node server with single AMD EPYC processors per node.

TwinPro is an economical yet balanced system, which provides adequate density and the processing power that formerly required two processors, thanks to the AMD EPYC processor's high core counts. The single processor per node with the high core density keeps the cost low for large-scale datacenter and application licenses while providing more thermal headroom for high TDP, high-frequency CPUs. The systems also feature optimum airflow for energy-efficient cooling, easy maintenance, and high availability with hot-swappable nodes and redundant power supply modules.

OPPORTUNITIES FOR SUPERMICRO

To differentiate from larger competitors, Supermicro must continue to lead the way in terms of flexibility and agility in the development and management of its product portfolio. It means a hyperfocused execution on corporatewide strategic objectives, which are to:

- Maintain the ability to innovate and execute through rapid R&D and a large engineering organization
- Maintain better price performance and architectural advantages over its competitors as well as its own prior generations
- Focus on time to market for all innovative technologies and components
- Continue to collaborate closely with key architectural partners such as AMD to get the most out of the latter's advancements in their respective technologies

Like many of its competitors, Supermicro is not immune from the headwinds affecting the industry at large. Areas in which Supermicro may face challenges or risks to its business include:

- **Software solutions:** While Supermicro has its own systems management software and open standards such as Redfish API that the company offers in support of its hardware, it currently relies on partners for software solutions (e.g., software-defined storage). Supermicro must aggressively partner here.
- **Services offerings:** Supermicro can raise the awareness of its integration, migration, or consulting services with enterprises seeking to deploy PIC-I. Many other large server vendors leverage these types of practices to expand the total value of their sales and, therefore, achieve higher margins.
- **Expanding channel partnerships:** Supermicro has an extensive list of channel partners for whom it builds systems based on custom specifications. Supermicro should expand these partnerships to include strategic or long-term collaboration.

CONCLUSION

Building and maintaining a modern performance-intensive computing infrastructure are becoming critical success factors for enterprises in most industries. Business objectives and expectations have risen to a point where specific IT services like performance-intensive computing infrastructure are expected to gain strategic importance to key teams. Unexpected outages and downtime can have a direct impact on revenue and customer satisfaction, hence the need for a modern fit-for-purpose server environment. Up until recently, IT decision makers had a perilous situation when it came to the selection of such infrastructure that offered scalability but also did not break budgets. With AMD, and specifically AMD EPYC CPU-based servers from Supermicro, they now have this option. IDC encourages organizations to take a second look at this option. This provides IT staff the ability to optimize their infrastructure from a performance and security perspective while increasing their return of investment.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com