HYPERION RESEARCH

Technology Spotlight

Supermicro-AMD Partnership Creates Strong Momentum in HPC-AI Convergence Market

Sponsored by Supermicro and AMD

Steve Conway and Melissa Riddle
February 2022

HYPERION RESEARCH OPINION

HPC is nearly indispensable at the forefront of AI R&D, not only for established scientific and engineering applications but also for economically important new commercial uses including precision medicine, automated driving systems, cybersecurity, fraud and anomaly detection, business intelligence, affinity marketing, and IoT/edge/smart cities initiatives.

HPC-enabled AI is attracting more worldwide attention recently, because it indicates where the larger, mainstream AI market is likely headed in the future. The biggest gifts HPC is giving to the mainstream AI market are 40-plus years of experience with parallelism and the related abilities to process and move data quickly, on-premises and in more highly distributed computing environments such as clouds and other hyperscale environments. The HPC community is also an important incubator for applying heterogeneous architectures to the growing number of heterogeneous workflows in the public and private sectors.

Through the new commercial use cases, HPC has already entered the mainstream AI market and promises to become increasingly important there. In our newest worldwide study (August 2021), 80% of the surveyed 141 HPC sites reported that they are already running one or more of the commercial AI use cases. This confirms a trend Hyperion Research has been tracking for a decade, where Global 2000 firms are increasingly adopting HPC to support business operations in enterprise data centers (as opposed to the traditional use of HPC for upstream R&D in dedicated HPC data centers).

SITUATION OVERVIEW

The global HPC and AI markets are converging to create a projected $19.9 billion combined HPC server market opportunity in 2025, up from $13.7 billion in 2020. Adding storage, software and technical support revenue lifts the 2025 forecast to $37.8 billion. Hyperion Research forecasts that the AI portion of the HPC server market, including machine learning, deep learning, and other AI methods, will nearly triple (22.8% CAGR) from 2020 to 2025 (Table 1). This is nearly three times faster than the non-AI portion (8.0% CAGR).

TABLE 1

Forecast: Worldwide HPC-Enabled AI (ML, DL, & Other AI) Server Revenue ($M)

	2019	2020	2021	2022	2023	2024	2025	CAGR '20-'25
ML in HPC	667	719	806	1,018	1,213	1,368	1,569	16.9%
DL in HPC	209	263	341	501	692	899	1,133	33.9%
Other AI in HPC	42	57	70	98	129	162	204	29.0%
Total AI Server Revenue	918	1,039	1,216	1,618	2,034	2,429	2,905	22.8%

Source: Hyperion Research, 2021

HPC-Enabled AI Market Has Distinct Requirements

Some applications use analytics alone, but many HPC-enabled AI applications benefit from both data analytics and simulation methodologies. Simulation isn't becoming less important with the rise of AI. This frequent pairing of simulation and analytics indicates that HPC system designs need to be both compute-friendly and data-friendly. Newer designs are starting to reverse the increasing compute-centrism of recent decades and establish a better balance. Newer architectures combine fast processing, fast data movement, and highly responsive storage for a comprehensive and flexible solution.

Servers for HPC-enabled AI use require capable CPUs, paired with GPUs or other accelerators for data-intensive tasks. In 2020, x86-based CPUs represented 93% of all HPC base processors. Hyperion Research projects that x86-based CPUs will remain highly dominant for the foreseeable future, but alternatives, including Arm-based CPUs in the next few years and later RISC-V processors, will establish footholds in the global HPC market.

Overall, AI workloads demand HPC systems that are tightly integrated/dense (hardware-software-storage-packaging), highly available (fast builds), highly reliable (factory-assembled, factory-tested), customizable/flexible, environmentally friendly (Green Initiative), affordable and easy to use, with strong technical support available from the vendor. In addition to providing such support, vendors that flourish in this specialized space also tend to be customer-centric, with a long-term vision aimed at supporting customers' evolving requirements with early availability of state-of-the-art technology.

Key Selection Criteria and Return on Investment

HPC users consistently rank price performance on their desired applications as their top selection criterion for their next HPC server purchase, followed by application performance. Similarly, return on investment for an HPC system is most commonly defined as price/performance on the users' specific applications. Despite the lengthy list of requirements for AI systems, recent survey data shows that AI users within HPC also follow these trends when asked to prioritize their concerns.

Supermicro and AMD have had some significant wins recently because their solutions are designed to address these top requirements for varying workloads and deliver the highest ROI (price/performance on top applications). The Corona system at Lawrence Livermore National Laboratory, for example, features Supermicro's 4U 8 GPU servers, AMD Radeon Instinct MI50 GPUs and AMD EPYC 7002 CPUs. According to LLNL Deputy Associate Director of Computing Jim Brase, "the Corona system is a major advance in our capability for predictive biomedical modeling for COVID-19," and this architecture provided a "performance boost [that] will help the Corona system lead the way in accelerating pandemic response."

Goethe University also recently partnered with Supermicro and AMD when adding to their Center for Scientific Computing, ultimately selecting Supermicro's 4U 8 GPU A+ server with AMD Radeon Instinct MI50 GPUs and AMD EPYC CPUs. According to Goethe University Chair for HPC Architecture Dr. Volker Lindenstruth, this resulted in "a balanced system that can easily be utilized by a wide range of scientists who require a scalable and optimized cluster of fast servers for their research. Especially the design that allows the integration of eight GPUs and up to two 200Gb/s connectivity network cards without any PCI Express switch provides clear advantages for application performance."

The Supermicro-AMD Partnership Has Established Strong Momentum

Both Supermicro and AMD have shown strong growth recently in both HPC and the overall IT space, highlighted by promising developments such as contributions to the exascale race, AI-specific technologies, and GPUs.

AMD's share of the worldwide x86 CPU server market is at historically high levels, leaping to 22.5% in second-quarter 2021. The company has also recorded impressive gains in the HPC subset of that market. Several of the world's leading HPC sites have told Hyperion Research that the technical strength and cost-effectiveness of 3rd Gen AMD EPYC server processors make them hard not to consider seriously.

Many leading HPC sites have advanced from consideration to acquisition with AMD technologies now supporting some of the world's most powerful supercomputers. AMD's strong momentum in the HPC and HPC-enabled AI markets can be attributed to the company's technical prowess and vision that assigns strong importance to these markets. AMD's fabless relationship with TSMC gives the company added flexibility and the opportunity to focus on developing industry-leading technologies, while partnerships with Supermicro and other OEMs provide strong marketing-sales channels.

Supermicro's recent momentum has been just as impressive. Revenue for the firm's 2021 fourth fiscal quarter ended June 30 exceeded $1 billion for the first time and jumped 19% from the company's 2020 fourth quarter.

The partnership with AMD is centered around Supermicro's A+ servers, a broad portfolio of multi-node, blade, GPU server systems for HPC, cloud, and enterprise computing. These servers (e.g., the 4124GS and the high-density 8U 40-GPU SuperBlade system with 200 Gb/s InfiniBand switch) are

typically outfitted for HPC-enabled AI with 3rd Gen AMD EPYC series processors, along with AMD Instinct MI200 (or MI100) series accelerators or NVIDIA A100 GPUs. Specifically, Supermicro's newest Universal GPU system incorporates AMD Instinct MI250 accelerators and future models are expected to support AMD Instinct MI210 accelerators. The company isn't at liberty to name customers, but they already include one of the world's largest social networking companies, a major automotive firm, and a Global 500 manufacturer.

FUTURE OUTLOOK

Hyperion Research forecasts that the HPC market will continue to exhibit robust growth, propelled by the HPC users' perennial need for more computing power and newer factors including the global exascale race, HPC adoption by Global1000 firms, cloud computing and the rise of AI and other data-intensive methodologies. HPC is a bellwether technology for AI, nearly indispensable at the forefront of AI R&D today and showing where the mainstream AI market is likely headed in the not-distant future. AMD and Supermicro, separately and as collaborators, have established strong momentum in the overall HPC market and the superfast-growing AI portion of this market. These companies are well positioned to benefit substantially from the market growth that Hyperion Research projects.

About Hyperion Research, LLC

Hyperion Research provides data-driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology, and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). Hyperion Research provides thought leadership and practical guidance for users, vendors, and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com