Optimized AI accelerated computing solutions utilizing Supermicro's server family

August 7, 2024

Delivering lower total cost of ownership and increased efficiency for enterprises and edge data centers

El Dorado Hills, CA, August 7, 2024 —Blaize, Inc., the AI computing innovator revolutionizing edge computing solutions, is announcing two new AI server systems optimized for inference performance at a lower TCO (total cost of ownership) for enterprises and edge data centers. With these new systems, Blaize further enhances a diverse portfolio of sustainable AI solutions for enterprises looking to address the increasing energy demand of AI workloads in their data centers.

"As data produced at the edge grows exponentially, enterprises realize the benefits of bringing AI capabilities closer to the source, valuing high performance, low latency, and data locality," said Dinakar Munagala, CEO and co-founder of Blaize. "The close collaboration between Blaize and Supermicro has enabled the delivery of optimized AI servers that give enterprises the solutions they need, with increased efficiency and more performance at a lower total cost of ownership."

The Supermicro SuperServer SSG-121E-NES24R server unit has been qualified by Blaize to operate with up to 24 Blaize® Xplorer® X1600E EDSFF accelerator cards, enabling advanced AI inference workloads in a compact and efficient 1U form factor.

The Supermicro A+ Server AS -4125GS-TNRT1 server unit has been qualified to operate with up to 10 Blaize® Xplorer® X1600P-Q PCIe accelerator cards, providing an even more powerful system in a cost-efficient 4U form factor. This system can optionally host third-party training accelerators for enterprises requiring this functionality.

"These powerful and efficient systems from Supermicro and Blaize are the optimal platform to enable enterprises and service providers to deliver AI-enabled video analytics applications at scale," said Harry Woodrow, VP of Sales. "Our customers value performance and efficiency, and by using Blaize powered AI inference servers, systems manufacturers can deliver lower cost of ownership from the edge to the data center."

These new server systems come with Blaize's optimized complete software stack and AI toolkit, including Blaize® AI Studio® and Blaize® Picasso® Analytics. These tools enable fast data preparation and model evaluation, quick configuration and fine-tuning, and the development and tailoring of advanced analytics applications. The entire end-to-end lifecycle of AI/ML applications is covered in one unique platform via a no-code/low-code, visual, and intuitive user interface. Industry-standard APIs and frameworks are also supported, enabling rapid integration with existing customer systems and workflows. These AI server systems solutions are ideal for on-premises and edge data centers and are designed to enable applications such as large-scale complex video analytics, computer vision, and high-performance AI inference.

About Blaize

Blaize provides a full-stack programmable processor architecture suite and low-code/no-code software platform that enables AI processing solutions for high-performance computing at the network's edge and in the data center. Blaize solutions deliver real-time insights and decision-making capabilities at low power consumption, high efficiency, minimal size, and low cost. Blaize has raised over $330 million from strategic investors such as DENSO, Mercedes-Benz AG, Magna, and Samsung and financial investors such as Franklin Templeton, Temasek, GGV, Bess Ventures, BurTech LP LLC, Rizvi Traverse, and Ava Investors. Headquartered in El Dorado Hills (CA), Blaize has more than 200 employees worldwide with teams in San Jose (CA), Cary (NC), and subsidiaries in Hyderabad (India), Leeds and Kings Langley (UK), and Abu Dhabi (UAE).

Media Contact
Leo Merle
Blaize, Inc.
leo.merle@blaize.com