Compute-In-Memory APU Achieves GPU-Class AI Performance at a Fraction of the Energy Cost

October 20, 2025

SUNNYVALE, Calif., Oct. 20, 2025 (GLOBE NEWSWIRE) -- **GSI Technology, Inc. (Nasdaq: GSIT)**, the inventor of the Associative Processing Unit (APU), a paradigm shift in artificial intelligence (AI) and high-performance compute (HPC) processing providing true compute-in-memory technology, announced the publication of a paper led by researchers at Cornell University. Findings confirmed that GSI Technology's APU CIM (Compute-In-Memory) architectures can match GPU-level performance for large-scale AI applications with a dramatic reduction in energy consumption due to high-density and high-bandwidth memory associated with the CIM architecture.

Key findings include:

- **GPU-class performance** – The Gemini-I APU delivered comparable throughput to NVIDIA's A6000 GPU on RAG workloads.
- **Massive energy advantage** – The APU delivers over 98% lower energy consumption than a GPU over various large corpora datasets, underscoring its efficiency and sustainability.
- **Faster and more efficient than CPUs** – The APU's unique design allows it to perform retrieval tasks several times faster than standard CPUs, shortening total processing time by up to 80%.

"Cornell's independent validation confirms what we've long believed—compute-in-memory has the potential to disrupt the $100 billion AI inference market," said Lee-Lean Shu, Chairman and Chief Executive Officer of GSI Technology. "The APU delivers GPU-class performance at a fraction of the energy cost, thanks to its highly efficient memory-centric architecture."

Published on ACM and presented at the Micro '25 conference, the paper by the Cornell research team titled "Characterizing and Optimizing Realistic Workloads on a Commercial Compute-in-SRAM Device," represents one of the first comprehensive evaluations of a commercial compute-in-memory device under realistic workloads. The Cornell-led team benchmarked the GSI Gemini-I APU against established CPUs and GPUs, focusing on retrieval-augmented generation (RAG) tasks over datasets ranging from 10GB to 200GB.

The researchers' findings point to significant opportunities for GSI Technology as customers increasingly require performance-per-watt gains across various industries, including Edge AI for power-constrained robotics, drones, and IoT devices, as well as defense and aerospace applications where the APU can deliver high performance in environments with strict energy and cooling constraints.

Mr. Shu continued, "This tremendous work by Cornell highlights CIM advantages using the Gemini-I silicon. Our recently released second-generation APU silicon, Gemini-II, can deliver roughly 10x faster throughput and even lower latency for memory-intensive AI workloads, while further improving energy efficiency. Looking ahead, Plato represents the next step forward, offering even greater compute capability at lower power for embedded edge applications. The APU's unique combination of speed, efficiency, and programmability positions us to unlock high-growth opportunities across edge AI, data centers, defense, and other markets where energy efficiency is a critical strategic advantage."

The Cornell study also introduced a new analytical framework for general-purpose compute-in-memory devices, providing optimization principles that strengthen the APU's position as a scalable platform for developers and system integrators. A copy of the publication can be found on the GSI website at https://gsitechnology.com/characterizing-and-optimizing-realistic-workloads-on-a-commercial-compute-in-sram-device/.

ABOUT GSI TECHNOLOGY
Founded in 1995, GSI Technology, Inc. is a leading provider of semiconductor memory solutions. GSI's resources are focused on bringing new products to market that leverage existing core strengths, including radiation-hardened memory products for extreme environments and Gemini-I, the associative processing unit designed to deliver performance advantages for diverse artificial intelligence applications. GSI Technology is headquartered in Sunnyvale, California, and has sales offices in the Americas, Europe, and Asia. For more information, please visit www.gsitechnology.com.

About ACM

ACM publishes more than 50 scholarly peer-reviewed journals in dozens of computing and information technology disciplines. Available in print and online, ACM's high-impact, peer-reviewed journals constitute a vast and comprehensive archive of computing innovation, covering emerging and established computing research for both practical and theoretical applications. ACM journal editors are thought leaders in their fields, and ACM's emphasis on rapid publication ensures minimal delay in communicating exciting new ideas and discoveries.

Forward-Looking Statements

The statements contained in this press release that are not purely historical are forward-looking statements within the meaning of Section 21E of the Securities Exchange Act of 1934, as amended, including statements regarding GSI Technology's expectations, beliefs, intentions, or strategies regarding the future. All forward-looking statements included in this press release are based upon information available to GSI Technology as of the date hereof, and GSI Technology assumes no obligation to update any such forward-looking statements. Forward-looking statements involve a variety of risks and uncertainties, which could cause actual results to differ materially from those projected. These risks include those associated with the normal quarterly and fiscal year-end closing process. Examples of risks that could affect our current expectations regarding future revenues and gross margins include those associated with fluctuations in GSI Technology's operating results; GSI Technology's historical dependence on sales to a limited

number of customers and fluctuations in the mix of customers and products in any period; global public health crises that reduce economic activity; the rapidly evolving markets for GSI Technology's products and uncertainty regarding the development of these markets; the need to develop and introduce new products to offset the historical decline in the average unit selling price of GSI Technology's products; the challenges of rapid growth followed by periods of contraction; intensive competition; the continued availability of government funding opportunities; delays or unanticipated costs that may be encountered in the development of new products based on our in-place associative computing technology and the establishment of new markets and customer and partner relationships for the sale of such products; and delays or unexpected challenges related to the establishment of customer relationships and orders for GSI Technology's radiation-hardened and tolerant SRAM products. Many of these risks are currently amplified by and will continue to be amplified by, or in the future may be amplified by, economic and geopolitical conditions, such as changing interest rates, worldwide inflationary pressures, policy unpredictability, the imposition of tariffs and other trade barriers, military conflicts and declines in the global economic environment. Further information regarding these and other risks relating to GSI Technology's business is contained in the Company's filings with the Securities and Exchange Commission, including those factors discussed under the caption "Risk Factors" in such filings.

Source: GSI Technology, Inc.

Contacts:
Investor Relations
Hayden IR
Kim Rogers
541-904-5075
Kim@HaydenIR.com

Media Relations
Finn Partners for GSI Technology
Ricca Silverio
(415) 348-2724
gsi@finnpartners.com

Company
GSI Technology, Inc.
Douglas M. Schirle
Chief Financial Officer
408-331-9802

Source: GSI Technology, Inc.