GSI Technology Reports 3-Second Time-to-First-Token for Edge Multimodal LLM Inference on Gemini-II

January 29, 2026

Benchmark Results Demonstrate Fast Multimodal Edge Inference with Up to ~300% Better Performance per Watt versus Competitive Solutions

SUNNYVALE, Calif., Jan. 29, 2026 (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 processing, providing true compute-in-memory technology, today announced preliminary benchmark results for the Gemini-II Compute-in-Memory processor. These results demonstrated 3-second time-to-first-token ("TTFT") performance for multimodal large language models operating at the edge with video and text inputs.

Using the Gemma-3 12B vision-language model on GSI's production Gemini-II processor, GSI achieved the 3-second TTFT while consuming approximately 30 watts at the AI sub-system, including the chip. To GSI's knowledge, this 3-second TTFT at approximately 30 watts at the AI sub-system is the lowest publicly reported result for a multimodal 12B model running on an embedded edge processor.

Independent third-party testing of the same workload on competitive embedded platforms reported TTFT measurements of roughly 12 seconds on Qualcomm Snapdragon X Elite with 30W power, and 3 seconds on NVIDIA Jetson Thor with over 100W power. With performance on par with or superior to competitive platforms at lower power usage levels, GSI concludes that Gemini-II offers a favorable responsiveness and power-efficiency profile for power- and thermally-constrained edge environments.

"These benchmark results highlight what compute-in-memory can enable for physical AI," said Lee-Lean Shu, President and Chief Executive Officer of GSI Technology. "Edge deployments require fast response under tight power and thermal limits. A 3-second TTFT means the system can generate an initial response every three seconds, which is generally fast enough to be useful in video-based applications without missing meaningful events. Gemini-II's ability to deliver low-latency multimodal inference at low power supports a broader range of real-time applications, from autonomous systems to intelligent machines operating outside the data center."

GSI believes this performance profile is well-suited to "physical AI" markets, including drones, smart city, and other edge systems where workloads are episodic and constrained by battery life, thermal design, and form factor. Faster TTFT at lower chip power can enable more responsive systems, longer duty cycles, and lower total system cost.

Edge physical AI represents a growing segment of AI compute as workloads shift from cloud-assisted models to local inference to improve latency, reliability and operational efficiency. GSI's proprietary compute-in-memory architecture is designed to reduce data movement, which is a primary contributor to latency and power consumption in conventional architectures.

GSI's engineering team continues to work on further optimizing Gemini-II's responsiveness while collaborating with customers and partners, including G2 Tech, on system integration and proof-of-concept activity. Benchmark results are intended to support ongoing evaluation and do not guarantee future commercial outcomes.

ABOUT GSI TECHNOLOGY

GSI Technology is at the forefront of the AI revolution with our groundbreaking APU technology, designed for unparalleled efficiency in billion-item database searches and high-performance computing. GSI's innovations, Gemini-I® and Gemini-II®, offer scalable, low-power, high-capacity computing solutions that redefine edge computing capabilities. 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.

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, strategies, products, market opportunities and prospective customer engagements. 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 expected or implied.

GSI Technology's participation in a proof-of-concept is exploratory in nature and may not result in any commercial contract, extended engagement, or recurring revenue. There can be no assurance that the scope, performance, or findings of any proof-of-concept will meet customer expectations or commercial requirements, or that such activities will lead to further business opportunities, order volume, or deploy-at-scale implementations. Additional risks and uncertainties that could cause actual results to differ materially from those expected or implied include, among others: the preliminary and limited nature of benchmark results; differences in workloads, configurations, measurement boundaries, and methodologies that can materially affect TTFT and power measurements; variability in model architectures, versions and toolchains that may impact performance; the pace and extent of adoption of "physical AI" at the edge and the impact of safety, privacy, and security requirements; supply-chain constraints affecting semiconductors, components, or manufacturing partners; 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 its 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 its products; 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 its 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 its 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, export controls and other trade barriers, military conflicts, particularly in relation to Taiwan, and a challenging global economic environment. These risks are discussed in more detail in GSI Technology's most recently-filed Annual Report on Form 10-K, its Quarterly Reports on Form 10-Q and its other reports filed from time to time with the SEC. You are urged to review carefully and consider GSI Technology's various disclosures in this press release and in its reports publicly disclosed or filed with the SEC that attempt to advise you of the risks and factors that may affect its business.

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.