GSI Technology's Gemini-I® Demonstrates High Performance in Billion-Scale Similarity Search Benchmark Results

April 6, 2020

Gemini provides latency of 1.25 milliseconds with recall greater than 90% in a query-by-query search

SUNNYVALE, Calif., April 06, 2020 (GLOBE NEWSWIRE) -- **GSI Technology, Inc. (Nasdaq: GSIT)**, a leading provider of memory solutions for the networking, telecommunications and military markets, and developer of Gemini®, the Associative Processing Unit (APU), announced the publication of billion-scale similarity search benchmark results for Gemini-I, the first member of GSI Technology's Associative Processing Unit Gemini product line.

The paper published by GSI Technology introduces the Gemini® Associative Processing Unit (APU) and presents its role in an Approximate Nearest Neighbor (ANN) similarity search pipeline. Latency and recall numbers for query-by-query ANN searches are presented for the DEEP1B dataset, a public dataset consisting of 1 billion, 96-dimensional vectors with each dimension being a 32-bit floating point (FP32) number. The Gemini APU can efficiently handle either batch mode or query-by-query requests.

GSI Technology is the first chip vendor and solution provider to benchmark query-by-query latency on datasets as large as one billion items, such as DEEP 1B. The unique feature of this benchmark is the presentation of query-by-query results rather than batch mode numbers. In most real-world applications, online requests arrive and are processed one-by-one, otherwise known as query-by-query, where having the lowest possible latency is critical. The DEEP 1B dataset test with 4 Gemini APU boards running at a rate of 400 MHz delivered latency performance results of 1.25 milliseconds with recall of 92.5%. To GSI Technology's knowledge, this is the first published record of near 1 millisecond latency with this level of accuracy on such a large dataset.

Lee-Lean Shu, Chairman and Chief Executive Officer of GSI Technology, commented, "One of the key challenges in the era of big data is managing similarity search in applications where databases scale to over one billion items. Low latency and high accuracy are critical for many online applications working with very large datasets. Gemini APU has demonstrated its capabilities to improve an approximate nearest neighbor (ANN) similarity search solution for the billion-scale problem with low latency and very high recall, even for the difficult query-by-query task. Gemini-I also provides high-quality results with an exceptionally small system footprint and low power usage. GSI Technology is focused on applications, including e-commerce, recommendation, drug discovery and drug toxicity, facial recognition, signal classification and object detection, and cryptography, where Gemini-I can deliver significantly better outcomes than current solutions. We will begin offering remote cloud-based workshops to demonstrate Gemini-I functionality in these applications in the near future."

The Gemini APU complements existing hardware that performs similarity searches, such as the CPU, by offloading a significant portion of the similarity search pipeline. The Gemini APU performs vector management and partitions the database across multiple cards in a distributed architecture. The result is a similarity search solution that scales to databases of billions of items.
To efficiently perform similarity search at scale, the Gemini APU leverages techniques such as database clustering and data compression in such a way that yields low latency and high-quality results.

To receive a copy of the Company's publication of these results, please visit the GSI Technology website at https://www.gsitechnology.com/APU.

ABOUT GSI TECHNOLOGY

Founded in 1995, GSI Technology, Inc. is a leading provider of semiconductor memory solutions. GSI's resources are currently focused on bringing new products to market that leverage existing core strengths, including radiation-hardened memory products for extreme environments, and Gemini, the APU 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.

Contacts:
Investor Relations:
Hayden IR
Kim Rogers
385-831-7337

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

Source: GSI Technology, Inc.