In-Memory Acceleration for Big Data

By Linley Gwennap
Principal Analyst

July 2020

www.linleygroup.com

In-Memory Acceleration for Big Data

By Linley Gwennap, Principal Analyst, The Linley Group

Standard CPUs can easily perform complex calculations on small datasets, but they struggle to handle larger datasets, particularly for simple tasks such as search. GSI Technology has created a new type of processor that combines memory and compute units on a single chip. By placing the computation close to the data, the new Gemini accelerator achieves more than 100x the performance of a standard Xeon server on certain algorithms, even while using less power. GSI Technology sponsored this white paper, but the opinions and analysis are those of the author. Trademark names are used in an editorial fashion and are the property of their respective owners.

Introduction

Companies are increasing analyzing very large data sets to uncover new insights, a technique often referred to as big data. For example, once factory equipment can report status on line, manufacturers can sift through months of detailed data to track product defects, improve yields, and even predict when failures will occur. As these datasets become ever larger, however, traditional processors bog down, slowing the analysis.

These processors may have high compute capacity, but they are intended for working on smaller datasets. A large dataset quickly overflows the processor's local memory and must be stored in the large on-chip memory (cache). Unfortunately, the transmission rate (bandwidth) from this memory to the processor cores is relatively low, creating a bottleneck, as Figure 1 shows. The bottleneck slows down the processor cores, which can process data only as fast as they receive it.

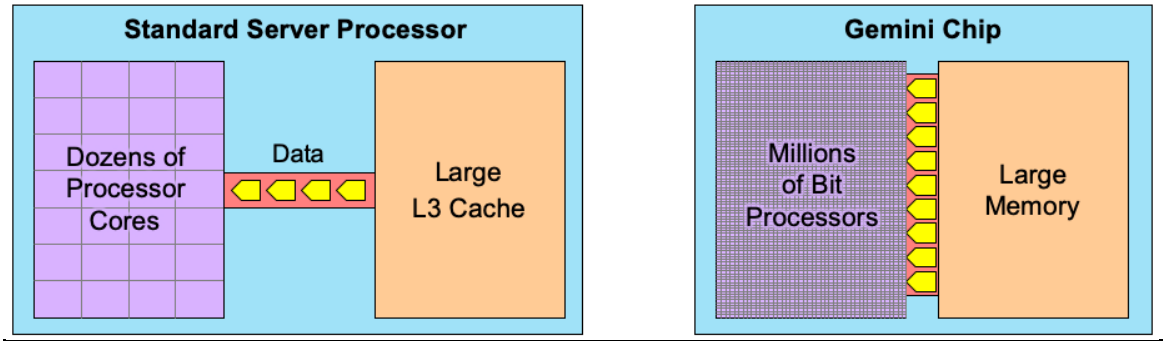

Figure 1. Breaking the memory bottleneck. Traditional processors (left) struggle with large datasets due to the narrow connection with the large on-chip memory. Gemini features millions of cores that can all access memory at once, allowing a much greater flow of data.

Although processor vendors haven't been able to break this bottleneck, one memory vendor has developed a unique solution: GSI Technology's Gemini, a dual-function processor with memory storage capability. As Figure 1 shows, this design includes millions of processors that can each load data directly from the on-chip memory. Dividing this memory into enough subunits to feed each processor allows for much greater data flow. Furthermore, this on-chip connection is extremely short, reducing the

power needed to transfer the data. As a result, Gemini outperforms standard processors by 100x or more on certain big-data workloads while reducing power by 70%.

GSI has two decades of experience in supplying high-performance memory chips, both SRAM and DRAM. To develop the Gemini chip, it combined a new processor design from its MikaMonu acquisition with its fast SRAM technology. To simplify deployment, GSI sells this chip as part of a full-height half-length PCIe board, called Leda-G, that plugs into standard servers. The company also provides a software stack that allows customers to run different types of search algorithms using the chip; they can also develop custom algorithms using C++ or Python. The rest of this paper explains the technology in more detail and how it might apply to your application.

Gemini Technology

A standard server processor, such as a Xeon chip, must be able to run a wide variety of tasks, resulting in a large and complex design. Gemini is designed for one task: analyzing large amounts of data. To accomplish this feat, GSI created a simple processor that could be inserted in bulk into the read-modify-write lines of an SRAM chip so that all of the processors can work in parallel without being starved for data. In fact, each Gemini chip has more than two million tiny bit-processing units. This approach creates a tremendous compute capacity: 105 trillion 8-bit logic operations per second (TOPS) at 400MHz. In contrast, the largest Xeon chip can perform nearly 10 TOPS at 2.4GHz.

For large datasets, Xeon can't come close to its peak performance because of the memory bottleneck. Gemini, however, interleaves the processing units with the memory cells; as Figure 2 shows, the chip looks like a standard memory chip, since the tiny processors aren't visible at this scale. This approach allows data to flow directly from the memory cells into the processing units, so each processor can access data every cycle. In fact, the total data rate into the processors is a massive 26 TB/s.

By operating so many processors in parallel, the chip can quickly analyze all the data stored in the chip, which holds 96Mbits (12MB). For example, to perform a similarity search using Hamming distance, the chip first loads the search pattern into all two million processors simultaneously. Then each processor loads one database entry from SRAM, calculates the Hamming distance between that entry and the search pattern, and compares it to the minimum distance found. In tens of microseconds, each processor tests 24 entries, and the entire database has been searched.

To handle larger datasets, the Leda-G card includes 16GB of DRAM that connects to the Gemini chip. Typically, the software stores an index to the database in the Gemini chip, performs a fast search to identify which section of the larger database holds the target item. Then the chip can load the appropriate section from DRAM and perform a second fast search. Even larger datasets can be divided among multiple cards so each card can search in parallel. For example, the company has successfully tested a billion-entry database using four Leda-G cards. GSI provides software that can coordinate several cards in a single server.

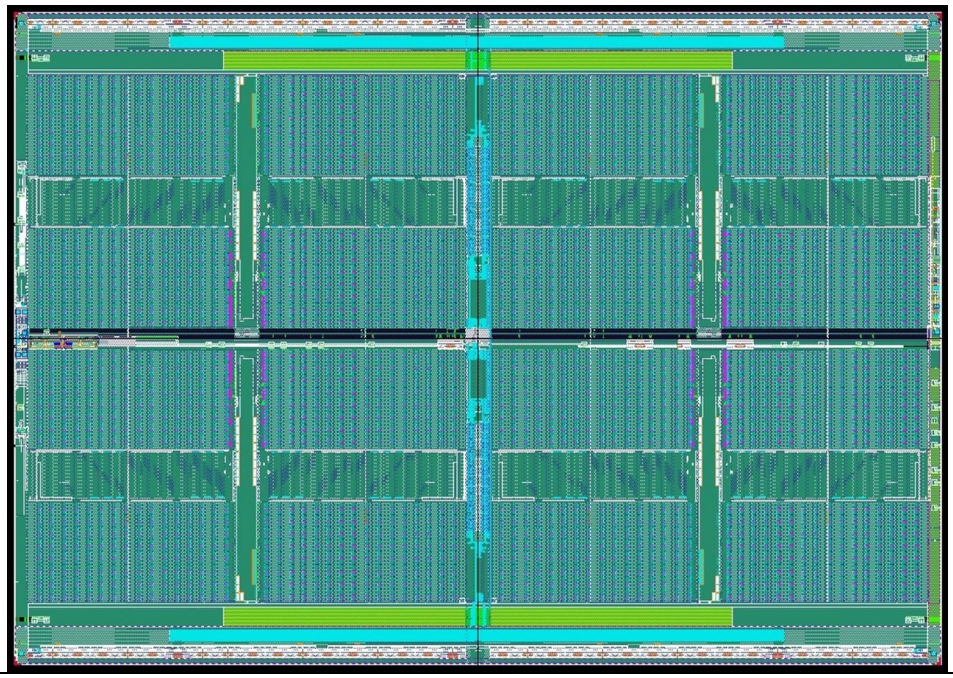

Figure 2. Gemini chip photo. Rather than combining a large processor and a memory on one die, the design interleaves more than 2 million tiny processing units among 48 million memory cells, creating a highly regular structure.

Advantages Over Competition

Gemini is classified as an APU (associative processing unit) to contrast with CPUs, GPUs, and other common processor types. By specializing its design and dividing the computation among many small cores, Gemini easily outperforms other types of processors. For example, a top-of-the-line Xeon 8280 processor has only 28 CPU cores, versus Gemini's 2 million, that generate 60% less performance. The Xeon has three levels of cache; the first generates about 10 TB/s of bandwidth (across all 28 cores) but holds less than 1MB. This is 60% less bandwidth to 90% fewer bytes than in Gemini. The x86 processor has a large L3 cache, but the bandwidth to this cache is even less: only 1TB/s, or 95% less than Gemini's. To add insult to injury, the Xeon 8280 consumes more than 200W. Mainstream Xeon processors have even less memory, compute performance, and bandwidth.

GPUs provide another alternative. For example, Nvidia's newest A100 "Ampere" chip includes 104 GPU cores that can each process 4,096 bits at a time, resulting in impressive performance of 624 TOPS. But this performance is achieved only for a specific math operation (matrix multiplication) used in neural networks and other scientific applications. For general operations, the chip produces 75 TOPS, as Table 1 shows. Furthermore, the data rate from the GPU's large L2 cache is 7TB/s, still only 27% of Gemini's. For rapid data searches, the GPU will bottleneck on this limited bandwidth. The Nvidia chip

requires even more power than the biggest Xeon chip. Nvidia's less expensive V100 chip provides less than half the compute power and bandwidth of the A100 with only 6MB of L2 memory.

	Gemini APU	Xeon 8280	Nvidia A100	Graphcore
Compute Cores	2 million x 1 bit	28 x 2x512 bits	104 x 4,096 bits	1,216 x 64 bits
Compute Speed	400MHz	2.7GHz	1.4GHz	1.6GHz
Peak Compute*	25 TOPS	10 TOPS	75 TOPS	16 TOPS
On-Chip Memory	12MB L1	38.5MB L3	40MB L2	300MB L1
Mem Bandwidth	26 TB/s	1 TB/s	7 TB/s	16 TB/s
Power	60W TDP	205W TDP	400W TDP	150W TDP

Table 1. Processor comparison. The APU provides more memory bandwidth than any other type of processor, allowing it to quickly analyze large datasets. *Trillions of 8-bit ADD operations per second (TOPS). (Source: vendors)

Startup Graphcore has developed a chip that has even more processors: 1,216 cores. Like the GPU, however, this chip is optimized for neural networks and only achieves its peak rate of 250 TOPS for matrix multiplication. For general-purpose calculations, the cores can reach 16 TOPS, as Table 1 shows, 40% less than Gemini. The Graphcore chip has a large amount of internal SRAM that can deliver 16TB/s to the cores; in this case, the bottleneck in a general-purpose application is the limited compute power.

A high-end FPGA such as the Xilinx Ultrascale+ contains a large number of DSP cores that can perform basic math and logic functions along with nearby "block RAM" that feeds data into the cores. The largest Ultrascale+ model (VU13P) has 12,288 DSP cores that can compute a total of 33 TOPS at 775MHz. This model also has 12MB of block RAM that delivers 17TB/s of data at that speed—34% less bandwidth than Gemini. Xilinx offers a PCIe card containing the VU13P for $12,995 that burns 225W TDP. The big disadvantage of an FPGA is that the end user needs to create a logic design to configure the chip for their particular task.

Performance Examples

GSI tested its design on an image-recognition database (Deep1B) containing one billion records totaling 96GB.[1] Each record contains 96 image features that GSI hashed into 768-bit vectors. This database is too large to store in Gemini's on-chip memory, so the company sorted the database into 512K groups of similar images, then created an index in which each entry represented the average (centroid) characteristics of a group. Each 12MB Gemini holds 128K 768-bit centroids, so the index requires four chips. Once the centroid index is loaded into these chips, the system can continuously perform searches.

To recognize a new image (face), the CPU sends the image parameters to the four cards, which calculate the Hamming distance for each index record. Hamming distance is simply the number of different bits within two binary values. In less than 100 microseconds, each chip returns the most-likely group among its portion of the index. The

[1]For a detailed description of this test case, refer to the GSI Technology white paper "Gemini APU: Enabling High-Performance Billion-Scale Similarity Search" at www.gsitechnology.com.

host processor then searches the 2,000 records in the indicated groups to find the best match. The entire process requires 1.25ms, enabling 800 queries per second.

Many other applications can use this approach. For example, biomedical and pharmaceutical research often involves identifying molecules (stored in Oracle Database) based on their characteristics. To accelerate this recognition process, GSI tested a database of 38 million molecular "fingerprints." This database can be represented in different ways, for example as 512 bits per fingerprint in 2.3GB, or 1,024 bits per fingerprint in 4.6GB. In either of these cases, a single Leda-G card can hold all 38 million records in DRAM. GSI's Oracle Database driver allows the customer to continue using the same application software while using the Gemini technology to find the *k* nearest neighbors (KNN) in the database. The Leda-G card can handle even larger fingerprints, such as 2,048 or 4,096 bits, that take too long to run on CPUs, providing a new capability in the field.

To perform a KNN search, the Gemini chip computes the Tanimoto similarity (a more complex operation than Hamming) for 12MB of data while simultaneously loading the next 12MB from DRAM. In this way, it can cycle through the entire dataset of 38 million fingerprints in 1.1 seconds for 512-bit fingerprints or 2.1 seconds for 1,024 fingerprints. To improve performance, the chip can process multiple queries at a time before reloading from DRAM; for example, 100 parallel queries take a total of 1.41 seconds, or 71 queries per second. Another test involved a 680 million-entry dataset containing 512-bit fingerprints. The dataset size was 43.5MB and split across four Leda-G boards. This search takes 4.5 seconds. By comparison, this dataset would require more than one hour when processed on an Oracle server without the GSI accelerators, as Figure 3 shows.

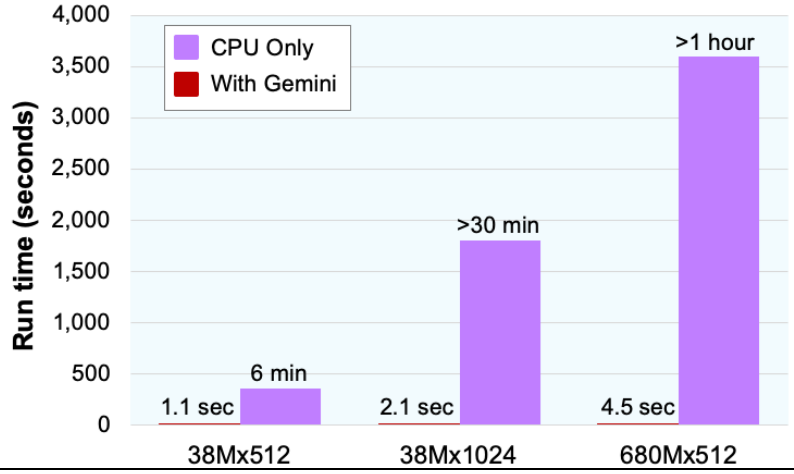

Figure 3. KNN search comparison. On this common search function, a dual-socket Xeon Gold 5115 server using Gemini accelerator cards is hundreds of times faster than a standard Oracle server. GSI tested datasets with different record counts (38 million and 680 million) and record sizes (512 and 1,024 bits). (Source: GSI Technology)

Gemini is also efficient for the SHA hash function, which combines many shifts and other bit operations with modular addition. For example, GSI tested a 1U server with 16 Gemini chips running the 256-bit SHA1 algorithm. This system performed 5.4 million hashes per second—greater throughput than a 4U server holding eight Nvidia V100 cards while using half the power. Applications that repetitively apply SHA, such as

Hashcat, can also take advantage of the chip's high memory bandwidth. Other applications of the Gemini technology include RF signal classification and identifying the likelihood of toxicity of medical compounds.

Conclusion

GSI has produced a new type of processor, called APU, that is optimized for performing simple tasks on large amounts of data. In contrast, a CPU is optimized for performing complex tasks on small amounts of data. The CPU has a small number of powerful general-purpose compute cores, but these cores become bottlenecked on workloads that require a simple computation (such as Hamming distance) on each data item. The Gemini technology implements millions of tiny bit-processors that can each load data directly from memory, resulting in about 100x more data per second than the CPU can achieve. This design can quickly search or analyze a large dataset.

Other types of processors, such as GPUs and AI accelerators, offer greater memory bandwidth than CPUs but still fall short of Gemini. Furthermore, these chips are optimized for matrix multiplication and other scientific computation; they don't perform as well for simpler database analysis. FPGAs also fall short of Gemini's compute and bandwidth capability, and they require specialized design skills. These alternatives all use more power than Gemini and are typically more expensive.

GSI has demonstrated Gemini's performance when running similarity searches on datasets as large as 1 billion records and expects the technology will scale to 40 billion records and beyond by combining multiple systems with multiple cards. Gemini delivers orders of magnitude better performance than a single processor, or it can deliver large cost and power savings relative to the number of processors required to match its performance.

In addition to the image-recognition and molecular-identification examples discussed above, applications such as e-commerce, natural-language processing (NLP), and visual search can take advantage of Gemini's ability to quickly search a large dataset. But Gemini is a fully programmable architecture that can handle a variety of tasks, and GSI offers a complete programming environment, including a Python/C++ library, to assist customers. Any big-data workload that is limited by memory bandwidth on a traditional processor might benefit from this technology, particularly if the compute requirements for each data item are modest or include many bitwise operations. For these workloads, an army of specialized processors are far better than a handful of muscular ones.

Linley Gwennap is principal analyst at The Linley Group and editor-in-chief of Microprocessor Report. *The Linley Group offers the most-comprehensive analysis of microprocessors and SoC design. We analyze not only the business strategy but also the internal technology. Our in-depth articles cover topics including embedded processors, mobile processors, server processors, AI accelerators, IoT processors, processor-IP cores, and Ethernet chips. For more information, see our website at www.linleygroup.com.*