Persistent Memory for Artificial Intelligence and Machine Learning Applications

Data centers should take advantage of persistent memory to eliminate bottlenecks and accelerate performance in Artificial Intelligence and Machine Learning technologies.

By Arthur Sainio, Director of Product Marketing, SMART Modular Technologies

In today's enterprise datacenters, limited memory capacity and the input/output (I/O) performance of mass storage are the two biggest causes of bottlenecks. These two pain points have historically been perceived as different computing concepts: memory is a temporary store of code and data to support a running application, while discs and other persistent storage hold data on a long-term basis. When an application needs to access data from disc (which happens frequently with large data sets that cannot be held in memory), the slow access imposes a significant penalty on the application's performance. The introduction of persistent memory has marked a turning point in the traditional data center memory and storage hierarchy through the possibility of a new unified hyper-converged architecture that dramatically accelerate enterprise storage server performance.

The Growth of AI and ML Applications

The explosion of data has resulted in huge growth in Artificial Intelligence (AI) and Machine Learning (ML) applications, but traditional systems are not designed to address the challenge of accessing these large data sets. The key hurdle for AI and ML applications entering the IT mainstream is reducing the overall time to discovery and insight based on data intensive ETL (Extract, Transform, Load); and checkpoint workloads. AI and ML create highly demanding I/O and computational performance for GPU accelerated ETL. Varying I/O and computational performance is driven by bandwidth and latency. The high-performance data analytics needed by AI and ML applications require systems with the highest bandwidth and lowest latency.

According to the International Data Corporation (IDC) Worldwide Artificial Intelligence Spending Guide, spending on AI and ML systems will reach $97.9 billion in 2023, more than two and a half times the $37.5 billion that will be spent in 2019. In turn, the data processing required needs to keep up with this expansion will be exponential in growth. Conventional memory solutions today lack the vital component to answer this push: non-volatility, even as parallel architectures are being designed to answer future data needs. However, while these architectures are being refined, power losses could cost data centers millions of dollars. Hence the immediate need for non-volatile memory.

Moving Non-Volatile Memory Closer to the CPU

Checkpointing is a process where the state of the net being trained is stored to ensure that the result of the learned data is not lost. Checkpointing is a particular challenge for AI and ML applications because it wastes processing capacity and burns a lot of power, without directly offering a benefit to the application itself. Processing in other nodes may also be halted when writing data to a central store. The operation is also write-intensive, compounding the problem in some situations as conventional storage such as hard drives are inefficient when data is written to them.

As checkpointing to a central memory can significantly re-duce the speed to insight in AI and ML applications, engineers are moving non-volatile memory closer to the CPU to minimize the impact of this essential process. This produces a better balance between data and compute, enabling the system to deliver the overall production needs.

NVDIMMs in AI and ML Applications

Persistent memory, in the form of NVDIMM (a Non-Volatile Dual-Inline Memory Module), is being used to increase the performance of write-latency sensitive applications, effectively providing a persistent storage model with DRAM performance. Data centers have a unique opportunity to take advantage of NVDIMMs to achieve the low latency and increased performance requirements of AI and ML applications without major technology disruptions.

When NVDIMMs are plugged into a server, they are mapped by the BIOS as a subsection of persistent memory within main memory. The application is then free to use this persistent memory for high-speed checkpointing. The alternative is the traditional approach in which the checkpointing data is transferred through the I/O stack, over NVMe and then saved to an SSD. This system incurs the latency penalty of the I/O stack and the NAND Flash.

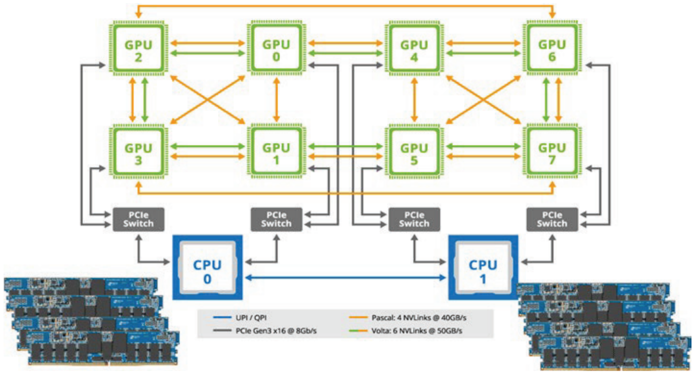
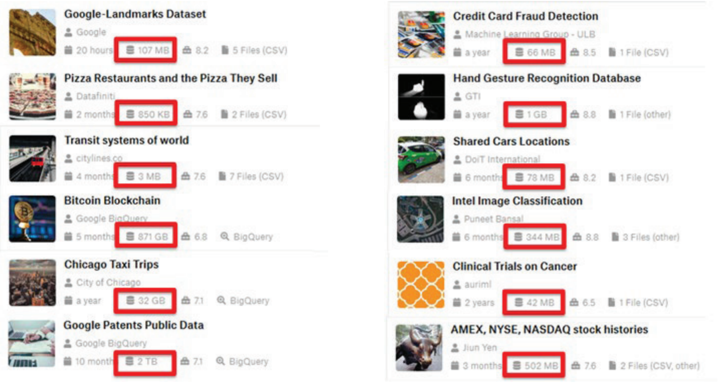


Figure 1. Four 32GB NVDIMMs are used for each CPU providing a total of fast byte-addressable persistent memory.

NVDIMMs are an ideal solution for high-performance AI and ML servers. Data intensive ETL and checkpointing workloads can use the persistent memory region of main memory, allowing them to operate at DRAM latencies (<100ns) and DRAM bandwidth (25.6GB/s).

While NVDIMMs are used to accelerate checkpointing for AI applications, they can also be used for ML to increase performance and protect data being collected by algorithms. GPU configured storage servers run algorithms which are part of simulation and ML. NVDIMMs are used to protect the GPU servers from losing simulation data. Typical algorithm data set sizes vary from Kilobytes (kB) to Terabytes (TB), and lost data would cause a need to restart work. When four servers are con-figured with NVDIMMs, dataset sizes up to 1TB can use persistent memory, as opposed to traditional storage, to dramatically improve performance without risk of losing data.

The most common method used to process AI, ML and simulation datasets (which all have similar characteristics) is for the datasets to come through the network via InfiniBand or Ethernet into the AI/ML server then cached into the SSD to eliminate the risk data loss. Portions of the datasets are then moved to DRAM by the GPU where the calculations can be performed. An example of this process would be performing calculations on a dataset to determine if the data represents a picture of a dog or cat. Once the calculation is completed the response is sent back out to the network. If there is a system crash during this process all calculations are lost. By switching to NVDIMMs, this process can be dramatically streamlined. There is no need to cache the incoming datasets into the SSDs. The datasets can be moved directly to DRAM where the GPU can immediately start its calculations. The response to determine if a specific dataset represents a picture of a dog or cat can occur magnitudes faster. At the same time, there is no risk of losing the datasets or the calculations because the NVDIMMs are persistent.



Google-Landmarks Dataset
Google
20 hours 107 MB 8.2 5 Files (CSV)

Pizza Restaurants and the Pizza They Sell
Datafiniti
2 months 850 KB 7.6 2 Files (CSV)

Transit systems of world
citylines.co
4 months 3 MB 7.6 7 Files (CSV)

Bitcoin Blockchain
Google BigQuery
5 months 871 GB 6.8 BigQuery

Chicago Taxi Trips
City of Chicago
a year 32 GB 7.1 BigQuery

Google Patents Public Data
Google BigQuery
10 month 2 TB 7.1 BigQuery

Credit Card Fraud Detection
Machine Learning Group - ULB
a year 66 MB 8.5 1 File (CSV)

Hand Gesture Recognition Database
GTI
a year 1 GB 8.8 1 File (other)

Shared Cars Locations
DoiT International
6 months 78 MB 8.2 1 File (CSV)

Intel Image Classification
Puneet Bansal
6 months 344 MB 8.8 3 Files (other)

Clinical Trials on Cancer
auriml
2 years 42 MB 6.5 1 File (CSV)

AMEX, NYSE, NASDAQ stock histories
Jiun Yen
3 months 502 MB 7.6 2 Files (CSV, other)

Figure 2. Examples of machine learning datasets ranging from 850KB to 2TB.

NVDIMMs are not only well suited for AI and ML applications, they can also be used in financial applications commonly referred to as FinTech. FinTech applications demand high performance (reducing latency and increasing transaction rates) because time is money. Processed transactions need to be logged synchronously before the next transaction can be started. This synchronous function, while critical for auditing, also creates a significant bottleneck for many systems, slowing the transaction velocity. By utilizing NVDIMMs the current process of logging data to SATA or NVMe SSDs can be eliminated. Instead of sending logging data through the I/O to the Flash SSD, the logging data can be put directly into high-speed DRAM made persistent with the use of NVDIMMs. The NVDIMMs enable the system to begin the next transaction with the confidence that the previous transaction is logged to a secure location with no risk of data being lost.

While NVDIMMs have been around for more than a decade, the benefits of using this type of persistent memory for AI and ML applications is still being investigated by various sectors from banking and retail to discrete manufacturing, process manufacturing, healthcare and professional services. The support ecosystem for NVDIMMs including the operating systems, hardware enablement and JEDEC standardization was the result of many companies working together to adopt persistent memory. NVDIMMs are intersecting with the growth of AI and ML to provide an ideal way to increase system performance.