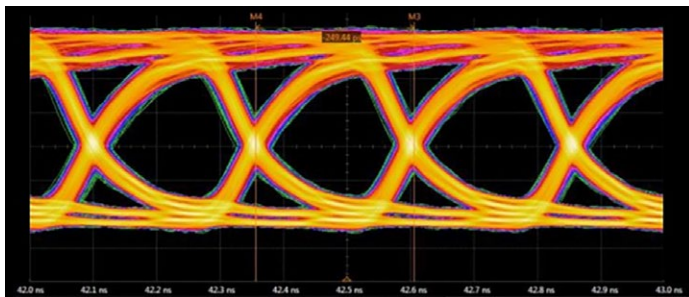


raised the HBM2E data rate to 3.6 Gbit/s, and demonstrated performance in silicon to 4 Gbit/s. At 3.6 Gbit/s 'per pin', a HBM2E memory device can achieve 461 GByte/s, and a next-generation AI accelerator with six such devices can achieve a whopping 2.765 TBytes/s of bandwidth.

Design challenges

The biggest challenges are the physical design between the interposer and the DRAM. To get the HBM2E system to work correctly requires investing a lot of time optimising that channel. The thousands of data lines running through the interposer are single-ended signals. So, they are very susceptible to crosstalk and insertion loss. This channel is also very resistive because it is going through silicon. The task becomes one of balancing all the parameters that affect signal integrity, such as managing the crosstalk, resistive channel, insertion loss, and reflections. Significant simulation work is required to make sure that those signals are going to be nice and clean.



bus' standard interface license, a reference design is provided which includes the interposer. Support is then provided to help the customer iterate through that design to meet actual system requirements.

The industry needs more speed – as well as addressable memory if it is to keep up with the exponential growth in workloads such as AI/ML training.

A little over a year ago, the largest training model had 10 billion parameters, now the largest is over 100 billion. For the past decade, the pace of increase has been 10X per year so AI/ML training models with a trillion parameters can be expected in 2021 or 2022. The training models of today can take days or weeks to run. More bandwidth and capacity are key to accelerating that process.

Taking HBM2E to 3.6 Gbit/s puts aggregate system bandwidth within reach of 3 TB/s, a huge step up over the best-in-class systems of today. The industry is working on designs that continue to push the limits of performance with innovative PHY, controller, and interposer designs.

Whatever the challenges, it is clear that new generations of HBM will emerge to deliver the bandwidth at the network edge and in the data center, to accelerate the next wave of technology.

Rambus understands these challenges well. It has long focused on signal integrity and has worked closely with numerous customers not only on integrating its interface (PHY and controller) designs but also assisting with interposer channel and package design. These elements are seen as a significant design risk as it is often the case that chip engineering teams do not have deep experience of HBM design. As part of Ram-

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 per-

formance 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 reduce the speed to insight in AI and ML applications, engineers are moving non-volatile memory closer to the CPU to minimise 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

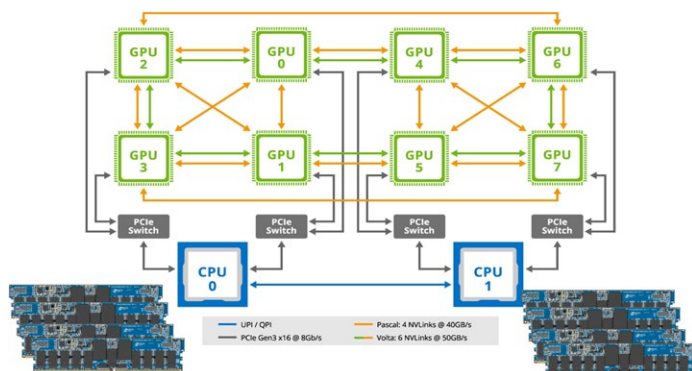


Figure 1. Four 32GB NVDIMMs are used for each CPU providing a total of fast byte-addressable persistent memory.

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.

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 configured 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.

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

Professional Power

InPOTRON™

Schaltnetzteile
Switch Mode Power Supplies



Industrial applications

Switch Mode Power
Supplies for industrial
requirements

- customized development
- efficiently
- smart



Customized Solutions
Made in Germany

www.inpotron.com

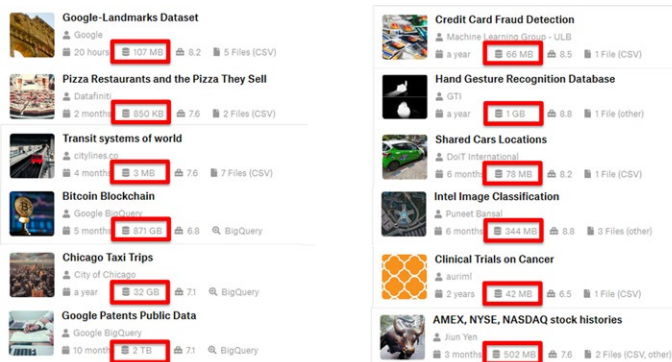


Figure 2. Examples of machine learning datasets ranging from 850KB to 2TB.

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 audit-

ing, also creates a significant bottleneck for many systems, slowing the transaction velocity. By utilising 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.

Multi-channel 10bit digitizer with up to 4 GS/s sampling

Flexible 10bit PCIe digitizer with software-configurable channel count and sampling rate up to 4Gsamples/s Teledyne SP Devices in Sweden has developed a modular data acquisition board with configurable channel count and sampling rate. The PCI Express digitizer complements the previously released ADQ8-8C by offering a higher sampling rate and software-selectable two- or four-channel mode of operation. The high channel density, flexible mode of operation, and open FPGA architecture make it suitable for large-scale physics installations and Original Equipment Manufacturer (OEM) product integration. The programmable analog front-end (AFE) supports multi-purpose operation and can therefore be used with a wide variety of detectors and in applications such as particle physics, scientific instruments, time-of-flight applications, and more. "ADQ8-4X extends our 10-bit product portfolio by offering a flexible analog front-end and faster sampling at two different rates. This helps our customers

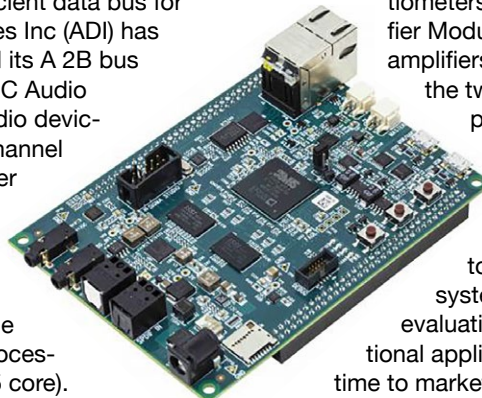


optimize channel count, sampling rate, and cost," said Jan-Erik Eklund, Digitizer Product Manager at Teledyne SP Devices. This allows large multi-channel systems to use a combination of different digitizer models with simultaneous acquisitions on a large number of channels distributed over many chassis with a timing alignment of better than 200 picoseconds and extended over time. Additional capabilities of the ADQ8-4X include an AFE with programmable channel count, sampling rate, DC-offset, and input voltage range, 10bit resolution with 2 GS/s sampling rate in 4-channel mode and 4 GS/s in 2-channel mode as well as 1 GHz analog input bandwidth. It supports open Xilinx FPGA with resources available for customized real-time digital signal processing along with 1 Gbyte of onboard acquisition memory and hardware trigger and highly accurate multi-channel synchronization capabilities. There is extensive software suite including easy-to-use evaluation/integration software Digitizer Studio

Flexible
[www](http://www.teledynesp.com)

Complete audio system opens up A2B bus

A2B inventor Analog Devices (ADI) has introduced an end-to-end prototyping system that drives the efficient data bus for applications beyond the car. Analog Devices Inc (ADI) has launched a full audio system based around its A 2B bus technology. The system features the SHARC Audio Module (SAM) for the creation of digital audio devices, including audio FX processors, multi-channel audio systems, MIDI synthesisers, and other DSP-based audio systems. In addition to the main SHARC audio module board, ADI offers daughter boards to provide added functionality to the main board and expand the audio system. SAM includes the dual-SHARC+ core ADSP-SC589 audio processor SoC (with an integrated Arm Cortex-A5 core). In addition to the main SHARC audio module board, the company offers daughter boards to provide added functionality to the main board and expand the audio system.



The Audio Project Fin board mates directly to the main board, providing MIDI input/output as well as pushbuttons and potentiometers to modify audio effects. The A 2B Amplifier Module features two high-efficiency Class-D amplifiers to output digital audio data received over the twisted-wire pair A 2B bus from PDM microphones and/or serial TDM sources on the main board (or another connected A 2B node). This complete audio system delivers high-fidelity multichannel digital audio with low and deterministic latency to a fully synchronised distributed audio system. The system is aimed at fast prototyping, evaluation projects, demonstrations, and educational applications. It enables users to realise a shorter time to market with a "ready to go" prototype system that provides a comprehensive hardware and software solution.

Analog Devices (ADI)
www.analog.com