AI Acceleration Drives Architectures to Focus on Memory Solutions

AI and the Evolution of Industrial Systems

Decades ago, an industry evolution was birthed by the application of statistical analysis, model generation, and digital/analog control systems. In the next few years, there will be yet another leap forward in industrial systems; this time, the leap will likely have several magnitudes greater impact than previous stages of industrial systems.

Myth: Faster compute will solve all AI bottleneck problems

Fact: Solutions must be architected with memory and storage in mind

The AI revolution isn't only a natural step forward in industrial systems, it is a necessary solution to an unexpected bottleneck in the emergence of the data economy: the inability to store volumes of data and timely process it to create intelligence to execute decisions. This stems from the growing abundance of data being captured and collected.

Where previous industrial systems relied on experts for every level of decision making and depended on advancements in data processing speeds and bandwidth, the innate parallelism of future AI architectures will place a greater burden on the design and performance of memory and storage than ever before. Keeping pace with these new requirements—for both sensors and AI systems used from the edge to the data center—will drive competitiveness in the data economy.

This progression has become a challenge for solutions providers as AI and machine learning technologies continue to morph, disrupting expectations of their requirements and use cases. Therefore, it is becoming critical for memory and storage hardware to be designed with AI in mind to account for these shifting requirements and enable the future of intelligent systems.

This white paper discusses the value and constraints of modern and next-generation AI. It's urgent to discuss how memory and storage systems can enable enhanced AI performance necessary for scaling the growth of the new data economy.

AI/Machine Learning: Value and Constraints

Currently, the abundance of data is an untapped resource, one in which AI systems could more efficiently tap than today's data processing systems. Data is being acquired in larger amounts, with faster speeds, at lower latencies, and at higher resolutions than ever before. AI and machine learning algorithms can be trained and learned from real data sets, as opposed to the last generation's theory and approximation approach to handling complex systems.

The new capabilities of AI systems remove the need for lengthy analysis and development time by a large pool of highly trained experts, instead enabling an organic learning system that improves through experience. Moreover, AI systems can be designed to make decisions rapidly based on real-time data and can find connections amongst vast amounts of seemingly unrelated information.

Given the sheer enormity of the information being generated and the need for data processing at every level, distributing AI systems with a wide range of requirements and capabilities is the only likely solution to handle the diversity of information in the modern world. Hence, it will become necessary to employ AI at the edge to process and determine the value of data before either sending the pre-processed information to the cloud or sending the raw data to a larger and more capable cloud AI for further processing and storage. At the edge, latency is key, whereas in the cloud, it will be more dependent on parallelism, mass storage, and processing large amounts of data.

In these new use cases for AI, memory and storage play an even more critical role than previous generations of data processing technologies. Past data processing focus has been geared toward high-throughput CPUs, as they were the bottleneck for earlier generation data processing. With AI systems, the speed and bandwidth of memory, along with the speed and volume of data storage, have become the limiting factors in leveraging AI and enabling more complex machine learning algorithms and intelligent systems.

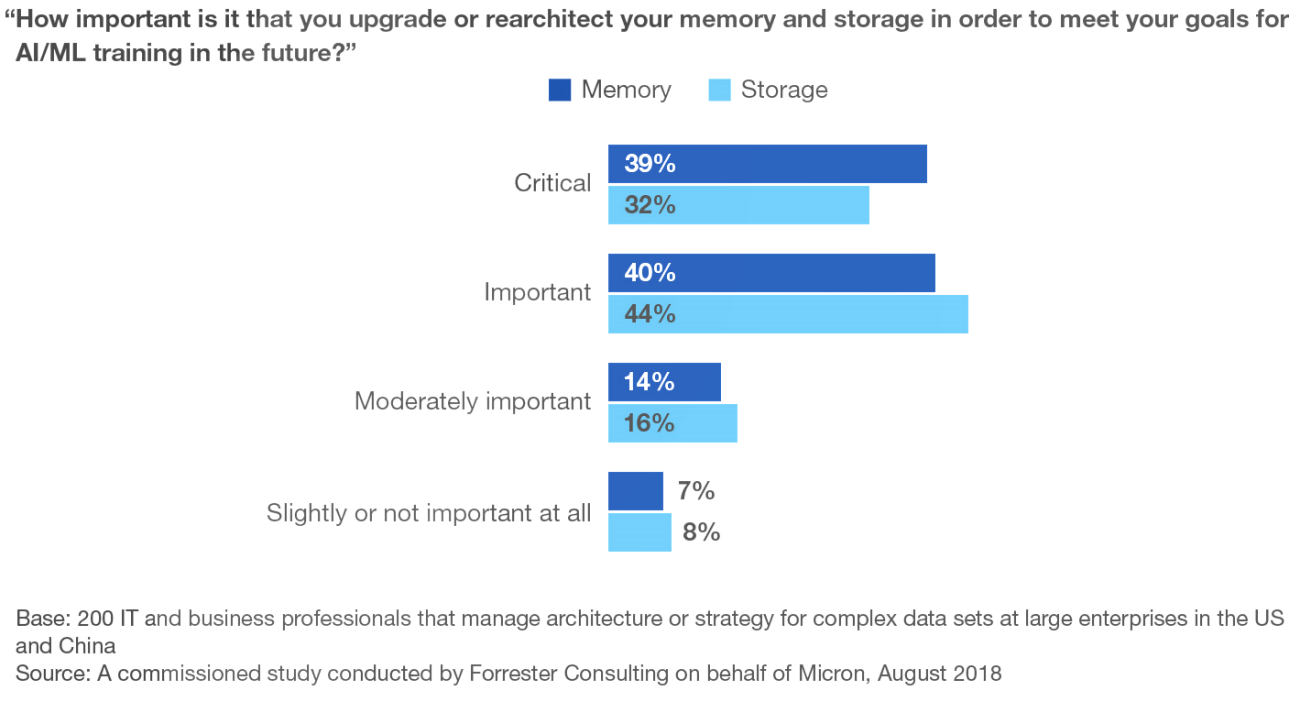

Figure 1: Importance of memory and storage in AI/ML training

Solution: Memory and Storage Enable AI

The diversity and evolving applications of AI are driving memory and storage requirements in a variety of ways; specifically, at the AI training phase, processing at the edge, and for AI implementations in the cloud. Key memory performance factors include bandwidth, density, latency, power, and cost. Each AI application will require a balance of these factors, also determined by the type of memory or storage being used.

Training AI requires a large amount of storage with fast read throughput. Faster storage read throughput enables improved training time, and higher storage density allows for larger training data sets to be used. These ultimately impact the accuracy and effectiveness of an AI system. Additionally, high bandwidth memory with high density is also an enabling factor for improving training time and enabling larger models to be used during training for higher precision computations and greater AI accuracy. The benefit of using larger datasets for training and the abundance of data, enhanced storage and memory density, along with throughput, directly relates to the fidelity of an AI's performance and its ability make decisions in nuanced scenarios.

On the other hand, the requirements for AI inference, and where the AI is in the information/decision hierarchy, place a different requirement on memory and storage. For real-time applications, such as autonomous vehicles, communications, security, and other edge applications, power and latency become more important than bandwidth and throughput. Battery powered and mobile devices have limited size and power to support powerful memory and storage. Moreover, not as much or as high performing storage is needed at the edge as compared to the cloud or for AI training.

For distributed AI applications at the edge, cost versus performance is a trade-off that also must be made. Higher cost and performance memory may not be viable for mass market AI applications. Conversely, cloud AI and critical infrastructure AI will likely warrant a higher cost and higher performance memory and storage.

Myth: AI workloads are focused only in the datacenter and cloud

Fact: AI is happening everywhere, from the datacenter to the edge to the endpoint, where the growth at the edge is seeing the largest ramp

Myth: All memory and storage are equal

Fact: Micron has the portfolio and expertise to help customers optimize AI solutions

As the AI landscape matures, there will be opportunities to optimize memory and storage solutions. From datacenter and cloud to smart edge and intelligent endpoint devices, the industry will experience a transformation in how AI resources are deployed. The concept of "data gravity" will pull AI computing that is typically done in the datacenter into the edge, closer to where data is collected. In Figure 2 below, Micron sees varying memory and storage requirements depending on where you are in the landscape and the AI task being performed.

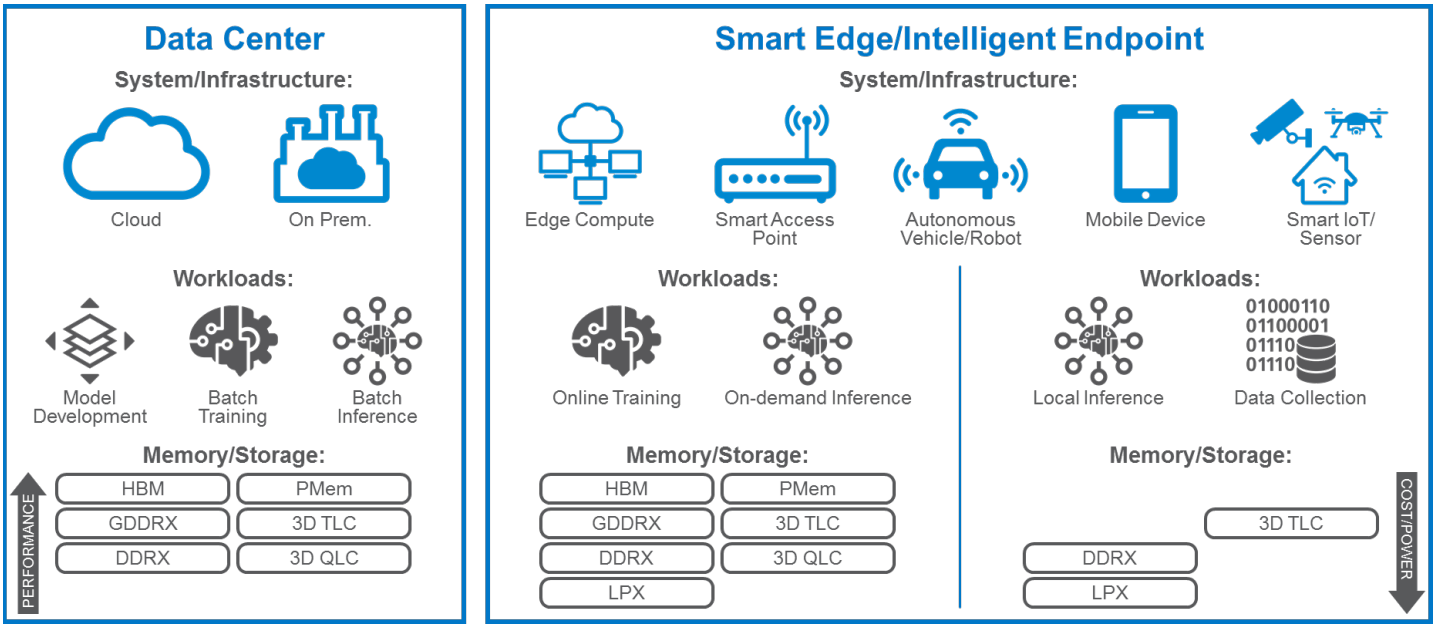

Figure 2: Different tasks require different memory and storage

Advancing Memory Architectures Provide Opportunities for AI

Currently, there is an evolution of memory architectures, each with their own place serving future AI applications.

DDR4 to DDR5

DRAM memories, such as DDR4, are a workhorse for datacenter and server applications. Given the incentive to leverage greater amounts of data, the transition from DDR4 to DDR5 is seen as a greater leap in performance and density over previous DRAM generations. It is likely that the majority of future datacenters, many of which will be powering AI and communications applications, will be leveraging DDR5 memory. The main reasons for this are the predicted performance enhancements of DDR5 over DDR4. Foreseeable DDR5 advantages include:

- Increased DDR channels from 12 to 16, enabling much higher density memory chips (64GB to 128GB).
- Higher operating frequencies and bus efficiency.
- Increased bank groups.
- Improved refresh schemes.
- Supporting features that enable reliable performance at higher operational modes.

LPDDR4 to LPDDR5

Similar to the transition to DDR5, the LPDDR4 to LPDDR5 transition is also set to boost memory bandwidth, while reducing runtime power use and utilizing features to reduce overall power consumption. Where DDR5 will be found mainly in datacenters and PCs, LPDDR5 will target edge AI applications and mobile devices where size, cost, and power are the most significant constraints. Performance and power advantages may drive LPDDR5 beyond traditional industries and applications.

GDDR5/5X to GDDR6

GDDR memory differs from DDR memory because graphics applications require much greater parallelism and lower latency as opposed to memory density. GDDR5 has been around for nearly 10 years and is now giving ground to GDDR6. GDDR6 is proving to be the near-term workhorse for AI, ready and enabled today.

Beyond raw speeds, the standard GDDR6 indicates speeds up to 16 Gb/s with bandwidth to 72 Gb/s per chip. GDDR6 is also featuring a reduced operating voltage of 1.3V compared to the GDDR5/5X 1.5V standard. GDDR6 will also feature 2 channels as memory sizes double that of GDDR5.

Compared to DDR5, GDDR6 is focused as a higher performance solution, which targets AI applications requiring low latency and very high bandwidth. Hence, GDDR6 will more likely be used for AI hardware, integrated into critical systems that need to reduce latency, and where cost is less of a constraint.

Near Memory: HBM/HMC to HBM2/HBM3

High bandwidth memory (HBM) and hybrid memory cube (HMC) are 3D DRAM technologies designed to overcome the memory bandwidth bottleneck of other 2D DRAM technologies. With stacked DRAM chips, HBM chips feature additional I/O ports, increasing the memory bandwidth almost proportionally with the number of stacked chips.

Increased memory bandwidth that is closer to the processing core (typically a GPU or FPGA) can further reduce overall latency compared to 2D memory technologies. HBM memories also benefit from lower power consumption and a smaller 2D footprint than 2D DRAM solutions.

The latest generation of HBM (HBM2) features improved memory speed, bandwidth, and density. A similar trend is likely for HBM3, although it will feature less maximum memory capacity compared to GDDR5/5x/6. Moreover, HBM chips, due to the uniqueness of the manufacturing process and additional silicon involved, are substantially more expensive than other DRAM technologies.

It is likely that AI applications with the greatest need for near-memory bandwidth, speed, low latency, and compact footprint will leverage HBM2/HBM3 memory. This could include complex AI training and critical inference applications that require the utmost in performance.

Conclusion

Although current AI systems are in the early development, prototype, and proof-of-concept phases, the next few years will experience the birth of mainstream AI in industrial systems. These upcoming AI systems will be the decision making and analysis services in cloud systems, at the edge, and those serving the critical networking operations in-between. AI solutions have traditionally focused on compute capability and now realize that architects must design AI systems with memory and storage in mind. The next generation of memory and storage technologies are key to alleviating the bandwidth, latency, density, power, and cost bottlenecks which would otherwise limit future AI applications.

micron.com