AI and Machine Learning Demand Fast, Flexible Infrastructure

From an academic subject which evolved into a set of technologies — many of which are now embedded and offered in commercial applications — AI has become a market force, transforming every sector of the economy. And yet, experts believe we are only in the early days of AI. To be clear, researchers have aspired to "general AI" since the 1960s. Hopeful business use cases in the 1980s got buzz but didn't gain much traction. Then, about 2010, three IT trends finally advanced mainstream analytics/AI as pragmatic enablers for commercial business:

1. New mathematics uncovered bolder algorithms, which enable machine learning systems to crunch through data sets, even defining their own data sets, efficiently learning as they go.
2. New affordable and effective data storage and processing, from 3D NAND flash, to GPUs, to cloud computing, capture and deliver the massive data required for training.
3. Fast, vast data moved off hard disk drives (HHDs) onto accelerated solid state drive (SSD) flash storage, so no longer does AI/ML require specialized labs; and organizations now prepare and monetize data streams for ML.

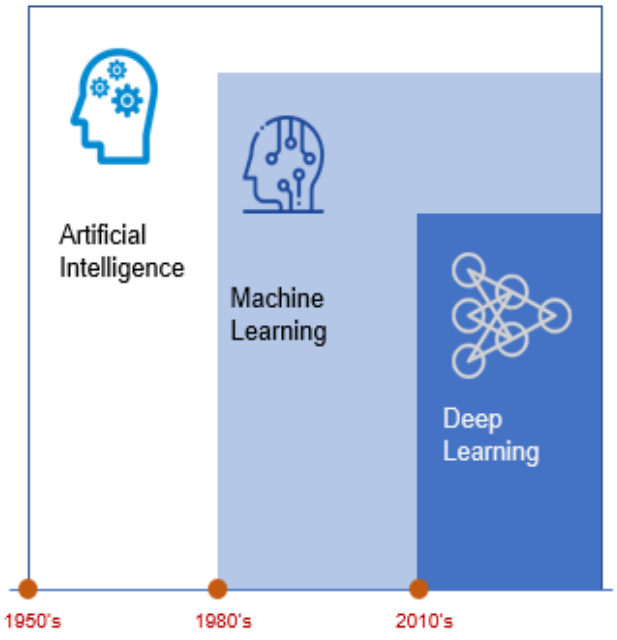


Figure 1: AI – ML – DL Timeline

Why Micron Technology for AI/ML/DL

AI enables insights from knowledge and uncovers value rapidly and at scale. Enterprise, healthcare, customer experience, and smart cars/homes are all implementing AI/ML/DL. Meanwhile, Micron memory and storage have been foundational in AI's transformation to these highly adaptable, self-training, ubiquitous, machine-learning systems for mainstream use.

Mainstream AI demands a new generation of faster, intelligent, global infrastructures. Micron's fast, vast storage and high-performance, high-capacity memory and multichip packages power AI training and inference engines, whether in the cloud or embedded in mobile and edge devices. Leaders look to Micron flash storage, DRAM, GDDR graphics memory, FPGAs, and other innovative memory technologies to help them serve up huge volumes of data in real time to training AI systems, accelerate inference with high-performance, low-power memory for AI chips, enable edge devices with memory and storage to keep them smart, and more!

Artificial Intelligence (AI) is about creating software that thinks about problems like a human does.

Machine Learning (ML), a type of AI, is a branch of data science that uses data sets to train machines on statistical methods of analysis.

Deep learning (DL), a subset of ML, is about the breakthrough algorithms based on multi-layered neural networks that can learn (the more layers, the more complexity they can capture) and have enabled AI machines to get smarter more autonomously.

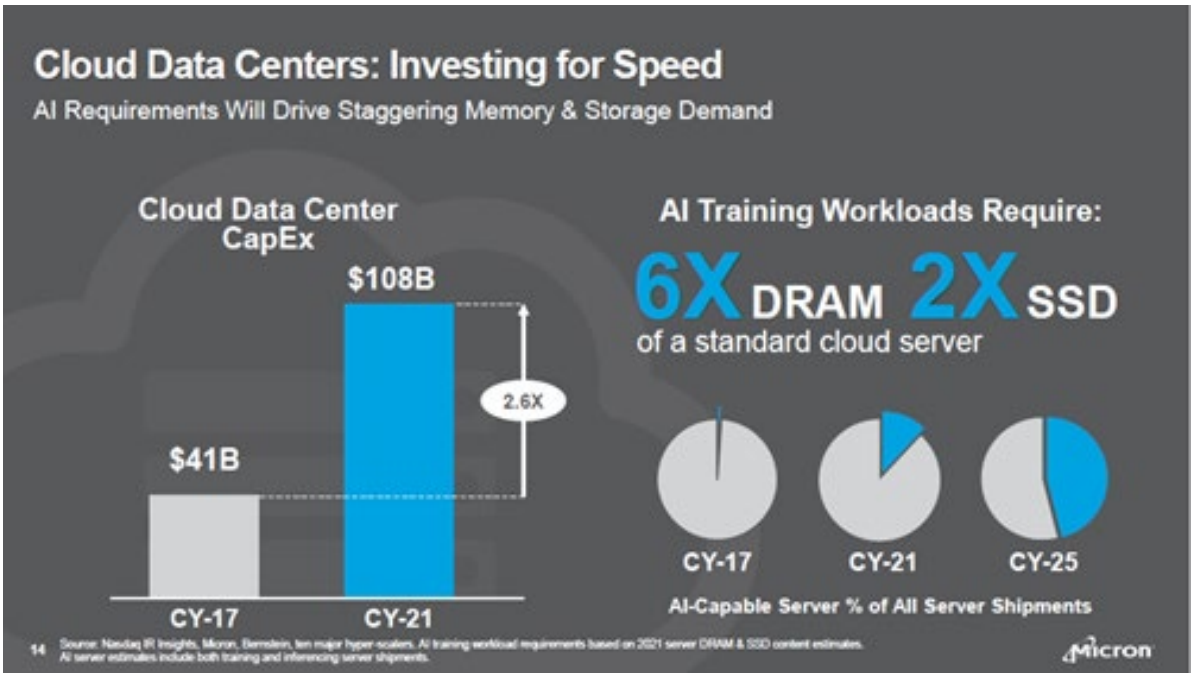


Figure 2: Computations from the Micron 2018 Investor Conference Anticipate Significant Growth from AI

AI servers will become their own category, Micron predicts. From negligible today, AI servers will be about 10% of the cloud infrastructure market by 2021. By 2025, about 50% of the total servers deployed in the cloud infrastructure will be AI servers. What differentiates these? They'll require 6X the amount of DRAM memory and 2.6X the storage capacity in the form of SSDs, compared to a standard cloud/data center server. Micron continues to innovate on memory and storage to enable faster data access and data processing for AI.

An AI study conducted by Forrester Consulting on behalf of Micron in August 2018, seems to agree. Is upgrading or rearchitecting memory and storage critical to meet future AI/ML training goals? Yes, almost 80% of the time according to respondents. The respondents were 200 IT and business professionals who manage architecture or strategy for complex data sets at large enterprises in the U.S. and China. In addition, they indicated that moving memory and compute closer together is essential for AI/ML success (per 90% of the firms).

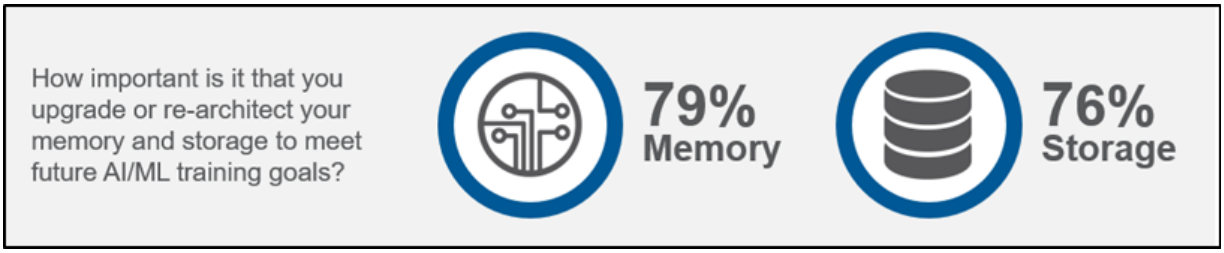


Figure 3: Results from 2018 Forrester Consulting Study on Behalf of Micron

Micron Infrastructure Activates the AI/ML Workflow

While advancements in data processing speeds and bandwidth have enhanced our ability to gain faster and better insights from data up until now, the innate parallelism of AI architectures places a greater burden on the design and performance of memory and storage than ever before. Let's consider some of those developments by looking at infrastructure needs by phase. AI/ML has a workflow that typically consists of four major components: ingest, transform, train, and production/execution. For each phase, Micron memory and storage can play a key role.

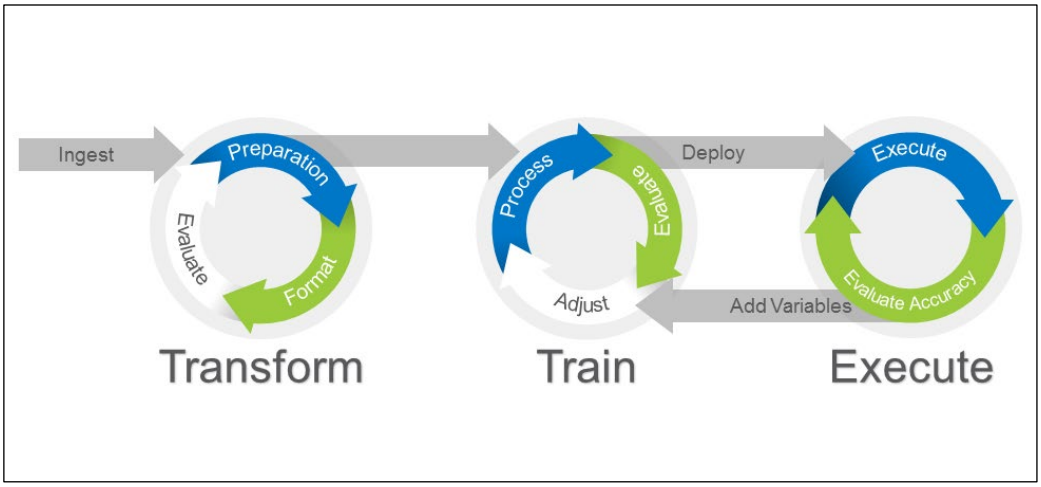


Figure 4: The AI/ML Workflow: Ingest, Transform, Train, and Execute

The ingesting of data and transformation process is one of the most important steps in how quickly your AI system will be able to provide value. In many AI solutions, these steps can represent up to 80% of the entire AI

execution process and has recently become a major focus of data science. To better understand why this is the case, it is important to understand what each of these steps entails.

Ingest

Ingest is exactly what it sounds like. Data must be collected from various sources — often incompatible with each other in format – and stored so the transformation process(es) can convert it to a usable form to train the system. The size of the ingested data can vary and is often in unstructured object or file forms such as videos, images, documents, or conversation transcripts, among others. It's typically also located in disparate data lakes and other data sources. AI processes, as well as the answers those processes provide, depend on massive amounts of data.

Ingest Infrastructure Requirements: The storage solution must be fast – otherwise the transformation of the data can take an exorbitant amount of time before it is usable. During the ingest process, data movement, relative to the repository, is 100% write, but it is typically "write once." It is this variability in the data format that requires the transformation process (normalization) that follows. The ingestion process relies on two major components: high-speed (high-bandwidth) network connections and very large, fast data repositories. While we need large capacity to collect this data, it is even more important for the storage solution to be fast. All of this enables the data to be usable to train the system. Training is what makes an AI system "smart" and useful in the real world.

Transform

The transformation process is the first of three iterative processes that make up an AI solution and is usually the most impactful to the AI development. Because the ingested data probably comes in a wide variety of sizes and formats, it is important that the data be normalized into a single format easily consumable by the training process. The format chosen and centered on from this transformation process must support the training and production engines selected. Today, an open source platform (such as TensorFlow™) or another AI framework works best.

To transform data into a standard format requires iterations. This process is broken into three major steps: Preparing the data for conversion, converting the data to the target format (e.g., TensorFlow data format), and evaluating the formatted data results to identify unusable records. Every step normalizes more of the data, and the steps are repeated on each set until all data is properly written in the target data format.

Transform Infrastructure Requirements: The speed of transformation of the data will depend on the quantity and quality of the memory installed in each compute node and the speed of the storage solution. The storage access during this phase is varied – unlike the previous ingest process — requiring both sequential and random access to the ingested data. This read-to-write ratio will vary depending on the target AI framework and what it takes to enable it for training. For typical transformation processes, the worst-case scenario would be 50% reads and 50% writes, though this depends somewhat on the data set. For example, when a data object is converted, every object is read and then written in the target format. If you are analyzing conversational data and pulling only the text of the data and removing all the metadata, then your read percentage will be more like 80%.

Train

The training step is typically extremely resource-intensive, though inference can also drive high resource use. The train phase of the workflow involves a repetitive set of steps executing a set of mathematical functions on the data ingested to identify high probability in getting the desired response/result. The results are then evaluated for accuracy. If the accuracy is not acceptably high — typically in the range of 95% accuracy or more — the mathematical functions are modified and then tried again by applying the updates to the same data set.

Simple image recognition is a classic example for an AI use case. In this example, the best-known model and data set for testing image recognition is called ImageNet with a set of functions called ResNet. The ImageNet training data set is 1.2 million images and takes around 145GB of data storage. ResNet has varying degrees of complexity, but the typical one used is ResNet-50 (there is also a ResNet-101 and -152). The number represents the how many neural network "layers" of different mathematical functions called "neurons" are being used (which also represents the complexity of the AI model).

Train Infrastructure Requirements: The training process — like the ingest/transform stage before it — can be a time-consuming and complex process. But unlike the ingest/transform stage, the train stage depends on high-performance compute to execute the mathematical functions. This is where really hefty hardware, typically in the form of graphics processing units (GPUs) with lots of fast memory, are used. The amount of fast storage and memory available to the solution directly impacts the amount of time it takes to complete a given training run (called an epoch). The faster we can complete each epoch, the more epochs we can execute and the more accurate we can make our AI system while keeping training time relatively low. Using HDDs for our training data storage is a problem because rotating media is really slow. Wasting time is wasting money in this case. The GPUs, an expensive resource, cannot get the data fast enough to complete the training epoch in a timely fashion and sometimes sit waiting for data to be fed. SSDs are typically orders of magnitude faster (in terms of IOPS and latency) than HDDs.

Increasing the size of the data being fed to each epoch — what we call a "batch" — also enables completing each epoch faster. But, while we could put 2TB or more DRAM in a server and call it a day, that can become very expensive. Most organizations are constantly balancing cost and efficiency. Remember, SSDs cost less than DRAM on a per-byte basis. Based on our testing, better results come from focusing on faster storage (SSDs) and doing so at a much better price point.

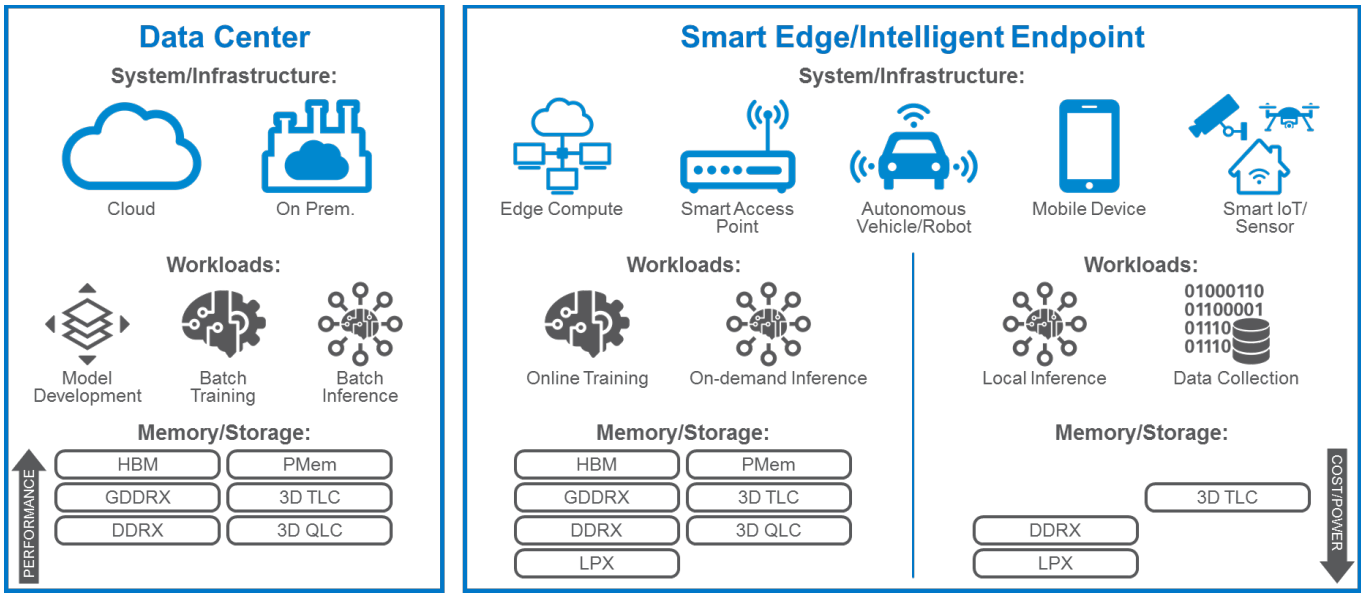


Figure 5: Memory and Storage from the Data Center to the Edge

Execute

The execution phase of the AI process is where the rubber hits the road. This phase is where you realize the useful benefits of AI. In the execution phase, our trained and refined AI model must be deployed, often to various edge devices (cameras, sensors, etc.), and used to execute decisions, also known in the AI world as "inference." As you are inferring, you will also usually need to continually evaluate your accuracy. This trend or feedback analysis is typically not done on the edge or on IoT devices, but as part of an analytics process in the data center or cloud that uses data captured by the devices and the inference results. The breadth of usage scenarios is virtually unlimited. Use cases can focus predominately on real-time analysis and decision making, or on scenarios requiring both real-time decision-making and post-real-time analytics.

Execute Infrastructure Requirements: Depending on the use case, there will be a different ratio of or dependence on memory and storage for the execute phase. Memory that provides a good balance of high performance and

low power consumption is the focus when executing inference on small remote/mobile devices. Micron offers low-power DRAM solutions in a variety of form factors suitable to mobile, automotive, and custom edge devices.

The inference process must constantly be reanalyzed to ensure that the AI inference engine is meeting expectations — a feedback loop for constant process improvement. The faster you can perform this analysis, the faster you can improve your real-time inference. Also, storage capacity and location depend on several factors, whether short-term, on-device retention versus long-term, off-device big data repositories. The data retention time also affects decisions about which storage solution to deploy.

On-device storage may be required — as it is in many automotive use cases — to meet regulatory requirements. While self-driving cars are doing massive amounts of real-time inference, they also need to retain all data from a wide variety of sensors (cameras, engine performance data, etc.) for a specific amount of time so that agencies such as the U.S. NHTSA can use it to analyze a crash. In fact, by 2020, the typical vehicle is expected to contain more than 300 million lines of code and will contain more than 1TB (terabyte) of storage!

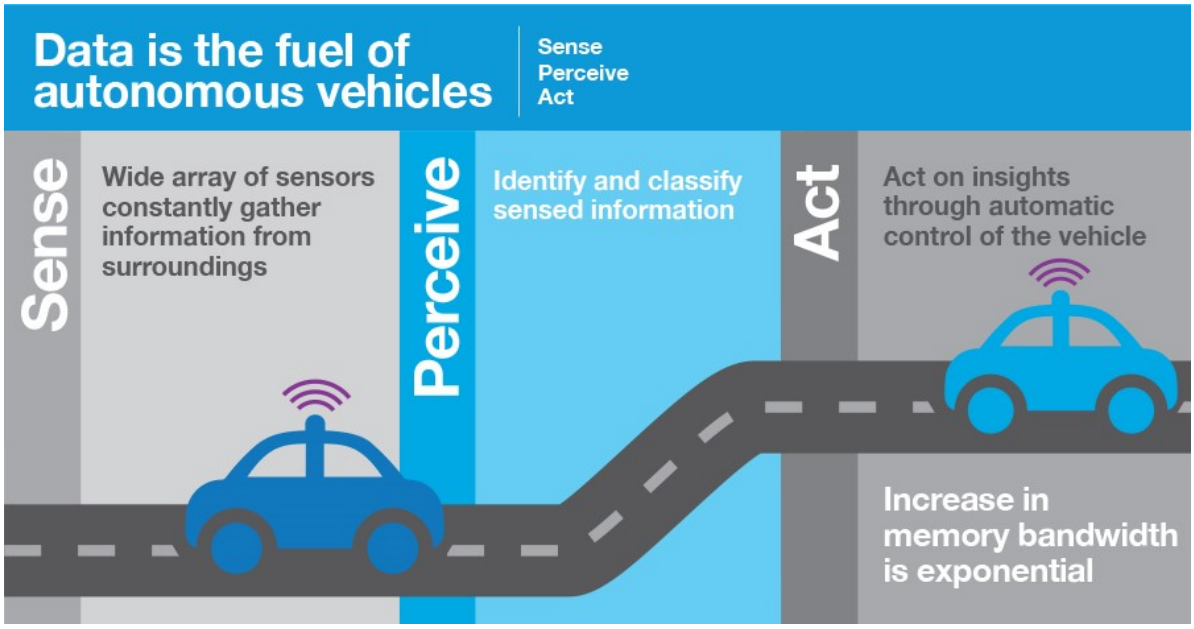


Figure 6: Data Storage and Compute on Edge Devices for Autonomous Vehicles

Micron Flash Storage Solutions for Your AI/ML

- Micron 5210 ION SSD is the first quad-level-cell-based SSD on the market. The emphasis on density means this is Micron's keenest price-for-performance SSD solution for read-intensive enterprise use cases. Replacing HDDs with cost-effective 5210 flash accelerated a TFRecord file creation transformation project about 8X compared to a simila- sized HDD, per a 2018 Colfax International machine learning QLC test.
- Micron 9300 PRO is our highest-performance, lowest-latency, class-leading, highest-capacity (15.36TB) commercially available SSD. The 9300 SSD with NVMe™ can accelerate data ingest and trim test and training cycle times for AI, ML, and DL. Often processes that were sequential can be performed in parallel because of the high capacity and high bandwidth storage the 9300 SSD delivers.
- Micron's latest GDDR graphics DRAM accelerates memory bit rates 16 GB/s.

- Micron's low-power, high-performance accelerators enabled by advanced programmable logic (FPGAs) for deep learning architecture feature the FWDNXT ML compiler, which makes DL easier by completely abstracting the hardware away.
- Micron 5210 and 9300 SSDs are often used together in hot- and warm-tiered storage, providing the cost-efficient storage back end for ML. We also offer storage class memory solutions that allow an additional layer of non-volatile storage performance that is 10X faster than current SSD solutions.
- Big data acceleration: Unlike HDDs, SSDs can support massive bandwidth. We've seen how adding a small amount of flash to an existing Hadoop cluster boosts performance by as much as 36%.

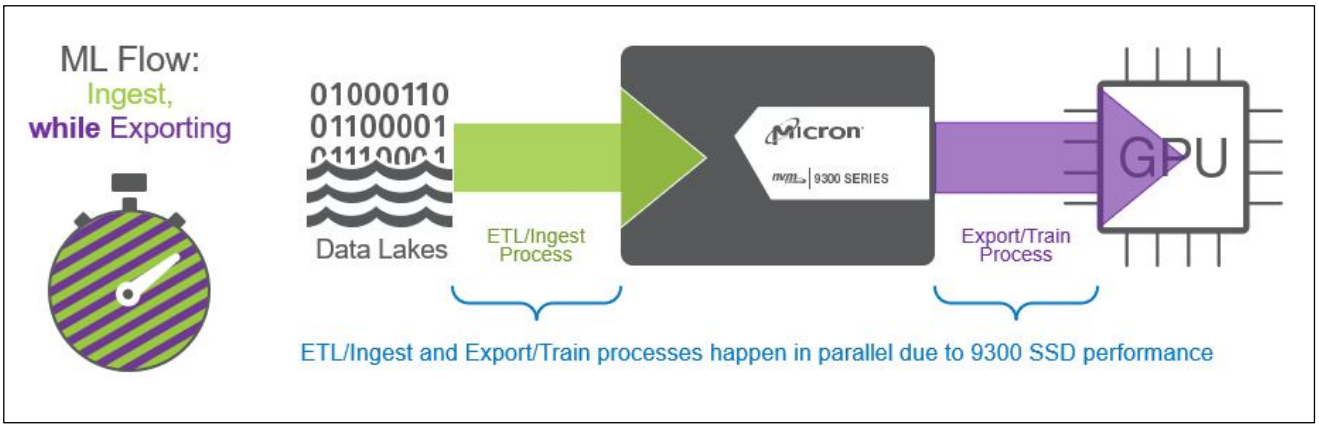
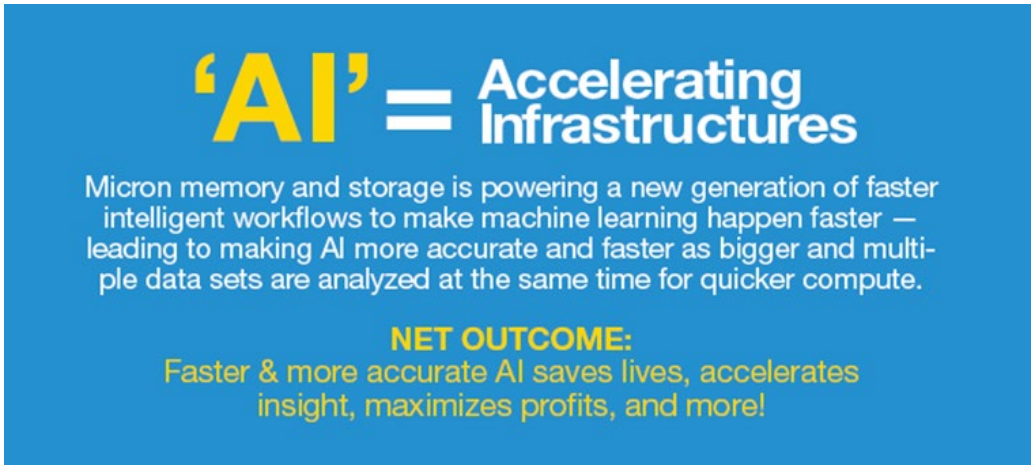


Figure 7: Parallelization: Machine Learning Ingest and Export Can Be Accelerated With the Micron 9300

Conclusion

Talk with Micron about your machine learning and deep learning projects as you develop effective, powerful and cost-efficient AI. Learn more online at micron.com/AI.



Micron for AI: Accelerating Infrastructure

1. Train your machines with data when and where you need it. Micron innovations in 3D NAND and high-performance, high-capacity storage and memory make our hardware a go-to solution for the fast and vast amounts of data it takes for effective machine learning.

2. Boost big data with AI and fast memory and storage. Solid state drives can support massive bandwidth. Fast memory and storage let you get more data closer to processing engines for faster analytics.

3. Smarten up edge devices. AI systems often need computing and data filtering at the network edge, pre-processing data prior to ingest. Intelligent edge-of-network devices need Micron's high-performance and low-power memory.

4. Get to the science of data management with Micron. Intelligent devices rely on more data to provide useful experiences, but they also create more data. Micron delivers the capacity and performance to handle the increasing amount of data across the AI landscape.

5. Supercharge your infrastructure. Micron has a broad portfolio of memory and storage for your AI workload-specific needs.

micron.com