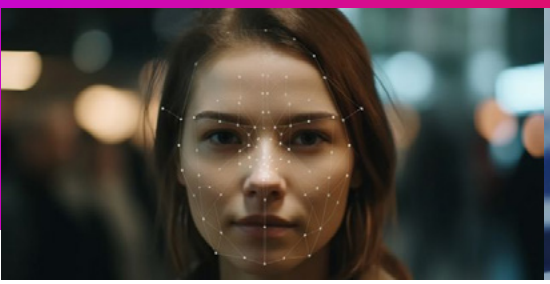
Your top seven AI architecture questions answered

Computer Vision | Natural language processing | Predictions & forecasting



Learn how to reduce AI training time, minimize compute costs and improve inferencing accuracy for key AI use cases.

Aren't the CPUs and GPUs doing all the heavy lifting in AI? Does it really matter which memory and storage products I use in my AI servers?

While CPUs and GPUs are essential components of AI servers, they aren't the only components to consider. Memory and storage products can substantially affect outcomes.

CPUs and GPUs (the "compute" parts of any AI server) don't directly manage the large datasets needed for AI. In fact, the CPUs and GPUs rely on fast memory and storage to store and manage training data, ensuring that it's fed to the compute elements quickly and consistently.

If the dataset ingest backs up, it can slow the training process, incurring additional costs due to CPU and GPU underutilization. Advanced memory and storage offer the capacity and throughput to help ensure AI runs continuously, efficiently, and smoothly.

How can I reduce the time to train my AI models?

Micron's AI product offerings are optimized for complex AI training workloads.

Fast Micron DDR5 memory delivers local data access to compute with double the bandwidth, burst length, number of bank groups, banks, and concurrent operations[1]. DDR5 enables 7x results for AI performance[2], faster image classification for computer vision, and improved define and recognize for low-level, mid-level, and higher-level categories. It also improves the speed of light/dark identification, accelerating face recognition[2].

Local SSD cache accelerates data access to GPUs. In addition, networked data lakes benefit from capacity-focused SSDs to help ensure that multiple AI servers are fed with the data they need. Micron offers the PCIe Gen4 SSD performance leader for AI, helping avoid storage bottlenecks and data backups that slow AI training time[3].

Large language models (LLM) with trillions of parameters can lead LLM inference to be memory bound. How do I overcome this challenge?

"Memory bound" refers to data processing backing up, which can be caused by not providing sufficient bandwidth in DRAM or storage. For example, Micron engineers analyzed the effects of storage performance in AI training and found that higher performance storage enabled better results[4]. Advanced memory and storage can enhance continuous data flow and help reduce wait times for expensive GPUs and CPUs.

CPUs and GPUs are the most expensive components in my AI servers. How can I get the most out of this investment and ensure I'm not underutilizing compute resources?

Feeding a steady stream of data from fast memory and storage utilizes CPUs and GPUs with the most efficiency. Choosing memory and storage without this consideration may result in bottlenecks and increased costs.

Training complex AI models may require many GPUs and CPUs which are becoming more power hungry with each generational update. What hardware optimizations can I make to reduce power consumption?

There are two ways to analyze power consumption: Overall consumption and consumption efficiency. The former is a simple measure of the total power consumed by the platform, while the latter analyzes power consumed by productive work. The latter is commonly referenced in a total cost of ownership (TCO) calculation[5].

Micron has designed advanced memory and storage products with energy efficiency to help reduce power consumption while providing high performance. This same solution addresses the increasing cost (and potential environmental impact) of huge electricity use. Micron incorporates a "feed-right balance" that helps ensure energy efficiency while powering the advanced needs of AI.

How can I improve the accuracy of AI inferencing?

Memory resources with high throughput enable better accuracy in AI inferencing. Performance-based CPUs are also required, but memory is essential to their optimal operation.

In general, how large of a dataset is needed to train an AI model?

According to NVIDIA (a widely-recognized leader in AI technology), "....Training any AI model requires carefully labeled and diverse datasets that contain thousands to tens of millions of elements, some of which are beyond the visual spectrum. Collecting and labeling this data in the real world is time-consuming and expensive. This can hinder the development of AI models and slow down the time to solution[6]...".

Learn which memory and storage solutions can solve your AI challenges

Memory and storage have important roles in the success of training and using AI. Having these basics in place will make a huge difference in your AI experience. Contact our salespeople to get the lowdown on what the best memory and storage solutions are for your datasets and the type of AI solutions you want to run. Prevent headaches down the road by taking advantage of Micron's expertise in memory and storage for AI.

See https://www.micron.com/about/blog/2022/november/boost-hpc-workload-performance-with-micron-ddr5-and-amd-zen-4-cpu for additional information.
Micron DDR5 with Intel AMX delivers 5-7x performances for recommender, training and vision workloads, see www.micron.com/intel. Also see https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063 for additional information on deep learning model transitions among lower (letters), middle (words), and higher-level (sentences) categories.
The Micron 9400 SSD is ranked #1 in the MLCommons storage performance benchmark ranking: https://mlcommons.org/en/storage-results-05/.
See https://media-www.micron.com/-/media/client/global/documents/products/white-paper/micron_9400_nvidia_gds_vs_comp_white_paper.pdf for complete test results.
See https://www.snia.org/education/online-dictionary/term/total-cost-of-ownership.
See https://www.nvidia.com/en-us/omniverse/synthetic-data/.