Accelerating AI with GPUs

Unit 4 - Summary

- **GPU History:** GPUs have been used to process pixel data since the 1970s, with processing capability increasing dramatically over the years.

- **GPU Architecture:** GPUs have specific memory, cores, and cache memory to optimize their functionality.

- **GPU Acceleration:** GPU acceleration works by executing compute-intensive functions, while the rest of the code is executed by the CPU.

- **GPU Server Systems:** GPU server systems are specialized computer systems designed to accelerate complex computations through high performance computing, parallel processing, and workload optimization.

- **GPU vs CPU:** GPUs are optimized for executing highly parallel tasks, while CPUs are better suited for general-purpose computing and handling complex logic.

What is the history of GPUs?

Since the advent of computers in the 1970s, specialized electronic circuits have been used to process pixel data. Over the years, the processing capability of GPUs has increased dramatically as images have become photorealistic.

What are the key architectural components of a GPU?

The key architectural components of a GPU are GPU cores, cache memory, and GPU-specific memory.

What is the primary role of a GPU in computer systems?

The primary role of a GPU in computer systems is to execute highly parallel tasks.

What are the key differences between CPUs and GPUs?

CPUs are designed for general-purpose computing and handling of complex logic, while GPUs are optimized for high parallel processing capability.

What are GPU server systems?

GPU server systems are specialized computer systems with GPUs designed to accelerate complex computations through high performance computing, parallel processing, and workload optimization.

───

What are the benefits of using GPUs to run AI workloads?

GPUs are designed to execute complex instruction sets required by AI models and provide the compute power required by growing data and AI model sizes.