CPU Vs GPU in the Data Center

The rapid progress in deep learning applications has been enabled by major advances in computing power. The use of GPUs (Graphic Processing Unit) capable of executing matrix multiplications in a highly parallel fashion accelerated deep learning training and led to a breakthrough in the early 2010s. With the development of increasingly complex and deeper neural networks, training tasks required more hardware acceleration. This need for added speed fueled the proliferation of AI hardware acceleration startups as well as new and more powerful GPUs being offered by the incumbents. With the recent advent of generative AI workloads NVIDIA's position in AI training strengthened resulting in a virtual monopoly characterized by high prices and low accessibility of the AI training hardware.

AI training remains a time-consuming process with large networks requiring uninterrupted GPU-based pro-cessing. Without a doubt, AI training is a very costly task in terms of hardware cost and overall data center infrastructure and operational costs (including electrical power, real estate, water for cooling, and other cost components). When a model is ready for deployment in production the training process is not necessarily over as the model needs constant adjustments and retraining in the face of new data, new requirements, accuracy improvements, etc.

When it comes to inference, in most cases, the model deployed will not need GPU support to run at the required performance level. CPUs equipped in data center servers are extremely powerful, multi-core processors, GPU-based servers, with large on-chip caches as well as hundreds of gigabytes of DRAM at their disposal. Compared to GPUs, the CPUs are significantly less costly to operate both in terms of dollars per CPU-hour and power consumption for the same workload. In addition, given the higher investment cost required by GPU based servers, CPU-based servers represent a more accessible and available resource in the cloud. This makes CPUs the best solution for inference for most cloud-based applications. CPUs represent a dominant share of the AI inferencing market in the Data Center.

AI inferencing alone does not represent the complete application. Incoming data needs pre-processing, formatting, and scheduling. The resulting predictions need to be translated into actions either in terms of human readable conclusions or some sort of actuation of or signaling to an industrial equipment. These tasks are all performed by CPUs. Therefore, the CPU becomes a self-contained, flexible, and adaptable tool for AI inference workloads.

CPUs also have a performance advantage over GPUs where massive parallelization is not always possible. This can best be seen in use cases where latency is critical, and the incoming data cannot be grouped in large batches. Many use cases require a batch of one or at most a few samples. In those situations, CPUs will have similar or lower latency than a GPU accelerated system. Given that deploying CPUs is a significantly less expensive proposition, their use in situations where they meet or exceed performance requirements makes using GPUs an inferior choice. Additionally, CPUs outnumber GPUs in the data-center approximately three to one, so availability is not an issue like there is with GPUs. Therefore, using CPUs can reduce the Total Cost of Ownership (TCO) in AI Inference workloads while comfortably meeting latency and throughput requirements.

When choosing the right hardware for their applications, customers look to achieve a target performance while minimizing the costs. The environmentally minded customer will look for the best performance at the lowest possible energy level. The industry faces the challenge of keeping emissions in check with the rapid growth of data centers worldwide. Cloud Native CPUs help lower data center carbon footprint, handle the broadest range of workloads with good performance, can be deployed quickly, and scale well. Since AI applications are undergoing rapid changes a flexible inference accelerator based on optimized software acceleration on CPU is a cost-effective choice, and one that significantly reduces the risk of incurring runaway future costs. CPUs are the prevailing choice for AI inference because they meet AI application needs while limiting power costs.

Servers based on Ampere® Altra® and Altra® Max® CPUs deliver an added performance boost while keeping power and cost below competing legacy CPU architectures. The Ampere Altra family of processors offers highly competitive inference performance at the best price. As the first Cloud Native Processors in the market, they provide more than double the core count on a single chip. Running single threaded operations, they scale linearly and eliminate noisy-neighbor inference problems providing predictable and reliable perfor-mance. Ampere Computing has quickly gained market share with instances offered by cloud service providers (CSPs) such as Oracle Cloud Infrastructure (OCI), Microsoft Azure, and Google Cloud. These providers give end-users powerful resources with the best cost-to-performance ratio and perfor-mance-per-watt currently available in the market. The competitive advantage of Ampere processors makes them the best choice for all users looking to scale up their AI operations. Ampere has an optimized inference engine using Ampere's CPU instruction set that runs underneath popular AI frameworks such as TensorFlow, PyTorch and ONNX providing superior performance over competing legacy CPUs. Ampere AI software requires no API changes and works out of the box with these frameworks facilitating a fast and seamless deployment. With AI inference calling for low latency and optimized power consumption, Ampere Altra processors are the leading product on the market.