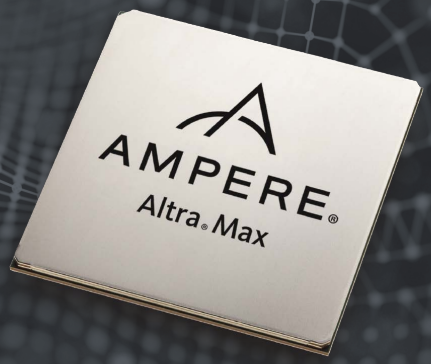
The Secret to Exceptional AI Inference Performance: Ampere Cloud Native CPU Architecture Augmented with Ampere Optimized Frameworks



Introduction

With the introduction of its Cloud Native processors, Ampere Computing has opened a whole new category of products that are ideal for running cloud applications. The 80 core Ampere® Altra® and 128 core Ampere® Altra® Max are being embraced by an increasing number of hyperscalers to leverage the advantages of this innovative yet easy-to-deploy product family with broad software support and a rich application catalog growing rapidly.

By exclusively focusing on cloud workload requirements and eliminating unnecessary legacy features, Ampere offers a CPU family having twice the number of cores running at 60% lower power compared to the x86 architecture under high workload utilization. The higher number of single-threaded cores enable users to flexibly configure their workloads based on the exact needs of their applications and reduce or eliminate noisy neighbor effects. The result is higher and shows more predictable performance at lower power consumption and a lower TCO (Total Cost of Ownership).

AI Workloads on Ampere Cloud Instances

One of the major workload categories in the cloud includes Machine Learning and Artificial Intelligence (AI) applications. We differentiate between ML (Machine Learning) and AI based on the nature of the applications and the software development packages they require. Indeed, ML consists predominantly of big data analysis using more mature methodologies, while AI consists primarily of deep learning as applied to Computer Vision, Recommendation Engines, and NLP (Natural Language Processing).

Running inference effectively in the cloud means primarily meeting the application's performance requirements. An essential criterion for real-time inference is the latency with which a prediction is made. The latency, i.e., the elapsed time between the arrival of the data and the delivery of the prediction, is critical for real-time applications. Similarly, throughput is another parameter that is essential for inferencing applications. Throughput is improved by batching inferencing tasks together and hence leveraging the effect of pipelining. The easily scalable multi-core architecture of Ampere Altra and Altra Max delivers unprecedented flexibility in adapting compute resources to the workload in hand for best results while minimizing $/CPU hours. Latency is improved quasi-linearly by deploying more cores. On the other hand, throughput for small batch sizes (1 – 4) is easily maximized with only four to eight cores. Sustained throughput performance is achieved by adding more cores as the batch size grows.

An example of inference in a Computer Vision application would be monitoring the traffic at an intersection and detecting vehicles, counting them, finding traffic violations, and recording their license plates. This task needs to execute in real-time and requires excellent performance. Ampere's Altra and Altra Max processors can handle such a workload without assistance from any sort of hardware accelerator by simply running on one of its industry-standard Ampere optimized frameworks such as TensorFlow, PyTorch, or ONNX-RT.

Figure 1: Ampere Altra Max Normalized CV Performance

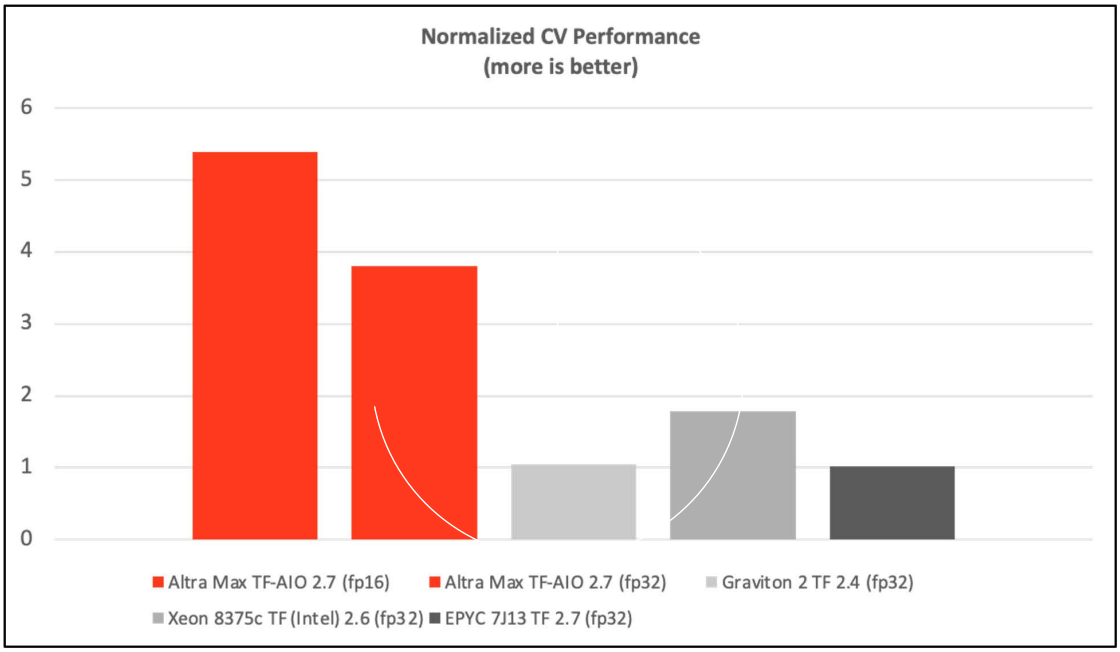


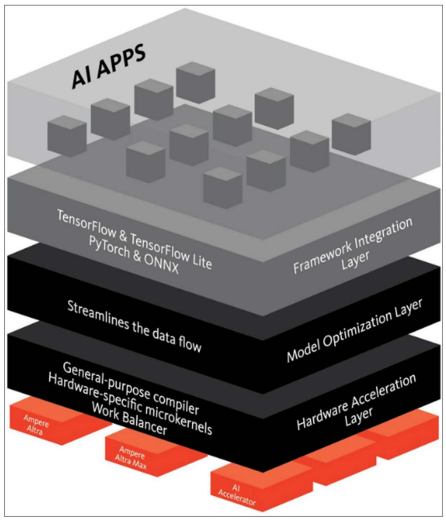
Figure 1 provides an overview of Ampere Altra Max CPU's performance compared to competing CPUs typically available on the cloud. It shows the composite normalized latency and throughput figures of Ampere's Altra Max and its competitors for a Computer Vision task. Ampere Altra Max running here on Ampere optimized TensorFlow outperforms its competitors on both accounts by 3-5 times.

Ampere Optimized Frameworks

Ampere's performance advantage in AI inference workloads in the cloud is made possible by its software stack that maps the popular AI development frameworks such as TensorFlow, PyTorch, and ONNX-RT to Ampere's Cloud Native architecture to achieve the highest level of operational efficiencies and execution speeds.

The optimization stack between the open standards AI development frameworks and Ampere CPUs' hardware execution layer consists of three different layers (see Figure 2):

Figure 2: Ampere AI Optimized Framework Software Stack



- Framework Integration Layer: Provides full compatibility with popular developer frameworks. Software works with the trained networks "as is." No conversions or approximations are needed.
- Model Optimization Layer: Implements techniques such as structural network enhancements, changes to the processing order for efficiency, and data flow optimizations without accuracy degradation.

- Hardware Acceleration Layer: Includes a "just-in-time" optimization compiler that uses a small number of micro-kernels optimized for Ampere processors. This approach allows the inference engine to deliver high-performance while supporting multiple frameworks.
- Running Ampere optimized frameworks, Ampere's AI team has benchmarked many popular neural network models in Computer Vision and NLP domains and has consistently outperformed the legacy x86 CPUs and AWS's ARM-based Graviton processors. These benchmark results are available on Ampere's AI resources page.

The Ampere AI team also provides a model library (AML). Our customers can download these models ready to run on Ampere platforms, reproduce the published benchmark results, and run their benchmarks based on their specific constraints. They can also easily incorporate them into their applications. More models will be added to Ampere's Model Library in the future.

Accessing Ampere AI Tools

Customers interested in getting started with Ampere AI have many choices for accessing Ampere Altra and Altra Max CPU resources. They can either open an account with one of Ampere's cloud partners or download Ampere's optimized frameworks from the Ampere AI's solutions page and install them onto their Ampere servers or workstations.

Access to AML (Ampere Model Library) is also enabled on the Ampere AI solutions page, with other information on running AI inference with Ampere CPUs.

Conclusions

Ampere's Altra and Altra Max Cloud Native processors deliver an exceptional value proposition for AI inferencing tasks. Combined with Cloud Native architecture's intrinsic benefits in scalability, predictability, and energy efficiency while delivering higher performance than its competitors, Ampere AI's optimized software tools deliver best-in-class performance with a significantly lower TCO.