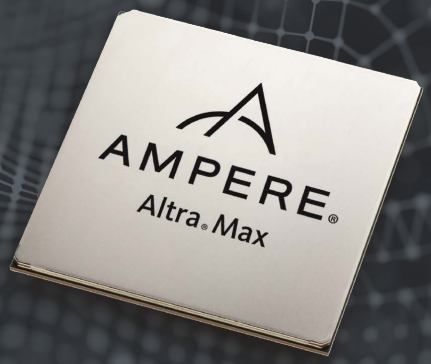
AMPERE®

Solution Brief

AI Inference on Ampere Altra MAX



Ampere—Empowering What's Next

The Ampere Altra Max processor is a complete system-on-chip (SOC) solution that supports up to 128 high-performance cores with innovative architecture that delivers predictable high performance, linear scaling, and high energy efficiency. Running AI inference is a rapidly growing production workload in the cloud. While training deep neural networks require a significant amount of GPU or similar hardware acceleration infrastructure, running inference on fully trained, deployment ready AI algorithms can be handled by CPUs in most situations. We demonstrate that Ampere Altra Max is ideal for running AI inference in the cloud, not only meeting latency and throughput requirements, but also outperforming CPUs based on x86 architecture as well as other ARM based processors currently used in the Cloud.

x264 on Ampere Altra Max

Ampere Altra Max is designed to deliver exceptional performance and power efficiency for applications like video transcoding. We use libx264 which implements the H.264/ MPEG-4 AVC standard that is the most widely used today. Ampere Altra Max uses an innovative architectural design, operating at consistent frequencies with single-threaded cores that make applications more resistant to noisy neighbor issues. This allows workloads to run in a predictable manner with minimal variance. Additionally, the processors are designed to be highly power efficient. Together, this gives Ampere Altra Max outstanding performance and power efficiency running x264.

Benchmarking Configuration

The benchmarks were performed using TensorFlow on bare metal single socket servers with equivalent memory, networking, and storage configurations for the x86 platforms shown. Processors tested include AMD EPYC 7J13 "Milan" with TF2.7 ZenDNN, Intel Xeon 8375C "Cascade Lake" with TF 2.7 DNNL, Intel Xeon 8380 "Ice Lake" with TF 2.7 DNNL and Ampere Altra Max M128-30 with Ampere Optimized TF 2.7. ARM-64 based "Graviton 2", available exclusively through AWS (c6g shape), was tested in 64-core configuration.

Key Benefits of running AI Inference on Ampere Altra Max

- **Cloud Native**: Designed from the ground up for 'born in the cloud' workloads, Ampere Altra Max delivers up to 2x higher inference performance than the best x86 servers and 5x better than similar ARM based processors.
- **Industry Standard Platforms:** Ampere Altra Max runs AI inference workloads developed on TensorFlow, PyTorch or ONNX without modifications. Customers can run their applications by simply using our optimized frameworks, available free of charge from Ampere or our cloud partners.
- **Support for fp16 format**: Ampere Altra Max is the only broadly available cloud CPU that natively supports the fp16 data format. Quantizing fp32 trained networks to fp16 is straightforward and results in no visible accuracy loss.

Inference Performance

Having run various AI workloads according to MLCommons benchmarking guidelines, we present some of our results below.

In Computer Vision using SSD ResNet-34 for a typical Object Detection application Ampere Altra Max outperforms in latency, Intel Xeon 8375C by 2x, AMD EPYC7Ji3 and Graviton by 4x in fp32 mode. In fp16, Altra Max extends its lead by an additional factor of two while maintaining the same accuracy. See Figure 1.

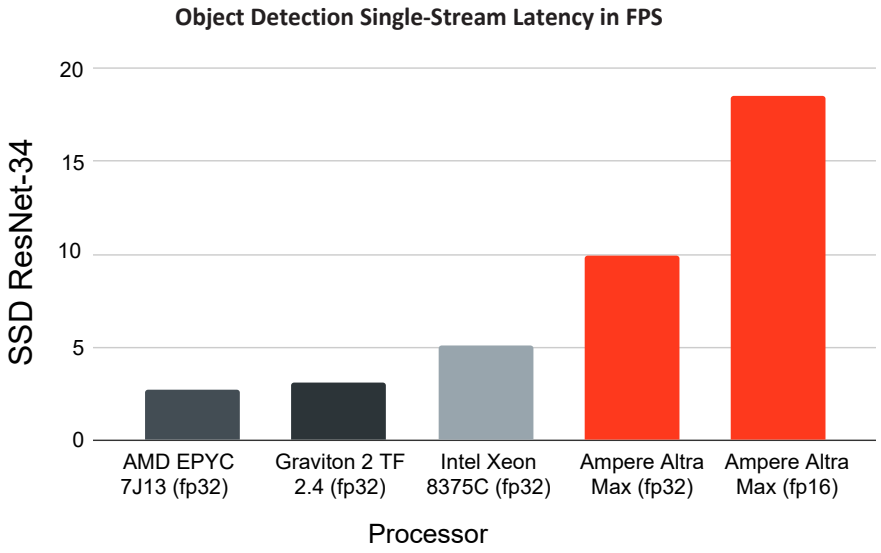


Figure 1. Throughput Performance

Ampere Altra Max also has a significant advantage in performance/watt over its competitors. In fp16 resolution Altra Max is around 5x more power efficient than Intel Xeon and AMD EPYC. In fp32 resolution Altra Max maintains a 2x advantage over the same Intel and AMD devices. See Figure 2.

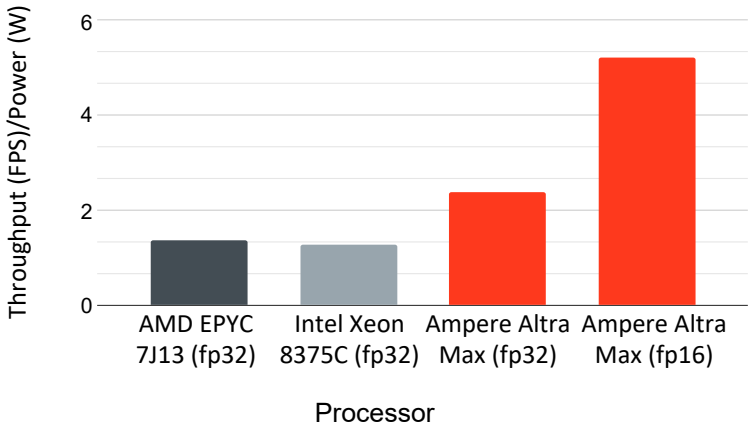


Figure 2. Power/W Advantage over Intel and AMD

- **Scalable**: With an innovative scale-out architecture, Ampere Altra Max processors have a high core count with compelling single-threaded performance. Combined with consistent frequency for all cores Ampere Altra Max delivers consistent performance at the socket level greater than the best x86 servers. This leads to much higher resistance to noisy neighbors in multitenant environments.
- **Energy Efficiency:** With up to 128 energy-efficient Arm cores, Ampere Altra Max has a 60% performance/watt advantage over leading x86 servers with better performance. Industry leading performance and high energy efficiency results in Ampere Altra Max having a smaller carbon footprint and reduces Total Cost of Ownership (TCO).

Ampere Altra Max

- 128 Armv8.2+ 64-bit cores at 3.0GHz
- 64KB i-Cache, 64KB d-Cache per core
- 1MB L2 Cache
- 16MB-32MB System Level Cache
- Coherent mesh-based interconnect

Memory

- 8x72 bit DDR4-3200 channels
- ECC and DDR4 RAS
- Up to 16 DIMMs (2 DPC) and 4TB/socket

Connectivity

- 128 lanes of PCIe Gen4
- Coherent multi-socket support
- 4x16 CCIX lanes

System

- Armv8.2+, SBSA Level 4
- Advanced Power Management

Performance

- SPECrate®2017_int_base:350

Summary

Ampere Altra Max processors are a complete System on a Chip (SOC) solution built for Cloud Native workloads, designed to deliver exceptional performance and energy efficiency for AI inferencing. Ampere Altra Max has up to 4x faster performance compared to Intel® Xeon® Platinum 8375c and AMD EPYC 7J13. In power efficiency Ampere Altra Max also leads its competitors by consuming 60% less power at equivalent throughputs.

Additional Benchmarking Details
We used the latest ffmpeg version available, N-105446-g2e82c61055, git commit 2e82c610553efd69b4d9b6c359423a19c2868255 and the latest libx264, git commit obb85e8bbc85244d5c8fd300033ca32539b541b7. We used the vbench configurations specified in "vbench: a Benchmark for Video Transcoding in the Cloud, a benchmark for the emerging video-as-a-service workload", Andrea Lottarini, Alex Ramirez, Joel Coburn, Martha A. Kim Parthasarathy Ranganathan, Daniel Stodolsky, and Mark Wachsler (2018), available at http://arcade.cs.columbia.edu/vbench.
Upload configuration: numactl -m 0 -C $CORE ffmpeg -i $INPUT -c:v libx264 -threads 1 -y -loglevel quiet -crf 18 $OUTPUT
VoD configuration: The bitrate is set per input file as described in the vbench paper.
numactl -m 0 -C $CORE ffmpeg -i $INPUT -c:v libx264 -threads 1 -y -loglevel quiet -passlogfile ffmpeg2pass -pass 1 -f null -an -sn -b:v $BITRATE -preset medium /dev/null
numactl -m 0 -C $CORE ffmpeg -i $INPUT -c:v libx264 -threads 1 -y -loglevel quiet -passlogfile ffmpeg2pass -pass 2 -b:v $BITRATE -preset medium ${OUTPUT}

AMPERE®

Ampere Computing® / 4655 Great America Parkway, Suite 601 / Santa Clara, CA 95054 / amperecomputing.com