Ampere® Computing for 2CRSI

March 2023
Disclaimer

All data and information contained in or disclosed by this document are for informational purposes only and are subject to change. This document may contain technical inaccuracies, omissions and typographical errors, and Ampere® Computing LLC, and its affiliates ("Ampere®”), is under no obligation to update or otherwise correct this information. Ampere® makes no representations or warranties of any kind, including express or implied guarantees of noninfringement, merchantability or fitness for a particular purpose, regarding the information contained in this document and assumes no liability of any kind. Ampere® is not responsible for any errors or omissions in this information or for the results obtained from the use of this information. All information in this presentation is provided “as is”, with no guarantee of completeness, accuracy, or timeliness.

This document is not an offer or a binding commitment by Ampere®. Use of the products and services contemplated herein requires the subsequent negotiation and execution of a definitive agreement or is subject to Ampere’s Terms and Conditions for the Sale of Goods.

This document is not to be used, copied, or reproduced in its entirety, or presented to others without the express written permission of Ampere®.

The technical data contained herein may be subject to U.S. and international export, re-export, or transfer laws, including “deemed export” laws. Use of these materials contrary to U.S. and international law is strictly prohibited.

© 2022 Ampere® Computing LLC. All rights reserved. Ampere®, Ampere® Computing, Altra and the Ampere® logo are all trademarks of Ampere® Computing LLC or its affiliates. SPEC and SPECInt are registered trademarks of the Standard Performance Evaluation Corporation. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.
Ampere’s Disruptive Value
Ampere Altra is the World’s First Cloud-Native Processor

Ampere’s Architecture is Optimized for the Cloud

- **High Performance**
- **Power Efficient**
- **Scalable**
- **Predictable**

Multi-Threaded Client Core
Inconsistent Operating Frequency

Limited Core Counts
Power and Area-Inefficient

Smaller Private Caches

Single-Threaded Cloud Core
Consistent Operating Frequency

Maximum Core Counts
Power and Area-Efficient

Larger Low Latency Private Caches

Legacy Architectures

Ampere Architecture

Ampere® Altra®
80 Cores

Ampere® Altra® Max
128 Cores
Innovation Delivered Annually

Powerful Multi-Year Roadmap Execution to Meet Industry’s Pace of Innovation

Ampere® Altra® Max
128 Cores

Ampere® Altra®
80 Cores

Ampere Future Processors
Custom-designed Ampere Cores*

AmpereOne Family
Custom-designed Ampere Cores*

Founded

Q4 2017

2020

2021

2022

2023

2024+

Innovating upon licensed IP to rapidly scale

Utilizing custom-designed Ampere IP to accelerate innovation

*Remain Arm ISA Compliant
Data Center Power Consumption is Rising

Data Centers are increasingly unwelcome neighbors:

Recent Limits & Moratoriums on DC Expansion

Notes:
1. International Energy Agency, Data Centres and Data Transmission Networks, Nov 2021
2. New perspectives on internet electricity use in 2030 – Andrae June 2020
# Server Efficiency is Fundamental to Sustainable Growth

<table>
<thead>
<tr>
<th>Projected</th>
<th>Legacy Approach (x86)</th>
<th>Cloud Native Ampere Approach $^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2025 Server</td>
<td>↑ 2.0x</td>
<td>↓ 0.8x</td>
</tr>
<tr>
<td>Power</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2025 DC Real</td>
<td>↑ 1.6x</td>
<td>↓ 0.7x</td>
</tr>
<tr>
<td>Estate</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Ampere: Sustainability at the Core
- Industry Leading Performance
- Industry Leading Power Efficiency
- Building Sustainable Data Centers

**Notes:**
1. Ampere internal models and analysis to identify total compute demand, power consumption numbers, and real estate footprint for legacy and Ampere processors in 2025.
Customers on Ampere® Altra®
Cloud Native Processors
Ampere: The Performance & Power Efficiency Leader

Why Cloud Native?
- High Performance → No Compromises
- Scalable → Linear in Socket Performance
- Predictable → Sayonara Noisy Neighbors
- Low Power → Rack Efficient, Sustainable

Cloud Native Processor Architecture is **Both** High Performance and Power Efficient

Est. SPECrate®2017_int_base Performance

Higher is Better →

Lower Is Better

Twice the Performance

Half the Power

Legacy x86 Processors

Intel Xeon “Icelake”

AMD EPYC “Milan”

Ampere® Altra® Max

Cloud Native Processors

gcc 10.2, flags used: -O3 -flto=32 -m64 -march<neoverse-n1/icelake-server/znver2>
Ampere Rack Value Proposition

Based on 42U rack @12.8 kW

Performance per Rack\(^1\)

<table>
<thead>
<tr>
<th>Workload</th>
<th>Intel</th>
<th>AMD</th>
<th>Ampere</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIR2017 Est.</td>
<td>1X</td>
<td>1.4X</td>
<td>2X</td>
</tr>
<tr>
<td>Redis</td>
<td>1X</td>
<td>1.5X</td>
<td>2.6X</td>
</tr>
<tr>
<td>NGINX</td>
<td>1X</td>
<td>1.7X</td>
<td>3.5X</td>
</tr>
<tr>
<td>x.264(^2)</td>
<td>1X</td>
<td>1.7X</td>
<td>2.25X</td>
</tr>
<tr>
<td>Cassandra</td>
<td>1X</td>
<td>1.1X</td>
<td>1.8X</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cores</th>
<th>1200</th>
<th>1792</th>
<th>4864</th>
</tr>
</thead>
<tbody>
<tr>
<td>Servers</td>
<td>15</td>
<td>14</td>
<td>38</td>
</tr>
</tbody>
</table>

Get 2-3X Better Performance for equivalent power

Notes:
1. Ampere internal models and analysis to identify total compute performance and system usage power consumption numbers, in standard 42U 12.8kW rack, see end notes
2. Data point uses data taken on M128-30 whereas all other data points use the M128-26.
## The Cloud Native Processor Value Proposition

### Web Services (NGINX)³

<table>
<thead>
<tr>
<th>Platform</th>
<th>Performance/W</th>
<th>Performance/W</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel Ice Lake</td>
<td>3.8x</td>
<td></td>
</tr>
<tr>
<td>AMD Milan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ampere Altra Max</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Memory Caching (REDIS)³

- **Redis Throughput**
  - **Noisy Neighbor Entries & Exits**
  - **Consistent. Predictable.**

### Video Services (H.264)³

- **H.264 Encoding Performance**
  - **Linearly Scalable**
  - **Stop Stranding Capacity**

### Data Services (Hadoop)³

- **Workload Usage Power**
  - **Hadoop Terasort**
  - **Low usage power!**
  - **Less Power is the New Power**

### Performance

- **3-4x Better Perf. & Efficiency**
- **No Compromises**
- **ZERO Loss to Noisy Neighbors**

(3) For more details, please refer to End Notes
Ampere: Leadership Performance for Cloud Workloads

Highest Performance and Power Efficiency Across Key Cloud Workloads\(^{(1)}\)\(^{(2)}\)

**Notes:**
1. Based on Company benchmarking
2. Intel Ice Lake represents Intel 8380 SKU; AMD Milan represents AMD 7763 SKU
3. Percentages represent AMD Milan and Ampere® Altra® Max indexed against Intel Ice Lake
4. Percentages represent Intel Ice Lake and Ampere® Altra® Max indexed against AMD Milan

---

**Web Services (NGINX)**\(^{(3)}\)
- Intel Ice Lake
- AMD Milan
- Ampere® Altra® Max

**Database (MySQL)**\(^{(3)}\)
- Intel Ice Lake
- AMD Milan
- Ampere® Altra® Max

**In-Memory Caching (Redis)**\(^{(3)}\)
- Intel Ice Lake
- AMD Milan
- Ampere® Altra® Max

**Media Transcoding (h.264)**\(^{(3)}\)
- Intel Ice Lake
- AMD Milan
- Ampere® Altra® Max

**AI Inference**
Image Classification (ResNet-50)\(^{(4)}\)
- AMD Milan
- Intel Ice Lake
- Ampere® Altra® Max

---

**Performance/W**

**Performance**

---

\(^{(1)}\) Throughput (Higher is Better)

\(^{(2)}\) Latency (Lower is Better)
Solution « cloud native » de Ampere : en pratique

- De manière générale, les benchmarks plutôt orientés *single thread* ne donne pas l’avantage à Ampere, la faute à de « petits » cœurs, mais aussi l’architecture Neoverse de première génération qui n’est pas la plus à jour.
Solution « cloud native » de Ampere : en pratique

- En revanche, sur des applications qui tirent parti du nombre de cœurs, la solution se place souvent devant les AMD EPYC Milan ou Intel Icelake.

---

C-Ray 1.1
Total Time - 4K, 16 Rays Per Pixel

- **Ampere Altra Q80-33 2P**
  - SE ++: 0.014, N = 7
  - Temps : 4.837 secondes

- **Xeon Platinum 8280 2P**
  - SE ++: 0.045, N = 4
  - Temps : 14.745 secondes

- **EPYC 7763 2P**
  - SE ++: 0.035, N = 7
  - Temps : 5.665 secondes

- **EPYC 75F3 2P**
  - SE ++: 0.028, N = 5
  - Temps : 8.969 secondes

- **EPYC 7713 2P**
  - SE ++: 0.039, N = 6
  - Temps : 6.590 secondes

---

1. (CC) gcc options: -lm -lpthread -O3
Solution « cloud native » de Ampere : consommation électrique

- Sur la suite de tests Phoronix, les Ampere Altra et Altra Ultra ont démontré des consommations plutôt contenues comparées aux solutions x86 Intel et AMD
Ampere® AI Value Prop

AI inference: Ampere® Altra® processor family with Ampere Optimized Frameworks

Easy to use out-of-the-box and no charge

Up to 5X better inference performance over Intel, AMD & Graviton

Native support for FP16 boosts performance without accuracy tradeoff

Optimized, pretrained models available for AI developers to use & for demos

AI training: Ampere Altra Systems with Nvidia GPUs

Platforms available with Nvidia GPUs for training

On-par performance with Intel and AMD
Ampere® AI: Leading Computer Vision Performance

Tensorflow CV Workloads

Example Use Cases:
- Image and video analytics
- Face or object recognition
- Autonomous vehicles and automation

High Perf
- Up to 2X faster than Intel optimized TF
- Up to 4X faster than AMD with TF- ZenDNN
- Up to 5X faster than Graviton with TF

fp16 is natively supported in Ampere Altra family
- Up to 2x faster than fp32
- Accuracy is on par with fp32, simple conversion

The benchmark is the mean performance ratio for latency (MLPerf single stream) and throughput (MLPerf offline) workloads across a set of typical computer vision models (ResNet 50 v1.5, DenseNet-169, SSD-ResNet-34).
Ampere® AI: Natural Language Processing

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Natural Language Processing (NLP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• CPU better in handling unbatched real-time NLP tasks (compared with GPU)</td>
<td></td>
</tr>
<tr>
<td>• Model reduction can further improve CPU performance.</td>
<td></td>
</tr>
<tr>
<td>• Altra / Altra Max can readily take advantage of FP16 with simple conversion</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Models</th>
<th>BERT Large Cased WWM Squad</th>
<th>RoBERTa Base Squad</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU Perf</td>
<td>Ampere optimized frameworks enhances BERT performance on Altra Max.</td>
<td>Ampere Altra/Altra Max delivers strong performance on HuggingFace RoBERTa Model.</td>
</tr>
<tr>
<td>Ampere Altra Max vs IceLake</td>
<td>1.25x over Intel IceLake.</td>
<td>1.25x over Intel IceLake</td>
</tr>
</tbody>
</table>

**BERT Large Cased Squad**
Throughput (QPS)

```
<table>
<thead>
<tr>
<th></th>
<th>Altra Max TF-2.7 AIO (FP32)</th>
<th>Xeon 8358 TF-2.7 w/ DNNL (FP32)</th>
</tr>
</thead>
</table>
```

**RoBERTa Base Squad**
Throughput (QPS)

```
<table>
<thead>
<tr>
<th></th>
<th>Altra Max TF-2.7 AIO (FP32)</th>
<th>Xeon 8358 TF-2.7 w/ DNNL (FP32)</th>
</tr>
</thead>
</table>
Ampere® AI: Training

Standard Nvidia’s software packages work on Ampere Altra out-of-the-box

• TensorFlow-GPU
• TensorRT
• CUDA

Same performance as x86 + Nvidia

CPU+GPU primarily used for training, high throughput inference can take advantage

Seamless pathway from Nvidia GPU to Ampere CPU with fp16
Ampere Altra Max 1P 128 cores (218W), Intel 8358 32 cores (250W), EPYC 7J13 64 cores (280W). MLPerf Offline benchmark for ResNet-50 v1.5 model and BERT, blended result. QPS= Queries Per Second
High Performance Computing

HPC is a large vertical with room for many architectures

Ampere offers a large core count with efficient SIMD units
- Great for compute bound workloads
- Enable more researchers simultaneously
- Scale up core count within existing power footprint

Workloads under investigation:

GROMACS, Weather Research Forecasting, OpenFOAM, NASPB, Ansys

More information and workload brief coming late Q2/Q3 2023
# HPC Performance on Altra Family

<table>
<thead>
<tr>
<th>Workload</th>
<th>Use Case</th>
<th>Altra</th>
<th>Altra Max</th>
<th>Unit</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPL Linpack</td>
<td>Scientific Performance</td>
<td>1369</td>
<td>1597</td>
<td>GFLOP/s</td>
<td>Higher</td>
</tr>
<tr>
<td>High Performance Conjugate Gradient</td>
<td>Scientific Performance</td>
<td>---</td>
<td>21.3</td>
<td>GFLOP/s</td>
<td>Higher</td>
</tr>
<tr>
<td>OpenFOAM Motorbike – 6 processes</td>
<td>Mesh time</td>
<td>---</td>
<td>65</td>
<td>Seconds</td>
<td>Lower</td>
</tr>
<tr>
<td></td>
<td>Execution time</td>
<td>---</td>
<td>131</td>
<td>Seconds</td>
<td>Lower</td>
</tr>
<tr>
<td>OpenFOAM drivaerFastback - small</td>
<td>Mesh time</td>
<td>---</td>
<td>33</td>
<td>Seconds</td>
<td>Lower</td>
</tr>
<tr>
<td></td>
<td>Execution time</td>
<td>---</td>
<td>107</td>
<td>Seconds</td>
<td>Lower</td>
</tr>
<tr>
<td>GROMACS</td>
<td>Molecular Dynamics</td>
<td>72.576</td>
<td>---</td>
<td>ns/day</td>
<td>Higher</td>
</tr>
<tr>
<td>Weather Research Forecasting 4.4</td>
<td>Weather</td>
<td>1.15</td>
<td>---</td>
<td>s/ts</td>
<td>Lower</td>
</tr>
<tr>
<td>Quantum Espresso</td>
<td>Quantum Chemistry</td>
<td>1513</td>
<td>---</td>
<td>Seconds</td>
<td>Lower</td>
</tr>
<tr>
<td>SpecFEM3D</td>
<td>Seismic</td>
<td>89.94</td>
<td>---</td>
<td>Seconds</td>
<td>Lower</td>
</tr>
</tbody>
</table>

1. GROMACS tests were performed using armclang and armpl
2. Other workloads tests were performed with gcc and BLIS ([https://github.com/flame/blis](https://github.com/flame/blis))
# Ampere Altra Family Overview

## Processor Subsystem
- Up to 128 Armv8.2+ 64-bit CPU cores at up to 3.3 GHz sustained frequency
- 64 KB L1 I-cache, 64 KB L1 D-cache per core
- 1 MB L2 cache per core
- Up to 32 MB system-level cache
- 2 x 128-bit SIMD units
- Hardware coherency supports 2P configurations
- Coherent mesh-based interconnect with distributed snoop filtering

## Memory Subsystem
- 8x 72-bit DDR4-3200 channels, 2DPC
- Up to 16 DIMMs and 4 TB/socket
- ECC, Symbol-based ECC, and DDR4 RAS features

## System Resources
- Full interrupt virtualization (GICv3)
- I/O virtualization (SMMUv3)
- Enterprise server-class RAS

## I/O Subsystem
- 128 lanes of PCIe 4.0 (1P)
- 192 lanes of PCIe 4.0 (2P)

## Technology & Architecture
- TSMC 7 nm FinFET
- Arm v8.2+, SBSA Level 4

## Power
- 40 W – 187 W Usage Power*
- 65 W – 250 W TDP
- Advanced Power Management

## Performance
- Estimated SpecRate2017_int_base: Up to 359*

*Performance and usage power data are based on estimated SPECrate®2017_int_base (GCC10) and are subject to change based on system configuration and other factors. Usage Power is defined as average power consumed over time by a given workload.
Ampere Altra Family Processor Complex

CPU Cores
- Up to 128 Armv8.2+ 64-bit CPU cores @ up to 3.3 GHz sustained frequency
- Four-wide superscalar aggressive out-of-order execution
- Dual 128-bit wide SIMD execution pipes
- 48-bit logical and physical addressing

L1 Cache
- 64 KB 4-WSA Icache and Dcache with 64-byte cache lines
- Dcache ECC protected
- Fully associated ITLB supporting 4KB, 16KB, 64KB, 2MB, & 32MB pages sizes
- Fully associated DTLB supporting 4KB, 16KB, 64KB, 2MB, & 512MB pages sizes

L2 Cache
- 8-WSA 1MB L2 cache w/ 64B lines and data ECC protection per 64 bits.
- The DSU interfaces with the mesh over a 256 bit wide CHI-B compliant interface
- SECDED ECC protection for all RAM structures except victim array
- Strictly inclusive with L1D and L1I data caches (I and D hardware coherency)
- Dynamic biased replacement policy
- MESI coherency protocol

System Level Cache
- Up to 32 MB distributed on-chip cache shared between all processors
- Memory-side cache for processor evictions providing caching of larger data and instruction structures for overall performance enhancements
- Mostly exclusive with L2 cache
- 256 bit data buses all around
- 16 ways, ECC protected

Cache Protection
- L1 Dcache, L2 cache, and System Level Cache ECC- protected

System MMU and GIC
- Arm SMMUv3.1
- Arm GICv3
Ampere Altra Family Memory Subsystem

Bandwidth and Capacity
• Eight 72-bit DDR4 channels
• Up to DDR4-3200
• Up to 20PC
• Up to 4 TB of memory

Supported Devices, Modules, and Configurations
• Support for UDIMMs, RDIMMs, LRDIMMs, and 3DS
• Support for x4 and x8, and for 8Gb and 16Gb devices
• Production support for 4, 6, and 8 active channels

Additional Features
• Hashed memory interleave across active channels
• DRAM throttling

PCIeGen4x128
Coherent Mesh Fabric
Up to 128 Arm Cores
64KB L1-Inst, 64KB L1-Data
1MB L2 Cache
Up to 32MB System Level Cache
8x72 ECC DDR4-3200
Ampere Altra Family High Performance I/O Subsystem

**PCIe 4.0**
- Compliant to PCIe Base Specification 4.0 v1.0

**Supported PCIe Controller Bifurcations**
- 128 lanes of PCIe 4.0
- Ampere Altra Max
  - 4x16 with CCIX support (bifurcates down to x4)
  - 4x16 (bifurcates down to x4)
  - x4 hot plug support
- Ampere Altra
  - 4x16 with CCIX support (bifurcates down to x4)
  - 8x8 (bifurcates down to x2)
  - x4 and x2 hot plug support
Ampere Altra Family Low Speed I/O

**SMPro Control Processor**
- Cortex-M3 Arm Processor (400 MHz)
- Responsible for wide range of system management:
  - System booting
  - Power fail detection
  - Error handling
  - BMC interface
  - Interface to CPUs/PMPro (doorbell interrupts, messaging)
  - Monitors memory accesses and asserts side band signal if access to secure memory range

**PMPro Control Processor**
- Cortex-M3 Arm processor (400 MHz)
- Responsible for wide range of power and thermal management
  - Power management
  - Temperature control
  - Dynamic voltage frequency scaling (DVFS)
  - Max Frequency mode
  - ACPI and logic
  - Sensor logic
  - Interface to CPUs/PMPro (doorbell interrupts, messaging)

**Low Speed I/O**
- Nine I2C controllers up to 1 MHz (master/slave)
- Two QSPI up to 30 MHz for SPI flash and TPM
- Five UARTs
  - One 4-pin
  - Four 2-pin
  - No function or I/O sharing between five UARTs
- Three sets of 8 GPIOs (secure/non-secure)
- One set of 8 GPs

**Device Timers**
- Two watchdog timers
- Four system timers

---

**PCIeGen4x128**

**Coherent Mesh Fabric**

**Up to 128 Arm Cores**
- 64KB L1-Inst, 64KB L1-Data
- 1MB L2 Cache

**Up to 32MB System Level Cache**

**8x72 ECC DDR4-3200**
Software, System Firmware
Platform Tools & Design
Collaterals
Verified Linux Operating Systems

- AlmaLinux 8.5
- Debian 11
- Fedora 35
- Oracle Linux 8.5
- Red Hat RHEL 8.5
- Rocky Linux 8.5
- SUSE SLE SP3
- Ubuntu 20.04
  - SOC Certified
Ampere’s Expanding Software & Provider Ecosystem

Broad Developer Ecosystem with 165+ Software Applications Undergoing Daily Automated Functionality and Performance Testing

Notes:
1. As of December 31, 2021
# Software Stacks Actively Tested & Regressed for Ampere Altra Family

<table>
<thead>
<tr>
<th>Languages</th>
<th>Software Stacks Actively Tested &amp; Regressed for Ampere Altra Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenJDK, Java, PHP, Python, Ruby, C/C++, Lua, Perl, PyPy, Go, Rust</td>
<td></td>
</tr>
</tbody>
</table>

## Orchestration & Containers
- Kubernetes, Docker, K3s, Rancher, OpenStack, Mesos, Loom, Ansible, Terraform, ...

## Web Services
- Apache httpd, NGINX, Tomcat, WordPress, Drupal, node.js, ...

## DevOps & Tools
- Grafana, Telegraf, Travis CI, Jenkins, Prometheus
- DataDog, TARS, ...

## Many Other Apps
- Anbox Cloud, Solr, Genymotion, Elasticsearch, Gradle, Joomla, Maven, ...

## Databases: MySQL, MongoDB, MariaDB, Memcached, Redis, KeyDB, Cassandra, InfluxDB, CouchDB, Postgres, Scylla, Zookeeper

## Frameworks
<table>
<thead>
<tr>
<th></th>
<th>Caffe</th>
<th>ONNX</th>
<th>TensorFlow</th>
<th>PyTorch</th>
<th>Apache Storm</th>
<th>CUDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Middleware</td>
<td>DPDK</td>
<td>ISA-L</td>
<td>OvS</td>
<td>SPDK</td>
<td>Ceph</td>
<td>x.264/265, AV1, VP9</td>
</tr>
</tbody>
</table>

## OS¹, VMM’s and Compiler Support
- CentOS, Debian, GCC, KVM, LAVA, Oracle Linux, Red Hat, SUSE, Ubuntu

## BIOS/UEFI and BMC
- American Megatrends, AptioV, MEGARAC, LinuxBoot, OpenBMC

## Languages
- OpenJDK, Java, PHP, Python, Ruby, C/C++, Lua, Perl, PyPy, Go, Rust

## Orchestration & Containers
- Kubernetes, Docker, K3s, Rancher, OpenStack, Mesos, Loom, Ansible, Terraform, ...

## Web Services
- Apache httpd, NGINX, Tomcat, WordPress, Drupal, node.js, ...

## DevOps & Tools
- Grafana, Telegraf, Travis CI, Jenkins, Prometheus
- DataDog, TARS, ...

## Many Other Apps
- Anbox Cloud, Solr, Genymotion, Elasticsearch, Gradle, Joomla, Maven, ...

## Database: MySQL, MongoDB, MariaDB, Memcached, Redis, KeyDB, Cassandra, InfluxDB, CouchDB, Postgres, Scylla, Zookeeper

## Frameworks
<table>
<thead>
<tr>
<th></th>
<th>Caffe</th>
<th>ONNX</th>
<th>TensorFlow</th>
<th>PyTorch</th>
<th>Apache Storm</th>
<th>CUDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Middleware</td>
<td>DPDK</td>
<td>ISA-L</td>
<td>OvS</td>
<td>SPDK</td>
<td>Ceph</td>
<td>x.264/265, AV1, VP9</td>
</tr>
</tbody>
</table>

## OS¹, VMM's and Compiler Support
- CentOS, Debian, GCC, KVM, LAVA, Oracle Linux, Red Hat, SUSE, Ubuntu


¹ Support for Linux OS's ➔ https://github.com/AmpereComputing/ampere-centos-kernel/wiki
AArch64 is fully supported by major Linux distros

**Commercial Options**
- RHEL 7.4
- RHEL 8.4
- RHEL 8.5

**Community Options**
- Fedora 32-35
- Centos 7.8
- Centos 8.2
- Centos 8.4

---

**Commercial Options**
- SLES 15 SP2
- SLES 15 SP3
- SLES 15 SP4

**Community Options**
- openSUSE Tumbleweed & Leap 15

---

**Commercial Options**
- Ubuntu 18.04 LTS
- Ubuntu 18.04 HWE
- Ubuntu 20.04 LTS
- Ubuntu 20.04 HWE

---

**Commercial Options**
- Oracle Linux 7.9
- Oracle Linux 8

**Community Options**
- Debian 10
- Debian 11

---

**Commercial Options**
- Ubuntu 18.04 LTS
- Ubuntu 18.04 HWE
- Ubuntu 20.04 LTS
- Ubuntu 20.04 HWE
Ampere Altra Family Platform Overview

Ampere Altra Family Platform Configuration Info found at https://solutions.amperecomputing.com/systems/altra
Ampere Developer Program
Application Architects
Build Engineers
Deployment Engineers
Data Scientists
OS / Kernel Engineers
Ampere Ready Software
Ampere Developer Center

https://developer.amperecomputing.com

50+ pieces of developer Specific Content

Refreshing with one new piece / week in 2023

Focusing on Tutorials and Transition & Tuning Guides

- Updated UI
- Latest News
- Curated Content by Persona
- Ampere Ready Software
- Action Bar (github, newsletter, blogs, systems, youtube)
## The Ampere Developer Program

developer.amperecomputing.com

### How to connect

**Ampere Developer Program**

https://developer.amperecomputing.com
developer@amperecomputing.com

### What it is

- Curated content for designing, building, deploying and optimizing on Ampere products
- Developer newsletter
- Sample code, documentation, examples, and videos
- System test drive options

### Who it’s For

<table>
<thead>
<tr>
<th>Designing Cloud Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information you need to implement the optimal software and services design.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Building Cloud Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Understand the spectrum of build tools to allow you to confidently move from test to production.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Deploying Cloud Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learn how Ampere products can provide consistent, reliable, well-managed deployments using tools and methods.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Using your data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Find the optimal Artificial Intelligence toolchains, frameworks, and library options</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Enabling the Open-Source Community</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learn how Ampere uses reference platforms and work with the open-source communities, including operating system vendors and cloud providers.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ampere Ready Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>See the spectrum of software running across Ampere-based instances.</td>
</tr>
</tbody>
</table>
End Notes
End Notes

Hardware Configuration
Ampere Altra® Q80-30, 80 cores, CentOS 8.0.1905
Ampere Altra® Max M128-30, 128 cores, CentOS 8.0.1905
Ampere Altra® Max M128-30, 128 cores, CentOS 8.0.1905
AMD EPYC 7763, 64 cores/128 threads, 2.25 GHz CPU, L1/L2/L3 = 32KB/512KB/256MB, DDR4@3200 – 32GB x 8 1DPC, cTDP=280W, CentOS 8.3
Intel® Xeon® Gold 6258R Processor, 28 cores/56 threads, 2.7 GHz CPU, L1/L2/L3 = 32KB/1MB/256MB, DDR4@2933 – 32GB x 6 1DPC, TDP=205W, CentOS 8.3
Common
1x Mellanox MT27800 ConnectX-5 NICs, 1x Intel Xeon 2679 v4 (Broadwell) load generators
Software Configuration
NGINX
NGINX v1.15.4 serving a 50KB static HTML file over HTTPS/TLS, Brotli for compression, LuaJIT to pre-process the URL string. Intel Xeon 2697 v4 Wrk load generator. Metric is throughput (requests/second) under an SLA – p.99 latency <= 10ms. Load was gradually increased till the SLA was violated.

Media Encoding
x264 v0.161.3027, clip used – Ducks Take off 1080p50
/x264 -preset medium -psnr -tune psnr -threads 1 -frames 100 -profile main
Multiple single-threaded x264 instances started up (1 per core/thread). The metric was aggregate of the FPS reported by each of the instances.

Encryption
OpenSSL v1.1.1g FIPS, run as follows: openssl speed -evp aes-256-gcm -multi <number_of_cores>

Core Count

Power Numbers (TDP)

TensorFlow Comparison
Ampere Altra® Max M128-30, 128 Core, Ubuntu 20.04, TensorFlow-AIO 2.4, ML Perf Version v1.1
Ampere Altra® Q80-30, 80 Core, Ubuntu 20.04, TensorFlow-AIO 2.4, ML Perf Version v1.1
AMD EPYC 7571 (AWS:m5s.24xlarge), Ubuntu 20.04, TensorFlow-AIO 2.4, ML Perf Version v1.1
Intel Platinum 8375C (AWS:m6i.32xlarge), Ubuntu 20.04, TensorFlow-AIO 2.4, ML Perf Version v1.1
Neoverse N1(AWS:m6g.metal), Ubuntu 20.04, TensorFlow-AIO 2.4, ML Perf Version v1.1
MLPerf Comparison
Ampere Altra® Max M128-30, 128 Core, Ubuntu 20.04, TensorFlow-AIO 2.4, ML Perf Version v1.1
Ampere Altra® Q80-30, 80 Core, Ubuntu 20.04, TensorFlow-AIO 2.4, ML Perf Version v1.1
Others
Performance Recap – Ampere Altra and Altra Max

Altra and Altra Max Performance – up to 3.2x Higher than x86

Altra and Altra Max Performance/Watt – up to 3.8x Higher than x86
# Industry Standard Benchmarks


<table>
<thead>
<tr>
<th>SPECrate® 2017 int_base Estimated</th>
<th>Altra® Q80-30</th>
<th>Altra® Max M128-30</th>
<th>AMD EPYC Milan 7763</th>
</tr>
</thead>
<tbody>
<tr>
<td>500.perlbench_r</td>
<td>305</td>
<td>461</td>
<td>318</td>
</tr>
<tr>
<td>502.gcc_r</td>
<td>201</td>
<td>200</td>
<td>285</td>
</tr>
<tr>
<td>505.mcf_r</td>
<td>114</td>
<td>93.2</td>
<td>163</td>
</tr>
<tr>
<td>520.omnetpp_r</td>
<td>135</td>
<td>144</td>
<td>177</td>
</tr>
<tr>
<td>523.xalancbmk_r</td>
<td>262</td>
<td>267</td>
<td>368</td>
</tr>
<tr>
<td>525.x264_r</td>
<td>738</td>
<td>1130</td>
<td>634</td>
</tr>
<tr>
<td>531.deepsjeng_r</td>
<td>365</td>
<td>560</td>
<td>354</td>
</tr>
<tr>
<td>541.leela_r</td>
<td>353</td>
<td>585</td>
<td>347</td>
</tr>
<tr>
<td>548.exchange2_r</td>
<td>899</td>
<td>1420</td>
<td>982</td>
</tr>
<tr>
<td>557.xz_r</td>
<td>166</td>
<td>211</td>
<td>217</td>
</tr>
<tr>
<td>Geomean</td>
<td>285</td>
<td>360</td>
<td>331</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SPECjbb® 2015 Estimated</th>
<th>Altra® Q80-30 1P</th>
<th>Altra® Max M128-30 1P</th>
<th>AMD EPYC Milan 7763</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max jOPS</td>
<td>135,311</td>
<td>173,633</td>
<td>193,086</td>
</tr>
<tr>
<td>Critical jOPS</td>
<td>119,550</td>
<td>157,237</td>
<td>176,283</td>
</tr>
</tbody>
</table>

![SPECrate® Integer Base Estimated & SPECjbb® 2015 Performance](image)

Industry-leading performance on Standardized Benchmarks using open-source compilers and JDK!
Ampere® Altra® Max Energy Efficiency

Ampere® Altra® Max maintains predictable core frequencies while consuming lower power (below TDP).

Power headroom means workload-driven power capping can lead to huge density improvements!

Compelling performance/Watt at competitive levels of performance.

<table>
<thead>
<tr>
<th></th>
<th>Performance</th>
<th>Usage Power (W)</th>
<th>Performance/Watt</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMD EPYC Milan</td>
<td>331</td>
<td>280W</td>
<td>1.0x</td>
</tr>
<tr>
<td>Ampere® Altra® Max</td>
<td>360</td>
<td>178W</td>
<td>1.71x</td>
</tr>
</tbody>
</table>

Ampere® Altra® Max

- Consistently running at max frequency
- Power headroom means workload-driven power capping can lead to huge density improvements!