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.

Ampere’s Disruptive Value

Ampere Altra is the World's First Cloud-Native Processor

Ampere's Architecture is Optimized for the Cloud

Multi-Threaded Client Core

Inconsistent Operating Frequency

Limited Core Counts

Power and Area-Inefficient

Smaller Private Caches

Legacy Architectures

Ampere Architecture

Single-Threaded Cloud Core

Consistent Operating Frequency

Maximum Core Counts

Power and Area-Efficient

Larger Low Latency Private Caches

High Performance

+

Power Efficient

+

Scalable

+

Predictable

Ampere® Altra®

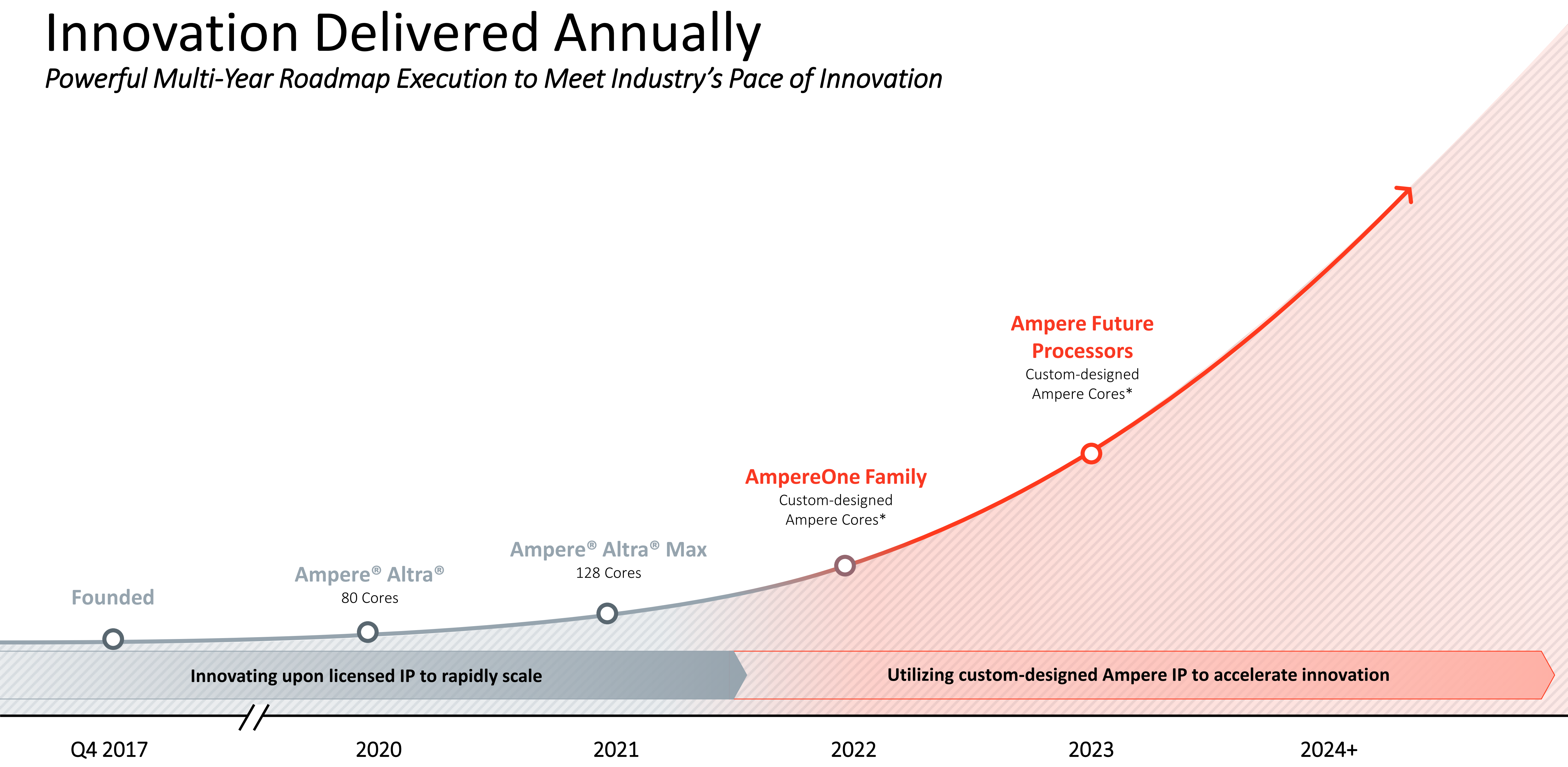
80 Cores

Ampere® Altra® Max

128 Cores

Innovation Delivered Annually

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



*Remain Arm ISA Compliant

Data Center Power Consumption is Rising

2020

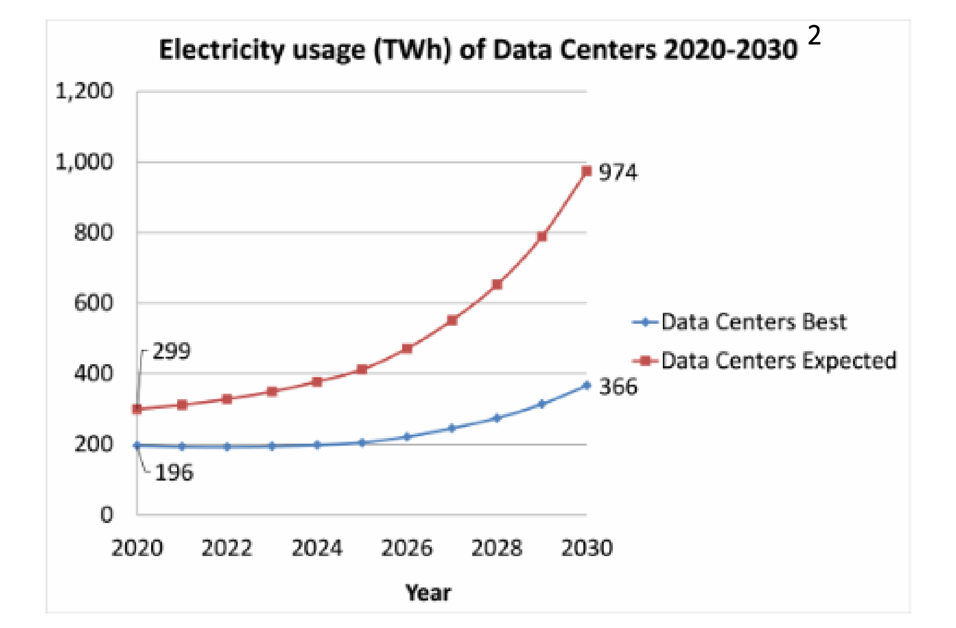
1-2%

Global Electricity Demand[1]

2030

Increasing

2-4X ↑



Data Centers are increasingly unwelcome neighbors:

Ireland Amsterdam Singapore Frankfurt London

Recent Limits & Moratoriums on DC Expansion

AMPERE.

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

Projected	Legacy Approach (x86)	Cloud Native Ampere Approach [1]
2025 Server Power	↑ 2.0x	↓ 0.8x
2025 DC Real Estate	↑ 1.6x	↓ 0.7x

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®

Microsoft

ORACLE Cloud Infrastructure

Tencent Cloud

ByteDance 字节跳动

CLOUDFLARE

EQUINIX

金山云

美团

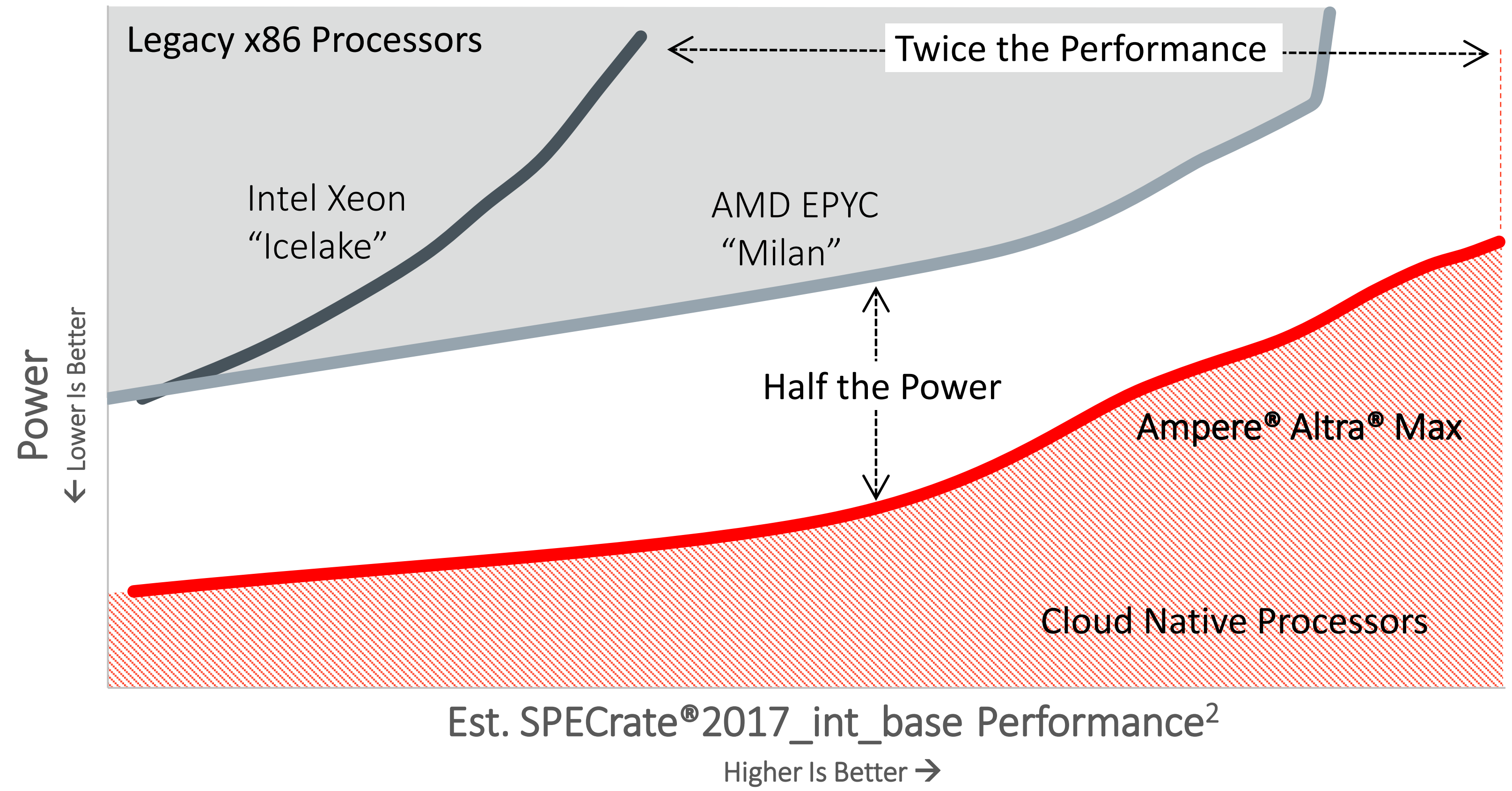
Scaleway
The cloud that makes sense

UCLOUD优刻得

AMPERE.

Cloud Native Processors

Ampere: The Performance & Power Efficiency Leader



Why Cloud Native?

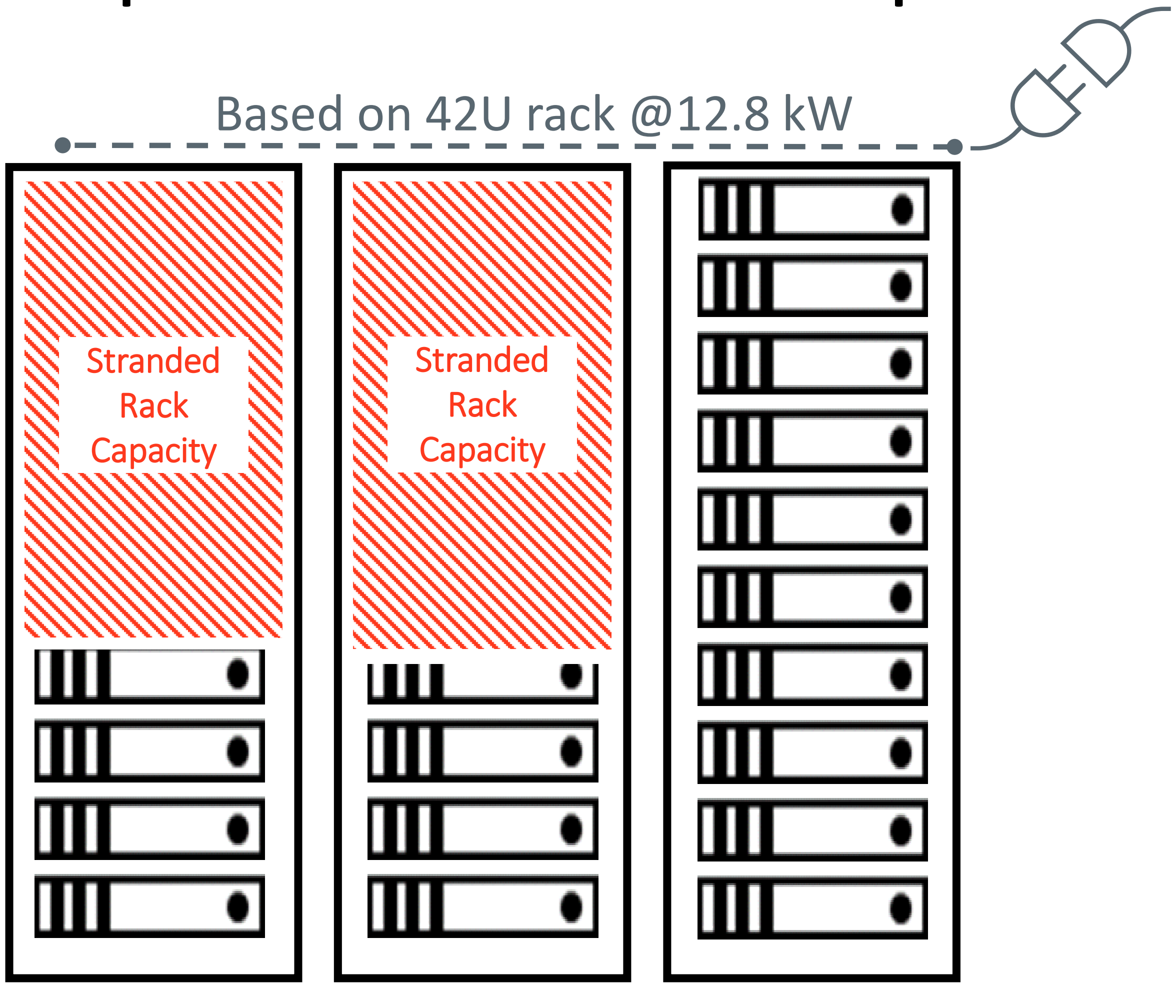
- ✓ High Performance ➡ *No Compromises*
- ✓ Scalable ➡ *Linear in Socket Performance*
- ✓ Predictable ➡ *Sayonara Noisy Neighbors*
- ✓ Low Power ➡ *Rack Efficient, Sustainable*

Cloud Native Processor Architecture is Both High Performance and Power Efficient

AMPERE.

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

Ampere Rack Value Proposition



Intel Ice Lake (2S) 8380 | AMD Milan (2S) 7763 | Ampere Altra Max (1S) M128-26

Performance per Rack[1]

Workload	Intel	AMD	Ampere
SIR2017 Est.	1X	1.4X	2X
Redis	1X	1.5X	2.6X
NGINX	1X	1.7X	3.5X
x.264[2]	1X	1.7X	2.25X
Cassandra	1X	1.1X	1.8X

Cores	1200	1792	4864
Servers	15	14	38

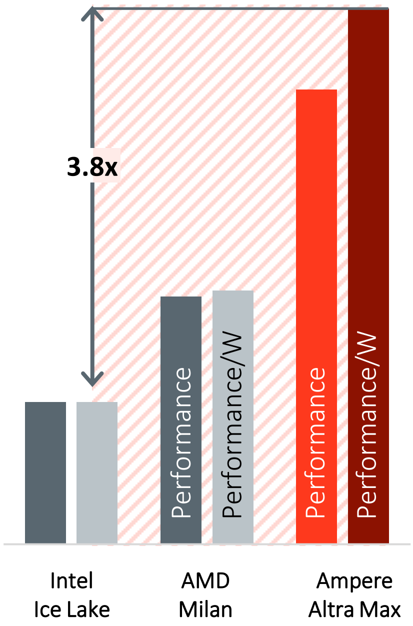
Get 2-3X Better Performance for equivalent power

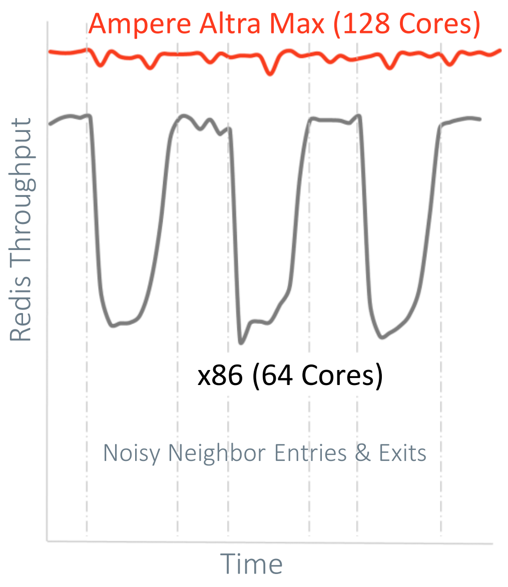
AMPERE.

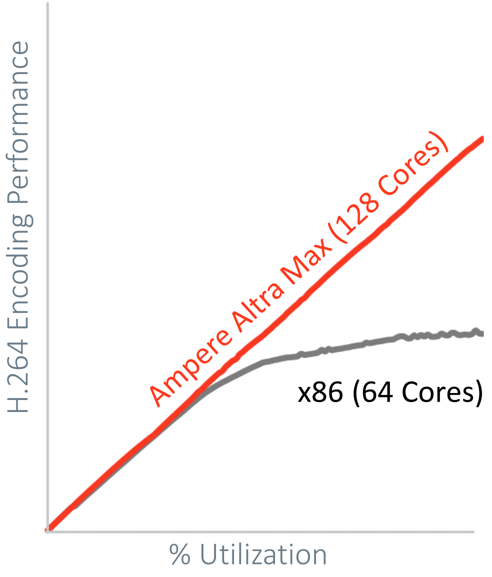
Notes:
1. Ampere internal models and analysis to identify total compute performance and system usage power consumption numbers, in standard 42U 12.8kW rack, see end notes
2. Data point uses data taken on M128-30 whereas all other data points use the M128-26.

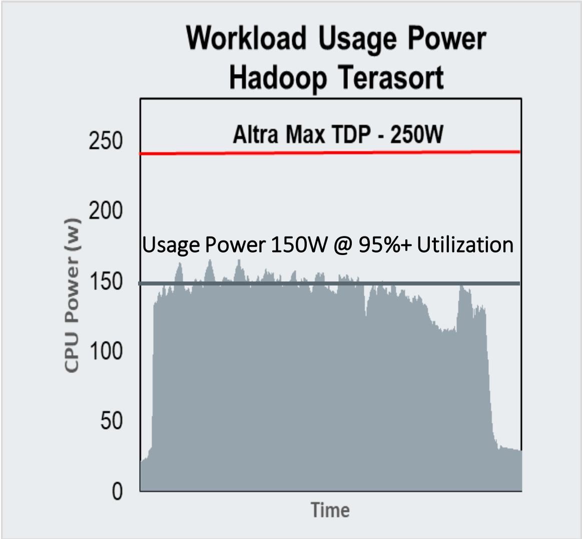
The Cloud Native Processor Value Proposition

Web Services (NGINX)[3]	Memory Caching (REDIS)[3]	Video Services (H.264)[3]	Data Services (Hadoop)[3]









3-4x Better Perf. & Efficiency	Consistent. Predictable.	Linearly Scalable	Low usage power!
No Compromises	*ZERO Loss to Noisy Neighbors*	*Stop Stranding Capacity*	*Less Power is the New Power*

AMPERE.

(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)

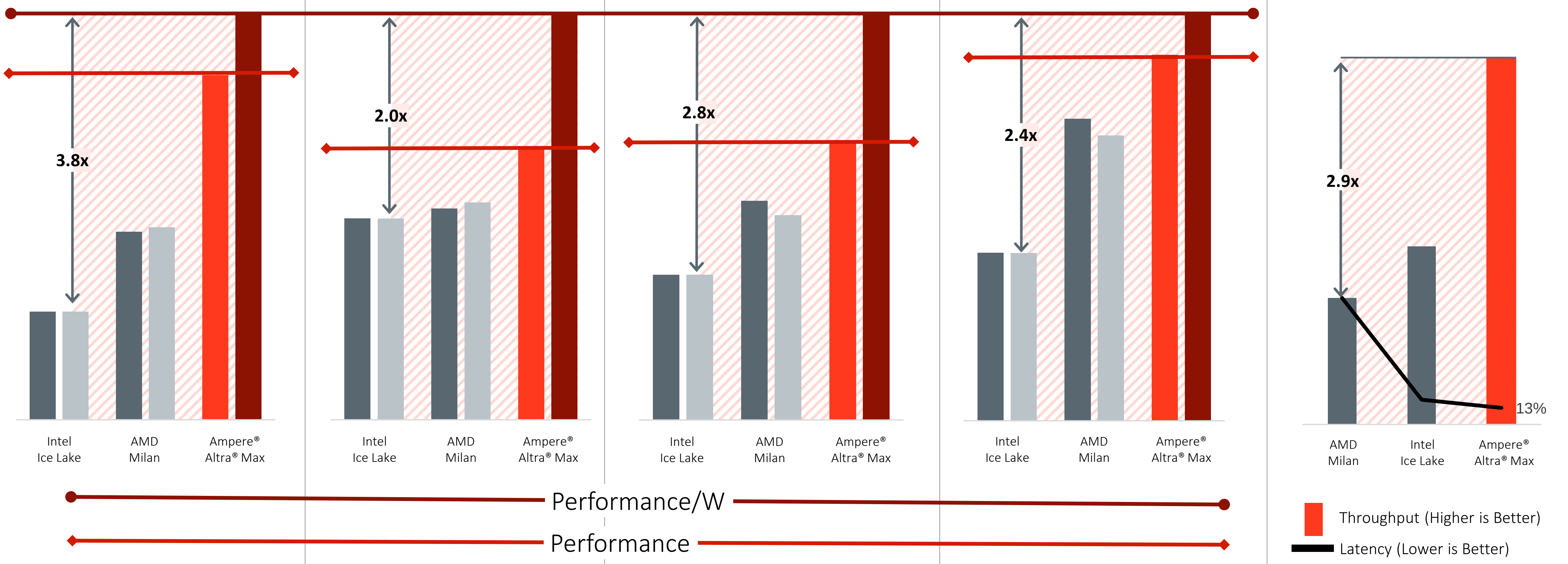
Web Services (NGINX)(3)

Database (MySQL)(3)

In-Memory Caching (Redis)(3)

Media Transcoding (h.264)(3)

AI Inference Image Classification (ResNet-50)(4)

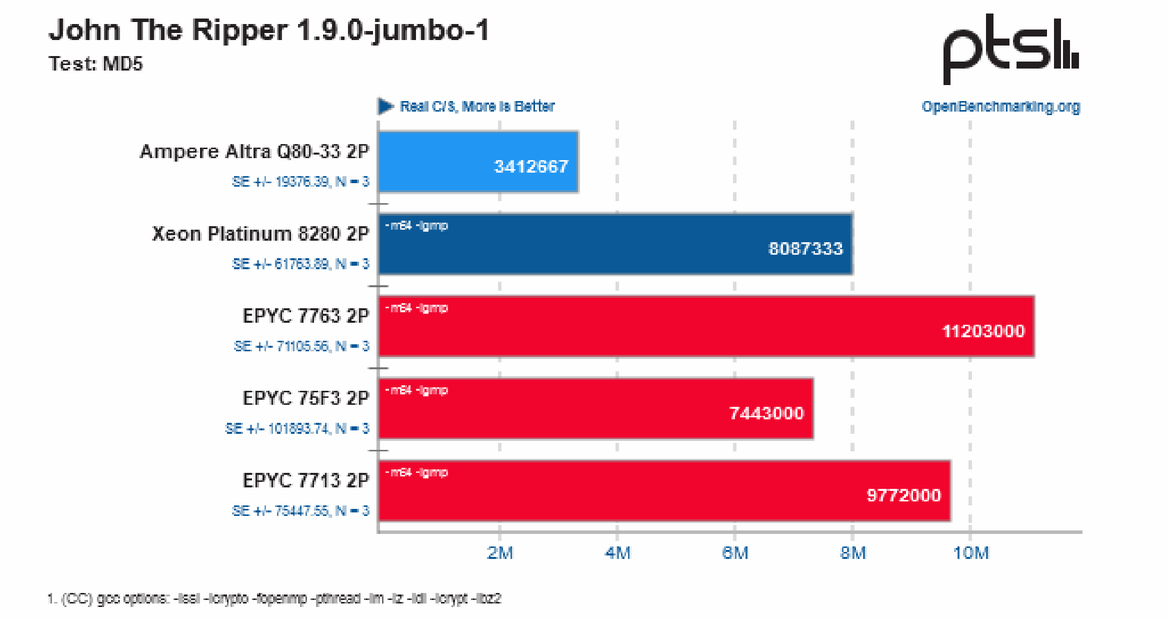


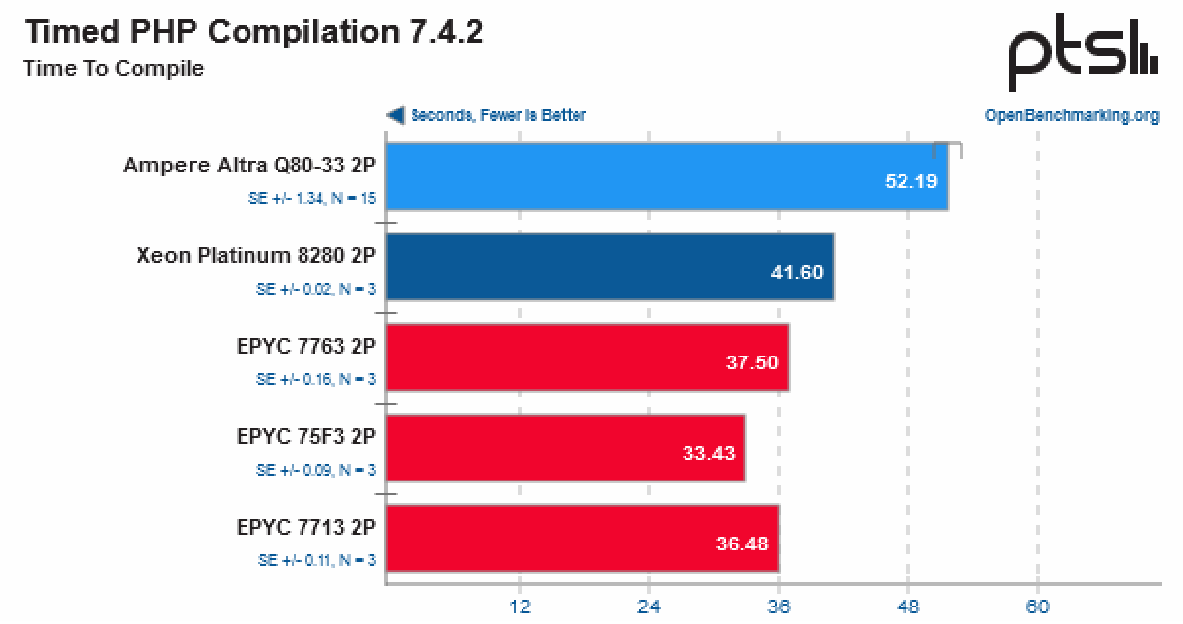
Notes:
1. Based on Company benchmarking
2. Intel Ice Lake represents Intel 8380 SKU; AMD Milan represents AMD 7763 SKU.
3. Percentages represent AMD Milan and Ampere® Altra® Max indexed against Intel Ice Lake
4. Percentages represent Intel Ice Lake and Ampere® Altra® Max indexed against AMD Milan

AMPERE.

Solution « cloud native » de Ampere : en pratique

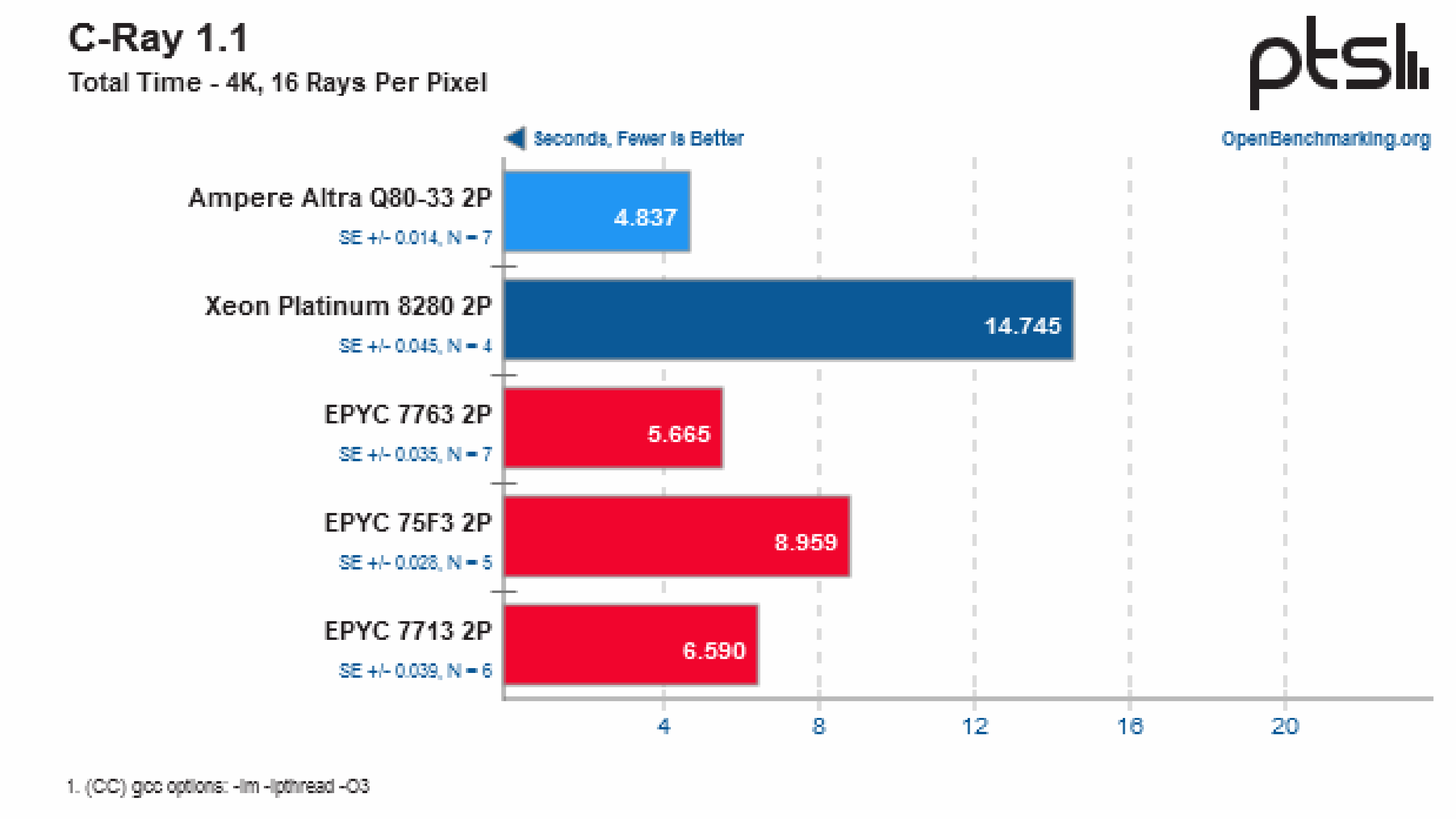
- De manière général, les benchmarks plutôt orientés *single thread* ne donne pas l'avantage à Ampere, la faute à de « petits » cœurs, mais aussi l'architecture Neoverse de première génération qui n'est pas la plus à jour.





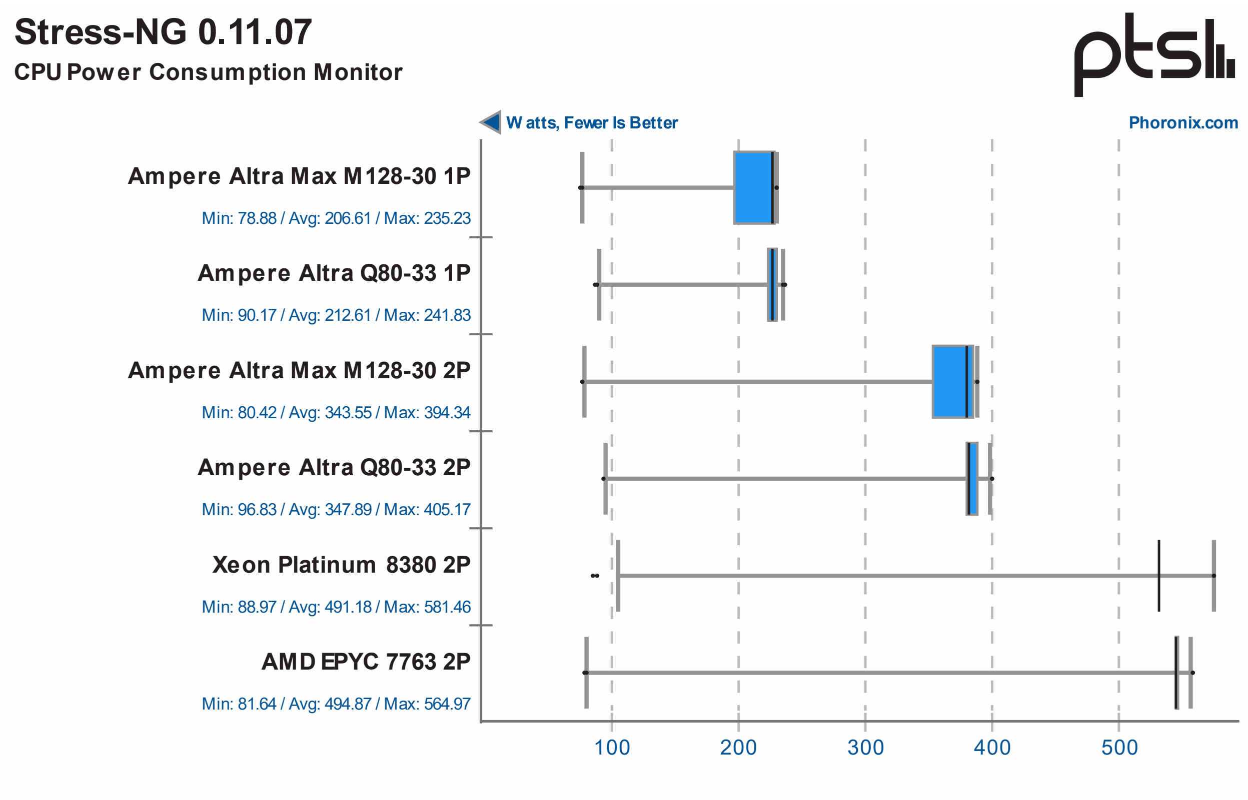
Solution « cloud native » de Ampere : en pratique

- 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.



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

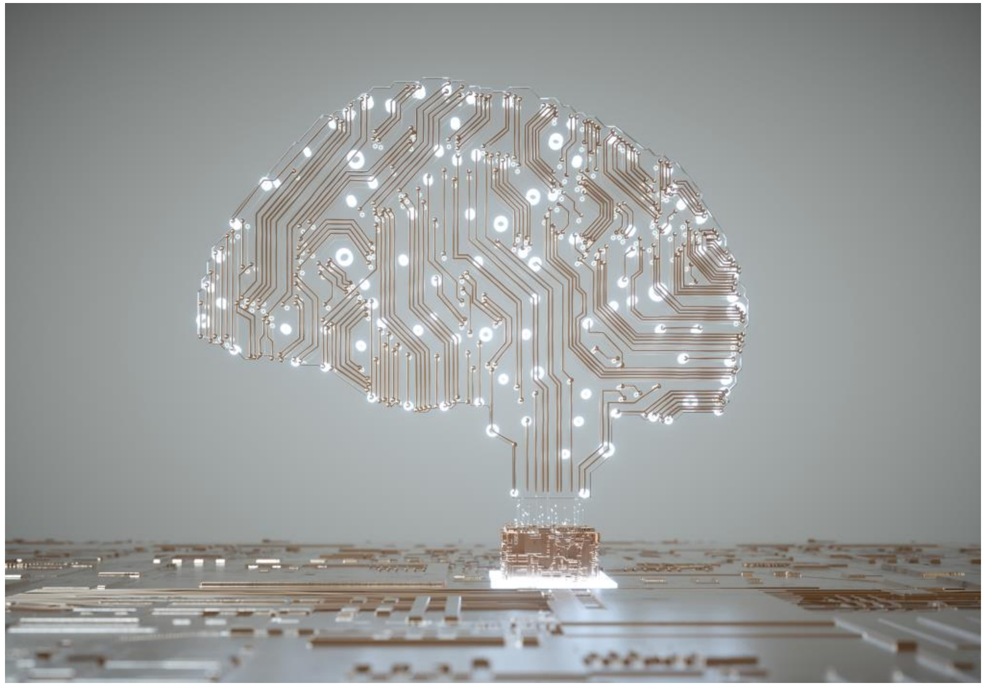
Native support for FP16 boosts performance without accuracy tradeoff

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

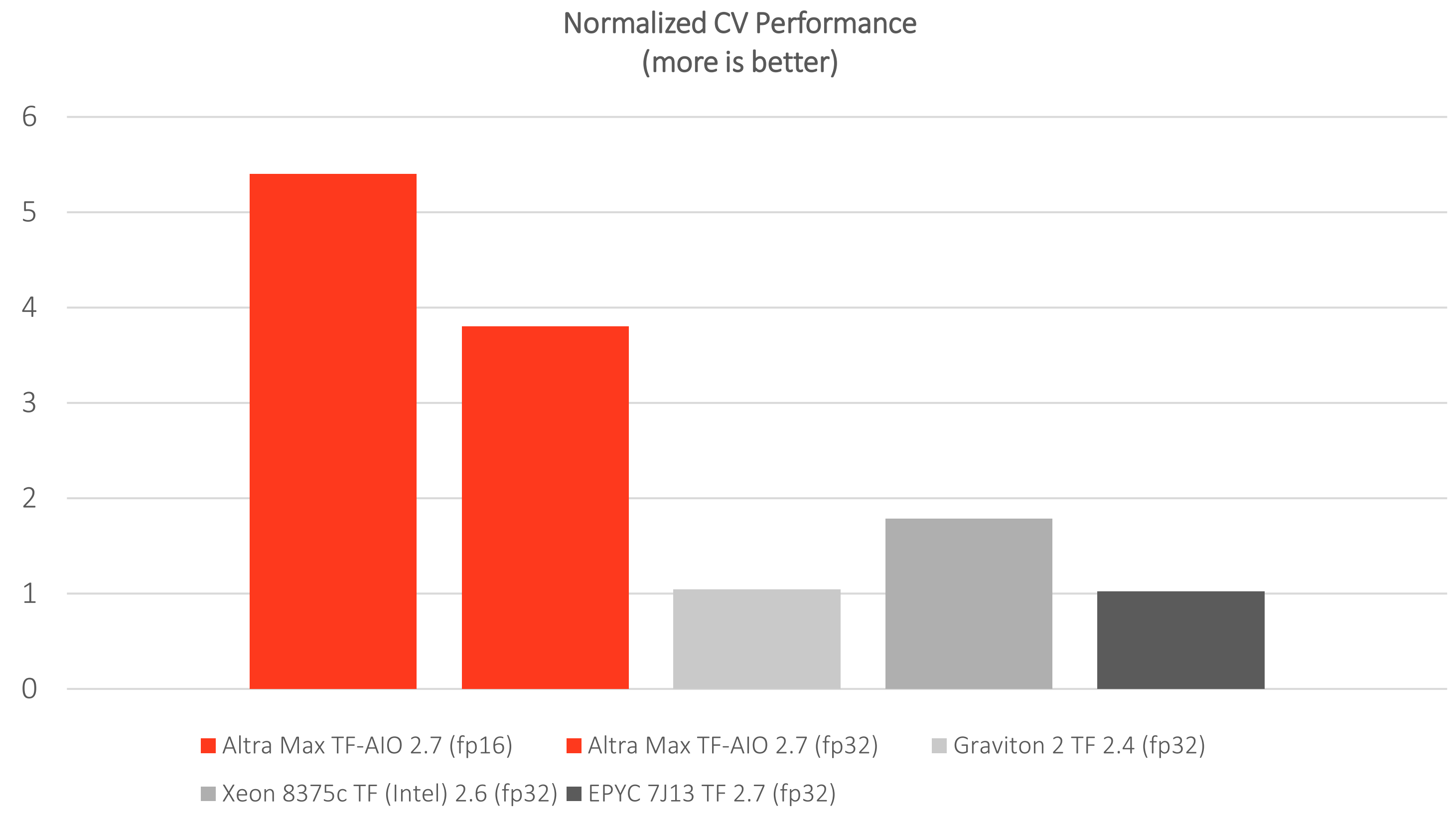
AI **training**: Ampere Altra Systems with Nvidia GPUs

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

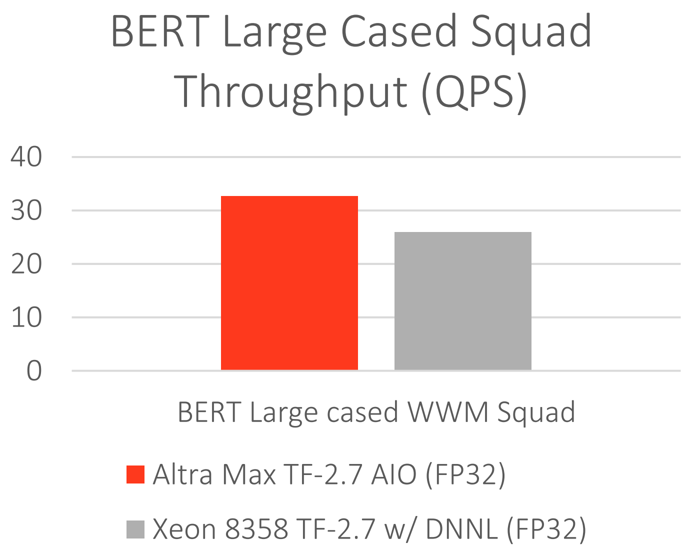
fp16 is natively supported in Ampere Altra family

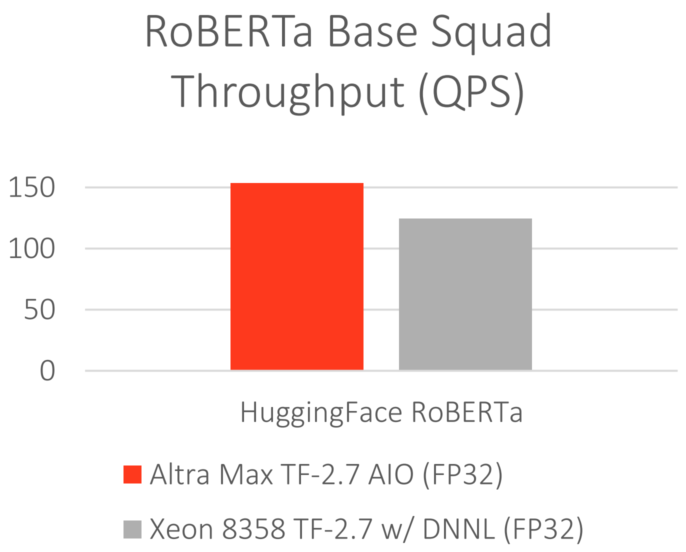
- Up to 2x faster than fp32
- Accuracy is on par with fp32, simple conversion

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

	Natural Language Processing (NLP)	
Characteristics	• CPU better in handling unbatched real-time NLP tasks (compared with GPU) • Model reduction can further improve CPU performance. • Altra / Altra Max can readily take advantage of FP16 with simple conversion	
Models	BERT Large Cased WWM Squad	RoBERTa Base Squad
CPU Perf Ampere Altra Max vs IceLake	• Ampere optimized frameworks enhances BERT performance on Altra Max. • 1.25x over Intel IceLake.	• Ampere Altra/Altra Max delivers strong performance on HuggingFace RoBERTa Model. • 1.25x over Intel IceLake





Ampere® AI: Training

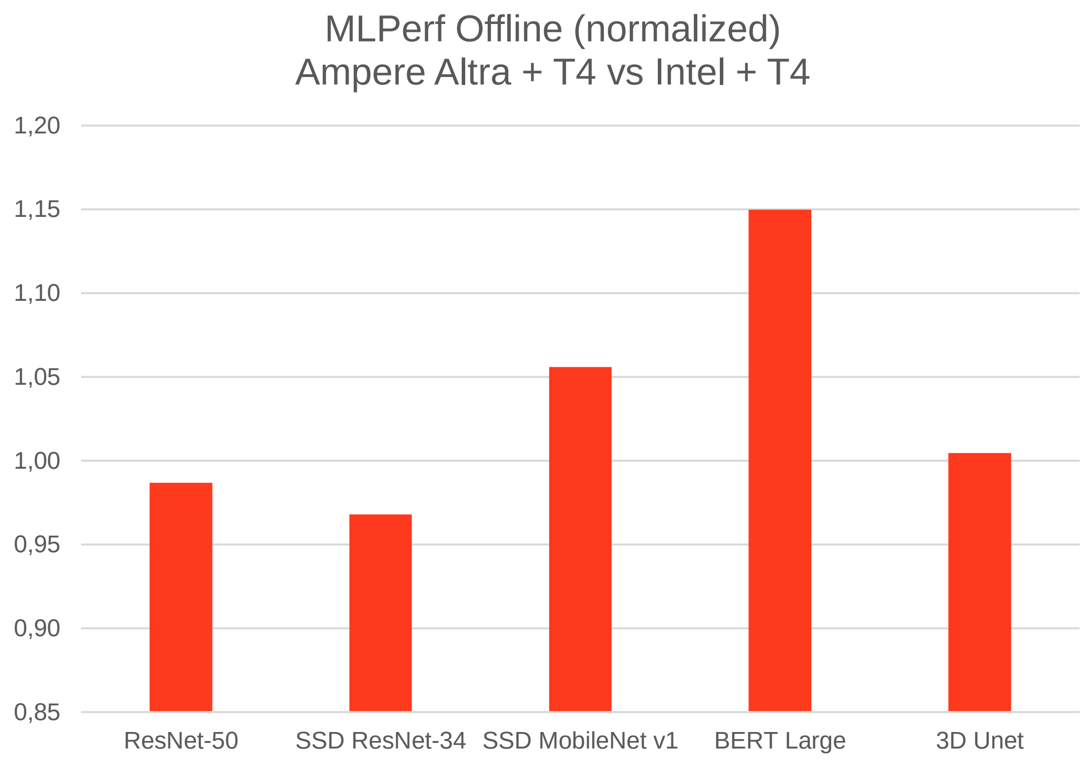
Standard Nvidia's software packages work on Ampere Altra out-of-the-box

- TensorFlow-GPU
- TensorRT
- CUDA

Same performance as x86 + Nvidia

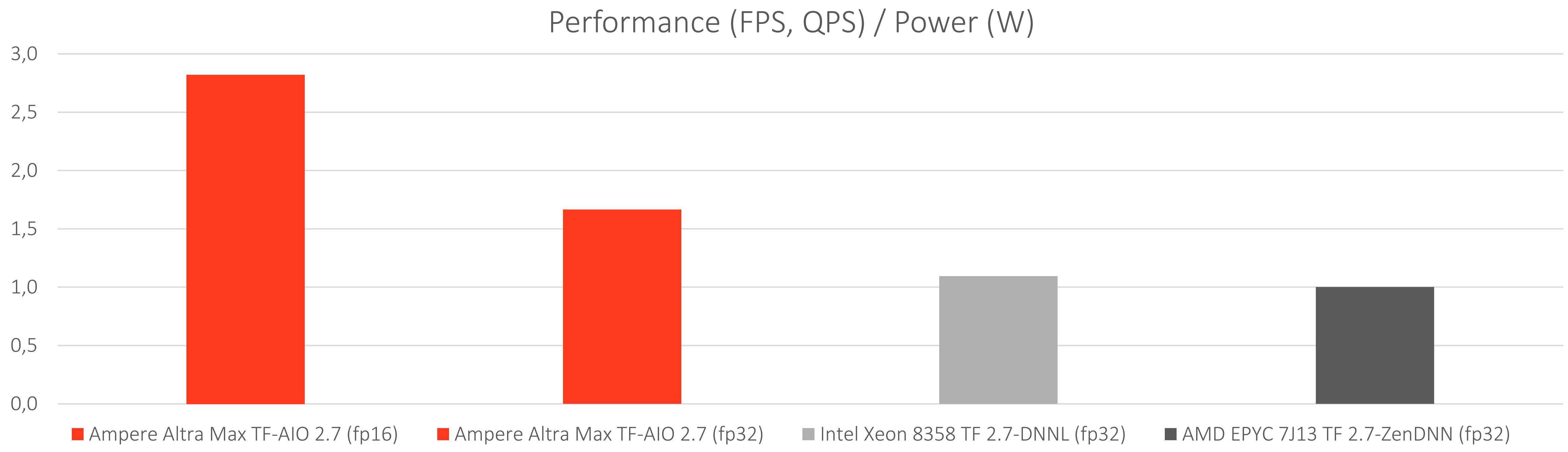
CPU+GPU primarily used for training, high throughput inference can take advantage

Seamless pathway from Nvidia GPU to Ampere CPU with fp16



TensorRT: same T4 GPU performance with Ampere Altra and x86

Ampere AI Performance / Power



Ampere AI has 60% perf/watt advantage on TensorFlow workloads over x86

With fp16 the perf/watt advantage increases to 180% over x86

TDP: 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

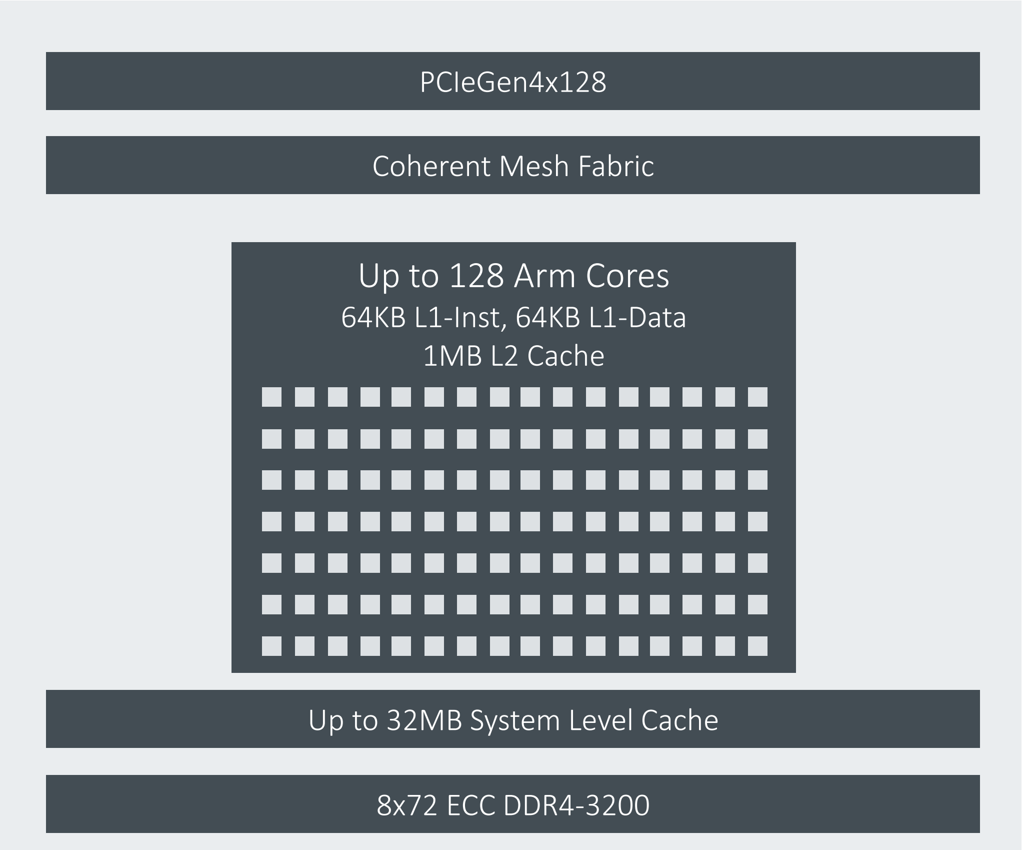
Workload	Use Case	Altra	Altra Max	Unit	Best
HPL Linpack	Scientific Performance	1369	1597	GFLOP/s	Higher
High Performance Conjugate Gradient	Scientific Performance	---	21.3	GFLOP/s	Higher
OpenFOAM Motorbike – 6 processes	Mesh time	---	65	Seconds	Lower
	Execution time	---	131	Seconds	Lower
OpenFOAM drivaerFastback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

1. GROMACS tests were performed using armclang and armpl
2. Other workloads tests were performed with gcc and BLIS (https://github.com/flame/blis)

Ampere Altra Family Overview

Ampere Altra Family Overview

Processor Subsystem	• Up to 128 Armv8.2+ 64-bit CPU cores @ 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 32MB 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 7nm FinFET • Arm v8.2+, SBSA Level 4
Power	• 40W – 187W Usage Power* • 65W – 250W TDP • Advanced Power Management
Performance	• Estimated SpecRate2017_int_base: Up to 359*

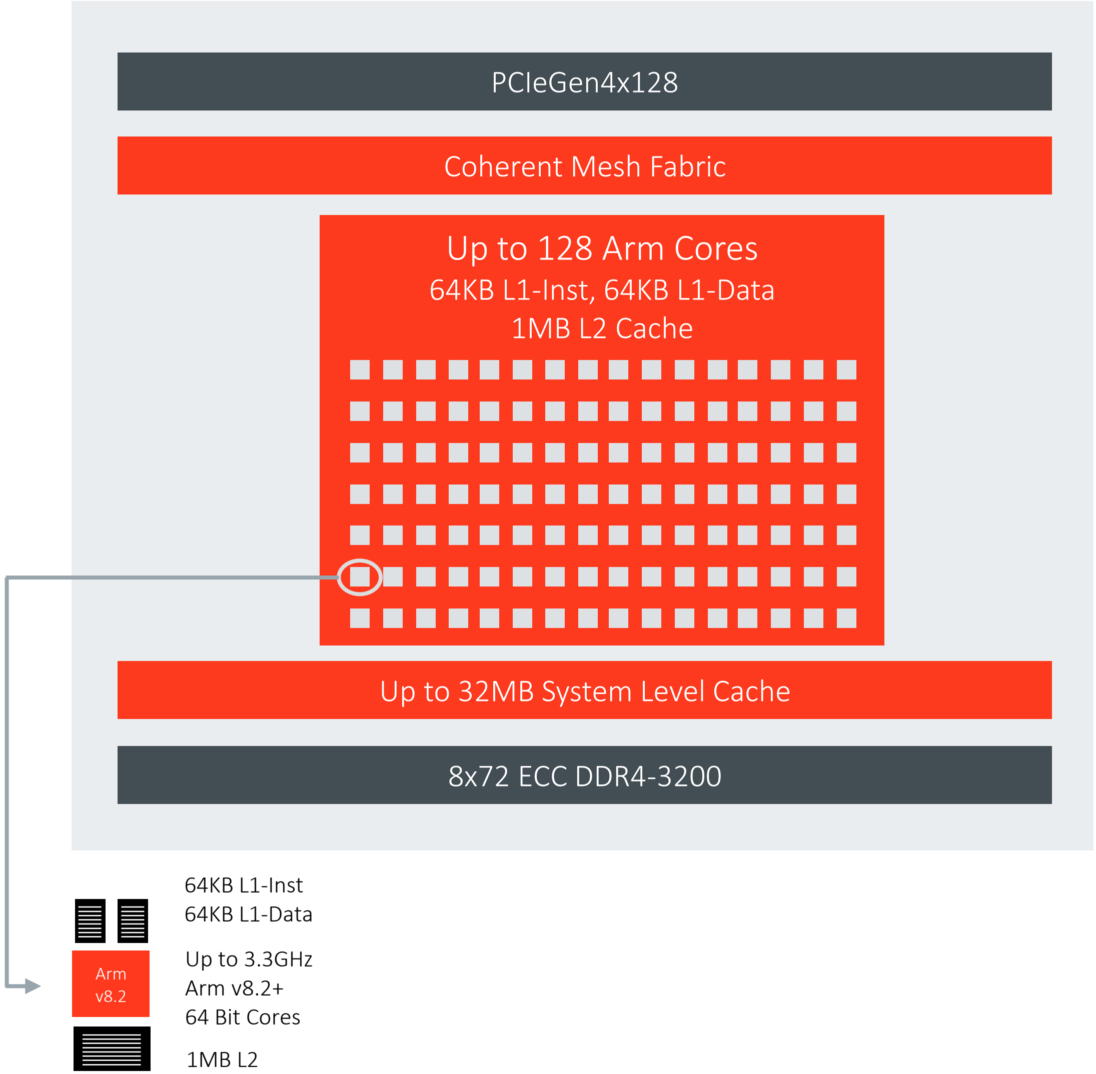


*Performance and usage power data are based on estimated SPECrate®2017_int_base (GCC10) and are subject to change based on system configuration and other factors. Usage Power is defined as average power consumed over time by a given workload.

AMPERE.

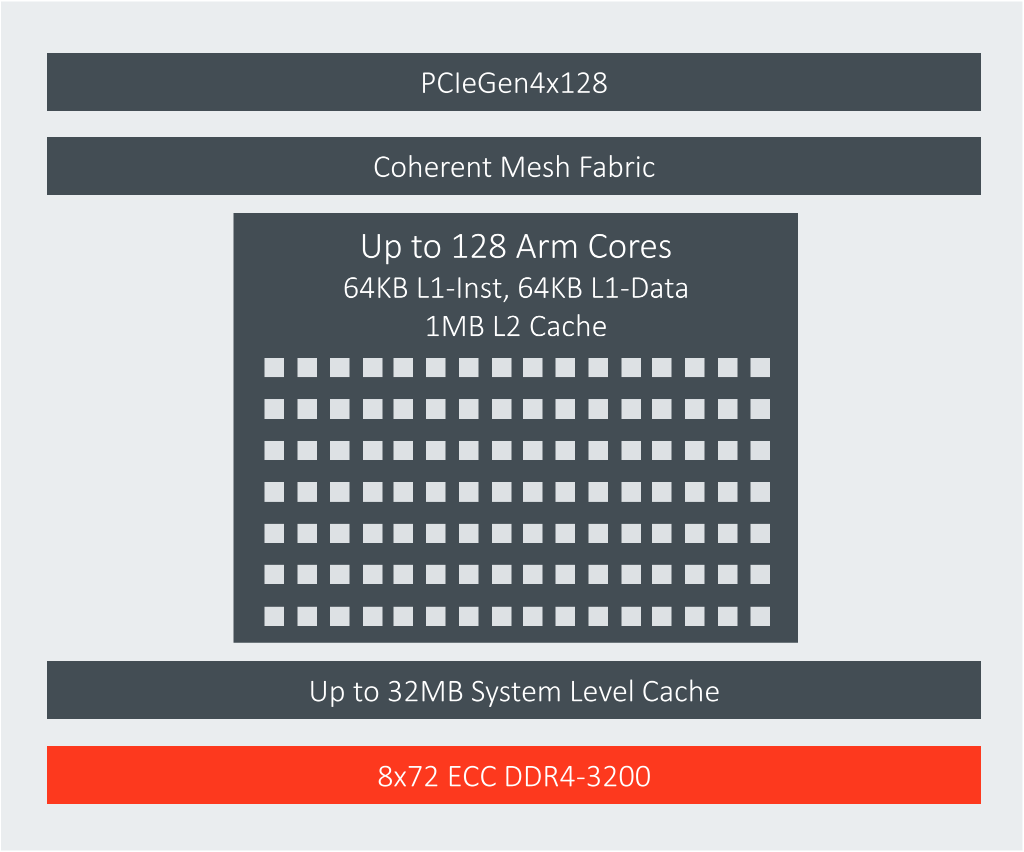
Ampere Altra Family Processor Complex

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

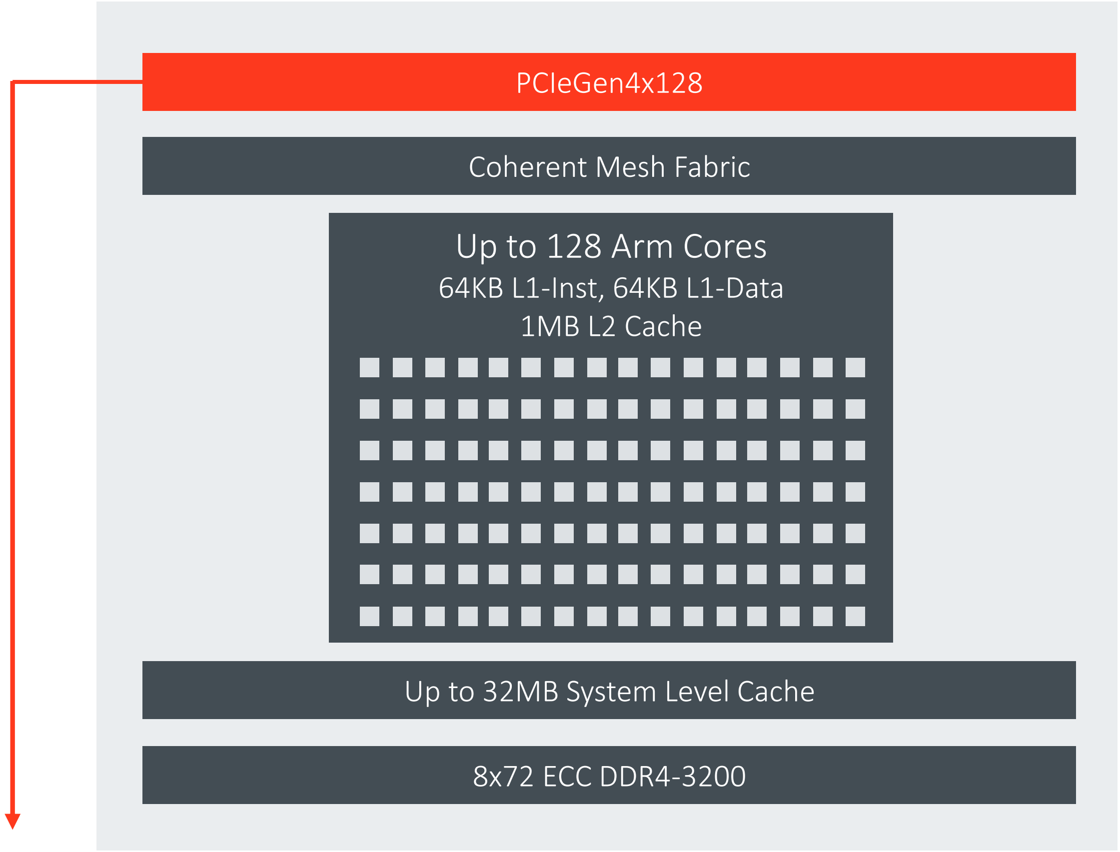
Bandwidth and Capacity	• Eight 72-bit DDR4 channels • UP to DDR4-3200 • Up to 2DPC • Up to 4 TB of memory
Supported Devices, Modules, and Configurations	• Support for UDIMMs, RDIMMs, LRDIMMs, and 3DS • Support for x4 and x8, and for 8Gb and 16Gb devices • Production support for 4, 6, and 8 active channels
Additional Features	• Hashed memory interleave across active channels • DRAM throttling



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



Supported on Altra Max only
Supported on both
Supported on Altra only

1x16	1x16	1x16	1x16	1x16	1x16	1x16	1x16
2x8	2x8	2x8	2x8	2x8	2x8	2x8	2x8
4x4	4x4	4x4	4x4	4x4	4x4	4x4	4x4
8x2	8x2					8x2	8x2
PCIE Gen 4	PCIE Gen 4	PCIE Gen 4 CCIX	PCIE Gen 4 CCIX	PCIE Gen 4 CCIX	PCIE Gen 4 CCIX	PCIE Gen 4	PCIE Gen 4

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

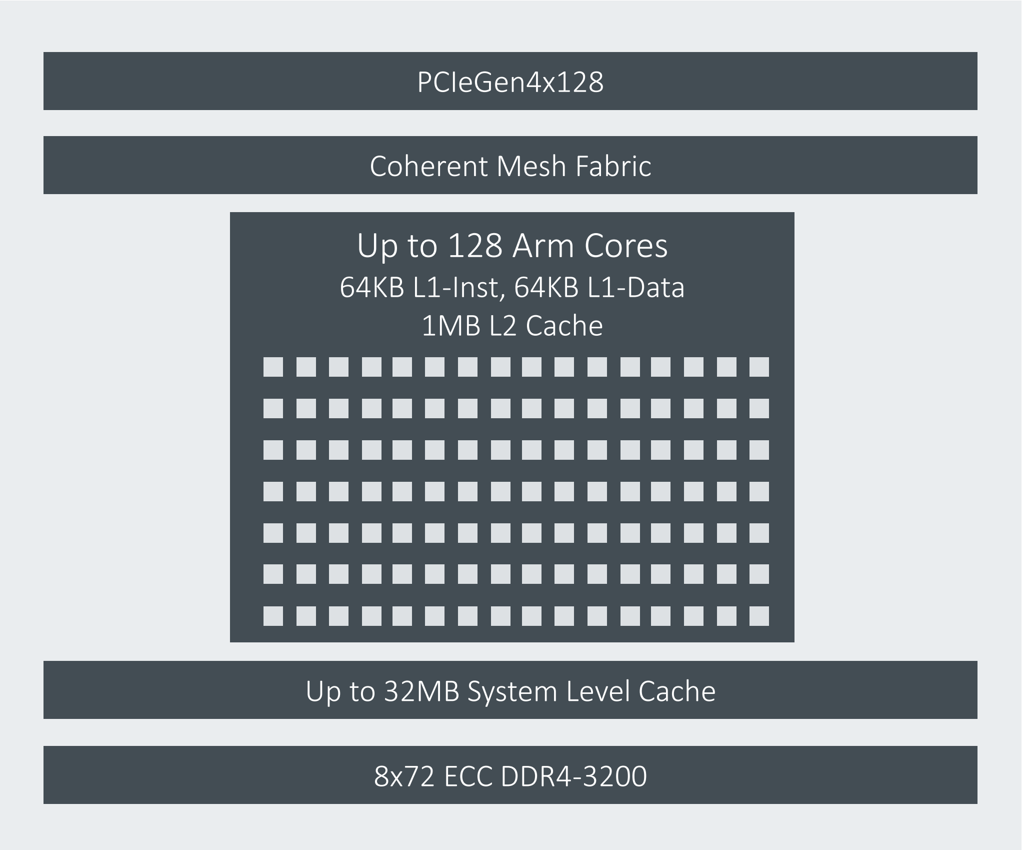
- Cortex-M3 Arm processor (400 MHz)
- Responsible for wide range of power and thermal management
  - Power management
  - Temperature control
  - Dynamic voltage frequency scaling (DVFS)
  - Max Frequency mode
  - ACPI and logic
  - Sensor logic
  - Interface to CPUs/PMPro (doorbell interrupts, messaging)

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 GPIs

Device Timers

- Two watchdog timers
- Four system timers



Software, System Firmware Platform Tools & Design Collaterals

Verified Linux Operating Systems

AlmaLinux

Alma 8.5

debian

Debian 11

fedora

Fedora 35

ORACLE Linux

Oracle Linux 8.5

Red Hat

RHEL 8.5

Rocky Linux™

Rocky 8.5

SUSE

SLE SP3

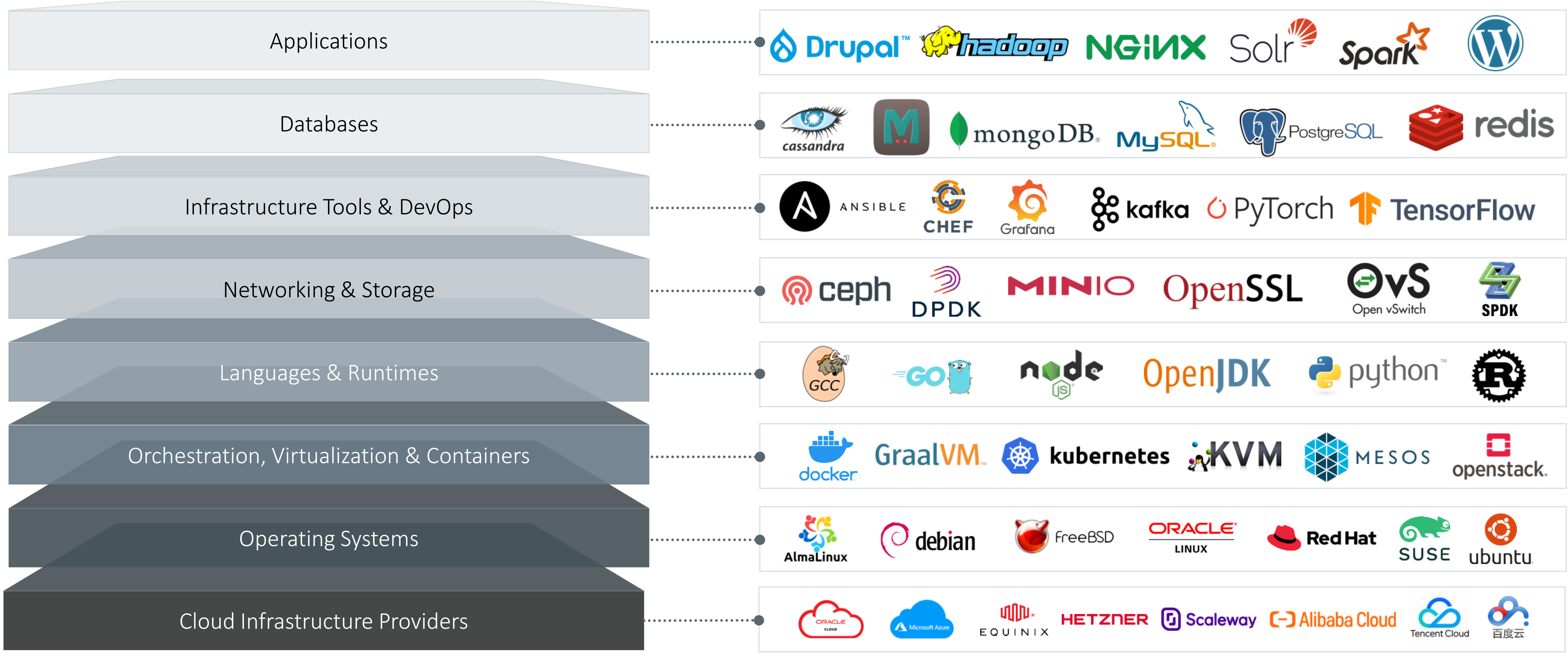
ubuntu

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

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

Orchestration & Containers
Kubernetes, Docker, K3s, Rancher, OpenStack, Mesos, Lokomotive, 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

Big Data: Hadoop, HDFS, Spark, Flink

Frameworks	Caffe	ONNX	TensorFlow	PyTorch		Apache Storm	CUDA
Middleware	DPDK	ISA-L	OvS	SPDK	Ceph	x.264/265, AV1, VP9	OpenSSL
OS[1], VMM's and Compiler Support	CentOS	debian	GCC	KVM	LLVM Compiler Infrastructure	ORACLE Linux, redhat	SUSE, ubuntu
BIOS/UEFI and BMC	American Megatrends	AptioV	MEGARAC	LinuxBoot	OpenBMC	tianocore	

Many more SW Stacks & Daily Regressions → https://solutions.AmpereComputing.com

[1]Support for Linux OS's → https://github.com/AmpereComputing/ampere-centos-kernel/wiki

AArch64 is fully supported by major Linux distros

debian

Community Options
Debian 10
Debian 11

ORACLE LINUX

Commercial Options
Oracle Linux 7.9
Oracle Linux 8

redhat

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

suse

Commercial Options
SLES 15 SP2
SLES 15 SP3
SLES 15 SP4

Community Options
openSUSE Tumbleweed & Leap 15

ubuntu

Commercial Options
Ubuntu 18.04 LTS
Ubuntu 18.04 HWE
Ubuntu 20.04 LTS
Ubuntu 20.04 HWE

Ampere Altra Family
Server Platform Overview

Ampere Altra Family Platform Overview

ADLINK LEADING EDGE COMPUTING

Production

Production

Foxconn Industrial Internet

Sampling

Mt. Bonnell

Production

Mt. Collins

GIGABYTE™

Production

Mt. Snow

Production

High Density Server

HUAQIN 华勤通讯

Production

P6410

inspur

Production

Aoqin

SUPERMICRO

Production

Mt. Hamilton

wiwynn®

Production

Mt. Jade

Ampere Altra Family Platform Configuration Info found at https://solutions.amperecomputing.com/systems/altra

Ampere Developer Program

Ampere Developer Center

Application Architects

Build Engineers

Deployment Engineers

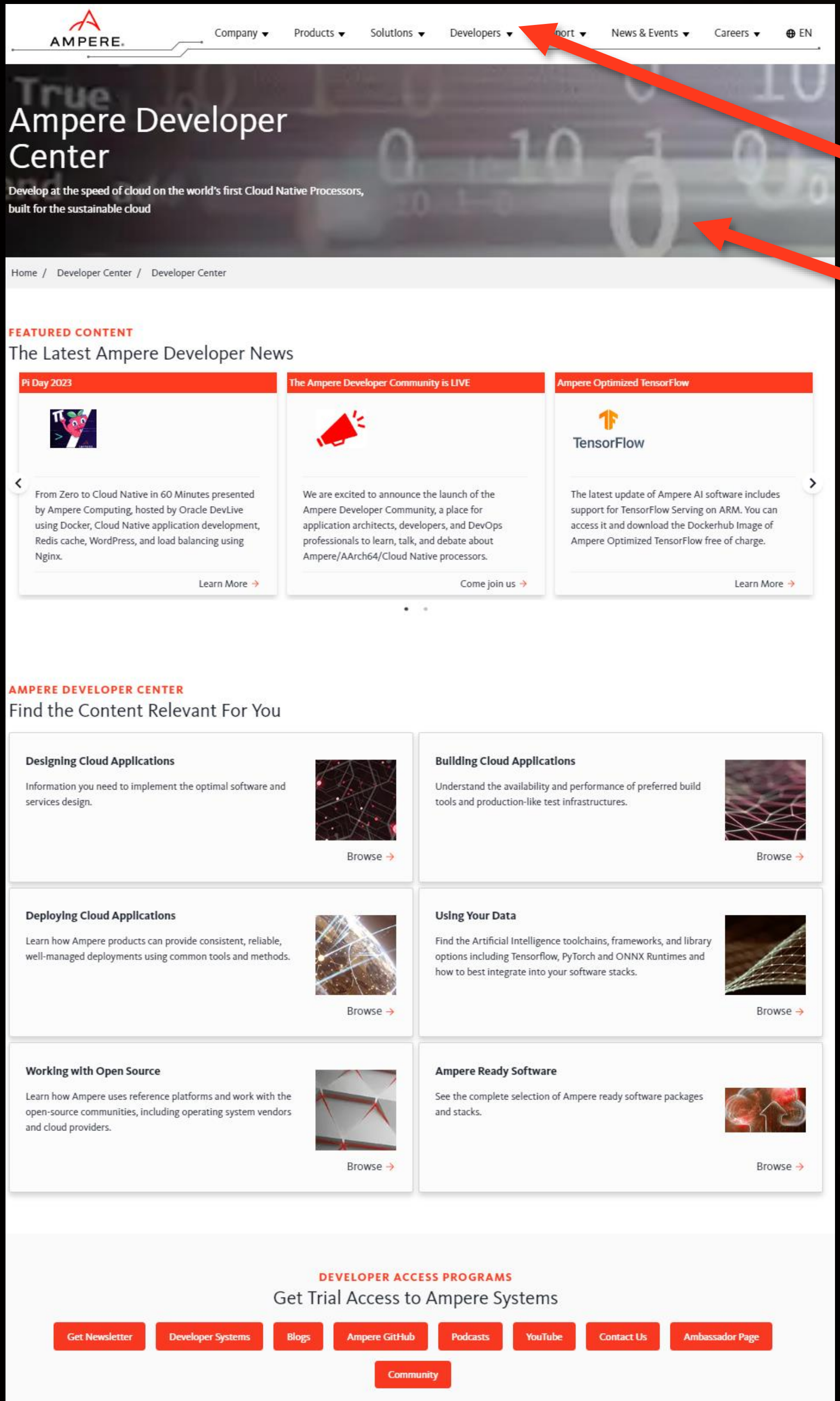
Data Scientists

OS / Kernel Engineers

Ampere Ready Software

Ampere Developer Center

https://developer.amperecomputing.com



Navigation from Corporate site

Updated UI

Latest News

Curated Content by Persona

Ampere Ready Software

Action Bar (github, newsletter, blogs, systems, youtube)

50+ pieces of developer Specific Content

Refreshing with one new piece / week in 2023

Focusing on Tutorials and Transition & Tuning Guides

The Ampere Developer Program

developer.amperecomputing.com

How to connect

Ampere Developer Program

AMPERE

https://developer.amperecomputing.com

developer@amperecomputing.com

Sign up for our Developer Newsletter

https://developer.amperecomputing.com/newsletter

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

Designing Cloud Applications
Information you need to implement the optimal software and services design.

Building Cloud Applications
Understand the spectrum of build tools to allow you to confidently move from test to production.

Deploying Cloud Applications
Learn how Ampere products can provide consistent, reliable, well-managed deployments using tools and methods.

Using your data
Find the optimal Artificial Intelligence toolchains, frameworks, and library options

Enabling the Open-Source Community
Learn how Ampere uses reference platforms and work with the open-source communities, including operating system vendors and cloud providers.

Ampere Ready Software
See the spectrum of software running across Ampere-based instances.

End Notes

End Notes

Hardware Configuration
Ampere Altra® Q80-33, 80 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
Cascade Lake Refresh – 28 Cores/Socket, Ice Lake – 40 Cores/Socket, AMD Rome – 64 Cores/Socket, AMD Milan – 64 Cores/ Socket, Ampere Altra® Q80 – 80 Cores/Socket, Ampere Altra® Max M128 – 128 Cores/Socket

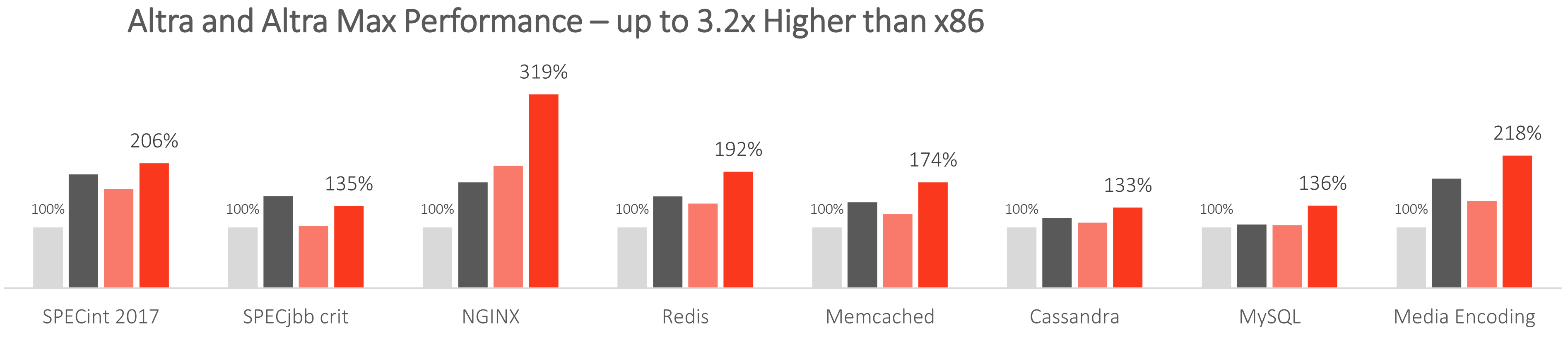
Power Numbers (TDP)
Cascade Lake Refresh – 205W, Ice Lake – 270W, AMD Rome – 240W, AMD Milan – 280W, Ampere Altra® Q80 – 250W, Ampere Altra® Max M128 – 250W

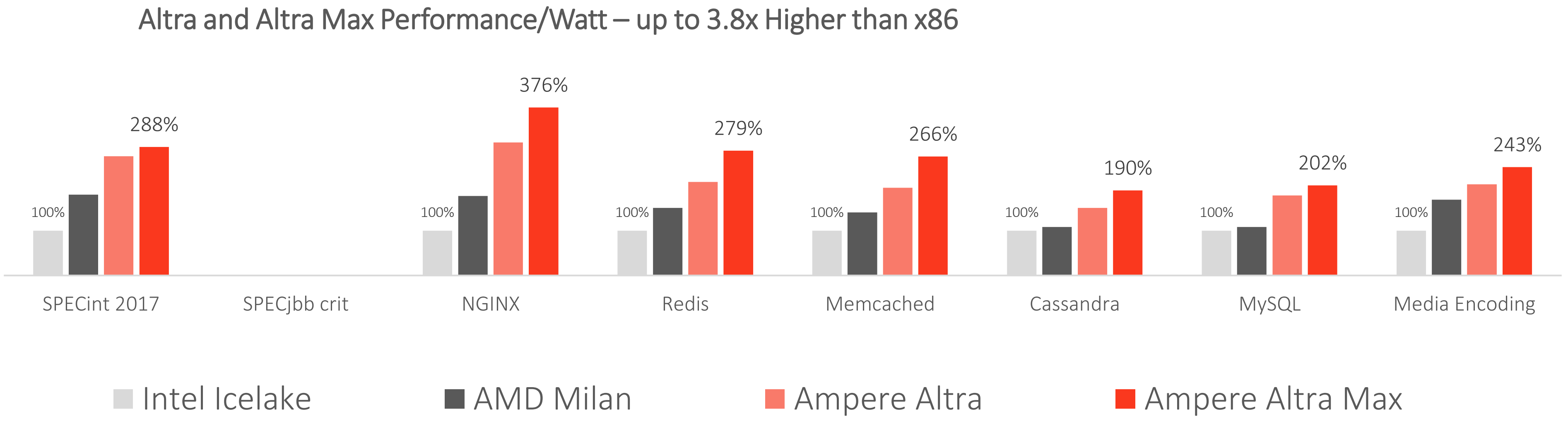
TensorFlow Comparison
Ampere Altra® Max M128-30, 128 Cores, Ubuntu 20.04, TensorFlow-AIO 2.4, ML Perf Version v1.1
Ampere Altra® Q80-30, 80 Cores, Ubuntu 20.04, TensorFlow-AIO 2.4, ML Perf Version v1.1
AMD EPYC 7571 (AWS:m5s.24xlarge), Ubuntu 20.04, TensorFlow-AIO 2.4, ML Perf Version v1.1
Intel Platinum 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 Cores, Ubuntu 20.04, TensorFlow-AIO 2.4, ML Perf Version v1.1
Ampere Altra® Q80-30, 80 Cores, Ubuntu 20.04, TensorFlow-AIO 2.4, ML Perf Version v1.1
Others
https://mlcommons.org/en/inference-datacenter-11/
https://mlcommons.org/en/inference-edge-11/

Ampere Altra Family
Workload Performance Overview

Performance Recap – Ampere Altra and Altra Max



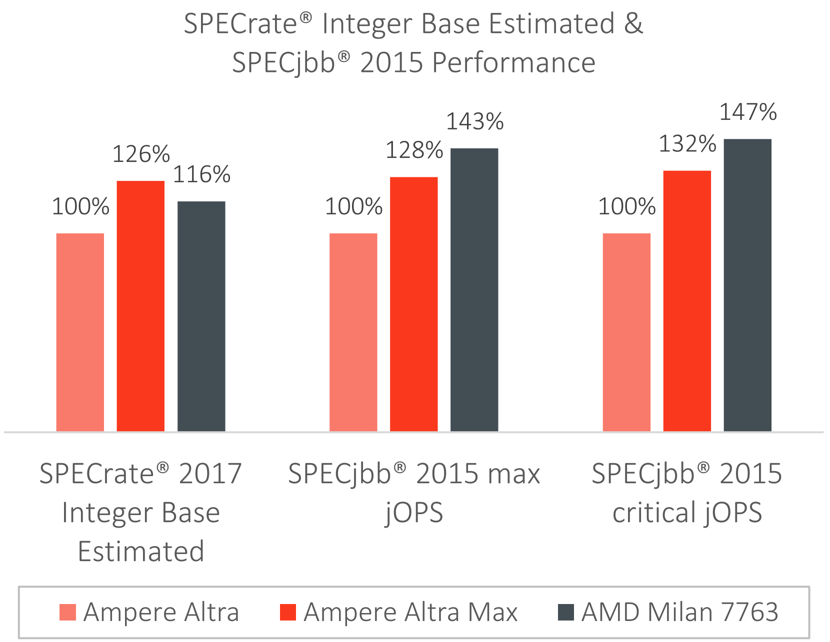


Industry Standard Benchmarks

Configuration: https://spec.org/cpu2017/results/res2021q3/cpu2017-20210811-28660.cfg

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

SPECjbb® 2015 Estimated	Altra® Q80-30 1P	Altra® Max M128-30 1P	AMD EPYC Milan 7763
Max jOPS	135,311	173,633	193,086
Critical jOPS	119,550	157,237	176,283



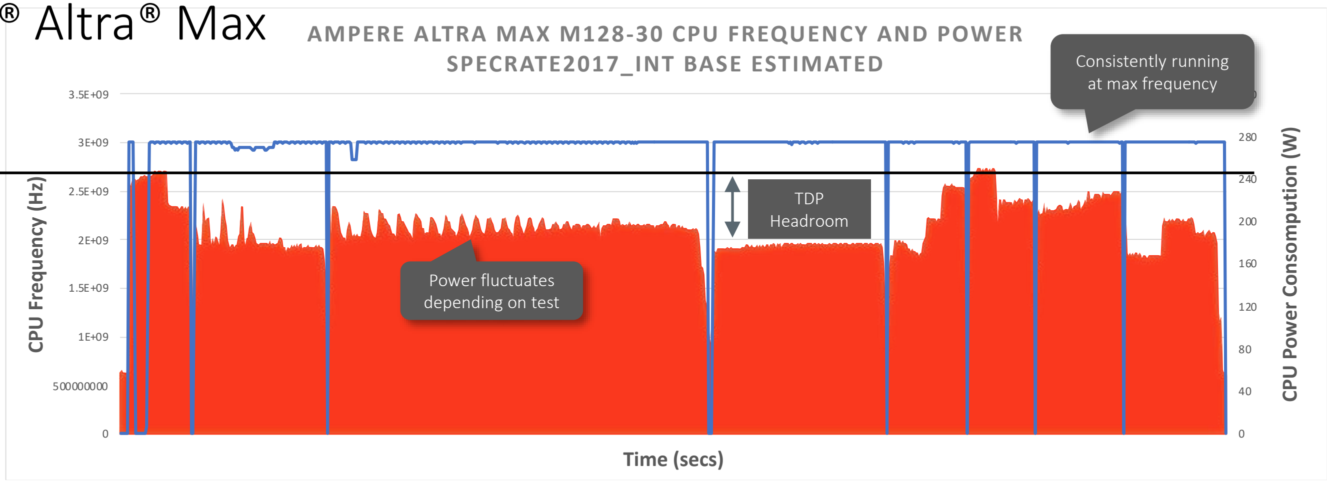
Industry-leading performance on Standardized Benchmarks using open-source compilers and JDK!

Ampere® Altra® Max Energy Efficiency

Power
CPU Frequency

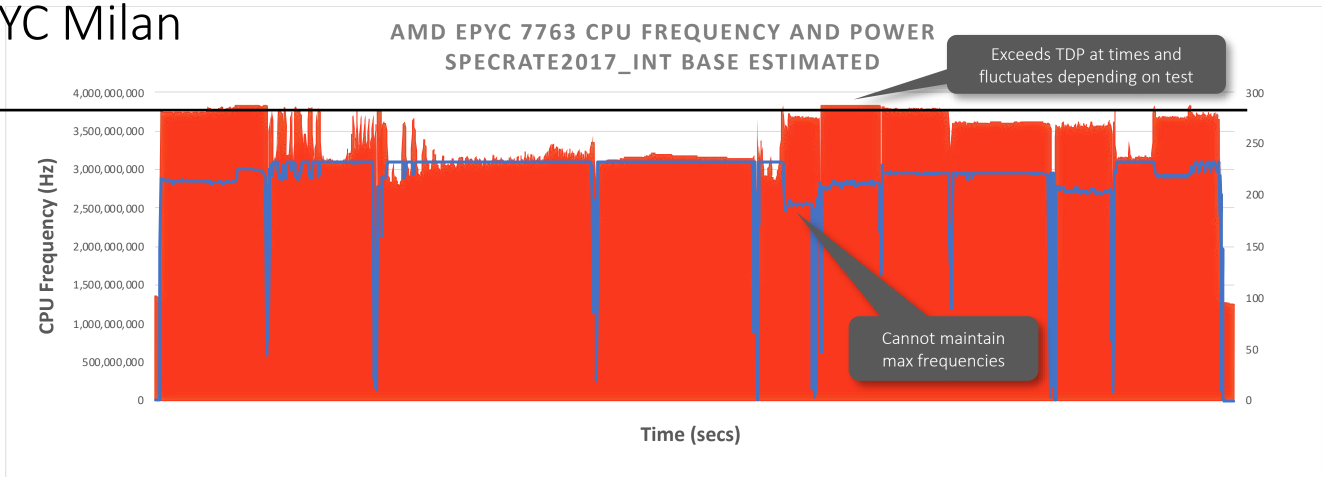
Ampere® Altra® Max

TDP=250W



AMD EPYC Milan

TDP=280W



	Performance	Usage Power (W)	Performance/ Watt
AMD EPYC Milan	331	280W	1.0x
Ampere® Altra® Max	360	178W	1.71x

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

AMPERE