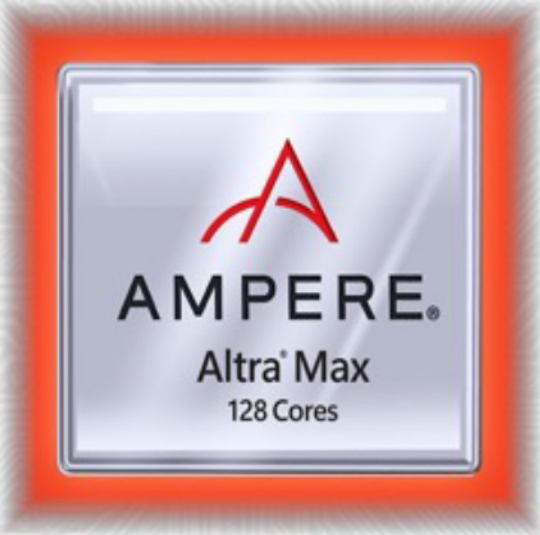
Cloud Native for the Edge

Scalable, High-Performance Compute for Embedded Applications



Introduction

We rely on edge devices to process large amounts of data in real time. Fast, consistent, and predictable compute is critical to applications such as industrial automation, self-driving cars, and content delivery. A challenge for these edge devices is that they usually must fit into a fixed size, weight, and power (SWaP) which cannot be increased to support the growth in the number of sensors, application demands, or networking bandwidth. In addition, 5G deployments with increased communications bandwidth and AI applications driving compute demand have made these challenges even more pronounced.

The edge compute market is projected to grow by 33% per year.[1] Our handheld devices, the sensors in our homes, and nearly every other convenience in our lives relies on local compute. With the number of connections to the internet growing by about a third each year, these devices will require even more computing capabilities over time.

With that increase in connections and data comes a corresponding increase in the energy needed to power the hardware, assuming it continues with legacy compute architectures such as x86. However, as the global energy crisis creates constraints on available and affordable power, developers and users need to maximize the *performance-per-watt* in edge devices more than ever before.

Even as the demand for high-performance computing at the edge continues to accelerate, so too does the need for more sustainable solutions that do not sacrifice performance for efficiency. Being able to deliver the required capabilities while not increasing size, weight, and power of the compute has quickly risen among the top priorities when developing new devices and solutions.

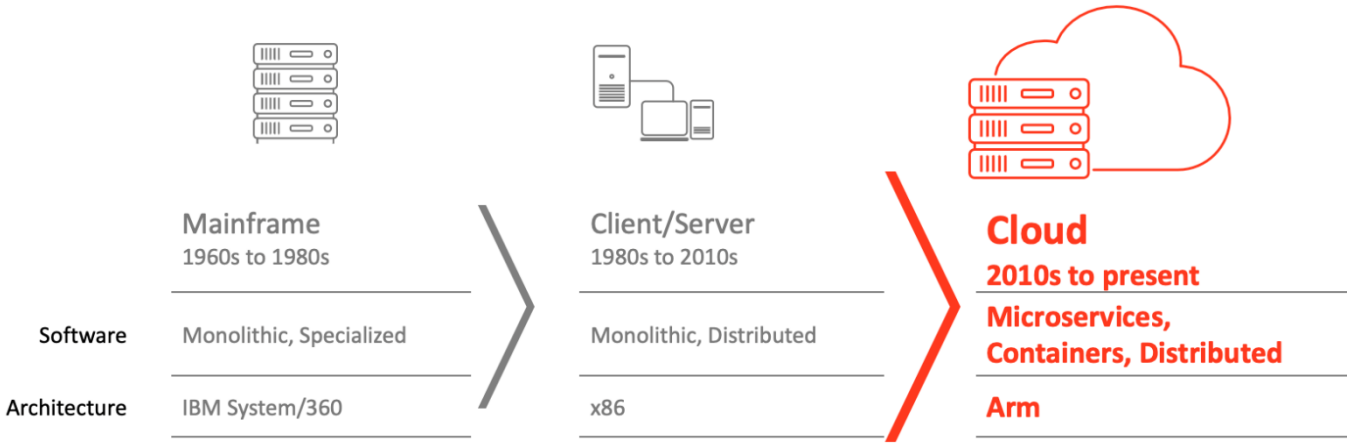
For example, the number of sensors for devices that process data for a variety of applications is increasing quickly. Even as their number grows, sensor bandwidth is also increasing, demanding higher resolution and faster run rates. Just five years ago, a 1 MP camera at 30 fps was the norm in manufacturing. Now it is common to have automatic optical inspection cameras that are 10 MP or higher running at 60 fps or faster. This twenty-fold increase in the data rate, ever faster networks, and higher sensor and network bandwidth are accelerating the adoption of higher performing and more efficient compute solutions.

This challenge of increasing the amount of compute with high speed and low latency while reducing power consumption requires a fresh approach and is the only way we can continue to enable those working in the embedded development communities, such as automotive, healthcare, smart cities, and more. Built for similar demands in cloud computing, Cloud Native Processors are outperforming legacy x86 processors and other Arm-based solutions in embedded applications and on the edge.

Cloud Native for Embedded and the Edge

Ampere Computing® launched the world's first Cloud Native Processors in response to compute needs shifting from the client-server model to today's modern era of cloud computing, where software is distributed and built for deployment and management at scale as shown in *Figure 1*. The core cloud native methodologies—elasticity, scalability, and flexibility—are now being built into edge platforms as well. In these types of deployments, microservices are typically loosely coupled using APIs and deployed using containers to deliver excellent resiliency, manageability, and observability through the entire stack, whether deployed on a single edge device or an entire fleet.

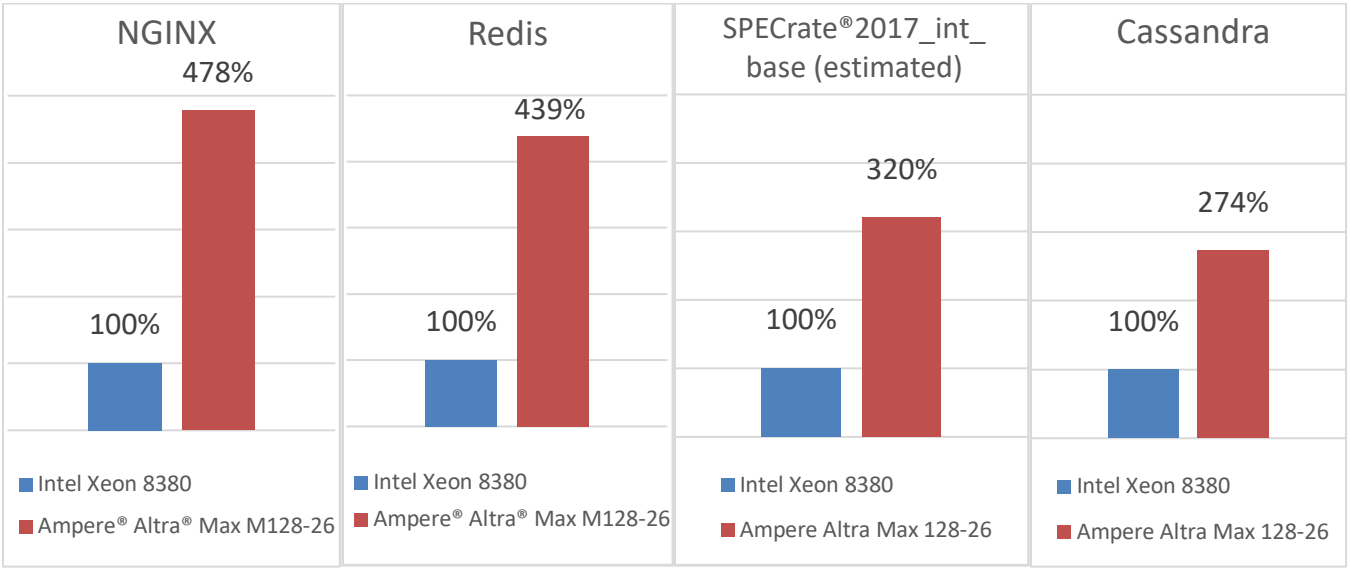
Figure 1: Compute Progression



Built using the Arm architecture, Ampere's Cloud Native Processors deliver significantly higher processing performance and greatly improved power efficiency compared with legacy x86 CPUs. Ampere's edge-focused processors range from 32 to 128 cores, drawing from 40 to 125 W under full load, referred to as "usage power".[2] The Ampere® Altra® family of Cloud Native Processors offer many times more cores and compute performance at the same or lower power consumption when compared to x86-based processors for a variety of popular workloads.

The high core density of the Ampere Altra family runs containerized applications more efficiently, maximizing the utilization of compute resources at the node and fleet-level as shown in *Figure 2*.

Figure 2: Processor Performance per Watt for Different Workloads

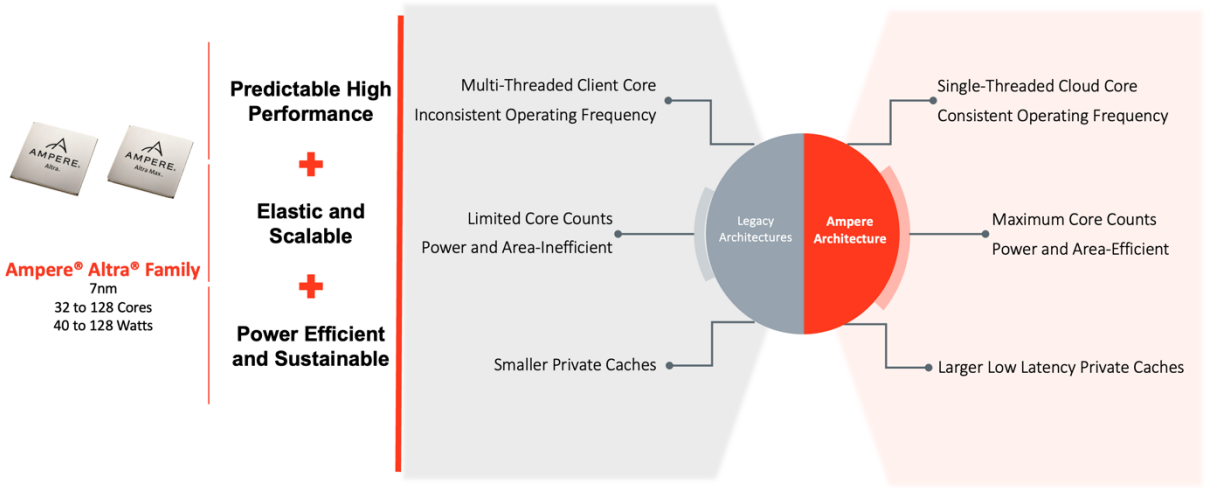


Ampere processors are excellent building blocks for developers of edge solutions. The Ampere Altra family of processors provides the flexibility needed to deploy low-power, high-throughput network devices, ultra-fast storage solutions, or low-latency signal processing, just to name a few. With the idle power as low as 11 W, there is an added power savings benefit for workloads that are not always operating at peak usage.

A New Architecture for a New Era

Ampere's Cloud Native Processor architecture produces remarkable performance and power efficiency with SKUs featuring unprecedented performance per watt for the edge. Ampere Altra processors also have up to 128 PCIe Gen 4.0 lanes, providing a large amount of I/O expansion for discrete GPUs, FPGAs, camera frame grabbers, network interface cards, storage, and other peripherals to support a multitude of use cases.

Figure 3: Ampere Altra Family Architecture Advantages



By utilizing the Arm ISA, the Ampere Altra family of processors provides an execution environment compatible with the various multi-core Arm embedded systems used in vehicles and other devices across the embedded and automation industries from robotics and satellites to Smart IoT.

The Ampere Cloud Native Processor architecture's three key features that make it ideal for the edge are:

1. **Maximum number of cores** that are power and area-efficient for your edge applications. Ampere Altra family processors range from 32 to 128 high-performance cores while providing large reductions in power consumption compared to legacy x86 processors.
2. **Single-threaded** execution within each of the cores allows for a consistent performance and operating frequency over time and across the processor. By contrast, x86-based processors with fewer cores try to increase their compute capacity by implementing multi-threading, where resources within the core are shared by multiple processes. A multi-threading architecture creates conflict between different workloads as they fight for the same resources and results in unpredictable performance. This forces platform and application providers to implement difficult workarounds or tolerate inefficient compute resources to maintain quality of service and determinism for all tenants and processes. At the same time, the operating frequency of x86 CPUs varies greatly over time, leading to unpredictable performance.
3. **Larger, high-speed, private, and low latency caches** when compared to legacy x86 CPUs[3]. The larger private caches accelerate the performance of each core's individual workload without creating conflict between various processes for the same resources. This is appropriate for cloud native edge computing, where nearly all processes are executed privately in each core. In contrast, in most x86 CPUs, a large, higher-latency shared cache supports the many shared functions in a client computing environment, and there is a correspondingly small provision of private cache.

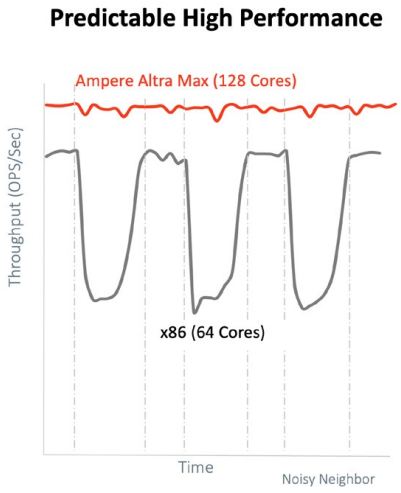
Density and Efficiency are the Keys to High-Performance Cloud Native Edge Computing

In an edge cloud native computing environment, a containerized workload or microservice is best executed as a single thread in its own core. This ensures predictable high performance and avoids allocation and prioritization conflicts when compute resources are shared. Hardware isolation of different workloads also reinforces the security protection in edge devices, especially those powering a variety of critical and high-fidelity use cases. Packing many more cores into the same silicon footprint offers multiple benefits in edge computing, including:

- Executing more concurrent workloads
- Delivering more predictable performance
- Lowering the power consumption across all workloads

Because of the ability of Ampere's Cloud Native Processor architecture to provide a consistent operating frequency and single-threaded cores that do not thermal throttle under load, the ability to drive full utilization with multiple heterogeneous workloads is far higher than with the x86-based alternative.

Figure 4: Ampere Altra Max Scalability and Predictability

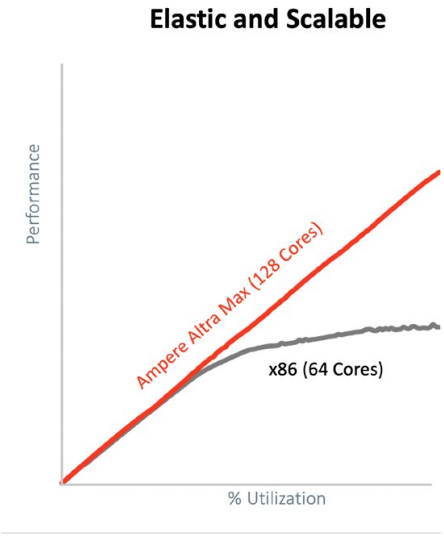


Eliminate the Noisy Neighbor with High-Performing Cloud Native Processors

Predictable performance is critical in multi-tenant cloud environments. Due to Ampere's single-threaded core architecture, individual processes draw from dedicated resources rather than contending for shared processing resources. The latter is often referred to as the "Noisy Neighbor" phenomenon, where existing workloads suffer from degraded performance due to additional workloads ramping up in the same shared core and memory complex.

Similarly, scaling predictability is highly sought after in modern compute environments. Most large-scale x86 deployments today target the 30% to 60% utilization range due to an incremental drop in performance at higher ranges. The Ampere Altra family demonstrates extremely predictable scaling for cloud native workloads.

Figure 5: Ampere Altra Max Performance and Utilization



Arm Native Scale Up and Out

Because the Arm architecture has been widely adopted for small, low-power embedded applications for a long time, there is an extremely large existing user and software base at the edge. But compared to the other Arm solutions, the Ampere Altra family of processors deliver more performance with a large number of high-performance and power-efficient cores, which gives users a way to quickly scale up and scale out their Arm applications.

For example, an experiment shown on YouTube (https://youtu.be/UT5UbSJOyog) showed that an edge system with an Ampere Altra 128-core processor delivers the performance of a 100 Raspberry Pi 4 cluster while being **22%** more energy efficient. There is more detail in the blog here.

Prior to Ampere, Arm developers with high performance needs had to cluster many devices together to get to the required performance. Solving the performance challenge by clustering can appreciably increase the size, weight, and power even when using the smallest compute unit. Ampere enables consolidating such workloads into a single compute node. It can also be used to consolidate many workloads to one node.

Another popular use for Ampere Altra processors in such an environment is for the development and testing of applications, containers, and OS images for Arm embedded systems, including industrial devices, smart cameras, and automotive ECUs. Low-power Arm devices may be powerful enough to run a particular application but developing and testing at scale usually benefits from having something that is architecturally compatible but much more powerful.

Open Standard Compute Solutions for the Edge

ADLINK and Ampere have developed a building block to tackle a wide range of challenging edge applications. Featuring the COM-HPC open standard for embedded computing applications, ADLINK's Ampere Altra compute module delivers Ampere's range of power-optimized processors in an embedded open systems standard with industry-wide acceptance. The flexibility of the COM-HPC Ampere Altra module makes it well suited for demanding edge and embedded use cases.

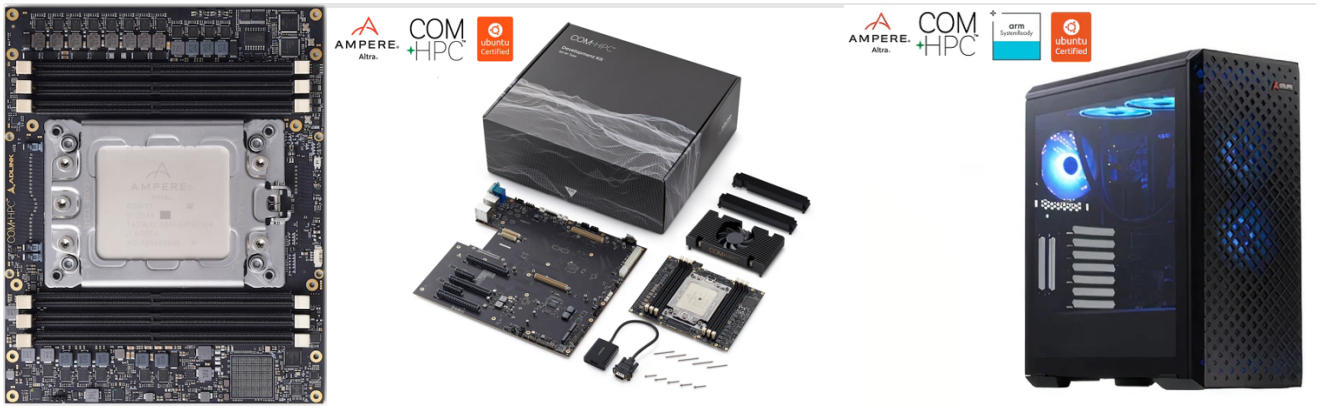
Utilizing ADLINK's COM-HPC Ampere Altra compute modules, platform developers can more quickly design and implement solutions tailored to their unique needs. The COM-HPC Ampere Altra compute modules also provide a great size, weight, and power advantage over legacy x86 processors.

ADLINK also developed a compatible COM-HPC reference carrier. This optional board provides various I/O and peripheral expansion for users to build platforms that suit their needs and fit into their size, weight, and power budget. ADLINK has validated the COM-HPC module and carrier for an operating range of -40°C to 85°C and MIL-STD shock and vibration.

Companies can also develop custom carriers for the Ampere Altra compute module to suit their unique requirements and build those into their products, such as vehicles, industrial automation, and telecom equipment. The collaboration between Ampere and ADLINK enables more scalability, flexibility, and performance than competitive offerings for the embedded systems community.

To facilitate evaluation and rapid development of solutions using Cloud Native Processors in a standardized compute module, ADLINK offers the Ampere Altra Dev Kit and Ampere Altra Developer Platform products.

Figure 6: ADLINK's Ampere Altra COM-HPC Module, Ampere Altra Dev Kit, and Ampere Altra Dev Platform



Using the family of Ampere Altra processors in your embedded system triples or quadruples the number of cores as compared to x86 solutions. For example, the performance per watt provided by the ADLINK Ampere Altra COM-HPC module is three to four times better than that of x86 embedded solutions.

Table 1: Ampere Altra Processors Used in ADLINK's COM-HPC Compute Module

PROCESSOR	CORES	PERFORMANCE[4]	POWER	PERF/WATT	PERF/$
Intel Xeon D-2776NT	16	85	117 W	0.7	–
Ampere Altra Q32-17	32	94 / **1.1x**	40 W	2.4 / **3.2x**	**3.1x**
Ampere Altra Q64-22	64	201 / **2.4x**	69 W	2.9 / **4.0x**	**5.4x**
Ampere Altra M96-28	96	299 / **3.5x**	128 W	2.3 / **3.2x**	**5.3x**
Ampere Altra M128-26	128	333 / **3.9x**	124 W	2.7 / **3.7x**	**4.8x**

Conclusion

Edge computing demands are being driven higher by adding more sensors, higher bandwidth sensors, and higher bandwidth communications. But the size, weight, and power available for the compute is limited and often fixed. Bringing the high density and energy efficiency of Cloud Native Processors into these situations offers relief. Using open standards compute modules speed adoption while future-proofing the solution. Ampere and ADLINK have partnered to bring these solutions to you.

End Notes

1. Allied Market Research: "Edge Computing Market", May 2019, https://www.alliedmarketresearch.com/edge-computing-market

2. Ampere Altra family processors typically run 30-40% lower than x86 competitors under common loads at high utilization. While running the industry-standard benchmark SPECrate®2017_int_base (estimated), Altra Max runs ~30% lower on average over the course of the benchmark run than rated TDP, while competitors run at their TDP or at times higher than their rated TDP.

3. Typical Altra family L2 (private) cache size is 1 MB per thread (1/core) as compared to x86 systems which range from 256 kB to ~512 kB per thread (2/core) based on publicly available specification data. Multi-threaded caches used in x86 systems are shared between the threads and, therefore, are not private.

4. Performance and usage power data are based on SPECrate®2017_int_base (estimated) (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 as compared to Xeon D per:
https://www.intel.com/content/www/us/en/products/sku/226239/intel-xeon-d2776nt-processor-25m-cache-up-to-3-20-ghz
https://www.spec.org/cpu2017/results/res2022q4/cpu2017-20221010-32556.html

Figure 2: Processor Performance per Watt for Different Workloads

Core count based on 12 kW rack with system usage power measured under SIR2017 load based on publicly available specification data.

- Intel Ice Lake – 40 Cores
- AMD Milan – 64 Cores
- Ampere Altra – 80 Cores
- Ampere Altra Max – 128 Cores

Figure 3: Ampere Altra Family Architecture Advantages

- Predictability
  - Performance of Redis combined with StressNG workload run on nonutilized cores/threads intermittently
  - Test run: Ampere Altra Max (128 cores) vs AMD Milan (64 core/128 threads)
  - Single socket tested
- Scalability
  - Performance of x.264 run on a single instance per thread or core
  - Ampere Altra Max – 128 Cores (single socket)
  - AMD Milan – 64 Cores and 128 threads (single socket)

Figure 4: Ampere Altra Max Scalability and Predictability

System configurations, components, software versions and testing environments that differ from those used in Ampere's tests may result in different measurements from those obtained by Ampere. The system configurations and components used in our testing were performed on bare-metal servers with one CPU socket with equivalent memory, storage, and networking options for all platforms referenced and then scaled linearly to the rack level using an 8 kW rack as the limit.

The processors used were AMD EPYC 7763 ("Milan"), Intel Xeon 8380 ("Ice Lake"), and Ampere Altra Max M128-30. Specific test configurations are noted below:

- Operating System: CentOS 8.0.1905 (kernel 4.18.0, 64k pages on Ampere Altra Max, 4k pages on Intel and AMD processors)
- Memory: 8x 64 GB DIMMs, DDR4-3200
- Networking: Mellanox ConnectX-5 100 Gb NIC
- Storage: 1-4 NVMe drives depending on the workload across all three platforms
- SMT: Enabled on the Intel and AMD platforms; not available on the Ampere Altra Max platform.

July 11, 2023

Ampere Computing reserves the right to change or discontinue this product without notice.

While the information contained herein is believed to be accurate, such information is preliminary, and should not be relied upon for accuracy or completeness, and no representations or warranties of accuracy or completeness are made.

The information contained in this document is subject to change or withdrawal at any time without notice and is being provided on an "AS IS" basis without warranty or indemnity of any kind, whether express or implied, including without limitation, the implied warranties of non-infringement, merchantability, or fitness for a particular purpose.

Any products, services, or programs discussed in this document are sold or licensed under Ampere Computing's standard terms and conditions, copies of which may be obtained from your local Ampere Computing representative. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Ampere Computing or third parties.

Without limiting the generality of the foregoing, any performance data contained in this document was determined in a specific or controlled environment and not submitted to any formal Ampere Computing test. Therefore, the results obtained in other operating environments may vary significantly. Under no circumstances will Ampere Computing be liable for any damages whatsoever arising out of or resulting from any use of the document or the information contained herein.

Ampere Computing
4655 Great America Parkway, Santa Clara, CA 95054
Phone: (669) 770-3700
https://www.amperecomputing.com

Ampere Computing reserves the right to make changes to its products, its datasheets, or related documentation, without notice and warrants its products solely pursuant to its terms and conditions of sale, only to substantially comply with the latest available datasheet.

Ampere, Ampere Computing, the Ampere Computing and 'A' logos, and Altra are registered trademarks of Ampere Computing.

Arm is a registered trademark of Arm Limited (or its subsidiaries) in the US and/or elsewhere. All other trademarks are the property of their respective holders.

About Ampere

Built for sustainable cloud computing, Ampere Computing's Cloud Native Processors feature a single-threaded, multiple core design that's scalable, powerful, and efficient.

Learn more:

See our solutions for a variety of demanding workloads: https://amperecomputing.com/solutions