AMPERE®

Switch to GPU-Free AI Inference with Ampere Cloud Native Processors

Right-size compute for resource and operational efficiencies without sacrificing performance.

Choose the Right Tool for the Workload

In the dynamic realm of AI deployments, ‘right-sized computing’ emerges as a crucial strategy, aligning computing resources precisely with AI application demands to **optimize performance, power consumption, and cost efficiency.**

Ampere® Cloud Native Processors stand out as a scalable and flexible solution, allowing agile resource adjustments for both edge and data center deployments.

> **This adaptability ensures optimal resource allocation, making Ampere ideal for a variety of AI applications, such as computer vision, natural language processing (NLP), recommendation systems, and generative AI.**

Computer Vision

Natural Language Processing (NLP)

Recommendation Systems

Generative AI

Table of Contents

01

A New Era of
AI Inference Efficiency

GPU-Free AI Inference

As the demand for AI solutions continues to grow, scaling AI effectively and sustainably requires right-sized compute solutions that balance high performance, hardware cost, and energy efficiency.

Additionally, Ampere's technology addresses the prevalent challenge of space constraints in large-scale AI inference deployments.

Ampere stands as a pioneer by offering GPU-free AI inference solutions. These help customers:

1. Minimize energy consumption
2. Optimize costs
3. Enhance the resource utilization rates

Wallaroo looks to meet the demand for AI computing with Ampere

“Wallaroo/Ampere solution allows enterprises to improve inference performance, increase energy efficiency, and balance their ML workloads across available compute resources much more effectively, all of which is critical to meeting the huge demand for AI computing resources today while also addressing the sustainability impact of the explosion in AI.”

Vid Jain

CEO, Wallaroo.AI

Wallaroo

Embrace AI's Transformative Power With Ampere

Ampere Computing is a modern semiconductor company pioneering sustainable AI inference capabilities. In our fast-paced, interconnected world, the ability to harness AI is a competitive necessity.

Ampere's AI inference technology is driving innovation across various sectors, from powering computer vision and natural language processing to recommender engines and generative AI. Designed for sustainable cloud computing, Ampere's products offer best in class performance per watt and rack.

Ampere's commitment is to:

1. Advance AI inference technology
2. Empower businesses with the necessary tools
3. Unlock higher savings and operational efficiency

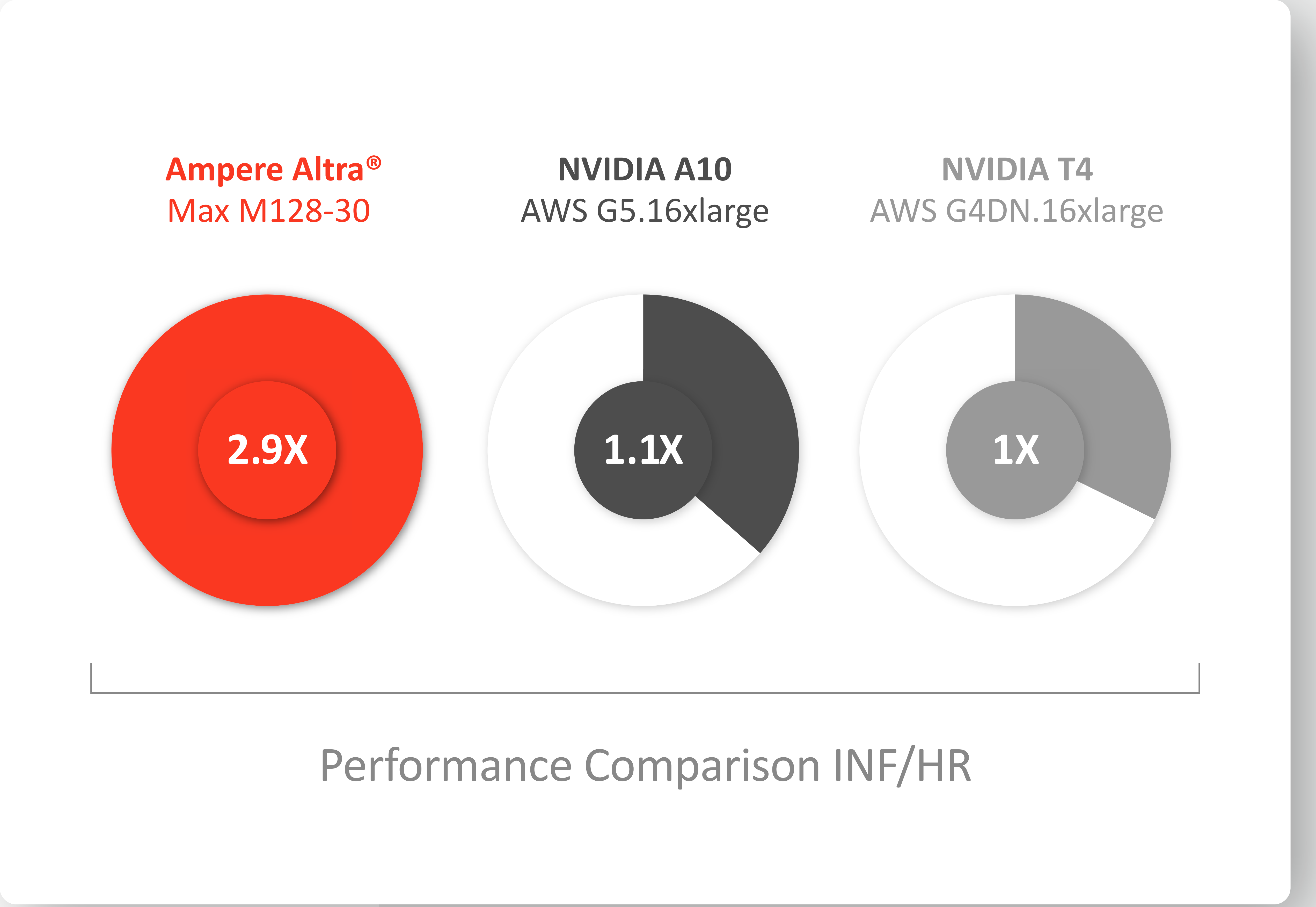
02

Maximize Throughput for Scalable AI Performance

Up To 2.9x Better Whisper Model Performance

Integrating Ampere Cloud Native Processors, particularly those utilizing OpenAI's Whisper model for speech recognition and transcription, businesses can achieve up to 2.9 times better performance compared to traditional GPU-based solutions.

> **The result is a dramatic enhancement in processing speech recognition tasks, offering businesses the dual advantage of superior performance and lower operational costs.**



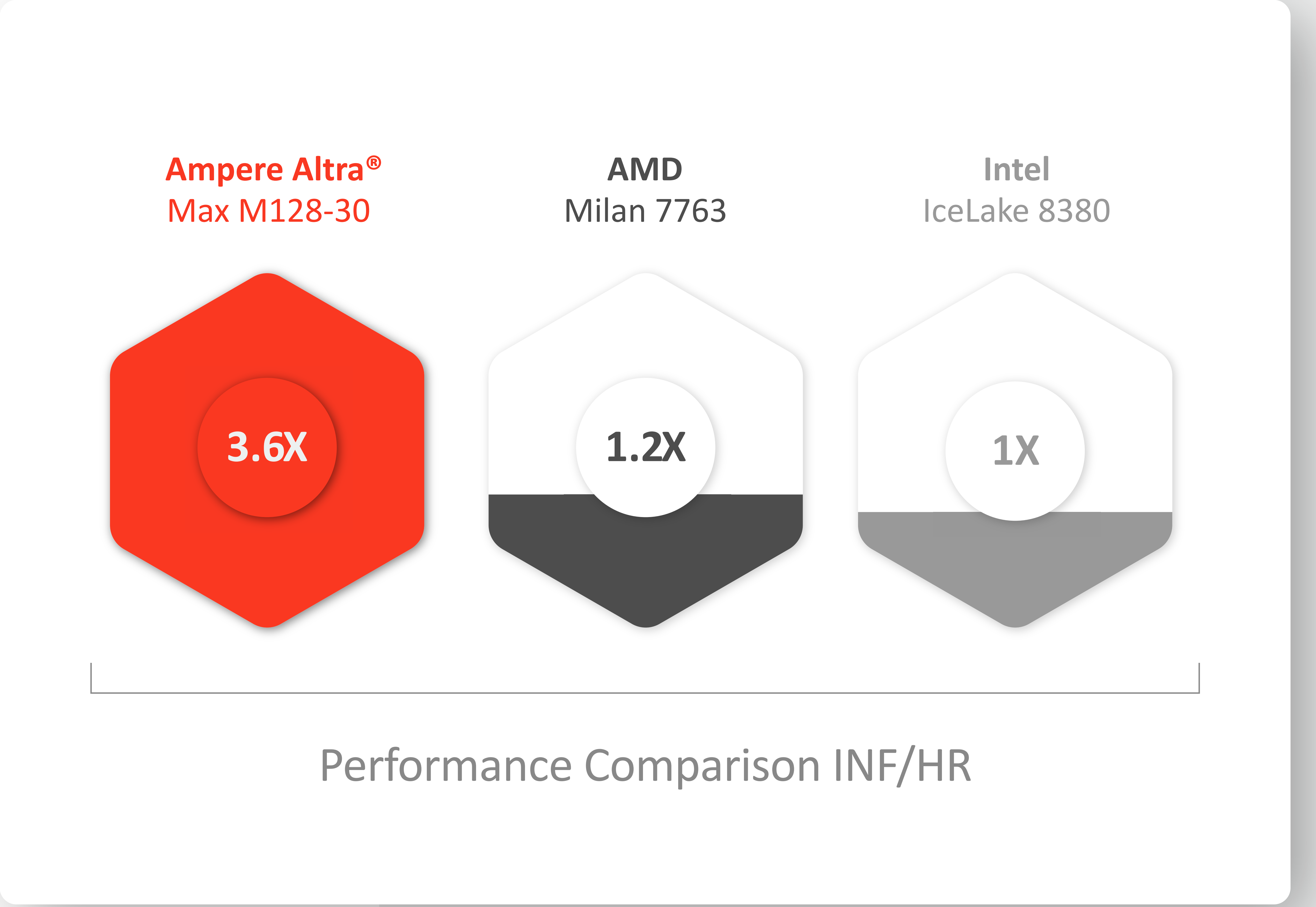
Footnote: The web services study in this eBook is based on performance and power data for many typical workloads using single-node performance comparisons measured and published by Ampere® Computing. Details are available at https://amperecomputing.com/home/efficiency-footnotes. For more benchmarks on specific AI models, visit https://amperecomputing.com/solutions/ampere.

Up to 3.6x Greater AI Inference Performance in On-Premise Deployments

Experience up to 3.6 times greater AI inference performance in on-premise or edge deployments with Ampere GPU-Free Cloud Native Processors.

When it comes to AI inference, most cases do not need expensive, resource-hungry GPUs to support the required performance level.

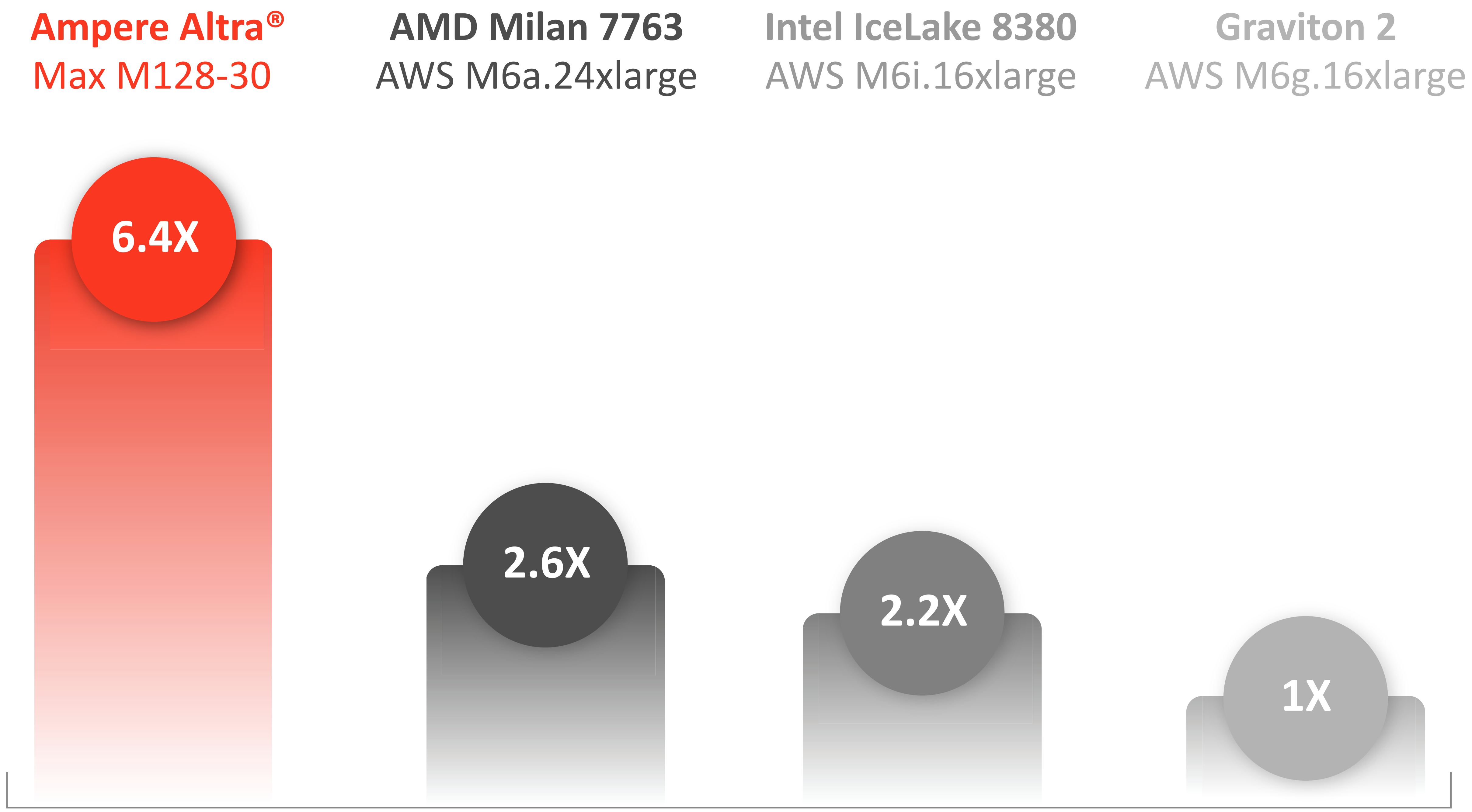
> **Ampere Cloud Native CPUs, with higher core counts, efficiently distribute computational tasks across multiple cores.**



Footnote: The web services study in this eBook is based on performance and power data for many typical workloads using single-node performance comparisons measured and published by Ampere® Computing. Details are available at https://amperecomputing.com/home/efficiency-footnotes. For more benchmarks on specific AI models, visit https://amperecomputing.com/solutions/ampere.

Up to 6.4x Greater AI Inference Performance in the Cloud

Designed for optimal AI inference in the cloud, with up to 6.4 times greater performance, Ampere Cloud Native Processors handle AI inference workloads concurrently with **higher throughput** while helping customers meet latency requirements.

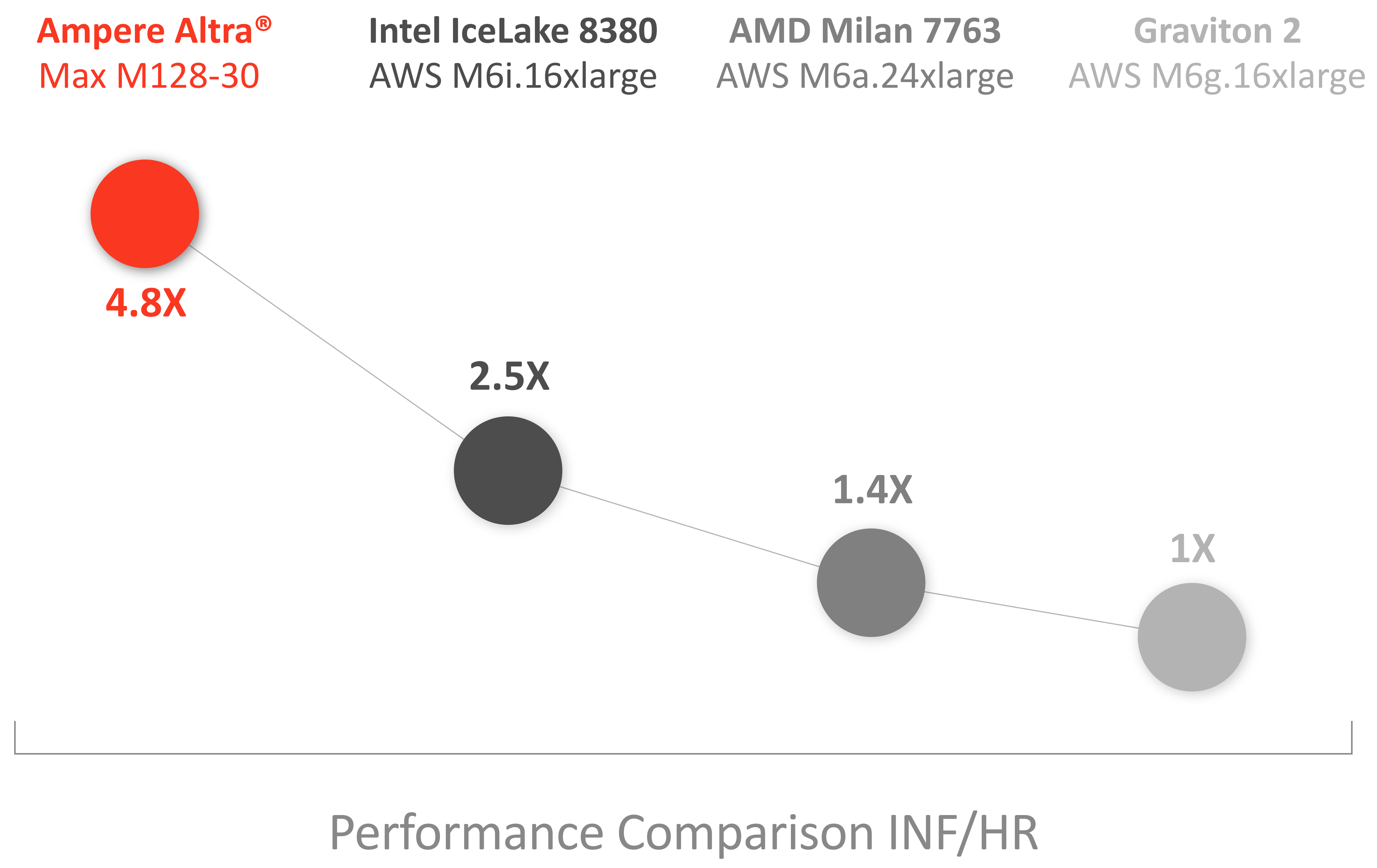


Performance Comparison INF/HR

Footnote: The web services study in this eBook is based on performance and power data for many typical workloads using single-node performance comparisons measured and published by Ampere® Computing. Details are available at https://amperecomputing.com/home/efficiency-footnotes. For more benchmarks on specific AI models, visit https://amperecomputing.com/solutions/ampere.

Up to 4.8x Better Recommender Engine Performance in the Cloud

Unlock efficiency and precision with Ampere's Cloud Native Processors, boosting your recommender engine's performance by as much as 4.8 times.



Footnote: The web services study in this eBook is based on performance and power data for many typical workloads using single-node performance comparisons measured and published by Ampere® Computing. Details are available at https://amperecomputing.com/home/efficiency-footnotes. For more benchmarks on specific AI models, visit https://amperecomputing.com/solutions/ampere.

03

Accelerate AI Inference Without Breaking The Bank

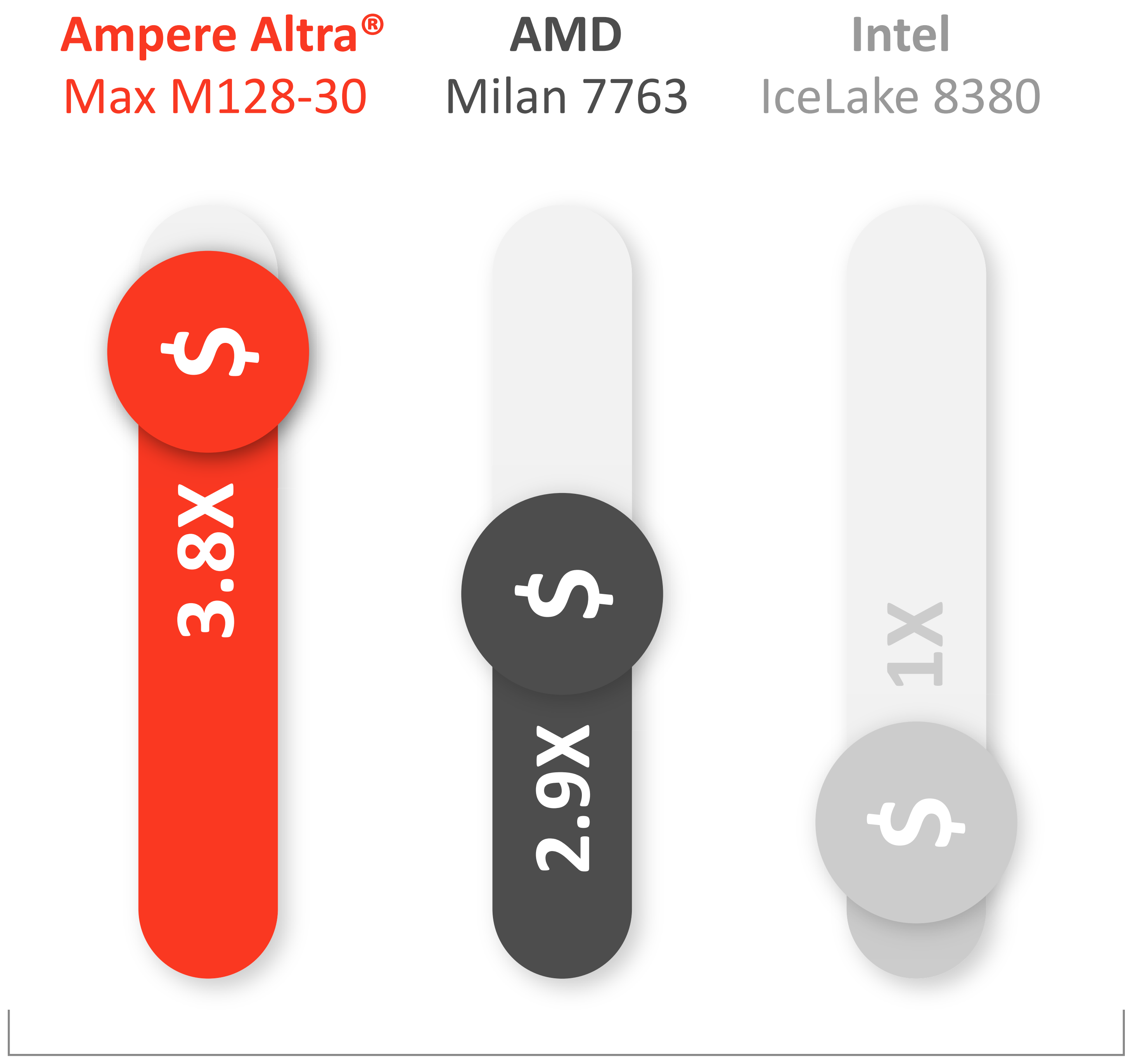
Up to 3.8x Better Recommender Engine Price-Performance in the Cloud

Ampere CSP partners offer a variety of high core-count instances with the best price-performance for many AI inference use cases. Specifically, in the case of recommender engines, they offer performance up to **3.8 times better than common GPU alternatives.**

CSPs can save money by taking advantage of the core density, lower energy consumption, and higher rack density passing these savings on to their customers through lower instance costs.

The higher performance of Ampere CPUs and lower instance cost yield the best in class price-performance in the cloud for many inference tasks.

> **Many different AI models or application use cases can be optimized through instance right-sizing to yield even better price-performance than many publicly available GPU-based instances.**



Price-Performance Comparison

Footnote: The web services study in this eBook is based on performance and power data for many typical workloads using single-node performance comparisons measured and published by Ampere® Computing. Details are available at https://amperecomputing.com/home/efficiency-footnotes. For more benchmarks on specific AI models, visit https://amperecomputing.com/solutions/ampere.

Lightly looks to lower costs for our customers with AmpereOne C3A

"Lightly.ai's customers can achieve over 3x cost reduction running on Ampere T2A instances on GCP using Ampere AI software solutions for AI Inference. At Lightly.ai we are always looking for ways to increase performance and lower costs for our customers, the next generation AmpereOne C3A instances on GCP will deliver on this continued value proposition."

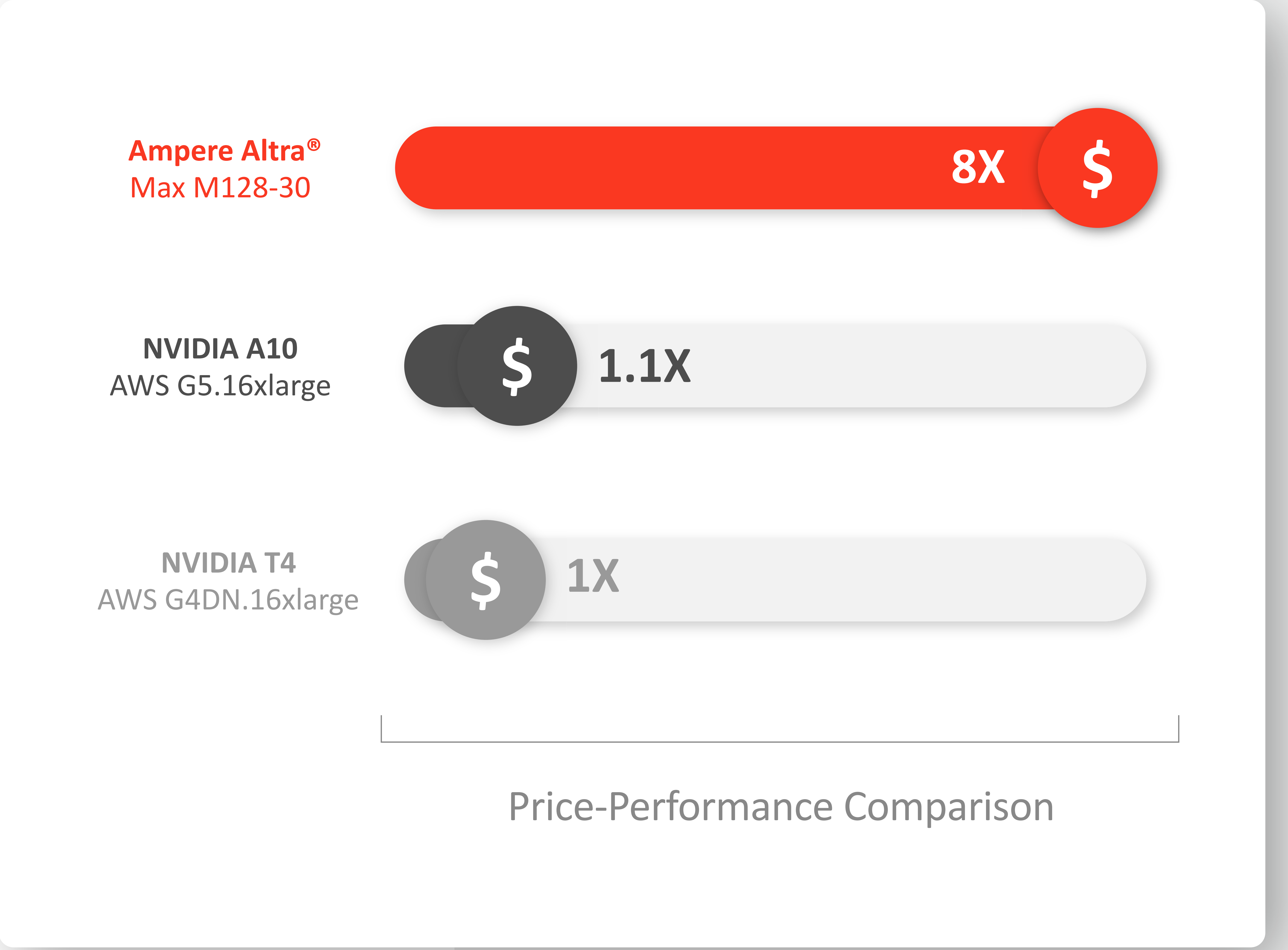
Igor

Susmeli, Co-Founder, Lightly.ai

LIGHTLY

Up to 8x Better Whisper Model Price-Performance in the Cloud

Ampere Altra Cloud Native Processors run OpenAI's Whisper model at a much lower cost, showcasing the value of Ampere GPU-free AI deployments.



Footnote: The web services study in this eBook is based on performance and power data for many typical workloads using single-node performance comparisons measured and published by Ampere® Computing. Details are available at https://amperecomputing.com/home/efficiency-footnotes. For more benchmarks on specific AI models, visit https://amperecomputing.com/solutions/ampere.

Elotl looks to reduce costs, while avoiding operational complexity with Ampere A1

“Using Ampere A1 instances on OCI with integrated Ampere Optimized AI library, we managed to right-size compute, providing price-performance advantage on deep learning inferencing relative to GPUs and to other CPUs. We found an order of magnitude or more reduction in cloud resource costs, measured at 4 operating points for 2 cloud vendors, while avoiding operational complexity for changes in model serving resource needs and cloud offerings.”

Madhuri Yechuri

CEO, Elotl

ELOTL

04

Building a Sustainable Future for AI

Less Energy and Space Required

Spend less on energy and reduce your data center footprint. Ampere's innovative architecture **decreases the amount of energy** needed to run AI Inference workloads. This efficiency is particularly beneficial at the data center level, where **lower power consumption** allows for **increased rack density.** The result is tangible savings across data center operations within a reduced footprint.

GPU-Free AI inference using Ampere platforms

1 Minimizes the additional hardware costs and energy consumption associated with underutilized GPU resources

2 Aligns AI inference with right-sized computing to balance performance, cost-effectiveness, and power consumption

Footnote: The web services study in this eBook is based on performance and power data for many typical workloads using single-node performance comparisons measured and published by Ampere® Computing. Details are available at https://amperecomputing.com/home/efficiency-footnotes. For more benchmarks on specific AI models, visit https://amperecomputing.com/solutions/ampere.

Matoha looks to achieve cost saving for training of algorithms with Ampere Altra

“Switching to Ampere Optimized Tensorflow running on OCI A1 instances has enabled us to achieve a 75 percent cost saving for the training of the algorithms for our plastics and fabrics identification machines, while lowering our CO2 emissions - thanks to Ampere Altra’s high energy efficiency.”

Martin Holicky

CEO, Matoha

Matoha

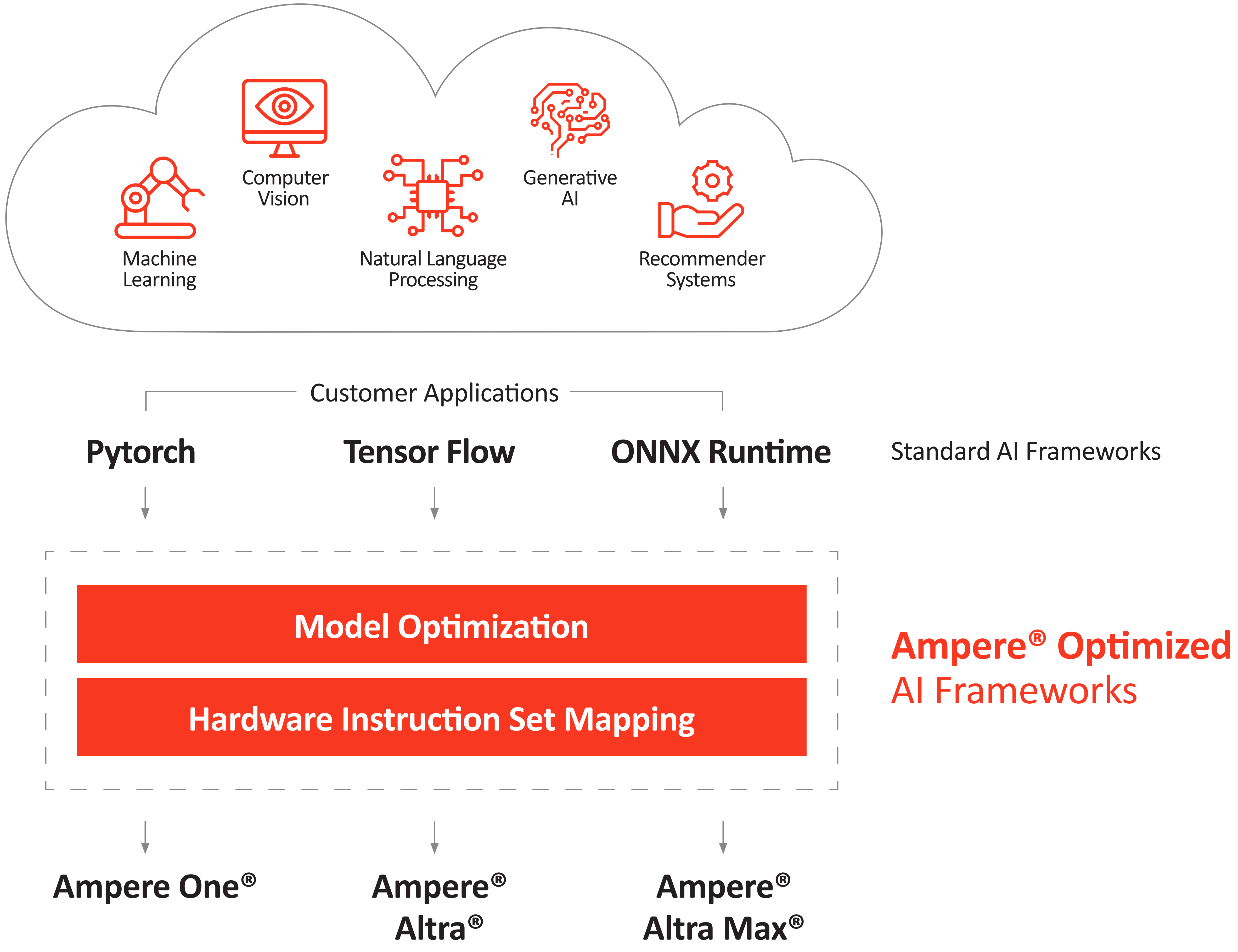
05

Ampere AI Software Acceleration and Ecosystem Support

Ampere Optimized AI Frameworks

Ampere focuses on delivering the **most efficient solutions** for AI inference, offering **seamless integration** with all AI applications via the support of all popular AI frameworks in the industry including Pytorch, TensorFlow, and ONNXRuntime.

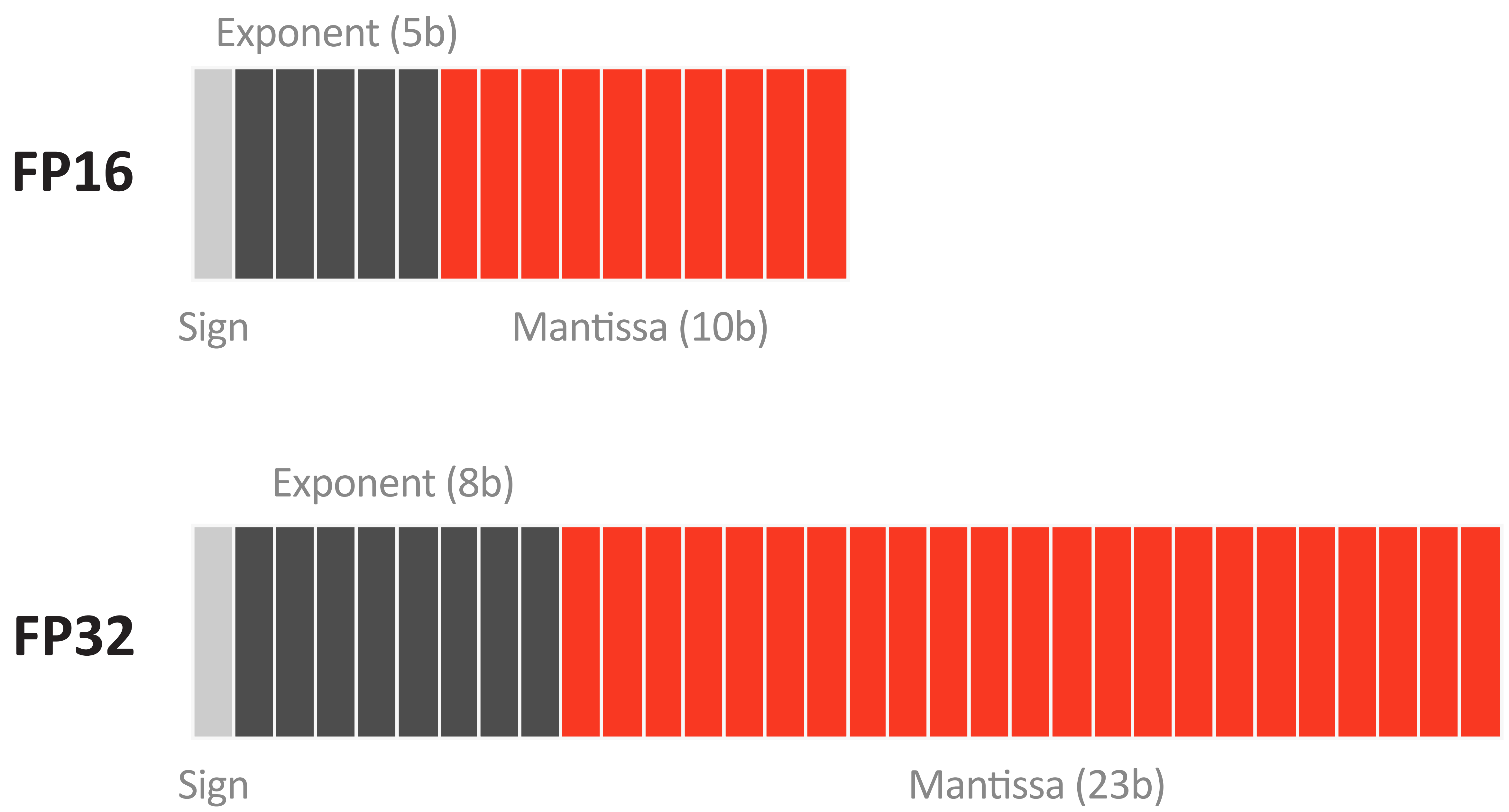
Ampere Optimized AI Frameworks work out of the box and do not require API changes or additional coding. This drop-in library supports all AI applications developed in the most popular frameworks and enables developers to seamlessly deploy their models across Ampere's GPU-Free AI platforms.



Footnote: The web services study in this eBook is based on performance and power data for many typical workloads using single-node performance comparisons measured and published by Ampere® Computing. Details are available at https://amperecomputing.com/home/efficiency-footnotes. For more benchmarks on specific AI models, visit https://amperecomputing.com/solutions/ampere.

Overview of Data Formats used in AI

Further **maximizing computational throughput** is Ampere's unique support for the FP16 data format that helps speed up AI inference operations, offering **accelerated computation** for real-time applications without compromising accuracy, underscoring our commitment to efficient and **high-performance AI solutions.**



Footnote: The web services study in this eBook is based on performance and power data for many typical workloads using single-node performance comparisons measured and published by Ampere® Computing. Details are available at https://amperecomputing.com/home/efficiency-footnotes. For more benchmarks on specific AI models, visit https://amperecomputing.com/solutions/ampere.

06

Succeed with Ampere GPU-Free AI Inference Solutions

Elevate AI Excellence with Ampere

Ampere® Altra® Family of Processors

Servers equipped with Ampere® Altra® and Altra® Max®processors deliver **best in class AI performance**, **cost-effectiveness, and power efficiency.** Tailored to diverse deployment needs, Ampere Cloud Native Processors help ensure optimal computing for AI inference. Beyond high performance, Ampere offers an **economically and environmentally sustainable solution**, helping to shield businesses against rising energy costs and providing a future-proof choice at the time of ever-growing scale of AI deployments.

As a leader in AI solutions, Ampere's commitment to hardware advancements empowers businesses for transformative change. Ampere Optimized AI Frameworks facilitate **fast and smooth transition**s from model development to deployment — making it easy to move existing workloads from legacy x86 architecture to Ampere Cloud Native Processors.

Drive innovation and success when you choose Ampere for unparalleled AI inference.

Footnote: The web services study in this eBook is based on performance and power data for many typical workloads using single-node performance comparisons measured and published by Ampere® Computing. Details are available at https://amperecomputing.com/home/efficiency-footnotes. For more benchmarks on specific AI models, visit https://amperecomputing.com/solutions/ampere.

Ampere GPU-Free AI Inference: Pioneering Efficiency, Performance, and Sustainability

Scale your AI inference power and performance efficiency by adopting Ampere's robust GPU-Free AI solutions. It's not just a crucial step for your performance and bottom line, it's a responsible step towards ensuring the sustainability of AI in the future.

- **Up To 2.9x Better Whisper Model Performance**
- **Up to 3.6x Greater AI Inference Performance in On-Premise Deployments**
- **Up To 6.4x Greater AI Inference Performance in the Cloud**
- **Up to 4.8x Better Recommender Engine Performance in the Cloud**
- **Up to 3.8x Better Recommender Engine Price-Performance in the Cloud**
- **Up to 8x Better Whisper Model Price-Performance in the Cloud**

Ampere leads the charge into the future of sustainable AI inference. Learn more

Footnote: The web services study in this eBook is based on performance and power data for many typical workloads using single-node performance comparisons measured and published by Ampere® Computing. Details are available at https://amperecomputing.com/home/efficiency-footnotes. For more benchmarks on specific AI models, visit https://amperecomputing.com/solutions/ampere.

AMPERE®

Ampere® Computing

Ampere is a modern semiconductor company designing the sustainable future of AI inference computing with the world's first processors optimized for the cloud.

Designed for efficient AI inference in the cloud, Ampere's GPU-Free AI solutions offer best in class performance and help maximize energy efficiency. Ampere's Cloud Native Processors stand out for their industry-leading performance, power efficiency, and scalability, making them a superior choice for robust and sustainable AI inference solutions.

Disclaimer

All data and information contained in or disclosed by this document are for informational purposes only and are subject to change. Your results may differ. This document is not to be used, copied, or reproduced in its entirety, or presented to others without the express written permission of Ampere®.

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. 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.

Footnote

The comparisons in this eBook are based on performance and power data measured and published by Ampere Computing.