ECS®
EQUUS COMPUTE SOLUTIONS

AI INFERENCING: ACCELERATING AI INNOVATION WITH CLOUD NATIVE PROCESSORS

Ampere and ECS provide infrastructure and innovation that enable AI.

AMPERE®

In today’s business landscape, Artificial Intelligence (AI) is a transformative force, promising unparalleled advancements in product enhancement and operational efficiency. The ubiquity of AI across diverse sectors underscores its indispensable role in driving future innovation. As businesses increasingly embrace AI, the focus shifts towards optimizing AI adoption strategies for maximum impact.

THE UBIQUITY OF AI

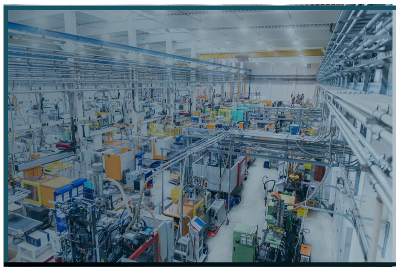
AI has permeated numerous sectors, including healthcare, manufacturing, retail, and transportation, revolutionizing operations and enhancing performance.

> **”Despite IT headwinds in 2023, businesses accelerated their exploration of GenAI to boost business transformation in 2023. In 2024, the shift to AI everywhere will enter a critical buildout phase with spending more than doubling to $40.1 billion and reaching $151.1 billion in 2027.”**
>
> - *Rick Villars, Group Vice President, Worldwide Research at IDC.*

This exponential growth drives more demand for computer capacity and power, outstripping the supply of affordable energy in many locations around the world. Ampere Computing® offers energy-efficient, high-performing, Cloud-Native Computing processors that will enable the industry to overcome many of these constraints and increase AI adoption rates.

ROLE OF AI IN BUSINESS INNOVATION

Humans and machines generate massive amounts of data every day. Organizations are grappling with how to extract value from this data and enhance competitiveness. Data analysis, pattern recognition, and real-time decision-making is essential for maintaining competitiveness in today's dynamic market. Companies are deploying AI across all industries:



Manufacturing

- Improve quality
- Improve production rates

Security

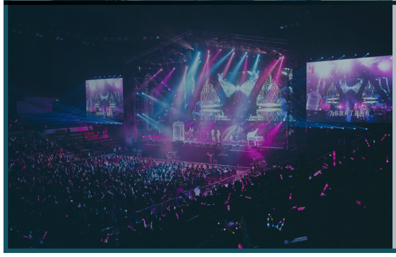
- Identify threats
- Monitor threats

Retail

- Improve customer service
- Improve loss prevention

Healthcare

- Lower costs
- Achieve better patient outcomes



E-commerce and Entertainment

- Improve user experiences

These are just a few examples of how AI is used today. There are a vast number of opportunities to leverage AI, and it is important to understand its functionality and how to extract the most value from your AI solution.

SIGNIFICANCE OF AI TRAINING AND INFERENCE

AI is comprised of two crucial components: training and inference. The training component is the oft-discussed concept of machine learning, which uses data and algorithms to teach the models how to make decisions. While training involves teaching models to recognize patterns, inference processes incoming data in real time, requiring specialized optimization for efficiency and performance. Most training workloads benefit from GPU-based solutions; however, the higher power requirements and costs associated with them make them a challenging option to use for inference. Up to 85% of AI silicon used today in data centers and at the edge is dedicated to AI inferencing. CPUs can efficiently handle most inference workloads. Thus, data centers and edge computing applications have options for right-sizing their infrastructure for the AI workloads at hand.

AMPERE:
ENABLING AI INFERENCING

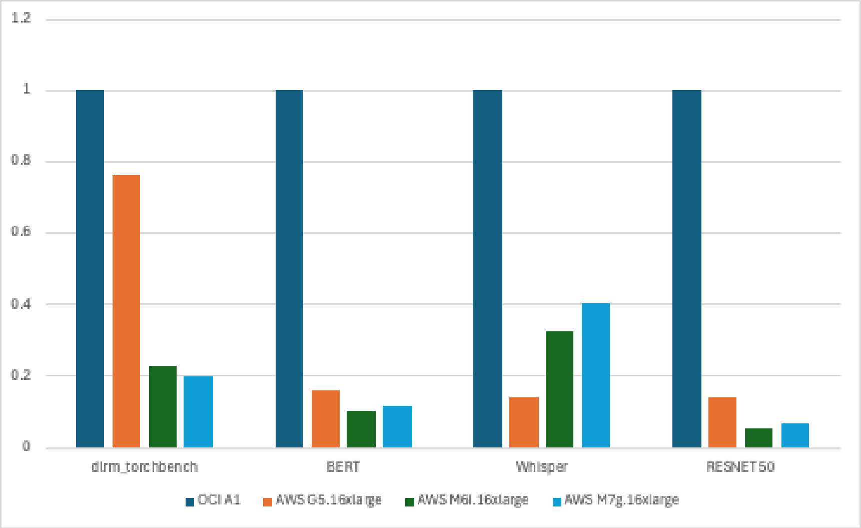
Ampere's approach to AI inference focuses on sustainability and cost-effectiveness, which are crucial as inferencing often consumes more compute cycles than training. Ampere designs Cloud-Native processors optimized for efficient AI inference. This is achieved through the high compute density of Ampere processors, equipped with up to 192 single-threaded cores and double the vector units per core, allowing for predictable performance and unprecedented computational scaling. In computational workloads such as AI, the reliance on these vector units is crucial to performance and is one reason the Ampere architecture is more efficient than legacy processors, such as the x86. AI is a workload that particularly benefits from high computational scale, making the additional vector units in each core a valuable resource for vector math operations. Ampere processors also include native support for FP16 data format, which boosts AI inference performance. FP16, also known as half precision, uses 16 bits for weighting in neural networks. The lower precision allows for faster processing speeds. Any associated loss of accuracy is minute and negligible in practical terms.

INDUSTRY IMPACT OF AI INFERENCE

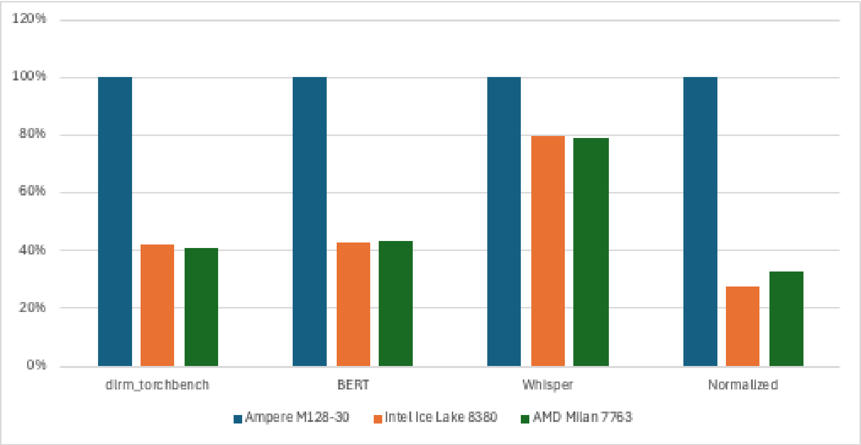
Ampere's technology powers real-world applications, from self-driving cars to intelligent customer service, driving innovation and efficiency. These applications require the ability to handle recommendation models, speech-to-text transcription and translation, and multiple video streams. Ampere’s CPUs were designed and optimized for these modern inference applications and often outperform legacy solutions based on x86 and other ARM instances.

Ampere AI:
GPU Free Leadership Inference Performance

Cloud Inference



Server Inference



ADVANTAGES OF AMPERE'S SOLUTIONS

Ampere offers comprehensive AI inference solutions, which balance performance, cost, and energy-efficiency. Along with offering CPUs with industry-leading AI performance and efficiency, Ampere offers optimized software frameworks, free of charge, that accelerate AI performance. Ampere Optimized AI Frameworks (AIO) support AI industry standard frameworks, including PyTorch, TensorFlow, and ONNXRuntime, seamlessly integrating with AI applications without the need for code changes or recompiling. Scalable Ampere Cloud-Native Processors empower businesses to navigate the challenges of burgeoning AI workloads effectively.

ECS: INNOVATION AS A SERVICE

ECS designs, builds and deploys the digital infrastructure that keeps companies relevant, viable, and growing. From personalized computing and data center infrastructure to liquid cooling to AI enablement, telecom systems and 5G management, ECS's customer-first approach delivers seamless solutions in form, fit, and function.

Within the ECS Innovation Center, ECS hosts on-prem AI and ML systems powered by Ampere processors. This setup allows organizations to test proof of concepts and demo AI projects either in the lab or via remote access. Allowing for a true try-before-you-buy environment.

TRANSFORMING BUSINESS WITH AMPERE AI SOLUTIONS

Ampere CPUs emerge as the optimal choice for AI workloads, delivering superior performance, cost-effectiveness, and energy efficiency. With Ampere's innovative hardware advancements and optimized software stack, coupled with ECS’s Innovation Lab, businesses can confidently embrace AI inference to drive transformative growth and success.

Together, Ampere and ECS empower businesses to elevate their AI capabilities and achieve new heights of innovation and success. By offering high-performance processors, optimized software libraries, and a unified inference model, Ampere and ECS enable seamless transitions from model development to deployment, ensuring maximum impact and efficiency in AI-driven endeavors.

Footnotes for GPU Free Leadership Inference Performance Charts, https://amperecomputing.com/home/efficiency-footnotes.

For more information on ECS’s platforms and services, visit equuscs.com.

For more information on Ampere’s portfolio of efficient processors optimized for artificial intelligence workloads, visit amperecomputing.com/solutions/ampere-ai.