AI Inference with Ampere® Cloud Native Processors

White Paper

AMPERE.

Introduction

As businesses increasingly acknowledge the significance of Artificial Intelligence (AI) and its potential to improve their products and operations, it has become clear that future innovation is inherently intertwined with AI adoption. The question is no longer whether to adopt AI, but how to best approach the task.

The Ubiquity of AI

AI has been seamlessly integrated into many aspects of the modern business landscape across many sectors, including healthcare, manufacturing, retail, transportation, and more. Ampere Computing recognizes the pervasiveness of AI and provides innovative Cloud Native solutions that enable efficient and high-performance AI inference.

Role of AI in Business Innovation

The ability to harness AI for tasks such as data analysis, pattern recognition, and real-time decision making is crucial for staying competitive in today's fast-paced world. Ampere's commitment to advancing AI inference technology empowers businesses to unlock new levels of innovation and efficiency.

Ampere: Enabling AI Inferencing

Ampere designs Cloud Native Processors with the efficient handling of AI inference workloads in mind. Since power is ultimately the limiting factor in how much compute can be deployed for AI, delivering the most inference operations within a given power envelope is the most critical factor to achieving scale.

The unique characteristics of the Ampere Family of Cloud Native Processors facilitate large-scale AI deployments, keeping operational costs in check—by minimizing energy consumption—and overcoming constrained space, weight, and power restrictions, which traditionally limit edge deployments. 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 so efficient vs. legacy x86 processors. 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.

The Significance of AI Training and Inference

AI consists of two critical components: training and inference. Training involves feeding vast amounts of data through models to teach them to recognize patterns. The inference phase, which involves processing incoming data in real time or in large batch jobs, requires specialized optimization of the models to gain efficiency and performance, which lowers the cost of deploying AI applications at scale.

Ampere's Approach to AI Inference

Ampere is the most sustainable and cost efficient compute platform for inference on the market. As models are deployed, inferencing often consumes many times more compute cycles than the initial training. As the scale of AI grows, it becomes critical to grow compute in a cost-effective and power-efficient manner.

Leveraging the unmatched computational scale of Ampere's processors, businesses can accelerate various data center and edge AI applications such as anomaly detection, customer service, and language translation. Ampere's commitment to a unified inference model allows seamless transitions from AI model training to inference on Ampere-powered servers.

CPU-Based Inference

Off-the-shelf servers featuring Ampere CPUs offer economical and high-performance AI inference due to Ampere's superior scalability and efficient approach to CPU design. CPU-based inference with Ampere minimizes the additional hardware costs and energy consumption associated with underutilized GPU resources, aligning AI inference with right-sized computing to strike a balance between performance, cost-effectiveness, and power consumption.

While training typically requires more extensive computational resources, inference is execution optimized, making Ampere's processors ideal for deploying trained models in an efficient and sustainable manner, without any need to modify those models. As AI moves to mainstream programmatic usage in many cloud, hybrid and on-prem use cases, CPU-based AI inference with Ampere is ideal for a wide range of volume AI applications, including computer vision, natural language processing, and recommendation systems.

Industry Impact of AI Inference

AI inference is more computationally optimized than training workloads and constitutes the majority of AI work. Trained models are deployed in volume and perform the majority of the decision-making tasks in bulk and real-time inference use cases. Many of these inference operations are run on CPU based servers because of their ubiquity and the need to reduce inferencing costs below the cost of training on highly expensive multi-GPU platforms.

AI's Impact on Various Sectors

From medical applications detecting anomalies in radiology to manufacturing inspection of product quality, AI inference plays a pivotal role. AI-powered recommendation systems transform customer experiences in retail while natural language processing enhances communication in sectors such as finance and media.

Real-World Applications

Ampere's AI inference technology extends to a wide range of real-world applications, powering innovations such as self-driving cars, medical diagnostics, smart traffic control, and intelligent customer service—to name just a few. The adaptability and performance of Ampere's solutions enable businesses to deploy AI inference models anytime and anywhere—from the cloud to the edge.

Industry Impact of AI Inference

Ampere provides a set of comprehensive solutions for AI inference, ranging from products meant for data center deployment to solutions for edge deployments. The unique features of Ampere's Cloud Native Processors make them the product of choice for both cloud service providers and those looking for on-prem deployments. The low energy consumption and high-compute density of Ampere CPUs deliver unprecedented performance within the constraints of space, weight, and power on the edge.

Ampere's Unified Inference Model

Ampere offers a unified inference model that bridges the gap between AI model development and deployment. This model provides seamless integration between the accelerated hardware used for training and Ampere's processors used for inference. Unlike AMD and Intel, Ampere does not compete with Nvidia in the GPU market.

Unlike legacy x86 CPU vendors, Ampere is focused on delivering the best CPU for AI while partnering with those having built GPUs and other accelerators for AI training. For example, Ampere fully supports the integration of Nvidia GPUs with Ampere CPUs across many server platforms—enabling customers to benefit from exceptional performance while also making use of their accelerator of choice. Any time an accelerator is used in a server, pairing it with Ampere CPUs will save power and cost.

Frameworks and Software Stacks Support

Ampere Optimized AI Frameworks, the cornerstone of the Ampere AI software stack, supports all the most popular AI frameworks, including PyTorch, TensorFlow, and ONNX Runtime.

Ampere Optimized Frameworks are drop-in libraries supporting all AI applications developed within the frameworks. This enables developers to choose their preferred framework for model development and seamlessly deploy the same model across all of Ampere's hardware platforms.

AI Libraries and Tools

Ampere's software acceleration strategy goes beyond providing optimized frameworks; it offers an easy transition from AI model development to deployment. This approach streamlines the AI lifecycle, allowing businesses to swiftly operationalize their AI initiatives. Ampere's holistic software stack ensures that the efficiency gained during development is carried forward to real-world inference scenarios, enhancing overall productivity and ROI.

Ampere's competitive edge in AI software acceleration is driven by two factors: hardware innovation and sophisticated software optimizations. By leveraging these strengths, Ampere empowers businesses to harness AI's transformative potential with unparalleled efficiency and performance.

Ampere's Hardware Advancements

As AI workloads continue to advance, the need for optimized hardware solutions becomes paramount. Ampere offers cutting-edge hardware innovations that redefine the landscape of AI inference and deliver unprecedented performance, energy efficiency, and scalability. In this context, two crucial aspects emerge: the superiority of single-threaded cores over multi-threaded cores for AI inference tasks, and the strategic advantage of Ampere CPUs' higher core count over legacy x86 processors. These advancements not only optimize AI inference but also pave the way for new possibilities in real-time applications, large-scale deployments, and resource- constrained environments.

Right-Sized AI Computing

In the rapidly evolving landscape of AI deployments, the concept of "Right-Sized Computing" has emerged as a pivotal strategy for AI inference. This means precisely calibrating computing resources to match the demands of AI applications, focusing on achieving the optimal balance between performance, power consumption, and cost efficiency. As the volume of AI deployments continues to surge, the imperative to streamline infrastructure becomes increasingly pronounced, necessitating a comprehensive approach that meets latency and throughput requirements while meticulously managing costs related to procurement, data center infrastructure, real estate, energy consumption, cooling, and other operational overhead. Merely throwing more expensive, power hungry, and narrowly specialized hardware at AI will not meet business requirements at the required scale.

Within this context, Ampere's Cloud Native Processors offer an exceptional solution. The scalability of Ampere's processors stands out, facilitating the agile adjustment of resources to meet AI deployment needs, whether for a single application or a large-scale scenario. The inherent flexibility allows data centers and cloud providers to effortlessly adapt to evolving demands, ensuring resources are allocated optimally to maximize performance and minimize waste.

Furthermore, Ampere's CPUs offer a notable advantage in energy efficiency. While legacy x86 architecture processors and GPUs struggle to strike the right balance between performance and power consumption, Ampere's CPUs deliver superior energy efficiency. This advantage stems from Ampere's architectural innovations, enabling the execution of AI inference tasks with minimal energy consumption that results in tangible savings across data center operations.

In contrast, legacy x86 CPUs often grapple with escalating power consumption as computational demands intensify, leading to ballooning operational costs. GPUs, while adept at parallel processing, often fall short in terms of energy efficiency and versatility. Other AI accelerators, although designed for a high performance on a subset of specific AI tasks, lack the broad adaptability that Ampere's Cloud Native CPUs provide.

In the ever-expanding landscape of AI deployments, the principle of Right-Sized Computing, embodied by Ampere's scalable and energy-efficient Cloud Native Processors emerges as a strategic imperative. This approach empowers businesses to navigate the challenges posed by burgeoning AI workloads and evolving infrastructure needs. With Ampere CPUs, companies can confidently plan for a future where AI deployments are optimized for performance and cost efficiency, thereby delivering the best user experience and securing their leadership position in the AI-driven era.

Advantages of Single-Threaded Cores over Multi-Threaded Cores for AI Inference

For AI inference, single-threaded cores holds distinct advantages over the multi-threaded cores of legacy x86 processors and offer a tailored approach that optimizes performance and efficiency. The following are key benefits of employing single threaded cores for AI inference.

Simplified Resource Allocation

Single threaded cores focus computational resources on a single task, ensuring efficient allocation of resources without the complexity of managing multiple threads. All resources are devoted to the latency-sensitive AI inference task at hand. This streamlined approach ensures that the AI inference process receives dedicated attention, minimizing resource contention and waste.

Predictable Latency

Single-threaded cores provide predictable and consistent latency for AI inferencing since as the processing resources are exclusively dedicated to that specific task, eliminating the need to wait for critical resources to become free. Each task runs to completion in a predictable amount of time. This is crucial in real-time applications, such as autonomous vehicles or robotics, where consistent response times are essential.

Tailored to Resource Constraints

Single-threaded cores are especially well-suited for scenarios where computational resources are limited, such as edge devices or embedded systems. It allows developers to extract maximum performance from constrained hardware without the overhead of managing multiple threads. This reduces the occurrence of bottlenecks across the system and ensures the full system is being utilized for critical AI inferencing.

Advantages of Ampere's High Core Count CPUs

Leveraging a higher core count is a strategic choice that aligns with the demands of modern AI workloads, delivering significant benefits that contribute to superior AI inference performance. Ampere's high core count CPUs are well-positioned to outshine lower core count legacy x86 architecture processors.

Parallelism and Throughput

AI inference is inherently parallelizable, as it involves processing numerous data points simultaneously. The high core count of Ampere CPUs enables efficient parallel execution of AI inference tasks, resulting in increased throughput and faster results. This is particularly crucial for large-scale deployments where timely processing of data is essential.

Scalability

Ampere CPUs offer maximum scalability via higher core count, allowing organizations to effortlessly scale their AI inference infrastructure. This scalability is ideal for businesses experiencing growing workloads and requiring flexible solutions to accommodate increasing demands.

Workload Versatility

With the most cores of a CPU, Ampere processors can handle multiple AI inference workloads concurrently—allowing different cores to process different AI models simultaneously, which enhances efficiency and reduces bottlenecks. Ampere CPUs can also run other non- AI tasks within the same server without interference. Because Ampere CPUs are general purpose processors, they are versatile enough to run whatever models or workloads emerge in the future.

Cost-Effectiveness

Ampere CPUs' higher core count increases cost-effectiveness by delivering more performance per watt compared to legacy x86 processors. At the data center level, because power is the ultimate limiter to rack density, lower power consumption also translates into the ability to fit in more servers within a rack. For more information, see Learn How to Triple Your Data Center Efficiency with Ampere AI Solutions.

Modern AI models are complex and demand significant computational resources. Ampere CPUs with higher core counts ensure efficient utilization of available resources by distributing the computational load across multiple cores—preventing underutilization of hardware and maximizing overall efficiency to virtually eliminate the issue of underutilized resources bogging down GPU deployments.

Software Optimizations for AI Inference

The suite of Ampere AI software solutions ensures that AI models run efficiently and deliver optimal performance. Software optimizations directly contribute to achieving the desired efficiency and performance balance in AI inference.

Benefits of Ampere Optimized AI Frameworks

Ampere Optimized AI Frameworks are designed to leverage the unique capabilities of Ampere CPUs, unlocking a new level of AI inference and training capabilities that propel businesses into the future. Ampere's optimized AI frameworks stand as a testament to the synergy between hardware and software, offering many advantages that redefine the AI experience.

Peak Performance

Ampere's AI frameworks are finely tuned to extract every ounce of performance from Ampere CPUs, maximizing their computational prowess. This optimization ensures that AI workloads are executed with unparalleled speed and efficiency, making Ampere-powered systems stand out in the competitive AI landscape. The additional performance derived varies depending on the AI model, commonly between 2X to 5X performance gain.

Seamless Integration

Designed with seamless integration in mind, Ampere's AI frameworks provide a smooth transition from development to deployment. Developers can harness the power of Ampere CPUs without intricate configurations, reducing setup complexities and accelerating time- to-value for AI projects.

Robust Software Ecosystem

Ampere's processors are based on the broadly used Arm ISA and come fortified with an AI software ecosystem that enables a seamless transition from the legacy x86 software ecosystem. Mature cross-compatibility, native performance optimization, extensive industry collaboration, developer-friendly resources, demonstrated real-world achievements, and close-knit partnerships with AI software vendors ensure a smooth migration to Ampere for AI applications.

Native FP16 Data Format Support

Ampere's AI inference capabilities are further empowered by unique native support for the FP16 data format. By leveraging FP16 precision, Ampere CPUs accelerate computations, making them ideal for real-time applications while maintaining compatibility and flexibility across various AI frameworks. The additional performance over the FP32 data format does not significantly impact accuracy nor requires cumbersome work on format conversion. This advancement underscores Ampere's commitment to providing efficient and high-performance solutions for AI inference workloads, particularly in scenarios where speed, accuracy, and resource utilization are critical.

Performance Proof

Ampere's investment in AI has been validated by benchmarking results that demonstrate remarkable speedups across different AI workloads. In the domains of natural language processing, recommendation engines, and speech recognition, Ampere CPUs—boosted by Ampere Optimized AI Frameworks—consistently deliver high performance and a software acceleration layer under continuous optimization of newly emerging AI models. Contact Ampere with any inquiries as to the performance of the models you use in your applications.

Conclusion

Ampere CPUs are the best choice for AI workloads. They deliver the best performance, cost-effectiveness, and power efficiency when compared to any other CPU or GPU. Regardless of the deployment requirements or model characteristics, Ampere delivers right-sized compute for AI inferencing, and Ampere AI acceleration software stack delivers high performance and seamless adoption. This is an economically and environmentally sustainable solution that future-proofs businesses against the risks of increased energy costs, changing AI workloads, and underutilization of expensive compute resources.

Seamless Integration

Designed with seamless integration in mind, Ampere's AI frameworks provide a smooth transition from development to deployment. Developers can harness the power of Ampere CPUs without intricate configurations, reducing setup complexities and accelerating time- to-value for AI projects.

Leveraging Ampere for AI Inference

As AI continues to reshape industries and drive innovation, Ampere stands at the forefront of providing advanced solutions for AI inference. Ampere's commitment to hardware advancements, software optimizations, and a unified inference model ensures that businesses can harness AI's transformative potential efficiently and effectively.

Transforming Business with Ampere AI Solutions

Ampere empowers businesses to elevate their AI capabilities and drive business transformation. By offering high-performance processors, optimized software libraries, and a unified inference model, Ampere enables seamless transitions from model development to deployment. With Ampere AI solutions, businesses can embrace the power of AI inference to achieve new levels of innovation and success.