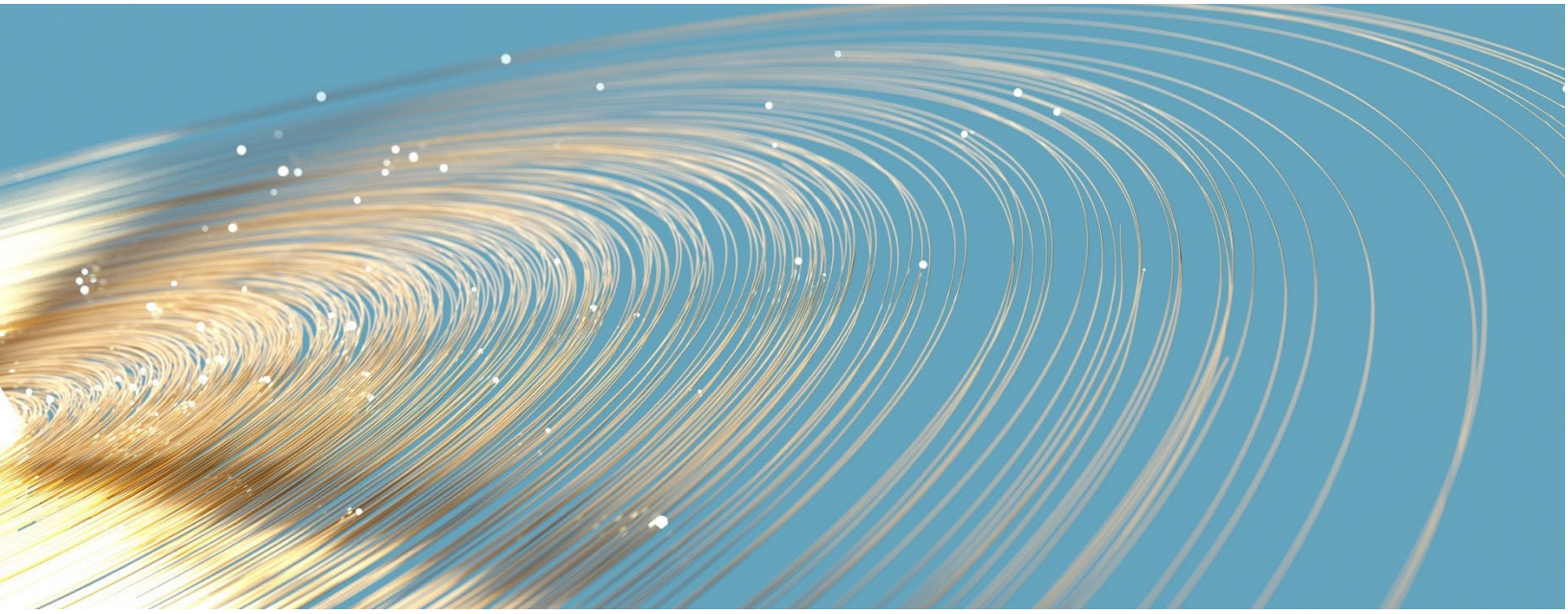
Everest Group®



PULSE | September 2024

Graphics Processing Units (GPUs) versus Central Processing Units (CPUs) in AI Processing

Insights into the Battle for AI's Future

CONTENTS

Author: **Richard Sear**, Managing Partner

Copyright © 2024, Everest Global, Inc. All rights reserved.

Why you should care

With 6-year-olds talking about it at the dinner table and 70-year-olds investing in it, AI will continue to impact our lives, from self-driving cars to life sciences and weather prediction. The infrastructure that powers these advances is becoming essential and will soon be as vital as any other infrastructure humans rely on. Processing approaches are key to this future, with two primary approaches battling for supremacy. The primary debate centers on Graphics Processing Units (GPUs) versus Central Processing Units (CPUs), with one or a combination likely defining AI's future over the next decade. The decisions we make today about which processing approach to use will have ripple effects across industries, determining innovation's pace, AI applications' feasibility, and tomorrow's competitive landscape. With that in mind, two companies hold significant influence in this space: NVIDIA and Intel Corporation, supported by many others.

Over the last five years, NVIDIA has outpaced Intel in the AI hardware market by focusing early on GPUs for Deep Learning (DL) and AI applications. NVIDIA's GPUs have become the industry standard for training and deploying AI models, leveraging the parallel processing capabilities essential for today's AI workloads. Alternatively, Intel has struggled to gain similar success despite significant investments in AI-focused CPUs and the acquisition of AI start-ups like Habana Labs (and earlier, Nervana, which was later replaced by Habana). While Intel's processors remain robust for traditional computing tasks, the company's AI strategy has produced mixed results, with its AI accelerators and GPUs failing to reach the same market adoption as NVIDIA's products. This scenario highlights Intel's challenges in transitioning from its stronghold in CPUs to competing in the AI hardware market. However, that market is changing, and there is a potential for novel approaches, especially in specialized learning models.

In this Pulse, we examine this debate's key points, identifying the technical issues, future directions, and any significant implications of these choices. This report will benefit observers interested in how AI develops, rather than the developers themselves. As an observer, you will need to understand how these decisions affect your product and technology choices, and how to ask the right questions of your teams.

What you need to know

The architectural differences

Understanding the fundamental differences in the architecture of CPUs and GPUs is essential to comprehending why these differences influence how we think about computing. As AI pushes traditional computing to its limits, knowing these differences is key.

- **CPUs** have been our computing foundation for decades. As an embedded principle, their architecture is designed for flexibility, with multiple powerful cores optimized for sequential tasks. CPUs manage everything from operating systems to executing complex algorithms. Each core processes a few threads at a time, making CPUs excellent for tasks that require single-threaded performance or complex but sequential decision-making processes. However, this same design limits their efficiency in AI applications, where massive parallel processing is needed
- **GPUs**, on the other hand, were initially developed to handle graphics rendering. Their architecture comprises thousands of smaller, simpler cores designed for high-throughput tasks like graphics processing. This design makes GPUs highly efficient in handling the massive parallelism required for AI tasks, especially in training deep neural networks. While each GPU core is less powerful than a CPU core, their large number allows them to process thousands of operations simultaneously, making them better suited for large-scale, matrix-like operations common in AI

Importance of processing power and parallelism

The debate over the relative strength of GPUs versus CPUs centers on how each manages processing tasks. AI's never-ending desire for data and high-speed computation has pushed traditional CPU architectures to their theoretical physical limits, resulting in high latency. GPUs, complementing CPUs, have become the driving force behind AI's current transformation.

- **CPUs** typically have 4 to 64 cores designed to handle complex instructions sequentially, excelling at tasks that require significant logic, control, and decision-making. They also feature large caches and advanced branch prediction to drive processing efficiency. However, CPUs struggle with tasks that benefit from parallelism, like large-scale data analysis or real-time AI inference. Their serial nature, combined with their relatively low core count, makes them slower than GPUs in these situations
- **GPUs**, on the other hand, feature thousands of cores that work simultaneously, offering significantly faster processing by handling multiple tasks simultaneously. For example, when training a neural network, a GPU can compute the gradients of thousands of weights across a large dataset at once, cutting a process that could take days on a CPU down to mere hours. However, GPUs are less efficient at handling tasks that require complex processing, an area where CPUs still dominate

Emerging concerns for energy efficiency

With rising energy costs and sustainability as a top corporate priority, computing infrastructure's energy efficiency has received significant attention, particularly when comparing the power consumption of CPUs and GPUs and their environmental impact.

- **CPUs** are designed with energy efficiency in mind, especially in mobile and server spaces where power consumption directly impacts battery life and operational costs. Advanced power management features like Dynamic Voltage and Frequency Scaling (DVFS) allow CPUs to optimize power use based on workload. This makes CPUs well-suited for energy-sensitive applications like mobile AI inference or large-scale cloud deployments that need cost-effective operations. Energy efficiency has been a focus in CPU design since the early days of silicon development
- **GPUs**, in contrast, are incredibly power-intensive due to their parallel processing capabilities, consuming far more energy per operation than CPUs. However, their speed can help offset this. For example, in DL model training, a GPU may use more power than a CPU, but it completes the task faster, ultimately reducing the energy consumed. This balance makes GPUs attractive in data centers where computation speed often outweighs higher energy costs

Escalating costs

Cost plays a vital role when deciding on an AI approach. Beyond the upfront hardware investment, the Total Cost of Ownership (TCO) includes power consumption, maintenance, and scalability. An increasing provocation is whether investing in GPUs is truly the best option or if CPUs still offer value.

- **CPUs** are generally cheaper and widely available for general-purpose computing. Being readily available means they can handle various tasks without specialized hardware, making them a cost-effective solution for most applications. However, for tasks requiring parallel processing, CPUs' cost-performance ratio diminishes. To achieve the same performance as a single GPU, a large cluster of CPUs may be necessary, driving up both hardware and power costs, ultimately reducing their value. Therefore, cost is a significant use case variable
- **GPUs**, particularly high-end models designed for AI (like NVIDIA's A100 or H100), can be exorbitant. However, as discussed, their ability to accelerate AI workloads can justify the investment, especially in industries where rapid time-to-market is vital

Examples in motion

Switching Bidirectional Encoder Representations from Transformers (BERT)'s training

There is no question that Google has been a leader in developing large-scale language models like BERT. The model's complexity and the vast datasets it processes means training BERT demands tremendous computational power, which, fortunately, was something Google had in abundance. While CPUs could handle the task, the training process would be painfully slow due to the need for

parallel processing. Instead, Google initially relied on GPUs, significantly reducing training times. However, as efficiency became more crucial, Google shifted to its custom Tensor Processing Units (TPUs), designed specifically for training Large Language Models (LLMs). By modifying GPUs to TPUs, Google achieved faster training times and enhanced energy consumption, highlighting the need for specialized hardware for large-scale AI tasks.

Object detection in Tesla's autonomous vehicles

Tesla's self-driving mode relies on real-time object detection, which involves processing vast amounts of visual data quickly to make micro-second decisions. It is a truly remarkable technology. While GPUs offer the processing power needed for real-time object detection, they adversely consume significant energy and generate heat – both major issues in the automotive world. Conversely, CPUs struggle to deliver the real-time performance needed for object detection. Initially, Tesla used NVIDIA GPUs but later developed its custom Full Self-driving (FSD) computer. This AI chip combines CPUs' efficiency with specialized AI processing units, balancing real-time performance with power and heat challenges, highlighting the trade-offs in hardware selection for specific AI applications.

AI-driven healthcare diagnostics at GE Healthcare

GE Healthcare has developed an AI-powered diagnostic tool to assess medical images like X-rays and MRIs. These tools assist doctors in detecting conditions like cancer or fractures. To be effective, these tools need to process large, high-resolution images quickly to be useful in clinical offices and hospitals. At first, CPUs were used for image processing, but the slow processing speeds made real-time diagnostics impractical, especially when handling large patient datasets. To overcome this, GE Healthcare shifted to GPUs, enabling faster image analysis through parallel processing. However, concerns over GPUs' costs and power needs prompted the exploration of specialized AI accelerators that combine GPUs' parallel processing power with CPUs' energy efficiency. This approach balanced performance with operational efficiency, demonstrating the complexity of optimizing AI hardware in healthcare.

Where is the pulse on GPUs versus CPUs heading?

A shift to hybrid approaches

As AI tasks become more complex, the future of computing will likely involve integrating GPUs and CPUs into a synergistic system, rather than choosing one over the other. Hybrid computing is emerging as the next step in AI infrastructure

evolution, offering a blend of processing capabilities to meet these expanding needs.

- **Hybrid systems:** Integrating CPUs, GPUs, and specialized hardware like TPUs is becoming the standard in high-performance computing. In this setup, CPUs handle control and logic tasks, GPUs manage parallel computations, and TPUs or Application-specific Integrated Circuits (ASICs) offer greater efficiency for AI-specific tasks. This approach maximizes overall performance, enabling organizations to tailor their infrastructure to suit their AI workloads

Emerging technologies will change the space

The AI hardware landscape is expected to evolve rapidly with advances in new technologies. While today's debate centers on CPUs versus GPUs, the future may look very different with emerging innovations like quantum computing.

- **TPUs and ASICs:** As discussed earlier, Google's TPUs and other custom-designed ASICs will see increased adoption for AI workloads, especially in cloud situations. These processors are optimized for operations required in ML like matrix multiplications and convolutional operations. Unlike GPUs, TPUs and ASICs are designed for efficiency and offer higher performance gains while consuming less power. However, they are less flexible than GPUs and CPUs, making them ideal for specific AI applications but less suited for general-purpose tasks
- **Quantum computing:** Although still in its early stages, quantum computing has the potential to transform AI processing sooner than expected. Operating on qubits, quantum computers can simultaneously represent multiple states, exponentially speeding up specific computations. While this might render the CPU versus GPU debate irrelevant for certain AI applications in the future, quantum computing still faces significant technical challenges. For now, GPUs and CPUs will continue to dominate the AI landscape. This is the topic for a Pulse I wrote earlier on "The Meaningful Path to Quantum Computing"

Specialized models versus LLMs

The reliance on large, general-purpose models like LLMs is gradually giving way to specialized models tailored for specific tasks. This shift in AI models will influence how we select processing units, CPUs, GPUs, or even specialized AI accelerators, driving a need for a more tailored approach to balancing power, cost, and efficiency.

- **LLMs:** Models like GPT-4 can handle diverse tasks, leveraging huge data amounts to offer generalized AI capabilities. These models need significant computational power for training and inference, making GPUs the preferred choice due to their ability to handle large-scale parallel processing efficiently
- **Specialized models:** Tailored for specific tasks, such as image recognition, natural language understanding in a particular area, or real-time decision-making, specialized models can run on less powerful hardware or custom AI accelerators like TPUs or ASICs. Their reduced complexity and targeted design

make them more suitable for CPUs or edge devices, where processing power is limited

As the trend shifts from LLMs to specialized models, the choice of hardware will become even more critical. GPUs will continue to dominate in large-scale applications, but specialized models will require hardware that balances power, cost, and efficiency. This shift could drive further innovation in AI hardware, leading to new processors optimized for specific AI tasks. Further blurring the lines between traditional CPU and GPU roles will likely lead to increasing acquisitions as companies seek to consolidate approaches. It may even create new opportunities for CPU-centric systems like Intel's to regain ground in the AI space.

Final word

The debate between GPUs and CPUs for AI processing is no longer binary. Today and in the future, it is an extraordinarily complex, long-term, and costly decision shaped by the specific requirements of each desired application. GPUs have redefined the boundaries of what is possible in AI, enabling breakthroughs in DL and complex simulations that were once impossible. However, CPUs continue to play a vital role, especially in areas that require versatility, energy efficiency, and complex decision-making.

As AI evolves rapidly, so will the hardware powering it. Hybrid systems that combine the strengths of CPUs, GPUs, and emerging technologies like TPUs and quantum computers will define AI infrastructure's future. The key takeaway for organizations, providers, and enterprises is to remain adaptable and continuously evaluate the trade-offs between different processing units to ensure their AI strategies are technically and economically optimized. The planning horizon seems to operate on a rolling three- to five-year basis, making today's decisions crucial for determining AI development's speed and efficiency.

As the AI revolution accelerates, understanding CPU and GPU architectures' nuances is essential for anyone looking to stay ahead of the curve in this exciting field.

Sources:

https://www.nvidia.com/en-us/data-center/gpus-for-ai/
https://www.intel.com/content/www/us/en/artificial-intelligence/ai-gpu-vs-cpu.html
https://cloud.google.com/tpu/docs/tpus
https://towardsdatascience.com/deep-learning-with-gpus-9ef7f9bb623
https://www.amd.com/en/technologies/instinct-mi100
https://developer.nvidia.com/deep-learning
https://openai.com/blog/ai-and-compute/
https://medium.com/analytics-vidhya/gpu-vs-cpu-for-deep-learning-5930f2924184
https://www.kdnuggets.com/2020/01/cpu-vs-gpu-deep-learning.html
https://www.zdnet.com/article/cpu-vs-gpu/
https://arxiv.org/pdf/2103.06137.pdf
https://machinelearningmastery.com/should-you-use-gpus-or-cpus-for-deep-learning/
https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html
https://developer.amd.com/resources/architecture-docs/
https://venturebeat.com/2020/12/28/tpu-vs-gpu-vs-cpu-what-should-you-use-to-train-your-ai-models/
https://www.datacenterknowledge.com/hardware/gpu-vs-cpu-ai-processing
https://www.techradar.com/news/gpu-vs-cpu-vs-tpu
https://www.nextplatform.com/2021/03/10/why-cpus-are-essential-in-a-gpu-driven-world/
https://www.nasdaq.com/articles/gpu-vs-cpu:-which-is-better-for-machine-learning-2021-03-18
https://www.forbes.com/sites/forbestechcouncil/2021/08/25/the-gpu-vs-cpu-debate-for-ai-continues/

Everest Group®
With you on the journey

Everest Group is a leading research firm helping business leaders make confident decisions. We guide clients through today's market challenges and strengthen their strategies by applying contextualized problem-solving to their unique situations. This drives maximized operational and financial performance and transformative experiences. Our deep expertise and tenacious research focused on technology, business processes, and engineering through the lenses of talent, sustainability, and sourcing delivers precise and action-oriented guidance. Find further details and in-depth content at www.everestgrp.com.

For more information about Everest Group, please contact:

+1-214-451-3000

info@everestgrp.com

For more information about this topic please contact the author(s):

Richard Sear,
Managing Partner
richard.sear@everestgrp.com

Notice and Disclaimers

Important information. Please review this notice carefully and in its entirety. Through your access, you agree to Everest Group's terms of use.

Everest Group's Terms of Use, available at www.everestgrp.com/terms-of-use/, is hereby incorporated by reference as if fully reproduced herein. Parts of these terms are pasted below for convenience; please refer to the link above for the full version of the Terms of Use.

Everest Group is not registered as an investment adviser or research analyst with the U.S. Securities and Exchange Commission, the Financial Industry Regulatory Authority (FINRA), or any state or foreign securities regulatory authority. For the avoidance of doubt, Everest Group is not providing any advice concerning securities as defined by the law or any regulatory entity or an analysis of equity securities as defined by the law or any regulatory entity. All Everest Group Products and/or Services are for informational purposes only and are provided "as is" without any warranty of any kind. You understand and expressly agree that you assume the entire risk as to your use and any reliance upon any Product or Service. Everest Group is not a legal, tax, financial, or investment advisor, and nothing provided by Everest Group is legal, tax, financial, or investment advice. Nothing Everest Group provides is an offer to sell or a solicitation of an offer to purchase any securities or instruments from any entity. Nothing from Everest Group may be used or relied upon in evaluating the merits of any investment.

Do not base any investment decisions, in whole or part, on anything provided by Everest Group.

Products and/or Services represent research opinions or viewpoints, not representations or statements of fact. Accessing, using, or receiving a grant of access to an Everest Group Product and/or Service does not constitute any recommendation by Everest Group that recipient (1) take any action or refrain from taking any action or (2) enter into a particular transaction. Nothing from Everest Group will be relied upon or interpreted as a promise or representation as to past, present, or future performance of a business or a market. The information contained in any Everest Group Product and/or Service is as of the date prepared, and Everest Group has no duty or obligation to update or revise the information or documentation. Everest Group may have obtained information that appears in its Products and/or Services from the parties mentioned therein, public sources, or third-party sources, including information related to financials, estimates, and/or forecasts. Everest Group has not audited such information and assumes no responsibility for independently verifying such information as Everest Group has relied on such information being complete and accurate in all respects. Note, companies mentioned in Products and/or Services may be customers of Everest Group or have interacted with Everest Group in some other way, including, without limitation, participating in Everest Group research activities.

Copyright © 2024, Everest Global, Inc. All rights reserved.