POV

PUT YOUR AI SOLUTION ON STEROIDS POINT OF VIEW

accenture

MATCH GPU PERFORMANCE AT HALF THE COST FOR AI INFERENCE WORKLOADS

Proven CPU-based solution from Accenture and Intel boosts the performance and lowers the cost of AI inferencing by enabling an easy-to-deploy, scalable, and cost-efficient architecture

AI INFERENCING—THE NEXT CRITICAL STEP AFTER AI ALGORITHM TRAINING

Artificial Intelligence (AI) solutions include three main functions—identifying and preparing data, training an artificial intelligence algorithm, and using the algorithm for inferring new outcomes. Each function requires different compute recourses and deployment architecture. The choices of infrastructure components and technologies significantly impact the performance and costs associated with deploying an end-to-end AI solution. Data scientists and machine learning (ML) engineers spend significant time devising the right architecture for all stages of the AI pipeline.

Once an AI computer/algorithm has been trained through traditional or deep learning techniques, it can deliver value by interpreting data (i.e., inferring). Through inference, an AI algorithm can analyze data to:

- Differentiate between various items
- Identify trends and patterns that can be leveraged during decision-making
- Reveal opportunities and possible solutions
- Recognize voices, faces, images, etc.

Revealing hidden possibilities—Accenture AIP, powering industry-leading Intel® infrastructure solutions

Accenture Applied Intelligence Platform (AIP) provides a complete continuum of tools for analytics and AI applications—enabling organizations to aggregate data, slice it as appropriate, and then plug it into purpose-built models. **The result—impactful business decisions based on complete, accurate data, enhanced through AI and automation.** Rather than offer a prescriptive solution to today's data analytics challenges, Accenture AIP delivers a flexible platform that can run on-premises or in the cloud of your choice. Fully tested to interoperate and wrapped in security, Accenture AIP can be tailored to support a variety of analytics workloads such as AI Inferencing, among others.

As we look to the future, AI inference will become increasingly important to businesses operating in all segments—from health care to financial services to aerospace. And as the reliance on AI inference continues to grow, so does the importance of choosing the right AI infrastructure to support it.

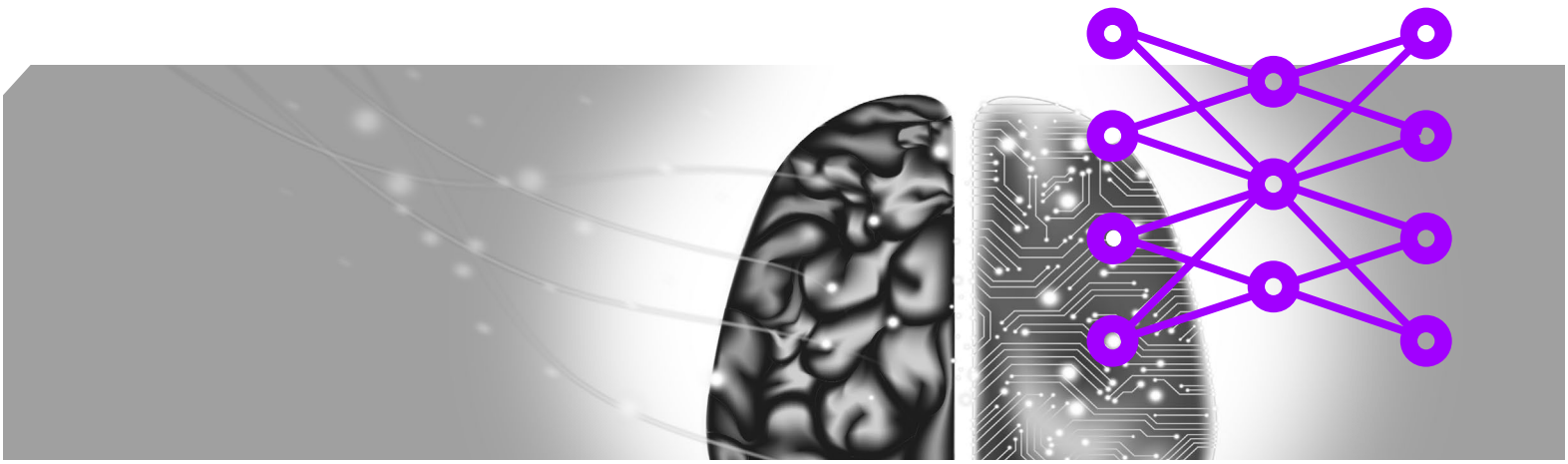
Today, Intel® and Accenture can help you make the right architecture choices to achieve a high-performance, cost-efficient, and multi-purpose AI infrastructure based on the latest generation Intel® Xeon® Scalable Processors. This solution delivers at-par performance as a GPU-based system but at up to 50% lower cost.

AI INFERENCE IN ACTION

While the use cases for AI inference are rapidly expanding, some of the most popular use cases include:

- Natural language processing (NLP)
- Chatbots and other smart web services
- Smart personal assistants
- Self-driving cars
- Fraud detection
- Supply chain modernization
- Sentiment analysis for marketing insights
- Traffic prediction for website or retail store demand
- Defect identification in manufacturing or in the field
- Content moderation in social media
- Advertising optimization spend to maximize revenue

Accenture implements a variety of AI algorithms in business, enabling our clients to do things differently and do different things. Our approach to embedding AI, automation, and analytics at the core of the enterprise helps break down silos and create more agile and adaptive processes. This approach enables better decision making and empowers businesses to identify and capture completely new opportunities. Our solutions can deliver these benefits at speed and scale thanks to an extensive suite of commercial solutions for industries and functions.



Accenture Applied Intelligence Platform is already redefining possibilities and powering new outcomes across business functions through:

- **Intelligent Customer Engagement**—To help unlock new insights and create more intelligent processes so you can provide excellent customer experiences at a decreased service cost per customer.

- **Intelligent Revenue Growth**—To reveal a real-time, 360° view of your customers and glean insights you couldn't see before—powering new strategies and capabilities for optimization across growth levers.

- **Intelligent Healthcare**—To deliver comprehensive patient service programs from clinical trials through treatment management, designed to improve patient care and outcomes.

- **Intelligent Supply Chain**—To increase visibility within and across all supply chain functions to create a customer-driven, integrated operating model that offers more efficiency, transparency, and agility.

- **Intelligent Financial Crime Detection**—To help financial institutions transform their organizational processes by streamlining operations and augmenting investigator performance to better detect threats and examine high case volumes quickly and thoroughly.

CPUS VS GPUS AND OTHER HARDWARE FOR AI INFERENCE

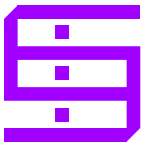
Training a deep learning model involves tuning weights in each layer of the network to reduce errors. This task is computationally intensive but can be parallelized. Graphical processing units (GPUs) are an excellent choice for running these massively parallel tasks.

Once a model has been trained, and its hyperparameters are fine-tuned, the compute requirement drops massively. This is where CPUs shine. They are just as performant for an overwhelming number of use cases, but a CPU costs only a fraction of a GPU. Therefore, for AI inference workloads, GPUs appear to be an unnecessarily high expense.

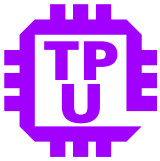
To provide an analogy, think of the GPU as an expensive race car, which is built to be fast and provide excellent performance on a racetrack. Now think of the AI inference process as driving a car in the city. The race car is poorly suited for this use case from both a performance and economic perspective; a cost-efficient compact car would be a better choice for the job.

For most of our clients, model training consumes only a very small fraction of total compute time. In contrast, AI inferencing is used very often and consumes a considerable amount of compute resources. Accenture's default recommendation is to use a scalable cluster of stateless CPUs for inferencing while utilizing GPUs where they perform best—algorithm training.

Other devices well suited for inferencing include:



- **Field-Programmable Gate Array (FPGA) chips**—Can be used to configure custom hardware and machine learning or data-related use cases (e.g., data centers, deep learning inference, etc.) with low power consumption—resulting in energy-efficient, lower-cost, low-latency computing. Intel's OpenVINO™ Toolkit provides the software stack required for deep learning inference using FPGAs.



- **Tensor Flow Processing Unit (TPU)**—AI Accelerator Application Specific Integrated Circuit (ASIC) designed by Google. The Edge TPU is a small, lightweight ASIC that provides high-performance machine learning inferencing for low-power devices. As AI use cases mature and become more specialized, we expect application-specific ASIC usage will grow rapidly, providing cost and performance benefits.

Table 1. GPU, FPGA, and ASIC comparison

Criteria	GPU	FPGA (Intel Arria)	ASIC (Google TPU-Edge)
Usage	Training at the core; inference at the edge or the core	Inference at the edge or the core	Inference at the edge or the core
Supported AI models	Most AI libraries and deep learning models	• Pre-trained open-source and OpenVINO models • Hardware and software optimization required for new models	Limited set of models
Model size	Large	Medium	Small
Latency	High	Medium	Low
Efficiency	Low	Medium	High

OUR AI INFERENCE BENCHMARKS

With the overall goal of reducing the cost of running inferencing workloads—while still maintaining the same level of high performance—Intel and Accenture have built an AI inferencing solution using a CPU-based architecture. To prove the effectiveness and efficiency of the solution, the Accenture | Intel Alliance team completed a series of benchmarking tests to compare the performance of running various AI inferencing workloads on CPU versus GPU architectures.

AI Inference Benchmarking

During the benchmark exercise, we used standard GPU and CPU instances on AWS EC2. We tested AI inferencing for some of the most popular algorithms, including NLP for sentiment analysis and XGBoost.

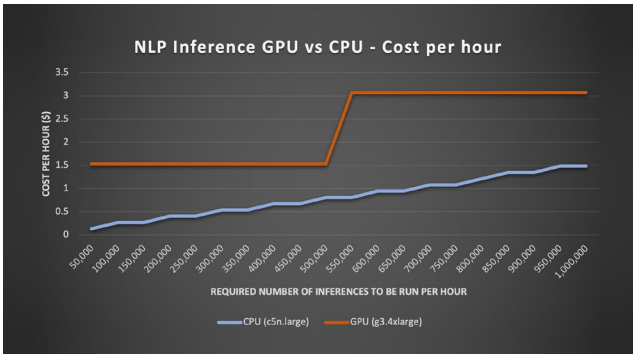
Benchmarking Results for NLP

The benchmark testing revealed that distributing NLP inference workloads across multiple CPU instances powered by 2nd Generation Intel® Xeon® Scalable Processors leads to the same inferencing performance at one-half the cost of a GPU instance running on the AWS public cloud.[1]

GPU VS CPU

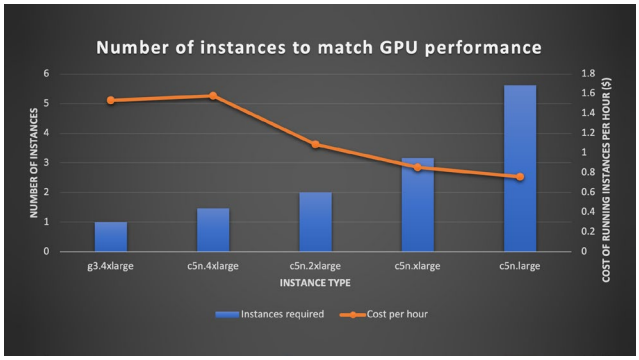
[1] Based on Sentiment Analysis Inferencing benchmark testing run on various AWS EC2 instances using pre-trained PyTorch models, March 2020

Figure 1. AI inference GPU vs. CPU cost analysis[1]



The chart shows how a scalable cluster of CPUs can respond to various levels of workload at significantly lower cost. When a smaller number of inferences are required, the GPU sits idle for a significant amount of time, i.e. the utilization is low. The advantage of using a CPU cluster with small instances is that the utilization factor is significantly higher than a GPU without compromizing performance.

Figure 2. AI inference GPU vs. CPU performance analysis



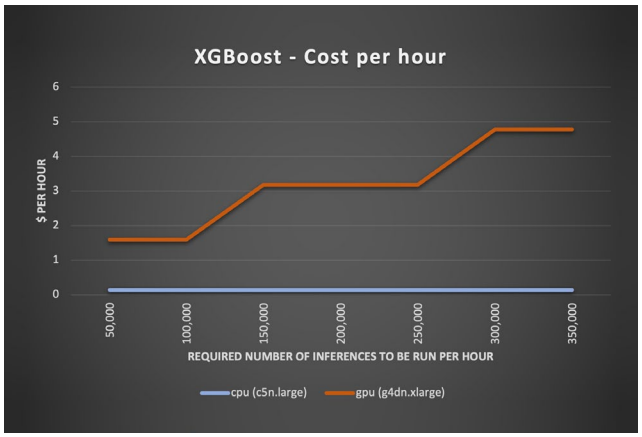
The chart shows the number of CPUs needed to match the performace of a GPU on the Stanford SST Sentiment Analysis Benchmark. Larger CPU instances can match the performance of a GPU but will not lead to a cost advantage. By spreading the workload across multiple smaller CPUs, c5n.large, we can achive the same performance of a GPU but at half the cost

When comparing benchmark run on physical cores, the results revealed performance gains of 12-18%, depending on the instance type, as shown above on Figure 1.

Benchmarking Results for XGBoost

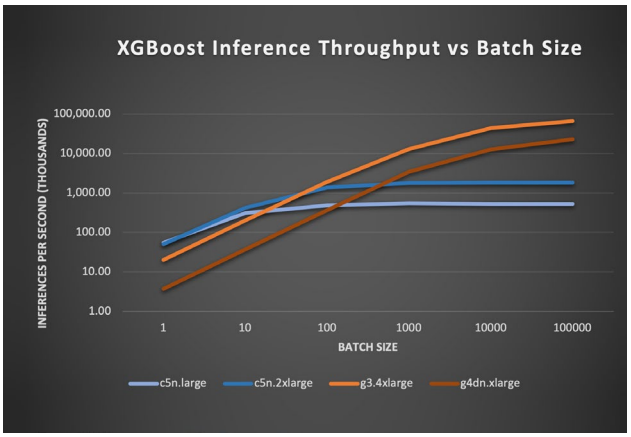
As illustrated in the summary tables below, CPUs offer higher throughput(10x) for most common batch sizes.

Figure 3 . Summary of XGBoost benchmarking test results

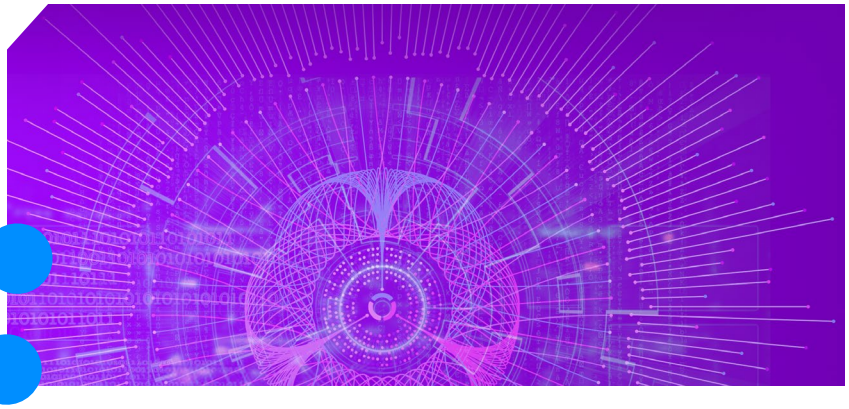


The chart shows how a cluster of CPUs compares against a GPU cluster. For most real world inferencing use cases, one CPU is sufficient. The GPU cluster however needs to spin up more instances as the load grows. Even though the CPU utilization is low in this use case, the cost is low as well. The CPU cluster outperforms the GPU cluster at least 10x in terms of cost for XGboost inferencing use cases.

Figure 4 . Summary of XGBoost throughput benchmarking results



The chart shows the comparison between throughput (inferences per second) achieved on a CPU vs a GPU. At lower batch sizes, where 99% of the inferencing use cases lie, the CPU beats the GPU performance 5x to 15x.



The Accenture | Intel recommended architecture uses a cluster of CPU instances to match the performance of a single GPU. Distributing the workload across the cluster is possible because each inference in the benchmark is stateless, so it does not depend on the results of an earlier inference. Running CPU inferencing in parallel enables this architecture to deliver the expected cost savings.

IN CLOSING

Experience has proven that GPUs are well-suited to support the deep learning required to train AI algorithms. For training purposes, massive parallel computations and high utilization justify the higher cost and greater complexity of a GPU-based architecture. During AI inferencing, however, utilization of each GPU instance is generally low and/or fluctuates often, making GPUs a poor fit for this type of workload.

Rather than assume the high cost of a GPU-based architecture for AI inferencing, Intel and Accenture recommend a scalable CPU-based cluster solution with auto-scaling. Such an architecture ensures high levels of utilization to significantly drive down costs, while also providing on-par performance to a GPU-based architecture.

Accenture and Intel work together to offer application-specific recommendations for optimized AI pipeline architecture. In addition to AI inferencing discussed in this document, Accenture and Intel can also assist you with hardware and software optimization projects. Regardless of your specific AI use cases and/or applications, the Accenture and Intel team has the expertise to provide custom solutions that meet your unique requirements.

Authors

Ramtin Davanlou,
AI Tech Arch Innovation
Senior Principal
Accenture

Neil Fernandes,
Data Science Consultant
Accenture

Tim Wu
Data Science Senior Manager
Accenture

accenture