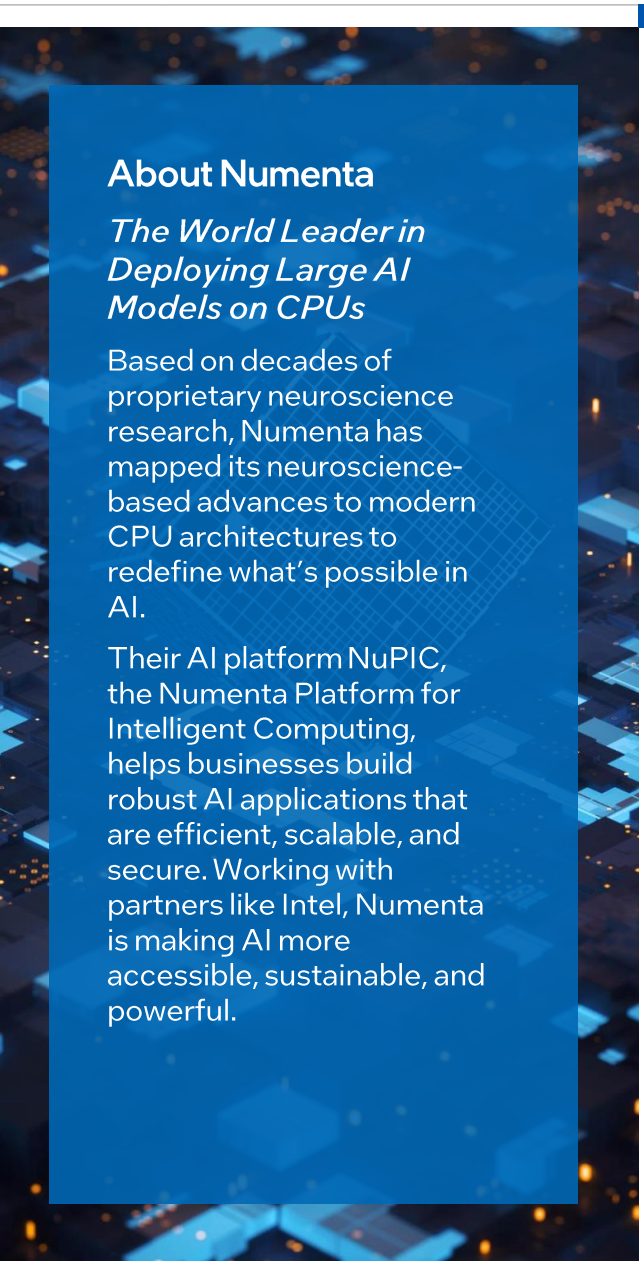
Solution Brief

AI Platform
Interactive Media
Financial Services, Gaming, Legal, Retail

Build and Scale Powerful AI Applications on CPUs with NuPIC

The Numenta Platform for Intelligent Computing marries neuroscience-based AI with innovative Intel® processors to offer efficient and secure deployment of large language models on CPUs for all verticals.

accelerated by intel



About Numenta

The World Leader in Deploying Large AI Models on CPUs

Based on decades of proprietary neuroscience research, Numenta has mapped its neuroscience-based advances to modern CPU architectures to redefine what's possible in AI.

Their AI platform NuPIC, the Numenta Platform for Intelligent Computing, helps businesses build robust AI applications that are efficient, scalable, and secure. Working with partners like Intel, Numenta is making AI more accessible, sustainable, and powerful.

The Challenges of LLM Deployment

In a short amount of time, large language models, or LLMs, and generative AI technologies have helped with the advancement and proliferation of artificial intelligence. By interpreting natural language and providing human-like responses, LLMs like GPT have garnered attention from all industries. In response, businesses are racing to experiment with new models and create new AI initiatives centered around LLMs and generative AI to stay competitive and maintain pace with the current evolution of technology.

Despite the growing popularity of LLMs, deploying them poses challenges that has limited their adoption. The complexity of LLMs' large-parameter, real-time interactions makes them prohibitively expensive and complex to deploy at scale. Up to this point, successful LLM deployments have relied on costly graphics processing units (GPUs) to meet their high computational requirements. However, deploying LLMs on GPUs can be extremely expensive and requires significant IT infrastructure and data resources. These challenges are compounded by the ongoing after-effects of the recent global GPU shortage, which may still force businesses to wait for extended periods of time to secure the hardware needed to start new AI projects or scale existing ones. Lastly, many businesses are hesitating to adopt LLMs due to the potential security concerns associated with sending sensitive data across online SaaS platforms.

Despite these deployment challenges and the concern around data privacy, businesses recognize that if they wait to adopt LLMs they risk being left behind by their competitors. Generative AI is here to stay, and the benefits can be substantial, which is why Numenta is helping businesses overcome these challenges with an AI platform that makes deploying LLMs on central processing units (CPUs) more simple, scalable, secure, and cost-effective.

Solution Overview: NuPIC™

The Numenta Platform for Intelligent Computing, or NuPIC, is a neuroscience-based AI platform that enables enterprises to build powerful and robust AI applications on CPUs. NuPIC empowers businesses to leverage the flexibility of CPUs, freeing them from a dependency on GPU-based hardware that is prohibitively expensive and complex for IT to scale.

At the heart of NuPIC is a highly optimized inference server that achieves high throughput and low latencies without compromising accuracy. Though not a requirement, NuPIC takes advantage of the Intel® Advanced Matrix Extensions (Intel® AMX) instruction set for maximum performance gains. Businesses can effortlessly kick-start their AI initiatives with NuPIC's library of pre-trained language models, customize the models to their business requirements, and run them directly in the NuPIC Inference Server.

NuPIC is designed to enable AI applications for a wide range of industries, including financial services, gaming, legal, and healthcare. For example, NuPIC makes it easy to streamline customer support by detecting customer sentiment, summarizing large amounts of text, automatically categorizing documents, and many other text analysis use cases. For businesses with unique requirements, Numenta offers the flexibility to bring your own model, or fine-tune NuPIC pre-trained models, and deploy them in the inference server. Whether a customer is deploying a new LLM or multiple existing models, Numenta enables them to build, enhance, and scale powerful language-based applications – no GPUs required.

Deployed via docker container, NuPIC is adaptable to any CPU-based server, whether it's on-premise, via private cloud, or even on a laptop. This deployment approach ensures that all data and models reside completely on the customers' private storage systems, giving customers complete control over their models, data, and data governance policies.

NuPIC Solutions

NuPIC allows customers to build CPU-based AI applications that extract insights, identify patterns, and generate summaries from unstructured data with the following set of solutions:

Document Retrieval:
Harness the power of retrieval augmented generation (RAG) to understand and locate relevant documents in real time.

Customer Support:
Sift through chatlogs and process live text to quickly identify, categorize, and address customer concerns.

Text Analysis:
Dive deep into your data, identifying trends and patterns to inform decisions.

What Sets NuPIC Apart: A New Approach to LLMs with Intel

Capitalizing on the innate benefits of CPUs, NuPIC enables businesses to run hundreds of different large language models in parallel and serve a large volume of asynchronous customer requests on a single CPU server.

Numenta leverages the **Intel® AMX instruction set on** Intel® Xeon® processors, high-power CPUs, to replace more expensive and complex GPUs in LLM operations. Although CPUs are more accessible than GPUs, their adoption for AI inferencing has traditionally been hindered by slower processing speeds. Numenta solves this challenge by optimizing a variety of NLP models for Intel Xeon processors that offer Intel AMX instruction sets designed to accelerate AI workloads. The result is LLMs that provide advantages in inferences per dollar and inferences per watt, making AI applications substantially lower in cost, consuming significantly less energy, and more sustainable and scalable overall.[1]

Sequence Length	Batch Size	Inferences Per Second on BERT-Large		Speed up
		NuPIC on 5th Gen Intel Xeon Scalable Processor	NVIDIA A100	
64	1	2891	166 *	17.4x
128	1	901	128	7.0x
128	8	-	495	1.8x

Table 1. *Inference results for NuPIC optimized BERT-Large on 5th Gen Intel® Xeon® Scalable processor vs. NVIDIA A100.[2] *estimated*

By optimizing LLMs on Intel® architecture enabled with Intel AMX, NuPIC delivers significant improvement in inference throughput compared to previous generations of Intel processors and significant speed ups compared to GPUs. For BERT-Large, a language model trained in natural language processing, NuPIC on Intel Xeon Scalable processors outperformed the NVIDIA A100 GPU by up to 17x.[2]

GPUs require higher batch sizes for best parallel performance. However, batching leads to a more complex inference implementation and introduces undesirable latency in real-time applications. In contrast, NuPIC does not require batching, making applications flexible, scalable, and simple to manage. NuPIC at batch size 1 still outperforms the batched NVIDIA GPU implementation by 1.8x.[2]

In addition to embedding models like BERT, NuPIC also enables businesses to run industry-standard generative AI models efficiently on CPUs. NuPIC's optimized generative AI model on AMX-enabled Intel Xeon Scalable processors outperformed a standard LLAMA27B model by 2.8x, making the platform well suited for applications that support a large customer base and serve many client requests while maintaining high throughput.[3] With NuPIC, applications can handle numerous requests without batching inputs, allowing each client to be fully asynchronous.[1]

In addition to serving customers asynchronously, NuPIC on Intel Xeon processors lets customers efficiently balance the computational requirements of different models. It empowers customers to run multiple embedding (e.g. BERT) and generative (e.g. GPT) models for different tasks concurrently on the same server.

This kind of efficiency is hard to replicate with GPUs, and saves time, energy, and costs. When customers' computational needs grow, CPUs like Intel Xeon processors are much easier to add to on-premise and cloud-based environments than integrating the complex infrastructures of additional GPUs.

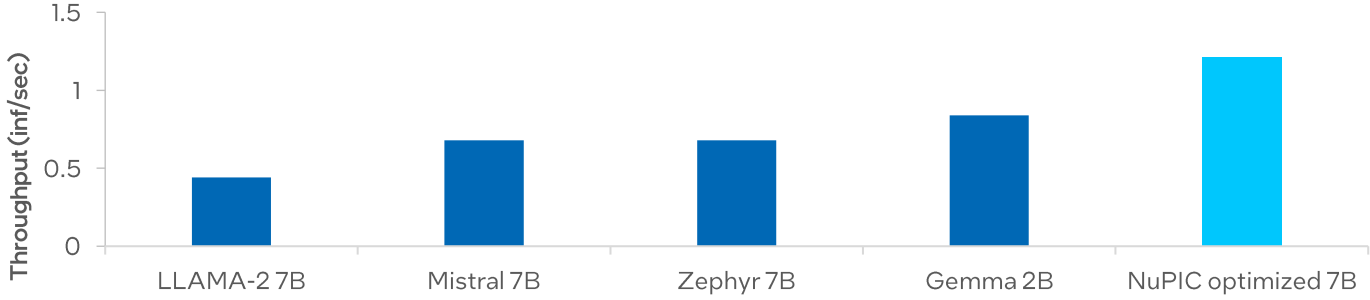


Table 2. *Inference results for NuPIC optimized GPT vs standard GPT models on 4th Gen Intel® Xeon® Scalable processor. Each inference request processes 512 input tokens and produces 64 output tokens.*

NuPIC's Performance Highlights

Realize the following performance improvements with NuPIC:

- 2-17X Acceleration of large language models compared to GPUs[2,3]
- 123X Increased inference throughput for short sequence-length tasks on Intel® Xeon® Scalable processors enabled with Intel AMX compared to AMD Milan[4]
- 20X Increased inference throughput for long sequence-length tasks on Intel® Xeon® CPU Max Series processors compared to previous-generation CPUs[5]

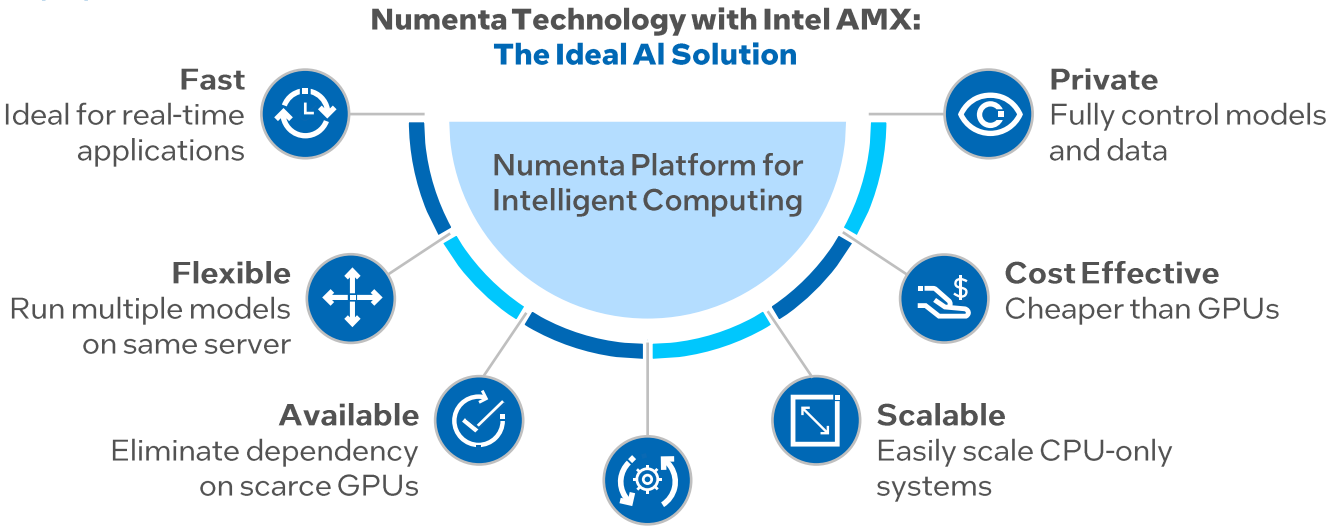
Advancing Sustainability Efforts

As LLM applications continue to increase, so do the alarming energy demands of AI. As models expand in size to accommodate more complex tasks, the demand for servers to process these models also grows exponentially. It is estimated that ChatGPT consumes approximately the same amount of electricity per day for inference as 33,000 US households do in the same time period.[6]

NuPIC offers a win-win solution: Businesses can invest in their future regulatory compliance while realizing cost savings. NuPIC-optimized models demand less memory and 5-21x less power,[7] and can lead to lower operational costs and a more efficient and sustainable AI solution.

> "Numenta and Intel are collaborating to deliver substantial performance gains to Numenta's AI solutions through the Intel Xeon CPU Max Series and 4th Gen Intel Xeon Scalable processors. We're excited to work together to unlock significant throughput performance accelerations for previously bandwidth-bound or latency-bound AI applications such as conversational AI and large document processing."
>
> **– Scott Clark, Vice President and General Manager of AI and HPC Application Level Engineering, Intel.**

Key Benefits



Numenta

Superior price performance on CPU-only infrastructure: Achieve the high-throughput, low-latency performance required for real-time applications while leveraging CPUs to keep costs down.

Unparalleled flexibility: Run hundreds of different models, handle a large number of client requests, add new models easily, and simplify IT management with a scalable CPU-based infrastructure.

Accelerated deployment and scaling of LLMs: Benefit from production-ready pre-trained LLMs that require no machine learning expertise or additional infrastructure investment.

Maintain ownership of models and sensitive data: Keep your models and data within your company walls, eliminating reliance on online APIs that collect data and are susceptible to data privacy risks. Numenta never sees your data.

Numenta Partners with Gallium Studios and Intel to Enhance AI Simulation Game

Challenge

Gallium Studios founders, Lauren Elliott (Co-creator of Carmen Sandiego) and Will Wright (Creator of The Sims) joined forces to create Proxi, an innovative AI simulation game. Proxi uses LLMs and other techniques to generate concepts, images, audio, and 3D scenes from memories players enter. These memories are then brought to life, like snow globes, in real time. Players can visit other players' Proxies in this virtual world.

The team at Gallium Studios initially relied on external APIs to power their LLMs. Because players can input anything for their memories, they needed a system that could compute data quickly in real time. However, the APIs were too slow and difficult to use. To address this issue, Gallium Studios brought vectorization in-house, but the escalating costs of GPU instances at scale prompted a search for alternatives.

Solution

Gallium Studios turned to Numenta and NuPIC to run LLMs on Intel processors. This transition eliminated the game's reliance on GPUs and the high costs associated with them while maintaining on-premise inferencing.

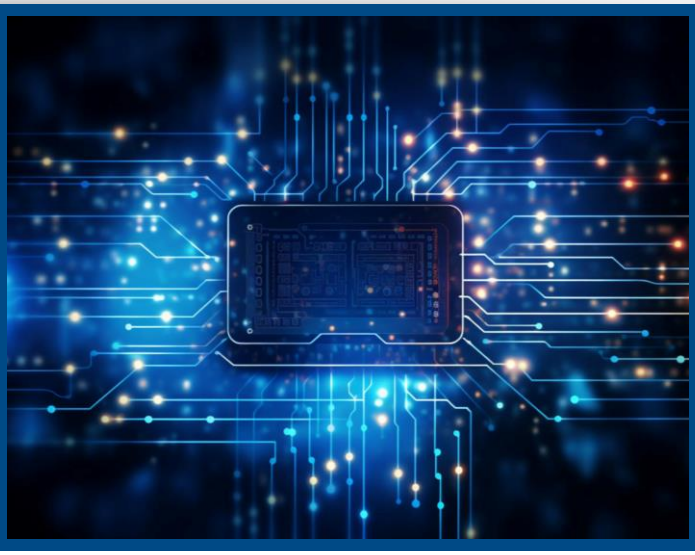
Results

Moving to NuPIC on Intel Xeon processors demonstrated an impressive 6.5X reduction in latency compared to the previous GPU-based system when creating each memory.[2] The move to NuPIC also contributed to enhanced user privacy, cost savings and facilitated simpler scalability of Intel architecture based on CPUs to support the game's growing virtual world.

In Summary

With NuPIC and Intel AMX, you get the versatility and security required by enterprise IT, combined with unparalleled scaling of LLMs on Intel architecture. This synergistic combination of software and hardware technologies makes running LLMs on CPUs more than a possibility – it becomes a strategic advantage.

Discover the power of CPU-based AI with NuPIC on Intel Xeon Scalable processors. Contact us for a free use case consultation to discuss how NuPIC can meet your business needs: https://numenta.com/get-started



- Numenta Website
- Ushering in a New Era of Accelerated AI on Intel® Processors
- Press Release: Numenta Achieves 123X Inference Performance Improvement for BERT Transformers on Intel® Xeon® Processor Family
- Intel® Xeon® Scalable Processors Product Page
- Intel® Advanced Matrix Extensions (Intel® AMX)

Sources

1. "Ushering in a New Era of Accelerated AI on Intel® CPUs." Intel, 2024. Link.
2. Embedding AI models results on NuPIC on 5th Gen Intel Xeon Scalable processor vs. NVIDIA A100: Intel® Xeon® CPU results were generated by Numenta on November 27, 2023 using a single-node, two-socket Intel® Xeon® Platinum 8592+ processor with 500 GB memory, Ubuntu 22.04 Kernel 5.15.0- 87, Numenta Platform for Intelligent Computing V1.0, NuPIC-Optimized BERT-Large, Sequence Length 64/128, BF16, and Batch Size 1. NVIDIA A100 results: https://github.com/NVIDIA/DeepLearningExamples/tree/master/TensorFlow/LanguageModeling/BERT#inference-performance-nvidia-dgx-a100-1x-a100-40gb
3. Generative AI models results on NuPIC on 4th Gen Intel Xeon Scalable processor were generated by Numenta on April 17th, 2024 using a single-node, two-socket Intel Xeon 8848C processor with 384 GB DDR5-4800, Ubuntu 22.04 Kernel 5.17, Numenta Platform for Intelligent Computing V1.1, 96 instances of Gemma 2B, Zephyr 7B, Mistral 7B, LLAMA2 7B, and NuPIC-Optimized GPT 7B.
4. Results with Numenta's BERT-Large model: Sequence Length 64, Batch Size 1, throughput optimized 4th Gen Intel Xeon. AMD Milan: Tested by Numenta as of 11/28/2022. 1-node, 2x AMD EPYC 7R13 on AWS m6a.48xlarge, 768 GB DDR4-3200, Ubuntu 20.04 Kernel 5.15, OpenVINO 2022.3, BERT-Large, Sequence Length 64, FP32, Batch Size 1. 3rd Gen Intel® Xeon® Scalable: Tested by Numenta as of 11/28/2022. 1-node, 2x Intel® Xeon®8375C on AWS m6i.32xlarge, 512 GB DDR4-3200, Ubuntu 20.04 Kernel 5.15, OpenVINO 2022.3, BERT-Large, Sequence Length 64, FP32, Batch Size 1. Intel® Xeon® 8480+: Tested by Numenta as of 11/28/2022. 1-node, pre-production platform with 2x Intel® Xeon® 8480+, 512 GB DDR5-4800, Ubuntu 22.04 Kernel 5.17, OpenVINO 2022.3, Numenta-Optimized BERT-Large, Sequence Length 64, BF16, Batch Size 1. [See P11 of Performance Index. Results may vary.] See also Intel: Numenta Delivers Powerful Inference Performance, 2023.
5. McQuate, Sarah. "Q&A: UW researcher discusses just how much energy ChatGPT uses". UW News, July 27, 2023. Link.
6. Power results: NuPIC-Optimized Bert-Large, seq_len 64 on 5th Gen Intel Xeon Scalable processor; estimated for Bert-Large seq_len 64 on NVIDIA A100 GPU Power consumption: estimations obtained from the Dell Tech Enterprise Infrastructure Planning Tool.

accelerated by intel

Accelerated by Intel® offerings take advantage of at least one Intel® technology, such as built-in accelerators, specialized software libraries, optimization tools, and others, to give you the best experience possible on Intel® hardware.

By taking advantage of acceleration technologies, such as Intel® Advanced Matrix Extensions (Intel® AMX), and others, our optimized solution helps accelerate time to innovation and insight.

With Intel technologies and capabilities, a vendor's optimized offering can go beyond the traditional compute and extend to accelerated networking, storage, edge, and cloud. It's all part of helping customers build an optimized infrastructure across the company.

Notices & Disclaimers

0724/CG/SR/361310-001US