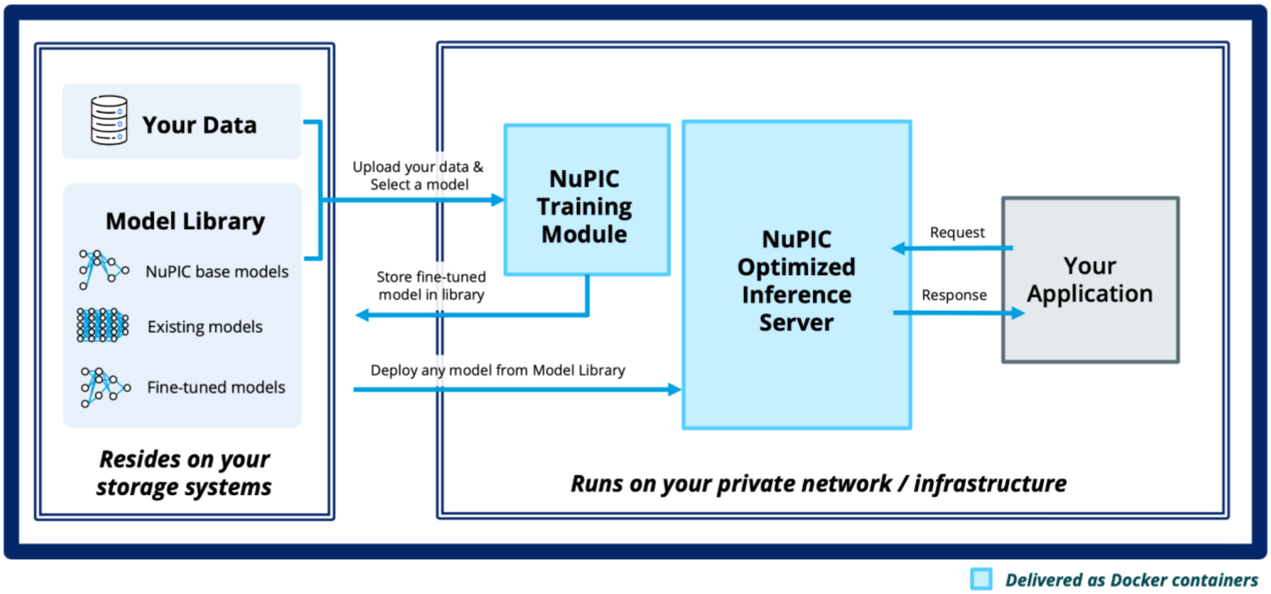
Numenta

Numenta has revolutionized AI with its groundbreaking developments, pushing the envelope with 10-100X cost and speed improvements across a wide range of applications, such as natural language processing. Built upon decades of neuroscience research, Numenta's unique, brain-based architectures and algorithms drive disruptive performance enhancements today and create a roadmap to truly intelligent computing.

NuPIC: unparalleled scaling of LLMs on CPUs. After beta engagements with numerous enterprises, NuPIC is now out of stealth. It is a robust solution that simplifies the scalable deployment of Large Language Models (LLMs). NuPIC brings higher throughput, lower latencies, and better accuracies.


Your Data
Model Library
NuPIC base models
Existing models
Fine-tuned models
Resides on your storage systems
Upload your data & Select a model
Store fine-tuned model in library
Deploy any model from Model Library
NuPIC Training Module
NuPIC Optimized Inference Server
Request
Response
Your Application
Runs on your private network / infrastructure
Delivered as Docker containers

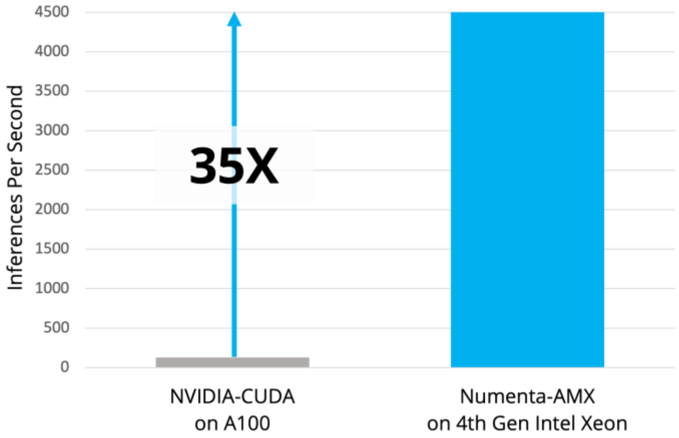
ORACLE Cloud Infrastructure · aws · Azure · Google Cloud · CoreWeave

Deployed as a Docker container, NuPIC operates entirely within a customer’s infrastructure, ensuring **models and data remain private** and fully under customer control. It runs on any cloud provider or on-premise and integrates into existing MLOps tools as a lightweight containerized solution.

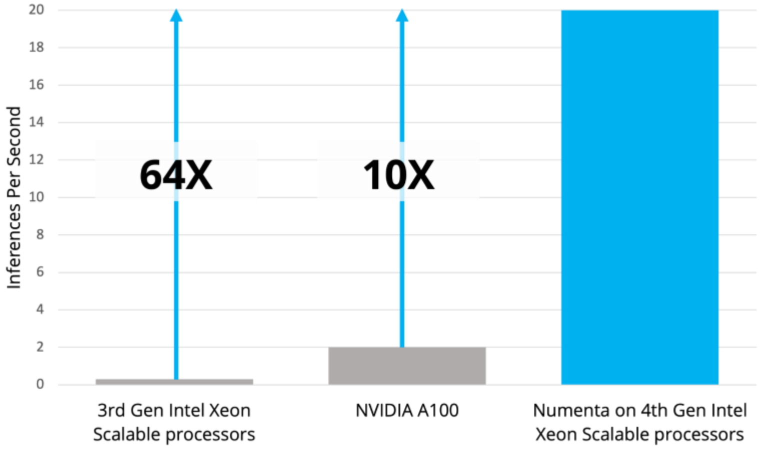
Results: In partnership with Intel, we implemented our technology on Intel's new 4th Gen Xeon servers. Combined with Intel's AMX instructions, NuPIC delivers **35X improvement in BERT-large throughput** compared to NVIDIA's A100 GPUs.

Unbeatable price / performance on CPUs

Throughput acceleration of BERT-large enabled by Numenta's technology in comparison to GPU technologies.



Numenta delivers a remarkable **64X improvement in GPT throughput** compared to current CPUs, and surpasses NVIDIA's A100 GPUs by **more than 10X**. Now customers can **deploy GPT models on CPUs,** substantially reducing cost and complexity.



Industry Leading GPT Acceleration

Throughput acceleration of GPT-J-6B enabled by Numenta's technology in comparison to other CPU and GPU technologies on the market.

Learn more:
numenta.com

Request a Demo:
numenta.com/demo

Why Numenta?	
Cost	Eliminate need for complex, expensive GPUs
Speed	Consistently high throughput and low latency
Stability	Model updates fully under your control
Scalability	Seamlessly integrate into standard MLOps tools
Accuracy	Control accuracy vs. speed, increased accuracy with custom fine-tuning
Training	Fine tune unlimited models on custom data
Data privacy	Maintain complete control over your data
SOTA	Get the latest cutting-edge AI technology