Numenta and Intel Deliver Cost-Effective, Powerful Inference Performance

The Challenge

High throughput, low latency technologies are required for a wide range of natural language processing (NLP) applications, from text classification to sentiment analysis to ChatGPT. Transformer models like BERT and GPT have become fundamental within the space because of their ability to handle complex text input and deliver accurate results. But as these large language models continue to grow in size and complexity, so do the costs of running them, making it increasingly difficult to deploy transformers in production.

To meet the rigorous throughput and latency demands today's sophisticated NLP applications require, customers typically resort to running transformers on GPUs (Graphics Processing Units) for deep learning inference, as GPUs are highly parallelized and can perform large-scale computations quickly and efficiently. However, GPUs can be more costly, and adding extra hardware can mean more maintenance for IT departments. Customers need a simpler solution that delivers superior performance benefits without breaking the bank.

Numenta Solution

Numenta has created a powerful AI platform based on its two decades of neuroscience research and breakthrough advances in AI technology, enabling customers to achieve from 10x to over 100x performance improvements in deep learning inference.[1,2,3]

In collaboration with Intel, Numenta combined their proprietary, neuroscience-based solution for dramatically accelerating transformer networks with the new Intel® Advanced Matrix Extensions (Intel® AMX) available in 4th Gen Intel® Xeon® Scalable processors.

A New Era of AI Computing with the CPU

ChatGPT has shown the world the power of transformers, and the demand for large deep-learning models continues to grow. As we look ahead to the many incredible possibilities, Numenta's results suggest a new era of deep learning with optimized models and CPUs. With the cost-effective, highly performant combination of Numenta solutions and Intel® CPUs, customers can get the high throughput and low latency inference results that their most sophisticated, highly complex NLP applications require.[3]

Numenta's dramatic acceleration of transformer networks on 4th Gen Intel® Xeon® Scalable processors brings several benefits, including:

- Avoiding the cost and complexity associated with GPUs for deep learning inference
- Allowing for more flexible and scalable deployment of transformer models
- Unlocking new possibilities for AI and NLP applications that can finally deploy transformer models in production

> "These breakthrough results make CPUs the best option for running transformers. Customers with performance-sensitive AI applications can use the combination of Numenta and 4th Gen Intel® Xeon® Scalable processors to deploy their real-time applications in a lightweight, cost-effective manner."
>
> **Subutai Ahmad**
> CEO of Numenta

Better Performance: Numenta on Intel® CPUs versus NVIDIA GPUs

Taking full advantage of Intel® AMX, Numenta observed a 35x throughput improvement versus NVIDIA A100 GPUs for BERT-Large inference on short text sequences and batch size 1.[2] Batch size 1 is ideal for low-latency applications as it provides the most flexibility in real-time scenarios where input data constantly changes.

GPUs typically perform better with higher batch sizes, but even with a batch size of 8 for NVIDIA A100, Numenta outperforms by 9x.[2]

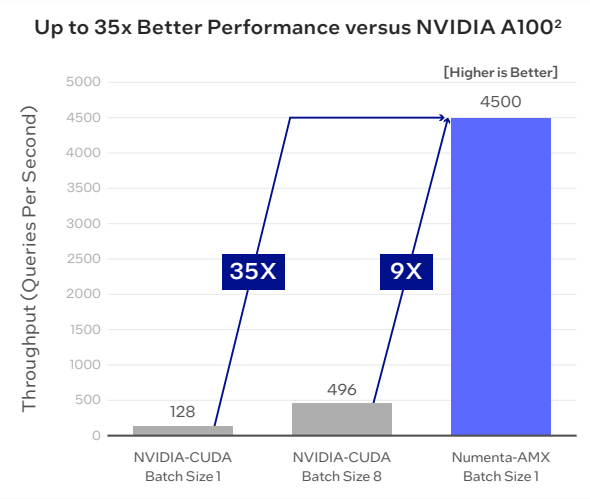


Figure 1: Inference throughput observed with Numenta's optimized BERT-Large model running on a 2-socket 4th Gen Intel® Xeon® Scalable processor, compared with standard BERT-Large models running on an NVIDIA A100 Tensor Core GPU.

This synergistic combination of algorithm and hardware advances led to unparalleled performance gains for BERT-Large inference on CPUs. With this dramatic performance acceleration, CPUs become not only a viable alternative to GPUs, but an ideal option for running transformers in production.

Turbo-charging CPU Inference Throughput

The combination of Numenta and Intel® technology has multiplicative effects. To tease out the effect of each component, Numenta broke it down by looking at the tradeoffs between throughput and latency in two different scenarios.

When optimizing for throughput, Numenta delivers more than 5,100 queries per second, which is a 70x throughput improvement versus current generation AMD Milan CPU implementations.[3] But how much of that improvement comes from Numenta? In this scenario, moving from 3rd Gen Intel® Xeon® Scalable processors to 4th Gen Intel® Xeon® Scalable processors without Numenta yields a 6.5x speedup.[3] Numenta adds an additional 9x throughput boost.[3]

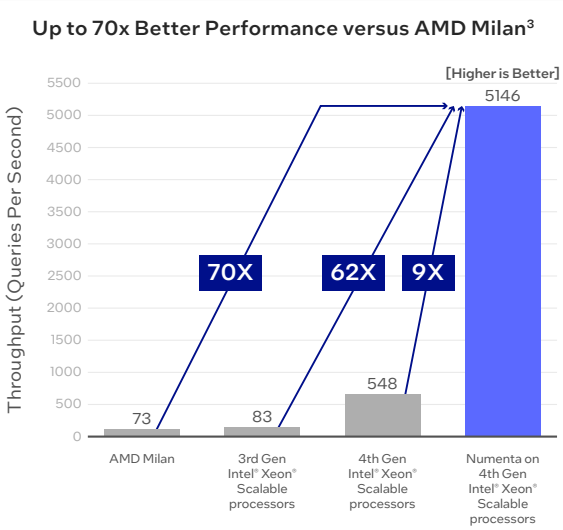


Figure 2: Inference throughput improvements observed when optimizing for throughput with Numenta's optimized BERT-Large model running on the latest 4th Gen Intel® Xeon® Scalable processors, compared to standard BERT-Large models running on a variety of other processor architectures.

If you're interested in leveraging these dramatic performance improvements within your deep learning networks and making GPU-less AI a reality for your organization, request a demo at www.numenta.com/demo

In situations where real-time applications have strict latency requirements, the objective changes: achieving peak throughput *while* respecting the minimum latency requirements. To demonstrate, Numenta imposed a 10ms latency restriction, which is often considered a key threshold for real-time applications.

As seen in the chart below, with a 10ms latency target, Numenta on 4th Gen Intel® Xeon® Scalable processors achieves 123x throughput improvement versus current generation AMD Milan CPU implementations.[3] In this scenario, Numenta's contribution is even more pronounced. Moving from 3rd Gen Intel® Xeon® Scalable processors to 4th Gen Intel® Xeon® Scalable processors without Numenta gives Intel a roughly 3x speedup.[3] Numenta gives an additional 19x speedup on top of the 4th Gen Intel® Xeon® processor acceleration.[3]

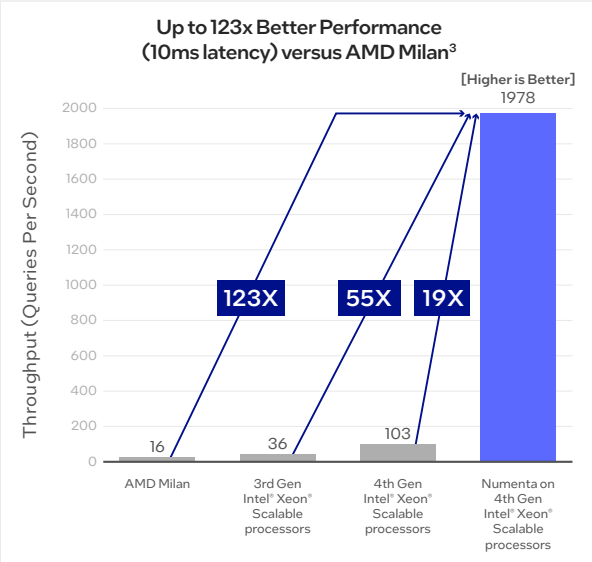


Figure 3: Inference throughput improvements observed with a 10ms maximum latency using Numenta's optimized BERT-Large model running on the latest 4th Gen Intel® Xeon® Scalable processors, compared with standard BERT-Large models running on a variety of other processor architectures.

Why Intel for Numenta?

"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

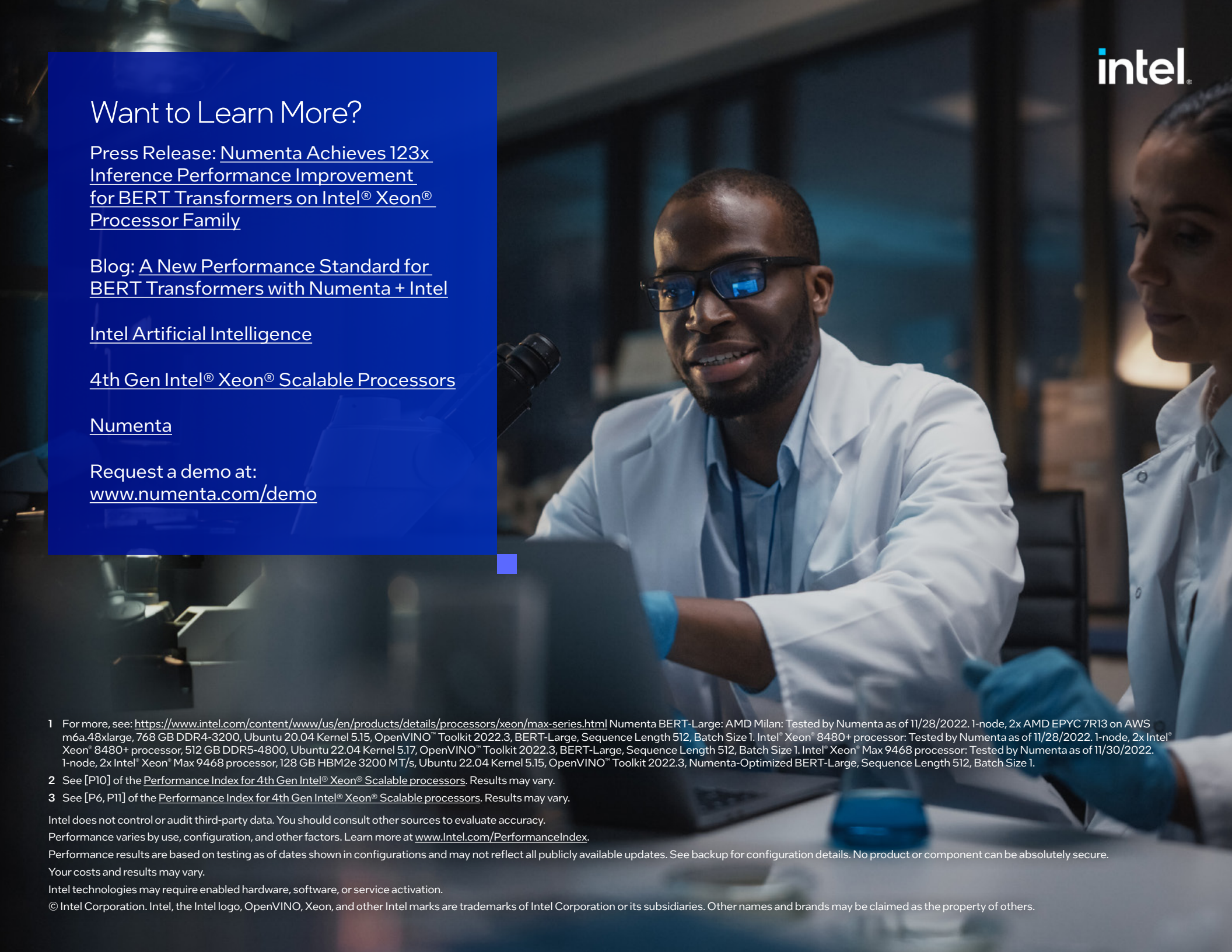
Vice President and General Manager of AI and HPC Application-Level Engineering, Intel

Numenta's AI technology also dramatically accelerates NLP applications that rely on analyzing extensive collections of documents.[1] For example, when applying transformers to document understanding, long sequence lengths are required to incorporate the full context of the document. These long sequences require high data transfer rates, and off-chip bandwidth thus becomes the limiting factor. Using the new Intel® Xeon® CPU Max Series, Numenta demonstrates it can optimize the BERT-Large model to process large text documents, enabling unparalleled 20x throughput speedup for long sequence lengths of 512.[1] This type of innovation is absolutely transformative for Numenta customers, enabling cost-efficient scaling for the first time.

About Numenta

Founded in 2005 by computer industry pioneers Jeff Hawkins and Donna Dubinsky, Numenta has two decades of research deriving proprietary technology from neuroscience. Leveraging the fundamental insights from its neuroscience research, Numenta has developed a cutting-edge AI platform that delivers disruptive performance improvements across broad AI use cases.

Numenta is engaged with several Global 100 companies to apply its platform technology across the full spectrum of AI, from model development to deployment – and ultimately enable whole new categories of applications.



Want to Learn More?

Press Release: Numenta Achieves 123x Inference Performance Improvement for BERT Transformers on Intel® Xeon® Processor Family

Blog: A New Performance Standard for BERT Transformers with Numenta + Intel

Intel Artificial Intelligence

4th Gen Intel® Xeon® Scalable Processors

Numenta

Request a demo at:
www.numenta.com/demo

1 For more, see: https://www.intel.com/content/www/us/en/products/details/processors/xeon/max-series.html Numenta BERT-Large: 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™ Toolkit 2022.3, BERT-Large, Sequence Length 512, Batch Size 1. Intel® Xeon® 8480+ processor: Tested by Numenta as of 11/28/2022. 1-node, 2x Intel® Xeon® 8480+ processor, 512 GB DDR5-4800, Ubuntu 22.04 Kernel 5.17, OpenVINO™ Toolkit 2022.3, BERT-Large, Sequence Length 512, Batch Size 1. Intel® Xeon® Max 9468 processor: Tested by Numenta as of 11/30/2022. 1-node, 2x Intel® Xeon® Max 9468 processor, 128 GB HBM2e 3200 MT/s, Ubuntu 22.04 Kernel 5.15, OpenVINO™ Toolkit 2022.3, Numenta-Optimized BERT-Large, Sequence Length 512, Batch Size 1.

2 See [P10] of the Performance Index for 4th Gen Intel® Xeon® Scalable processors. Results may vary.

3 See [P6, P11] of the Performance Index for 4th Gen Intel® Xeon® Scalable processors. Results may vary.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Performance varies by use, configuration, and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software, or service activation.