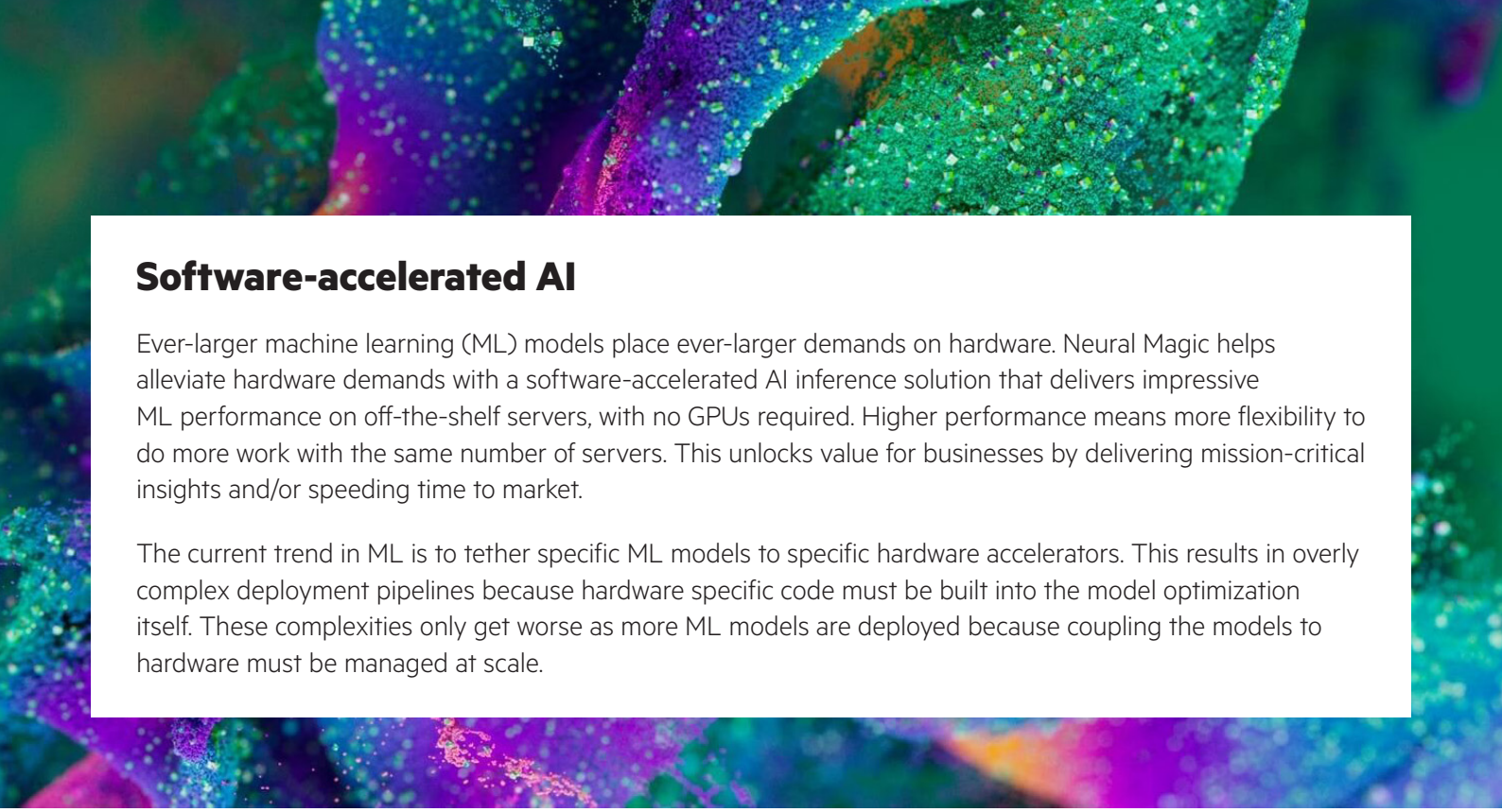
Hewlett Packard Enterprise | AMD

Unleash software-accelerated AI inference with Neural Magic on HPE ProLiant Gen11 servers powered by 4th Gen AMD EPYC processors

Unlock incredible levels of performance for AI inference workloads



Software-accelerated AI

Ever-larger machine learning (ML) models place ever-larger demands on hardware. Neural Magic helps alleviate hardware demands with a software-accelerated AI inference solution that delivers impressive ML performance on off-the-shelf servers, with no GPUs required. Higher performance means more flexibility to do more work with the same number of servers. This unlocks value for businesses by delivering mission-critical insights and/or speeding time to market.

The current trend in ML is to tether specific ML models to specific hardware accelerators. This results in overly complex deployment pipelines because hardware specific code must be built into the model optimization itself. These complexities only get worse as more ML models are deployed because coupling the models to hardware must be managed at scale.

Neural Magic

Neural Magic is on a mission to help customers unlock the full potential of their ML environment to accommodate the continuous growth of neural networks without added complexity or cost. Their products simplify ML deployments, so customers can use compute-heavy models in a cost-efficient and scalable way on existing CPU infrastructure.

Next-gen compute with leading edge technologies

HPE ProLiant Gen11 and 4th Gen AMD EPYC™ processors leverage leading-edge technologies, such as DDR5 DIMMs, PCIe® Gen5 I/O, CXL™ 1.1+ memory expansion, and AVX-512 and VNNI instruction extension support, to deliver generational Neural Magic performance uplifts of ~2.78x (pruned oBERT-Base) to ~3.89x (pruned YOLO v5s) compared to 3rd Gen AMD EPYC processors.

Comparing the performance of sparse quantized Neural Magic models to that of baseline dense models shows uplifts of ~2.57x (YOLO v5s) to ~6.66x (oBERT-Base) and ~3.50x (YOLO v5s) to ~9.19x (oBERT-Base) on 3rd Gen AMD EPYC and 4th Gen AMD EPYC processors, respectively.

HPE ProLiant: Compute engineered for your hybrid world

HPE ProLiant completes your hybrid environment wherever it lives—combining a cloud operating experience and built-in security while driving next-gen performance with the engineering leadership to power insights, innovation, and competitive advantage to drive your business forward.

Expect more from your infrastructure

Future-proof your business with HPE ProLiant Compute that delivers the efficiency, scalability, and economics to accelerate business outcomes while lowering TCO. Consolidate more workloads and increase ROI with breakthrough performance of next-gen compute.

An open approach to your most demanding workloads

Grow your business at scale with an open approach to industry standards for innovation of advanced technology in the data center and at the edge, and flexible management of cloud-native workloads.

What is BERT-Base?

Bidirectional Encoder Representations from Transformers (BERT) is a family of ML models for natural language processing (NLP). BERT helps computers create the context needed to understand ambiguous text by examining surrounding text. The pretrained BERT model referenced in this brief is trained on text from Wikipedia. Developers can further tune BERT using data sets for specific use cases, such as question and answer data sets like Stanford Question Answering Dataset (SQuAD).

Neural Magic has further optimized BERT by pretraining it and fine tuning it on user data for specific tasks. The result is oBERT-Base, which offers best-in-class performance and faster time to value when deployed in production. For example, retailers can utilize a chatbot to engage with customers for automated question answering and support. It also has the power to enable a better understanding of customer sentiment, enabling the answering of questions such as: what is the tone across social media channels, or how should requests be triaged according to urgency? You can also glean data insights from customer relationship management (CRM) systems in production environments.

What is YOLO?

You only look once (YOLO) is a convolutional neural network for accurate, real-time object detection using computer vision. A single neural network breaks a picture into parts, predicts bounding boxes, and assigns probabilities to each component based on a single run through the neural network.

YOLO is a fantastic enabler of production scenarios for in-store heat mapping, tracking foot traffic patterns, counting people, or even detecting when shelves need restocking. Delivering these applications on HPE ProLiant DL385 Gen11 Servers with AMD EPYC processors with Neural Magic software delivers superb performance and operational efficiency. For example, real-time video streaming can occur with frame rates of 30 frames per second when processing up to 70 camera streams on a single server.

System configuration and raw test results

The following tables show the system configurations used for these tests and the raw results. Each test was performed three times. The average of these three tests was then calculated to yield the final results displayed in the graphs.

System 1: 3rd Generation AMD EPYC processor

Table 1. 3rd Generation AMD EPYC system

CPU	2 x AMD EPYC 7763
Cores	64 (128 threads) per CPU
Base frequency	2.45 GHz
L3	256 MB
Socket	SP3
Memory	1.05 TB
Settings	SMT enabled
OS	Ubuntu® 20.04.2 LTS
Test results (Dense YOLO v5s 57.1% accuracy with mAP@0.50)	Latency < 33 ms with 7 streams Test 1: 32.825 Test 2: 32.9561 Test 3: 32.9866 Average: ~32.9226 ms
Test results (Pruned YOLO v5s 52.5% accuracy with mAP@0.50)	Latency < 33 ms with 18 streams Test 1: 32.8442 Test 2: 32.7946 Test 3: 32.7063 Average: ~32.7817 ms
Test results (Dense oBERT-Base 95.4% accuracy)	Queries/second Batch size: 256 Test 1: 306.3383 Test 2: 306.2958 Test 3: 306.3180 Average: ~306.3170
Test results (90% Pruned oBERT-Base, 93.7% accuracy)	Queries/second Batch size: 256 Test 1: 2032.7645 Test 2: 2039.6417 Test 3: 2052.2516 Average: ~2041.5526

System 2: 4th Generation AMD EPYC processor

Table 2. 4th Generation AMD EPYC system

CPU	2 x AMD EPYC 9654
Cores	96 (192 threads) per CPU
Base frequency	2.40 GHz
L3	384 MB
Socket	SP5
Memory	1.58 TB
Settings	SMT enabled
OS	Ubuntu 20.04.2 LTS
Test results (Dense YOLO v5s, 57.1% accuracy with mAP@0.50)	Latency < 33 ms with 20 streams Test 1: 32.9591 Test 2: 32.7182 Test 3: 33.1120 Average: ~32.9298 ms
Test results (94% Pruned YOLO v5s, 52.5% accuracy with mAP@0.50)	Latency < 33 ms with 70 streams Test 1: 32.7843 Test 2: 32.6904 Test 3: 32.6253 Average: ~32.7000 ms
Test results (Dense oBERT-Base, 95.4% accuracy)	Queries/second Batch size: 256 Test 1: 616.6593 Test 2: 616.4927 Test 3: 616.9794 Average: ~616.3771
Test results (90% Pruned oBERT-Base, 93.7% accuracy)	Queries/second Batch size: 256 Test 1: 5670.4560 Test 2: 5668.9001 Test 3: 5660.8382 Average: ~5666.7300

Generational performance uplifts

The graphs in this section illustrate the generational performance uplifts of the HPE ProLiant DL385 Gen11 Server with dual 96-core 4th Generation AMD EPYC 9654 processors over a prior-generation system powered by dual 64-core 3rd Generation AMD EPYC 7763 processors. The results in these graphs are normalized such that the 3rd Generation AMD EPYC results always equal 1.00x. The same dense and Neural Magic sparse quantized models are benchmarked on the same machines for each comparison. For all comparisons in this brief, higher uplift implies better performance.

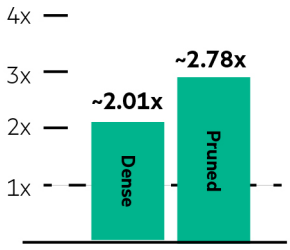
oBERT-Base generational uplifts

HPE ProLiant Gen11 servers with 4th Gen AMD EPYC 9654 processors show a ~2.01x uplift during dense (nonoptimized) testing and ~2.78x uplift on pruned (Neural Magic–optimized) testing over 3rd Generation AMD EPYC 7763 processors.

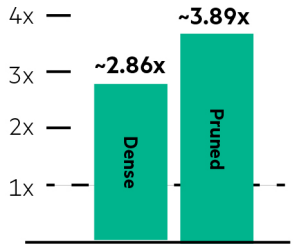
YOLO v5s generational uplifts

HPE ProLiant Gen11 servers with 4th Generation AMD EPYC 9654 processors show a ~2.86x uplift during dense (nonoptimized) testing and ~3.89x uplift on pruned (Neural Magic–optimized) testing over 3rd Generation AMD EPYC 7763 processors.

oBERT-Base Throughput
Batch Size=512
AMD EPYC 9654 vs. AMD EPYC 7763
Normalized to 3rd Gen AMD EPYC



YOLO v5s Concurrent Streams
Batch Size=1
AMD EPYC 9654 vs. AMD EPYC 7763
Normalized to 3rd Gen AMD EPYC



Neural Magic performance uplifts

The graphs in this section illustrate the performance uplifts offered by Neural Magic compared to running the same workload in dense, nonoptimized states.

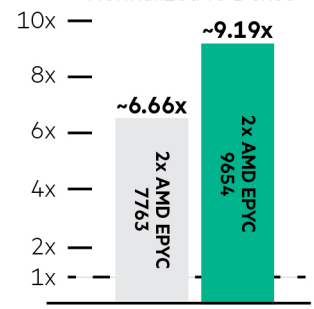
oBERT-Base Neural Magic uplifts

Using Neural Magic pruning and quantization improved performance by ~6.66x on the 3rd Generation AMD EPYC system and by a stunning ~9.19x on the HPE ProLiant DL385 Gen11 Server with 4th Generation AMD EPYC 9654 processors for oBERT-Base.

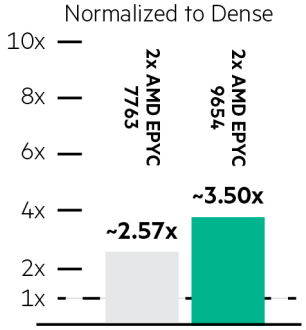
YOLO v5s Neural Magic uplifts

Using Neural Magic pruning and quantization improved performance by ~2.57x on the 3rd Generation AMD EPYC system featuring two 3rd Generation AMD EPYC 7763 processors and by an impressive ~3.50x on the HPE ProLiant DL385 Gen11 Server powered by dual 4th Generation AMD EPYC 9654 processors for YOLO v5s.

oBERT-Base
Dense vs. Pruned
AMD EPYC 9654 and AMD EPYC 7763
Normalized to Dense



YOLO v5s
Dense vs. Pruned
AMD EPYC 9654 and AMD EPYC 7763
Normalized to Dense



Key insights

HPE ProLiant DL385 Gen11 Server systems with 4th Generation AMD EPYC processors deliver robust performance gains versus comparable systems powered by 3rd Generation AMD EPYC processors running oBERT-Base and YOLO v5s. Pruning and optimizing these workloads with Neural Magic enabled performance uplifts from ~2.57x to ~9.19x compared to not using Neural Magic's sparse quantized models. Combining the workload pruning and quantization of Neural Magic with the higher core counts, per-core performance uplifts, and innovations found in 4th Generation AMD EPYC processors is a powerful way to harness the power of your ML workloads.

Top advantages for choosing HPE ProLiant Gen11 and AMD EPYC processor for optimized inference with Neural Magic

- Best-in-class performance for state-of-the-art ML inference on CPU infrastructure without the need for hardware accelerators, leveraging the same systems and processes as your typical CPU-based workload applications.
- HPE ProLiant Gen11 servers with AMD EPYC processors deliver 25% power savings while using 43% less rack space for the same performance envelope (compared to the previous generation)[1].
- Protect your infrastructure, workloads, and data from threats to hardware and risks from third-party software with a trusted edge-to-cloud security posture built on an HPE compute core hardened through a proven, zero trust approach to security.
- Unify compute management with a centralized console for self-serve operations and pivot IT resources from reactive to proactive with global visibility and insight of all compute devices with HPE GreenLake for Compute Ops Management.
- Securely bring cloud agility to distributed compute infrastructure with cloud-based management, meaning it's up to date with the latest features, security patches, and firmware versions.
- 4th Generation AMD EPYC processors introduce new instruction set extensions for enhanced AI capability, including AVX-512 VNNI.
- 4th Generation AMD EPYC processors now support PCIe Gen5 at up to 32 Gbps and up to 12 memory channels leveraging up to 6 TB of DDR5-4800 memory, alleviating the memory constraint for AI workloads.
- Other hardware features in 4th Gen EPYC processors include 1 MB of L2 cache vs. 512 KB in Zen 3, as well as up to four links of Gen3 Infinity Fabric™ at up to 32 Gbps.
- Have the freedom to run AI inference workloads anywhere with Neural Magic, from on-premises, to the edge, to the cloud.
- Extend additional operational efficiency to your ML solutions with horizontal and vertical scale on physical, virtual, containerized, and serverless deployment options with Neural Magic.

Your choice of compute matters

A new approach is needed to thrive in a cloud-native and data-first modernization era. You need a platform that can deploy and manage container clusters easily from edge to cloud with frictionless data access and right security measures to protect your organization. This is the foundation to successfully modernize your cloud-native environment to ease the process to modernize legacy applications, avoid siloed infrastructure, and prevent vendor lock-in. To realize that value and address the challenges, you need compute that powers the underlying infrastructure to deploy your cloud-native applications and workloads. And the right choice of compute—one that delivers a cloud operating experience built from the ground up with a fundamental foundation security approach—can set you apart from the competition.

[1] SPEC and the name SPECpower_ssj are registered trademarks of the Standard Performance Evaluation Corporation (SPEC). The stated results (#1169 and 1185) are published as of 2-16-23; see spec.org. All rights reserved.

HPE GreenLake for Compute Ops Management

Try it for free for 90 days.

Enable IT to easily monitor, manage, and update servers remotely through a cloud-based console—anywhere, anytime.

Start your trial today

Neural Magic DeepSparse for AI inference on CPUs

Learn more about Neural Magic's inferencing solution, DeepSparse, and start your free trial today.

Next steps

HPE ProLiant Gen11 servers with AMD EPYC processors and Neural Magic can bring best-in-class ML performance together with the flexibility, scalability, and efficiency to fast forward your business objectives.

Learn more at

HPE ProLiant solutions

HPE ProLiant servers

HPE and AMD solutions

Neural Magic

Chat now (sales)

a00131715ENW

Hewlett Packard Enterprise