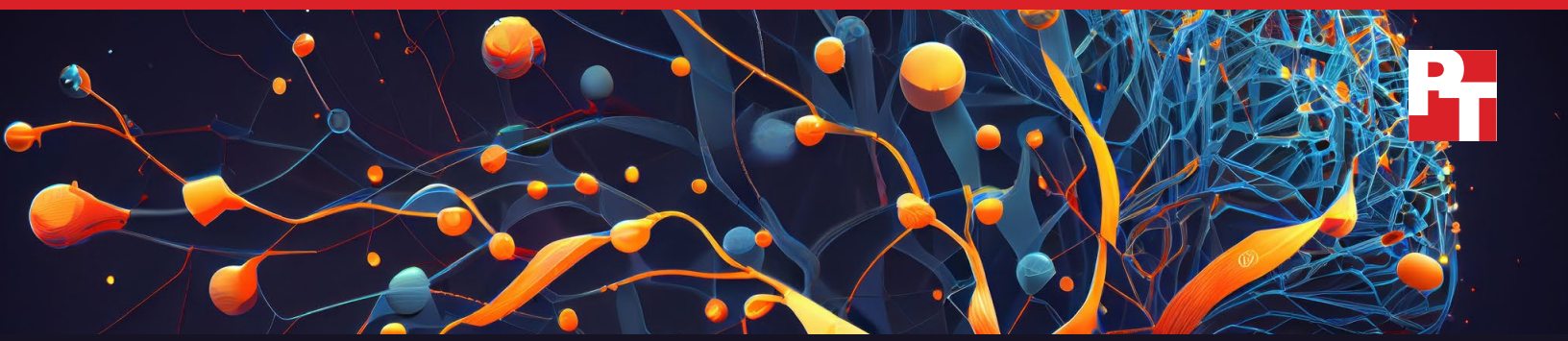


Improve AI inference performance with HPE ProLiant DL380 Gen11 servers, powered by 4th Generation Intel Xeon Gold processors

In ResNet-50 image-recognition testing, these servers handled significantly more samples per second than previous-generation HPE ProLiant servers while achieving lower latency

Companies across a wide range of industries are embracing AI inference applications. From healthcare to retail to manufacturing, these applications are solving business problems and providing functionality that would have been hard for most of us to imagine a few decades ago. Because machine learning is computationally demanding, decision makers might assume they must engage costly cloud solutions for these workloads. Or, if they want to run the applications in their own data centers, they might think it's necessary to invest in expensive servers backed by graphics processing units (GPUs). In fact, many organizations can meet the requirements of their AI applications on site with powerful HPE ProLiant DL380 Gen11 servers featuring 4th Generation Intel® Xeon® Gold processors. These processors include accelerators such as Intel® Advanced Matrix Extensions (AMX) to improve AI performance.

In the PT data center, we conducted a series of ResNet-50 tests to quantify the AI inference performance potential of these servers. First, we tested the Floating Point 32-bit (FP32) inference performance of the HPE ProLiant DL380 Gen11 server and compared it to that of its previous-generation counterpart with 3rd Generation Xeon processors. Next, we investigated performance of the new server with lower precision levels, which leverages the Intel AMX accelerator[1] in the 4th Generation Intel Gold processor to increase throughput.

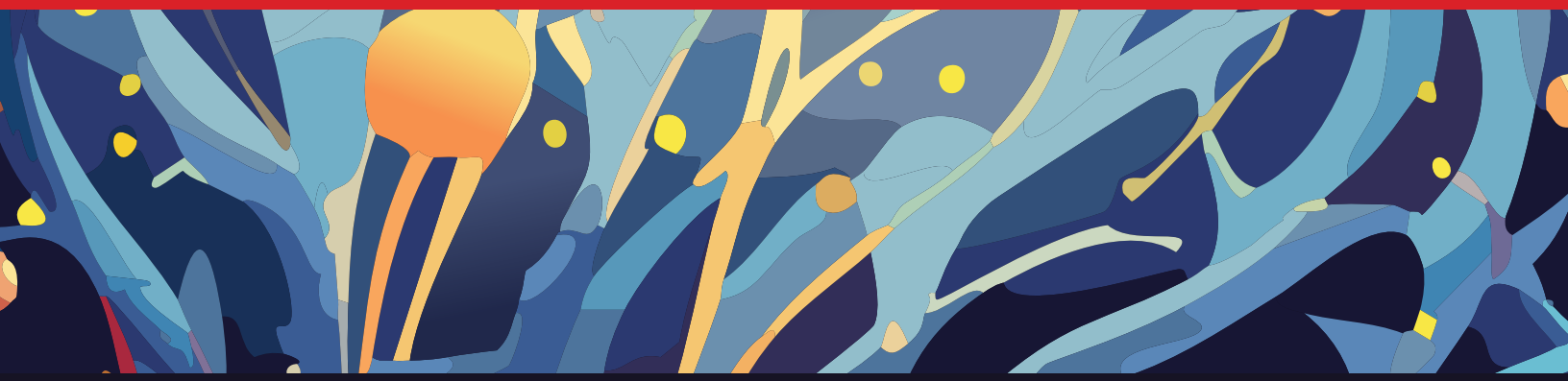
With an FP32 precision level, the HPE ProLiant DL380 Gen11 server processed 2.86 times as many images per second as the older server did. The HPE ProLiant DL380 Gen11 server also delivered strong performance at two additional precision levels that benefit from the new Intel AMX accelerators found in 4th Generation Intel Scalable processor. Based on our overall findings, the HPE ProLiant DL380 Gen11 server is an excellent candidate for running a wide range of image-recognition workloads for many organizations.

Process
2.86x
as many images per second at FP32 precision levels*

Reduce latency by
30.1%
at FP32 precision levels*

Enjoy
strong performance
at Int8 and bfloat16 precision levels

*HPE ProLiant DL380 Gen11 server featuring Intel Xeon Gold 6430 processors vs. HPE ProLiant DL380 Gen10 server featuring Intel Xeon Gold 6130 processors



Our test approach

We used the following two servers in testing:

HPE ProLiant DL380 Gen10 server

- Two Intel Xeon Gold 6130 processors
- 256 GB of RAM
- 1TB of SSD storage
- One dual-port 10GbE NIC

HPE ProLiant DL380 Gen11 server

- Two Intel Xeon Gold 6430 processors
- 256 GB of RAM
- 1TB of SSD storage
- One dual-port 10GbE NIC

We set up a bare-metal Linux® environment on each server. We captured ResNet-50 v1.5 performance (inference throughput) using the Intel Reference AI model, the Floating Point 32-bit level of inference precision, and a batch size of 116. The benchmark measured performance in terms of the number of images per second each server could process and latency. We conducted each test three times and report the median score. For complete details on our server configurations and step-by-step test methodologies, please see the **science behind the report**.

In addition to these comparisons, we tested two levels of inference precision, integer 8 (Int8) with a batch size of 116 and Brain Floating Point 16-bit (bfloat16) with a batch size of 80, on only the HPE ProLiant DL380 Gen11 server. These lower precision levels are appropriate for use cases with higher throughput demands and lower accuracy requirements than FP32. The processor in the older HPE ProLiant DL380 Gen10 server does not have the Intel® Advanced Matrix Extensions (AMX) extensions that accelerate these precision levels.

About the HPE ProLiant DL380 Gen11 server

According to HPE, the dual-socket 2U ProLiant DL380 Gen11 "delivers exceptional compute performance, expandability, and scalability for diverse workloads and environments at 1P economics."[2] The server is powered by powered by 4th and 5th Generation Intel® Xeon® Scalable Processors with up to 64 cores, has increased memory bandwidth, and offers high-speed PCIe Gen5 I/O. HPE says the server is particularly well-suited to "data-intensive workloads like software-defined storage, video transcoding, and virtualized apps that require large storage capacity, and high I/O and memory bandwidth."[3]

Learn more at **https://buy.hpe.com/us/en/compute/rack-servers/proliant-dl300-servers/proliant-dl380-server/hpe-proliant-dl380-gen11/p/1014696069**.

About 4th Generation Intel Xeon Gold 6430 processors

The HPE ProLiant DL380 Gen11 server we tested features Intel Xeon Gold 6430 processors, part of the 4th Generation Intel Xeon Scalable processor family. According to Intel, its strategy for these processors aligns processor "cores with built-in accelerators optimized for specific workloads and delivers increased performance at higher efficiency for optimal total cost of ownership."[4]

The processors deliver "a range of features for managing power and performance, making the best use of resources to achieve key sustainability goals. In addition, the Xeon CPU Max and the Max Series GPU add high-bandwidth memory and maximum compute density to solve the world's most challenging problems faster."[5]

Learn more at **https://www.intel.com/content/www/us/en/newsroom/resources/press-kit-4th-gen-intel-xeon-scalable-processors.html**.

What we learned

Figure 1 shows the number of images per second each server processed. The HPE ProLiant DL380 Gen11 server with Intel Xeon Gold 6430 processors handled 2.86 times as many images per second as the older server, which means it could do more work with a given number of servers or could perform a fixed amount of inference work using fewer servers, which has the potential to lead to savings.

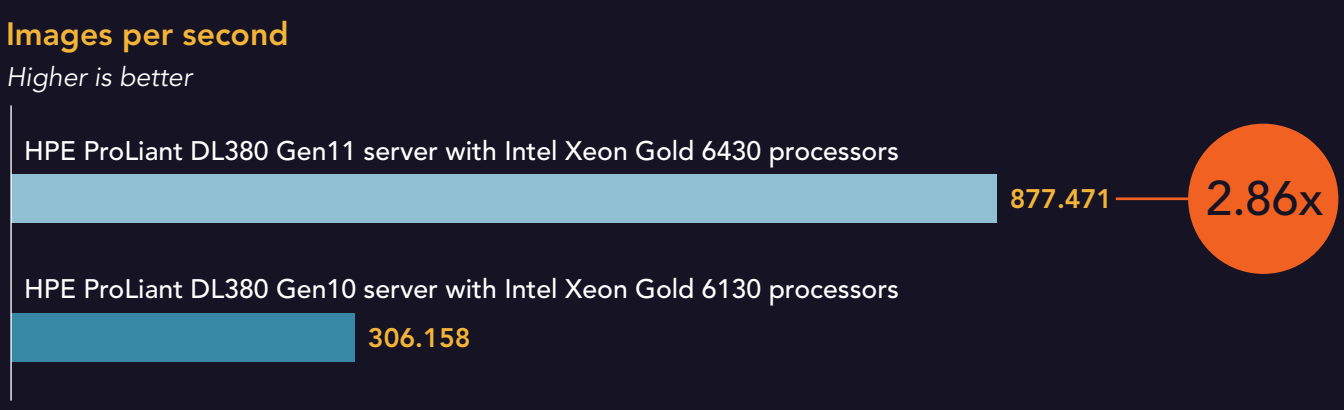


Figure 1: The number of images per second each server processed on the ResNet-50 v1.5 model with FP32 precision. Higher is better. Source: Principled Technologies.

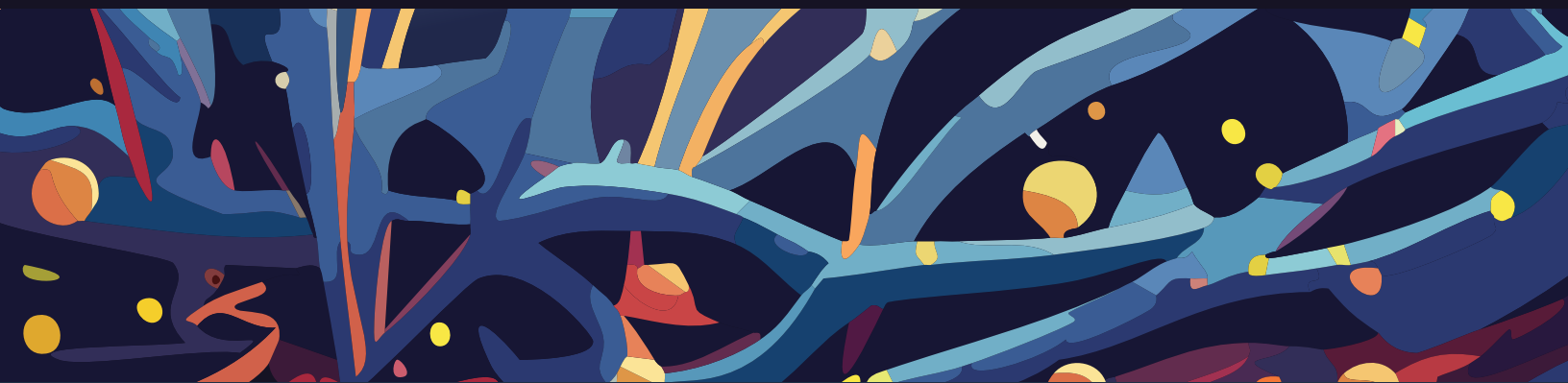


Figure 2 shows the latency each server achieved during testing. The HPE ProLiant DL380 Gen11 server with Intel Xeon Gold 6430 processors achieved 30.1 percent lower latency than the older server, which means it executed inference work more quickly.

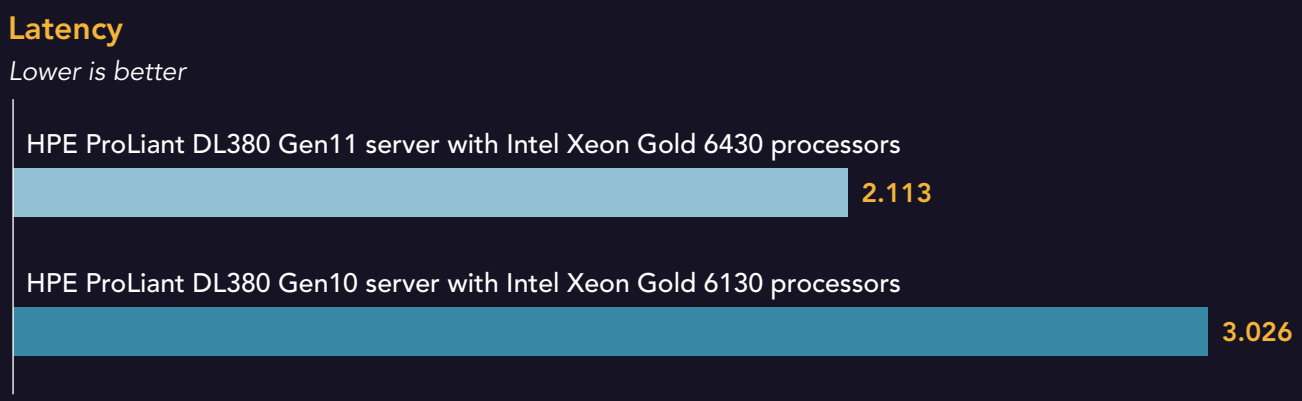


Figure 2: The latency in seconds each server delivered on the ResNet-50 v1.5 model with FP32 precision. Lower is better.
Source: Principled Technologies.

About ResNet-50

ResNet-50 is a model that organizations use for image classification, or the process of correctly identifying objects in images. Strong image classification performance may be vital for use cases in safety and security, retail, healthcare, manufacturing, and other markets. In our testing, we used the ResNet-50 v1.5 inference implementation from the Intel® AI Reference Models repository, which uses the ResNet-50 model and measures the number of (inference) samples per second a solution processes.

Learn more at **https://github.com/IntelAI/models/tree/master/quickstart/image_recognition/tensorflow/resnet50v1_5/inference/cpu**.

About the Intel® AI Reference Models repository

According to Intel, its AI Reference Models repository contains "links to pre-trained models, sample scripts, best practices, and step-by-step tutorials for many popular open-source machine learning models optimized by Intel to run on Intel® Xeon® Scalable processors and Intel® Data Center GPUs."[6]

Learn more at **https://github.com/IntelAI/models**.

What this could mean for your company

The potential business applications of image recognition AI are limitless, but to put our test results into context, let's look at a few industries.

Manufacturing

Image recognition use cases in manufacturing range from quality inspection to equipment maintenance to worker safety. Computer vision technologies can inspect for usage of personal protective equipment (PPE), such as masks and helmets, and can monitor how well workers follow safety protocols in factories or on construction sites.[7] By choosing the HPE ProLiant DL380 Gen11 server with Intel Xeon Gold 6430 processors over the previous-generation server to run these applications, companies can save by performing a given amount of work with fewer servers.

Agriculture

Image recognition is solving many business problems in agriculture, with use cases including animal monitoring, estimating crop yield by counting fruits and vegetables, and automatic harvesting, which detects when crops are ready for picking robots to harvest them. Farmers can even use computer vision systems to demonstrate compliance with animal welfare regulations.[8] For all these applications, running them on the HPE ProLiant DL380 Gen11 server with Intel Xeon Gold 6430 processors vs. its previous-generation counterpart could deliver reliable information faster.

Healthcare

In healthcare, AI is not only helping to detect disease, but also assists with logistical tasks such as inspecting hospitals for hygiene. Another application uses facial recognition to identify patients, which provides benefits such as "helping prevent medical identity fraud, streamlining the registration process, and preventing unauthorized access to sensitive information."[9] Hospitals and medical practices that select HPE ProLiant DL380 Gen11 server with Intel Xeon Gold 6430 processors rather than the older server we tested can get answers earlier.

A look at lower inference precision levels on the HPE ProLiant DL380 Gen11 server

As we noted earlier, we tested two lower levels of inference precision, Int8 and bfloat16, on the HPE ProLiant DL380 Gen11 server. Brain Float 16-bit is a condensed form of the single-precision floating-point format FP32 uses, where 8 exponent bits augment 8-bit precision, vs. 24-bit significance with 8 exponent bits. While this allows bfloat16 a wider dynamic range of numbers than would be available with 8-bit precision alone, it does so at the cost of accuracy. Int8 uses 8-bit integers instead of floating-point numbers, and integer instead of floating-point math, which significantly lowers the compute, memory, and storage overhead for machine learning applications. These lower precision levels grant higher throughput and lower latency, allowing for faster responses to queries, and potentially serving more concurrent users in a datacenter environment.

Figure 3 compares the performance in terms of images per second on the HPE ProLiant DL380 Gen11 server with the three precision levels. As it shows, the server processed from 3 to 5 times as many images per second using bfloat16 and Int8 precision levels as it did using FP32.

We attribute this strong performance in large part to the 4th Generation Intel Xeon Scalable processors in this server, which come with several built-in accelerators to help enhance workload performance. These include Intel® AMX, which "improves the performance of deep-learning training and inference on the CPU and is ideal for workloads like natural-language processing, recommendation systems, and image recognition."[10]

Images per second the HPE ProLiant DL380 Gen11 server achieved at different precision levels

Higher is better

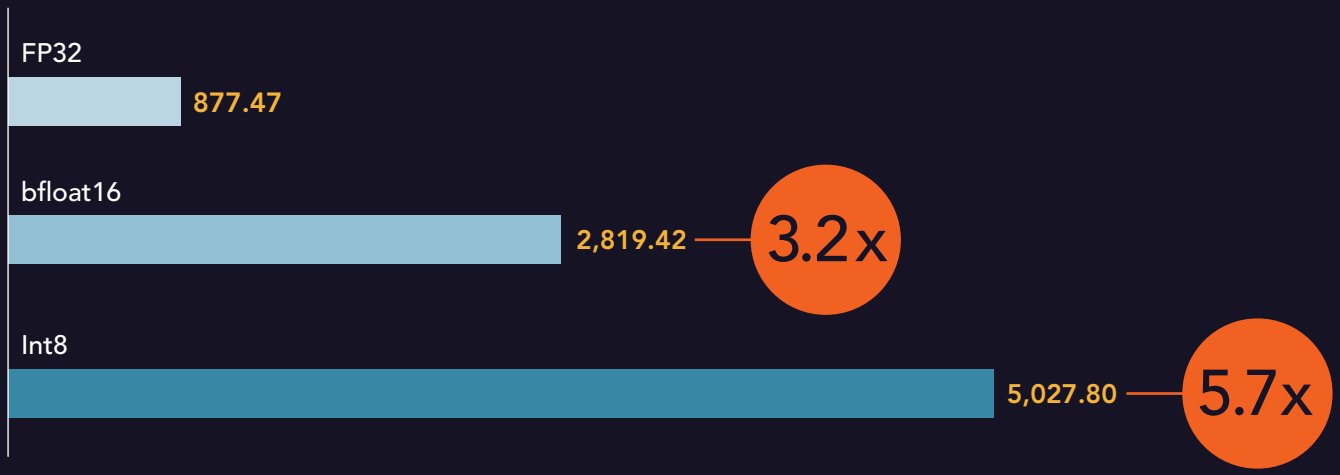


Figure 3: The number of images per second the HPE ProLiant DL380 Gen11 server processed on the ResNet-50 v1.5 model with FP32, bfloat16, and Int8 precision. Higher is better. Source: Principled Technologies.

Conclusion

Companies using AI inference to solve business problems have a range of choices for running these computationally demanding applications. We explored the potential of one solution, the HPE ProLiant DL380 Gen11 server featuring 4th Generation Intel Xeon Gold processors. We compared this server to its previous-generation counterpart on ResNet-50 tests using FP32 precision and found it delivered 2.86 times the inference performance while reducing latency by 30.1 percent. We also tested the HPE ProLiant DL380 Gen11 server at lower precision levels, which place greater demand on CPU resources, and found its performance to be strong with both Int8 and bfloat16 precision levels. Compared to potentially pricey pay-as-you-go cloud solutions and high-end GPU-based server solutions, the HPE ProLiant DL380 Gen11 we tested can be a smart option for businesses harnessing the power of AI imaging applications.

1. Intel, "Advanced Matrix Extensions Overview," accessed November 28, 2023, **https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/advanced-matrix-extensions/overview.html**.
2. HPE, "HPE ProLiant DL380 Gen11 server," accessed November 30, 2023, **https://buy.hpe.com/us/en/compute/rack-servers/proliant-dl300-servers/proliant-dl380-server/hpe-proliant-dl380-gen11/p/1014696069**.
3. HPE, "HPE ProLiant DL380 Gen11 server," accessed November 30, 2023, **https://buy.hpe.com/us/en/compute/rack-servers/proliant-dl300-servers/proliant-dl380-server/hpe-proliant-dl380-gen11/p/1014696069**.
4. Intel, "4th Gen Intel Xeon Scalable Processors," accessed November 28, 2023, **https://www.intel.com/content/www/us/en/newsroom/resources/press-kit-4th-gen-intel-xeon-scalable-processors.html**.
5. Intel, "4th Gen Intel Xeon Scalable Processors."
6. GitHub, Intel AI Models, accessed November 28, 2023, **https://github.com/IntelAI/models**.
7. Viso.ai, "The 100 Most Popular Computer Vision Applications in 2024," accessed November 28, 2023, **https://viso.ai/applications/computer-vision-applications/**.
8. Viso.ai, "Top Applications of Computer Vision in Agriculture," accessed November 28, 2023, **https://viso.ai/applications/computer-vision-in-agriculture/**.
9. Viso.ai, "Top 19 Applications Of Deep Learning and Computer Vision In Healthcare," accessed November 28, 2023, **https://viso.ai/applications/computer-vision-in-healthcare/**.
10. Intel, "Advanced Matrix Extensions Overview," accessed November 28, 2023, **https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/advanced-matrix-extensions/overview.html**.

Read the science behind this report at https://facts.pt/kh4Gyf2 ▶

Principled Technologies®

Facts matter.®

This project was commissioned by HPE.