Improve deep learning inference performance with Microsoft Azure Esv4 VMs with 2nd Gen Intel Xeon Scalable processors

Newer Esv4 VMs handled more images per second than Esv3 VMs with older processors

Using a subset of machine learning—deep learning—to classify images or make predictions from consumer data can help organizations put their mountains of data to good use. For these types of deep learning models, Microsoft Azure offers memory-optimized Esv4-series virtual machines (VMs). Azure Esv4-series VMs are based on Intel® Xeon® Platinum 8272CL processors, which include a feature, Intel Deep Learning Boost, that Intel designed to improve machine learning workloads.

At Principled Technologies, we used two inference benchmarks from the Model Zoo for Intel Architecture—ResNet50, which classifies images, and Wide & Deep recommendation system, which makes relationships between data—to compare the inference performance of older Azure Esv3-series VMs to newer Esv4-series VMs at various instance sizes. We found that for both deep learning benchmarks, the upgraded Esv4-series VMs offered significantly better inference performance, which shows that organizations seeking quick data insights can benefit from selecting Microsoft Azure Esv4-series VMs featuring 2nd Generation Intel Xeon Scalable processors.

Classify up to 8.40x more images per second

Get recommendations from data up to 3.48x as fast

How we tested

We purchased three sets of virtual machine instances from two memory-optimized Microsoft Azure VM series:

- Newer Esv4 series featuring Intel Xeon Platinum 8272CL processors (Cascade Lake)
- Older Esv3 series featuring Intel Xeon E5-2673 v4 processors (Broadwell)

We ran each VM in the East US region.

Figure 1 shows the specifications for the virtual machines that we chose. To show how businesses with different deep learning demands can benefit from choosing Esv4-series VMs, we tested small (8 vCPU), medium (16 vCPU), and large (64 vCPU) VM sizes.

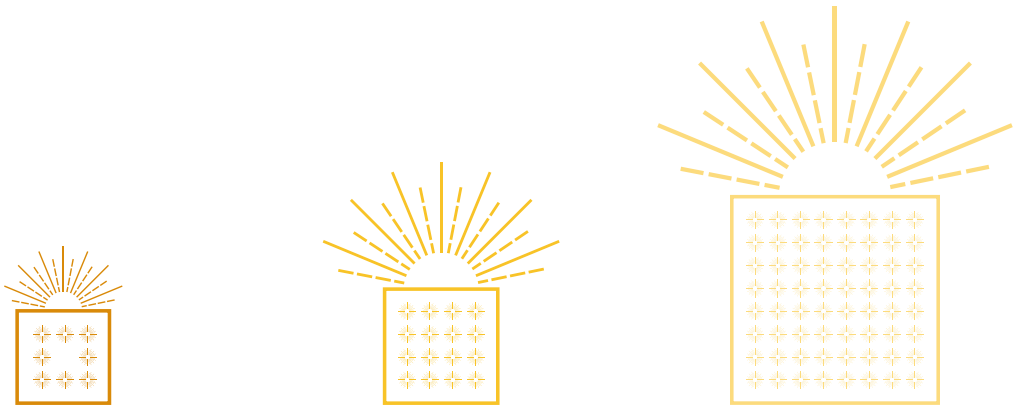

Figure 1: Key specifications for each VM we tested. Source: Principled Technologies.

About 2nd Generation Intel Xeon Scalable processors with Intel Deep Learning Boost

The 2nd Generation Intel Xeon Scalable processor platform—codenamed Cascade Lake—features a wide range of processor types, including Bronze, Silver, Gold, and Platinum, to support varying workload needs. To accelerate machine learning inference, 2nd Gen Intel Xeon Scalable processors offer Intel Deep Learning Boost (DL Boost). Intel DL Boost builds on Intel Advanced Vector Extensions 512 (AVX-512) instructions with Intel Vector Neural Network Instructions (VNNI), combining multiple processor instructions into one to improve machine learning inference performance through resource optimization.[1]

To learn more about Intel DL Boost built into 2nd Generation Intel Xeon Scalable processors, visit https://www.intel.com/content/dam/www/public/us/en/documents/product-overviews/dl-boost-product-overview.pdf.

What Esv4-series VMs can offer your organization

Compared to older Esv3 VMs, ESv4-series VMs offer:

- Up to 506GB RAM
- Guaranteed Cascade Lake processor with all-core Turbo clock speed of 3.4 GHz and Intel Vector Neural Network support (AVX-512 VNNI)
- The ability to support premium storage and premium storage caching

Image classification results – ResNet50

From Model Zoo for Intel Architecture, we chose the popular ResNet50 deep learning benchmark for testing. ResNet50 is a convolutional neural network that runs 50 layers deep and recognizes and classifies images. Using deep learning to classify images is useful for real-world applications such as self-driving cars or aiding in medical diagnoses. The benchmark reported throughput in images per second that the solutions handled using this model, with higher scores indicating better performance at this type of deep learning.

Small instances

If your deep learning needs are on the smaller side, selecting an Azure VM with 8 vCPUs could meet your image classification needs. We found that a new Azure Esv4-series VM with 8 vCPUs featuring 2nd Gen Intel Xeon Scalable processors (with INT8 precision) classified 8.40 times the number of images per second using the ResNet50 benchmark as the small-sized VM with previous-generation processors (with FP32 precision).

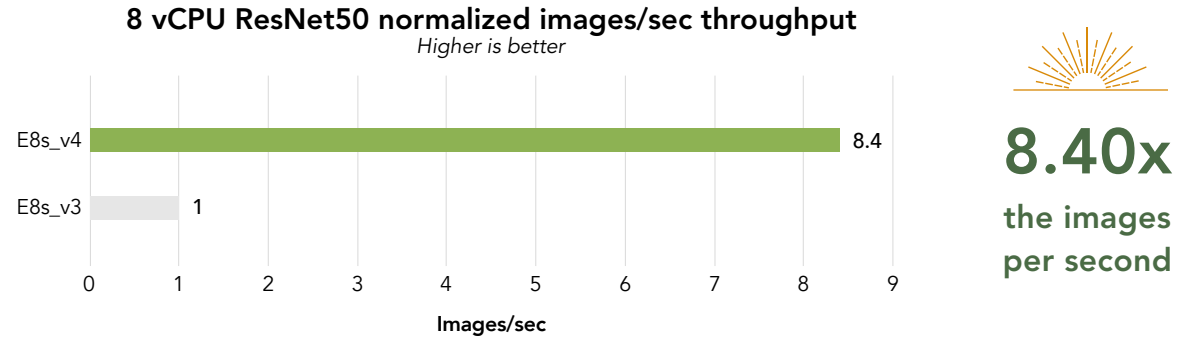

Figure 2: Relative number of images per second that the small-size VMs (8 vCPUs) classified using the ResNet50 benchmark. Higher numbers are better. Source: Principled Technologies.

Medium instances

Larger models or datasets may benefit from an increase to 16 vCPUs per virtual machine. We found that a new Azure Esv4-series VM with 16 vCPUs featuring 2nd Gen Intel Xeon Scalable processors (with INT8 precision) classified 6.67 times the number of images per second using the ResNet50 benchmark as the medium-sized VM with previous-generation processors (with FP32 precision).

Figure 3: Relative number of images per second that the medium-size VMs (16 vCPUs) classified using the ResNet50 benchmark. Higher numbers are better. Source: Principled Technologies.

Large instances

If your organization needs to run deep learning workloads to classify even larger datasets, VMs with 64 vCPUs can better tackle your needs. We found that a new Azure Esv4-series VM with 64 vCPUs featuring 2nd Gen Intel Xeon Scalable processors (with INT8 precision) classified 5.96 times the number of images per second using the ResNet50 benchmark as the large-sized VM with previous-generation processors (with FP32 precision).

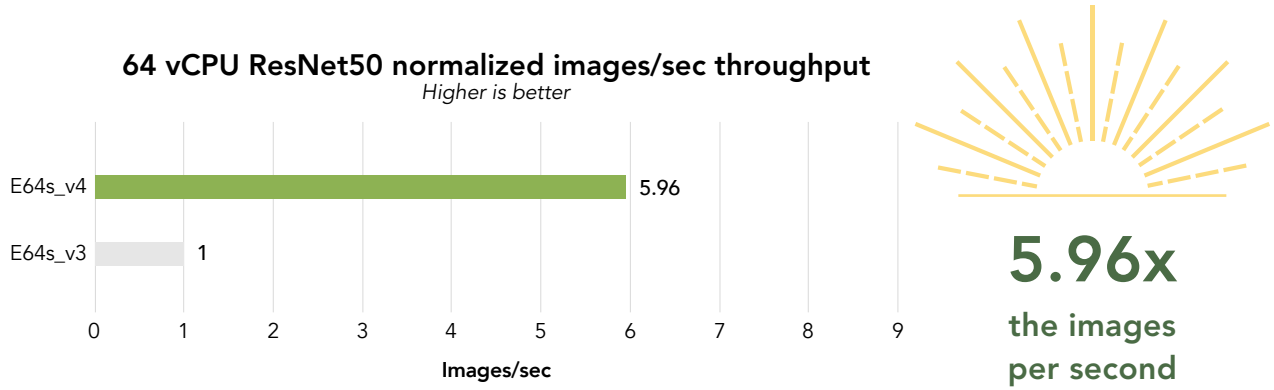

Figure 4: Relative number of images per second that the large-size VMs (64 vCPUs) classified using the ResNet50 benchmark. Higher numbers are better. Source: Principled Technologies.

Get more value from your cloud VMs

Budget considerations require weighing the cost of any performance improvements. Put simply: is the boost in performance worth the additional cost? We found that for deep learning performance on Microsoft Azure Esv4-series VMs, the answer is yes. Based on our test results, newer Esv4-series VMs can offer up to 8.40 times the deep learning performance at a lower (0.94x) overall cost. This means that upgraded Esv4-series VMs with 2nd Gen Intel Xeon Scalable processors can offer better overall value compared to older Esv3-series VMs.

Making recommendations based on data – Wide & Deep learning recommender

We used Model Zoo for Intel Architecture for Wide & Deep learning recommender testing. Wide & Deep uses wide linear models and deep neural networks to infer meaningful relationships between data and deliver recommendations based on that data. The benchmark reports the number of samples per second that the instance handled, with more samples indicating better performance.

Small instances

Smaller deep learning problems with smaller datasets may require VMs configured with 8 vCPUs. We found that a new Azure Esv4-series VM with 8 vCPUs featuring 2nd Gen Intel Xeon Scalable processors (with INT8 precision) handled 3.48 times the number of samples per second using the Wide & Deep benchmark as the small-sized VM with previous-generation processors (with FP32 precision).

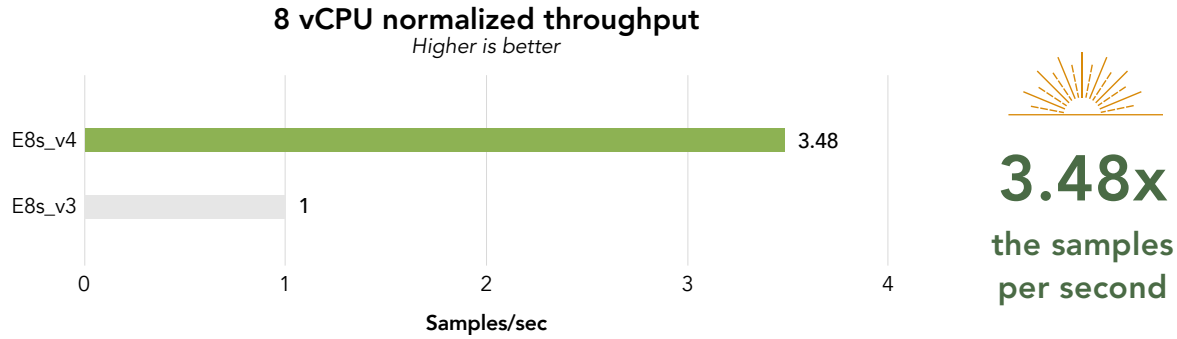

Figure 5: Relative number of samples per second that the small-size VMs (8 vCPUs) handled using the Wide & Deep benchmark. Higher numbers are better. Source: Principled Technologies.

Medium instances

For those seeking to make recommendations based on mid-sized datasets, 16-vCPU VMs may be more appropriate. We found that a new Azure Esv4-series VM with 16 vCPUs featuring 2nd Gen Intel Xeon Scalable processors (with INT8 precision) handled 3.23 times the number of samples per second using the Wide & Deep benchmark as the medium-sized VM with previous-generation processors (with FP32 precision).

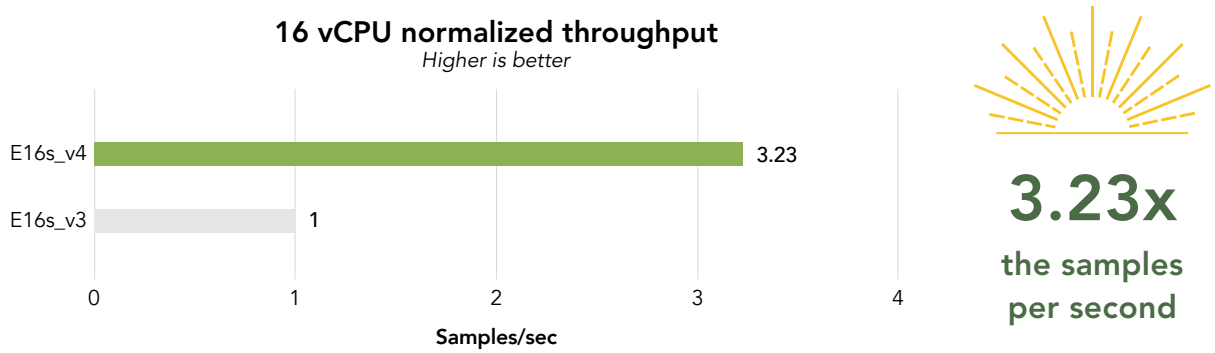

Figure 6: Relative number of samples per second that the medium-size VMs (16 vCPUs) handled using the Wide & Deep benchmark. Higher numbers are better. Source: Principled Technologies.

Large instances

VMs aren't one-size-fits-all, so large models and datasets may require more powerful virtual machines with 64 vCPUs. We found that a new Azure Esv4-series VM with 64 vCPUs featuring 2nd Gen Intel Xeon Scalable processors (with INT8 precision) handled 2.99 times the number of samples per second using the Wide & Deep benchmark as the large-sized VM with previous-generation processors (with FP32 precision).

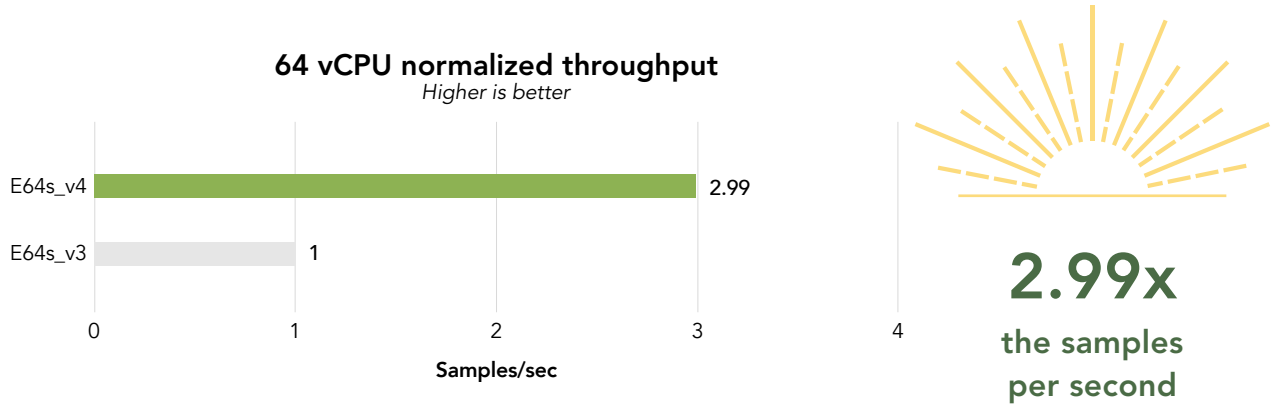

Figure 7: Relative number of samples per second that the large-size VMs (64 vCPUs) handled using the Wide & Deep benchmark. Higher numbers are better. Source: Principled Technologies.

Get insights faster with Azure Esv4 VMs featuring 2nd Gen Intel Xeon Scalable processors

While deep learning models and their applications can vary widely, getting insights from data faster is always the goal, to drive innovation or boost consumer sales. In our tests, we found that newer Microsoft Azure Esv4-series VMs featuring 2nd Gen Intel Xeon Scalable processors—which offer Intel Deep Learning Boost—improved deep learning inference performance for image classification and recommendations over older Esv3 VMs. And at just 0.94x the cost, the Esv4 series offers significantly better value per VM, which could mean your organization requires fewer VMs to support.

By choosing Microsoft Azure Esv4-series VMs with 2nd Gen Intel Xeon Scalable processors, your organization can get deep learning insights from data faster than with older Esv3-series VMs.

---

1 Intel, "Intel Deep Learning Boost," accessed July 29, 2021, https://www.intel.com/content/dam/www/public/us/en/documents/product-overviews/dl-boost-product-overview.pdf.

Read the science behind this report at http://facts.pt/lHyrW1n ►

Principled Technologies®

Facts matter.®

This project was commissioned by Intel.