Make Inferences From Data Up to 8.40x Faster with Microsoft Azure Esv4 VMs

ResNet50

8.40x more images per second with 8-vCPU Esv4 VMs

vs. Esv3 VMs

6.67x more images per second with 16-vCPU Esv4 VMs

vs. Esv3 VMs

5.96x more images per second with 64-vCPU Esv4 VMs

vs. Esv3 VMs

Improve ResNet50 Inference Performance with Azure Esv4 VMs featuring 2nd Gen Intel® Xeon® Scalable processors

Deep learning workloads can make inferences from data by sorting and classifying images and applying its knowledge to new data in real time. ResNet50 is a deep learning framework that enables this type of machine learning, creating a convolutional neural network running 50 layers deep to make accurate inferences. Selecting Microsoft Azure Esv4 VMs enabled by 2nd Gen Intel® Xeon® Scalable processors, which feature Intel Deep Learning Boost, can improve ResNet50 inference performance.

To determine which configuration offers better performance, independent third-party Principled Technologies tested ResNet50 performance across three different VM sizes. Azure Esv4 VMs featuring Intel Xeon Platinum 8272CL processors classified up to 8.40x more images per second than Esv3 VMs. With Esv4 VMs, organizations can make inferences from data faster to solve real-world problems in less time.

Improve Deep Learning Performance on Small Instances

The faster your cloud VMs can infer meaningful relationships between data, the faster you can put insights to use. As Figure 1 shows, 8-vCPU Esv4 VMs enabled by 2nd Gen Intel Xeon Scalable processors outperformed 8-vCPU Esv3 VMs in a deep learning ResNet50 benchmark test. The Intel Xeon processor-based VMs classified 8.40 times the images per second that the previous-gen VMs did, which means that each VM can do a staggering amount more work per VM to provide better value.

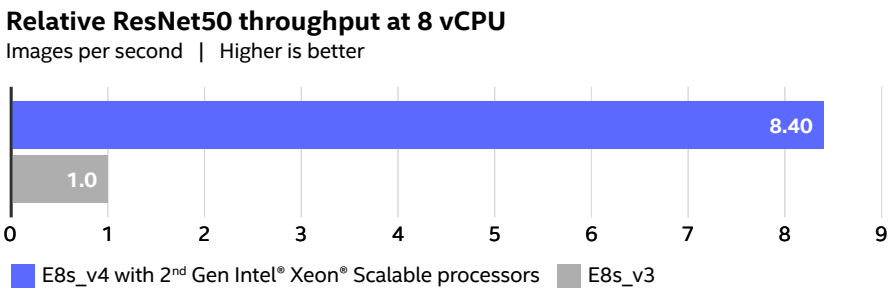

Figure 1. Relative results comparing the ResNet50 benchmark performance of small (8-vCPU) Esv4 VMs vs. Esv3 VMs.

Improve Deep Learning Performance on Medium Instances

Organizations with mid-sized datasets can also get improved deep learning inference performance by choosing VMs with newer processors. As Figure 2 shows, 16-vCPU Azure Esv4 VMs enabled by 2nd Gen Intel® Xeon® Scalable processors classified 6.67 times the images per second in ResNet50 tests compared to Esv3 VMs with previous-generation processors.

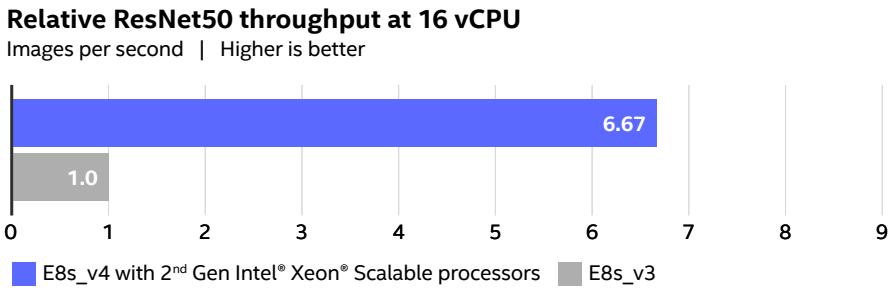

Figure 2. Relative results comparing the ResNet50 benchmark performance of medium (16-vCPU) Esv4 VMs vs. Esv3 VMs.

Improve Deep Learning Performance on Large Instances

Larger datasets that require larger VMs similarly benefit from choosing newer processor architecture for deep learning workloads. In tests, Esv4 VMs featuring 2nd Gen Intel Xeon Scalable processors classified 5.96 times the images per second using the ResNet50 benchmark test (see Figure 3).

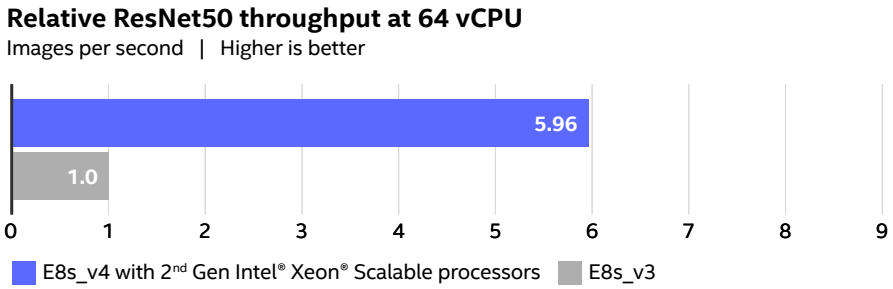

Figure 3. Relative results comparing the ResNet50 benchmark performance of large (64-vCPU) Esv4 VMs vs. Esv3 VMs.

For datasets small, medium, and large, selecting Azure Esv4 VMs with 2nd Gen Intel Xeon Scalable processors over Esv3 VMs with previous-generation processors can boost deep learning performance to form meaningful relationships from data and make inferences about new data to solve problems faster.

Learn More

To begin running your ResNet50 workloads on Azure Esv4 Instances with 2nd Gen Intel Xeon Scalable processors, visit http://intel.com/azure.

For complete testing results, visit http://facts.pt/YX3rsPQ.

Printed in USA 1021/JO/PT/PDF US001 ♻ Please Recycle