Classify Up to 6.17x More Images Per Second with 2nd Gen Intel® Xeon® Scalable Processor-Based AWS M5n Instances

ResNet50

6.17x more images per second with 8-vCPU M5n instances

vs. M4instances

5.78x more images per second with 16-vCPU M5ninstances

vs. M4instances

5.23x more images per second with 64-vCPU M5n instances

vs. M4instances

Boost ResNet50 Inference Performance with AWS M5n Instances Featuring 2nd Gen Intel® Xeon® Scalable Processors

As organizations amass data, they are turning to machine learning workloads to help make sense of it all so they can put the insight they glean to good use. One popular machine learning workload—a deep learning framework, is ResNet50. A convolutional neural network that runs 50 layers deep, ResNet50 analyzes data and recognizes and classifies images to make inferences. Tests show that choosing AWS M5n instances enabled by 2nd Gen Intel® Xeon® Scalable processors over M4 instances with previous-generation processors can improve ResNet50 inference performance. The 2nd Gen Intel Xeon Scalable processor family features Intel Deep Learning Boost, which improves deep learning performance. In third-party testing conducted by Principled Technologies, across three different instance sizes, M5n instances featuring Intel Xeon Platinum 8272CL processors classified up to 6.17x more images per second than M4 instances. With M5n instances, organizations can speed deep learning workloads and classify images in real-world applications such as diagnosing medical conditions faster.

Improve Deep Learning Performance on Small Instances

The faster your cloud instances can infer meaningful relationships between data, the faster you can put your insights to use. As Figure 1 shows, 8-vCPU M5n instances enabled by 2nd Gen Intel Xeon Scalable processors outperformed 8-vCPU M4 instances in a deep learning ResNet50 benchmark test. The Intel Xeon processor-based instances classified 6.17 times the images per second that the previous-gen instances did, which means you solve problems in less time.

Relative ResNet50 throughput at 8 vCPU

Images per second | Higher is better

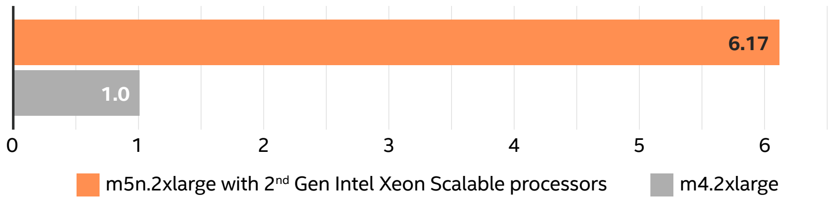

Figure 1. Relative results comparing the ResNet50 benchmark performance of` small (8-vCPU) M5n instances vs. M4 instances.

Improve Deep Learning Performance on Medium Instances

Organizations with mid-sized datasets can also get improved deep learning inference performance by choosing instances with newer processors. As Figure 2 shows, 16-vCPU AWS M5n instances enabled by 2nd Gen Intel® Xeon® Scalable processors classified 5.78 times the images per second in ResNet50 tests compared to M4 instances with previous-generation processors.

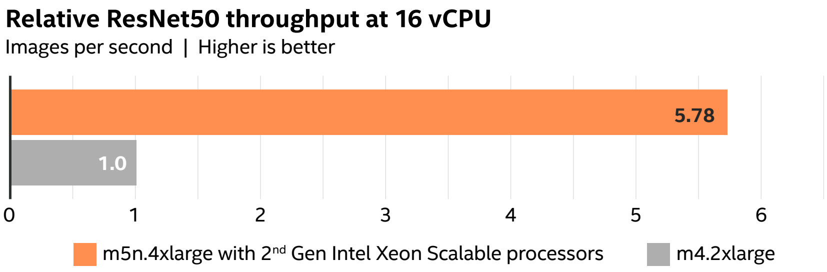

Figure 2. Relative results comparing the ResNet50 benchmark performance of medium (16-vCPU) M5n instances vs. M4 instances.

Improve Deep Learning Performance on Large Instances

Larger datasets that require larger instances similarly benefit from choosing newer processor architecture for deep learning workloads. In tests, M5n instances featuring 2nd Gen Intel Xeon Scalable processors classified 5.23 times the images per second using the ResNet50 benchmark test (see Figure 3).

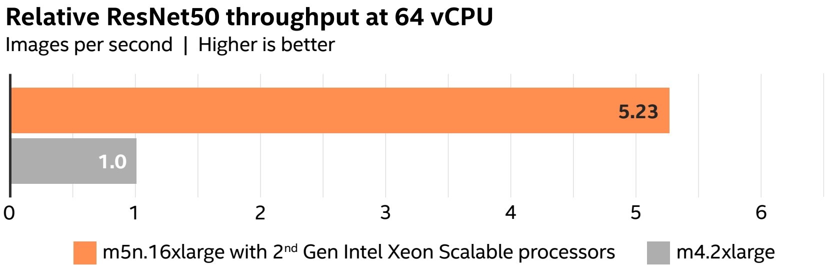

Figure 3. Relative results comparing the ResNet50 benchmark performance of large (64-vCPU) M5n instances vs. M4 instances.

Whether your datasets are small, large, or somewhere in between, selecting AWS M5n instances with 2nd Gen Intel Xeon Scalable processors instead of M4 instances with older processors can enhance deep learning performance. With improved deep learning performance, your organization can make sense of data in less time and speed responses to real-world problems.

Learn More

To begin running your ResNet50workloads on AWSM5nInstanceswith 2ndGen Intel Xeon Scalable processors, visit http://intel.com/aws.

For complete testing results, visit http://facts.pt/oOUDy0F.

Performance varies by use, configuration and other factors. Learn more at https://intel.com/benchmarks.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. Your costs and results may vary.

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

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