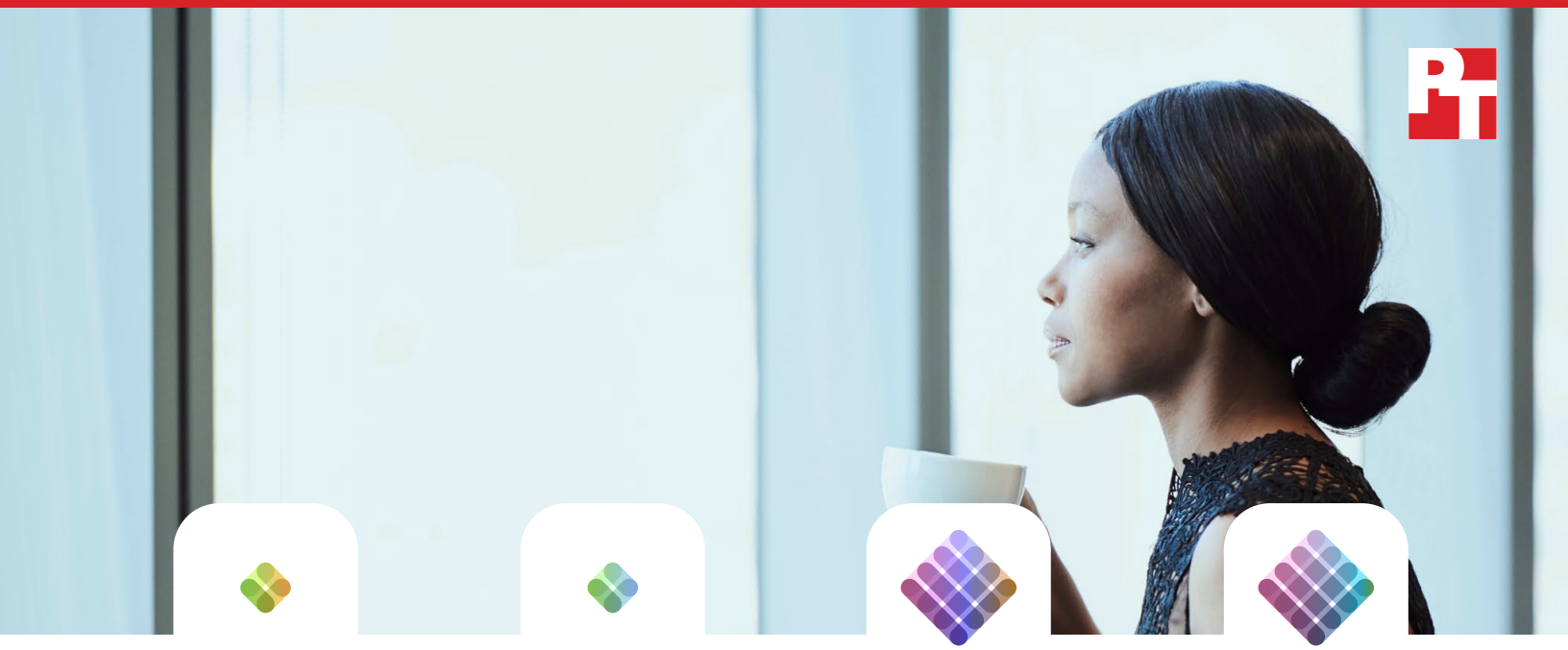


Classify up to 1.19x the frames per second

on small M6i instances with 16 vCPUs vs. M5n instances

Classify up to 2.49x the frames per second

on small M6i instances with 16 vCPUs vs. M6a instances

Classify up to 1.21x the frames per second

on large M6i instances with 96 vCPUs vs. M5n instances

Classify up to 2.94x the frames per second

on large M6i instances with 96 vCPUs vs. M6a instances

AWS EC2 M6i instances with 3rd Gen Intel Xeon Scalable processors accelerated image classification for machine learning workloads

At multiple instance sizes, M6i instances classified more frames per second than M5n instances with previous-gen processors or M6a instances with 3rd Gen AMD EPYC processors

Machine learning makes rapid image classification possible—giving us new ways to diagnose medical conditions, assess damage after natural disasters, and more. If your organization wants to run these types of machine learning workloads in the cloud, it's vital to choose a configuration that can deliver insights quickly so you aren't stuck waiting before you can act.

We compared the ResNet50 image classification performance of two Amazon Web Services (AWS) Elastic Cloud Compute (EC2) instance sizes with different processor configurations: M6i instances with 3rd Gen Intel® Xeon® Scalable processors, M5n instances with 2nd Gen Intel Xeon Scalable processors, and M6a instances with 3rd Gen AMD EPYC™ processors. Across instances with small and large vCPU counts, M6i instances with 3rd Gen Intel Xeon Scalable processors delivered better image classification performance than both the previous-gen M5n instances and the current-gen M6a instances. These results show that organizations that want to classify images quickly with machine learning workloads could get answers sooner when they select AWS M6i instances with 3rd Gen Intel Xeon Scalable processors.

How we tested

We purchased three sets of instances from three general-purpose AWS EC2 series:

- M6i instances featuring 3rd Gen Intel Xeon Platinum 8375C processors (Ice Lake)
- M5n instances featuring 2nd Gen Intel Xeon Platinum 8272CL processors (Cascade Lake)
- M6a instances featuring 3rd Gen AMD EPYC 7R13 processors (Milan)

We ran each instance in the US East 1 region.

Figure 1 shows the specifications for the instances that we chose. To show how businesses of various sizes with different machine learning demands can benefit from choosing M6i instances, we tested small (16 vCPUs) and large (96 vCPUs) instance sizes.

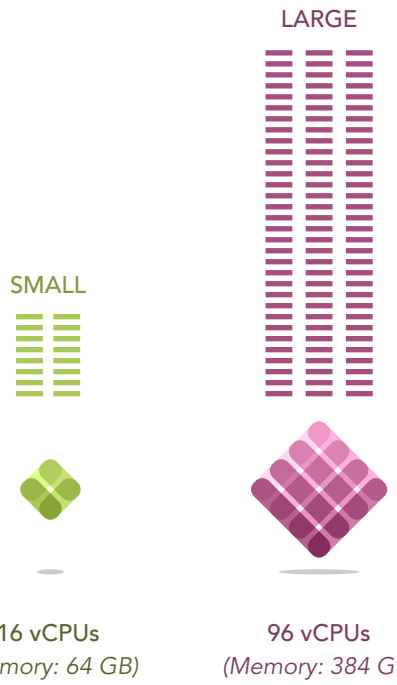


Figure 1: Key specifications for each instance size we tested. Source: Principled Technologies.

Key features of M6i instances with 3rd Gen Intel Xeon Scalable processors

New M6i instances with 3rd Gen Intel Xeon Scalable processors offer the following:[1]

- All-core turbo frequency of up to 3.5 GHz
- Always-on memory encryption with Intel Total Memory Encryption (TME)
- Intel Advanced Vector Extensions (Intel AVX-512) instructions and Intel Deep Learning Boost Vector Neural Network Instructions (VNNI) for demanding machine learning workloads
- Support for up to 128 vCPUs and 512 GB of memory per instance
- Up to 50Gbps networking

About 3rd Generation Intel Xeon Scalable processors

According to Intel, 3rd Generation Intel Xeon Scalable processors are "[o]ptimized for cloud, enterprise, HPC, network, security, and IoT workloads with 8 to 40 powerful cores and a wide range of frequency, feature, and power levels."[2] Intel continues to offer many models from the Platinum, Gold, Silver, and Bronze processor lines that they "designed through decades of innovation for the most common workload requirements."[3]
For more information, visit http://intel.com/xeonscalable.

Classifying images using ResNet50

Deep learning for image classification has many real-world applications, including urban planning, assessing impact from natural disasters, powering self-driving cars, and even aiding in medical diagnoses. From the TensorFlow framework, we selected ResNet50 to test image classification performance. ResNet50 is a convolutional neural network that runs 50 layers deep. The benchmark reported the rate of frames per second that the solutions were able to classify, with higher scores indicating better performance for this type of deep learning.

First, we compared the ResNet50 performance of new M6i instances with 3rd Gen Intel Xeon Scalable processors to previous-gen M5n instances with 2nd Gen Intel Xeon Scalable processors across two instances sizes. Then, we compared the performance of M6i instances to current-gen M6a instances with 3rd Gen AMD EPYC processors. In the following sections, we show results of testing using INT8 precision and batch size of 128 to better reflect real-world workloads. We also tested with a batch size of 1, and share those results in the science behind the report.

Small instances: M6i vs. M5n

Organizations with lower throughput demands can opt to host their deep learning workloads on instances with lower vCPU counts. In our first gen-over-gen processor comparison, M6i instances with newer processors had a strong performance advantage over the previous generation. Figure 2 compares the relative ResNet50 classification rate that the instances achieved on small, 16vCPU configurations. The M6i instances enabled by 3rd Gen Intel Xeon Scalable processors classified 1.19 times the frames per second on the ResNet50 benchmark as the M5n instances with 2nd Gen Intel Xeon Scalable processors.

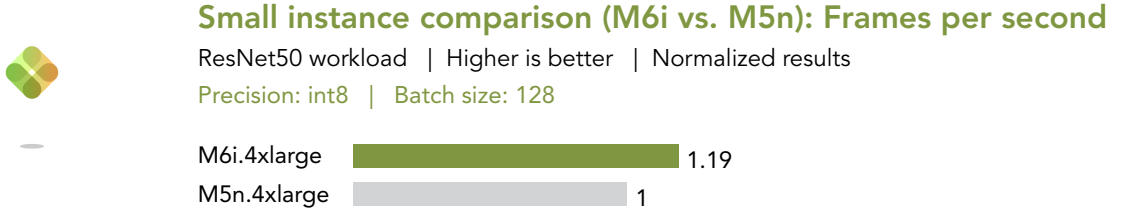


Figure 2: Relative number of frames per second that small M6i and M5n instances (16 vCPUs) handled using the ResNet50 benchmark. Higher numbers are better. Source: Principled Technologies.

Large instances: M6i vs. M5n

Larger instance sizes offered a similar performance increase for M6i instances over M5n, showing that at multiple instance sizes the newer processors enabled faster image classification performance. Figure 3 compares the relative ResNet50 classification rate that the instances achieved on large, 96vCPU configurations. The M6i instances enabled by 3rd Gen Intel Xeon Scalable processors classified 1.21 times the frames per second on the ResNet50 benchmark as the M5n instances with 2nd Gen Intel Xeon Scalable processors.

Large instance comparison (M6i vs. M5n): Frames per second

ResNet50 workload | Higher is better | Normalized results

Precision: int8 | Batch size: 128

Figure 3: Relative number of frames per second that the large M6i and M5n instances (96 vCPUs) handled using the ResNet50 benchmark. Higher numbers are better. Source: Principled Technologies.

Small instances: M6i vs. M6a

The deep learning comparison we tested of M6i instances with 3rd Gen Intel Xeon Scalable processors against M6a instances with 3rd Gen AMD EPYC processors showed an even larger performance gap then the gen-over-gen comparison, with M6i instances classifying more than twice the rate of images per second. Figure 4 compares the relative ResNet50 classification rate that the instances achieved on small, 16vCPU configurations. The M6i instances enabled by 3rd Gen Intel Xeon Scalable processors classified 2.49 times the frames per second on the ResNet50 benchmark as the M6a instances with 3rd Gen AMD EPYC processors.

Small instance comparison (M6i vs. M6a): Frames per second

ResNet50 workload | Higher is better | Normalized results

Precision: int8 | Batch size: 128

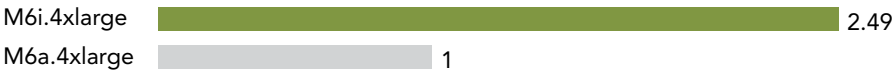


Figure 4: Relative number of frames per second that small M6i and M6a instances (16 vCPUs) handled using the ResNet50 benchmark. Higher numbers are better. Source: Principled Technologies.

Large instances: M6i vs. M6a

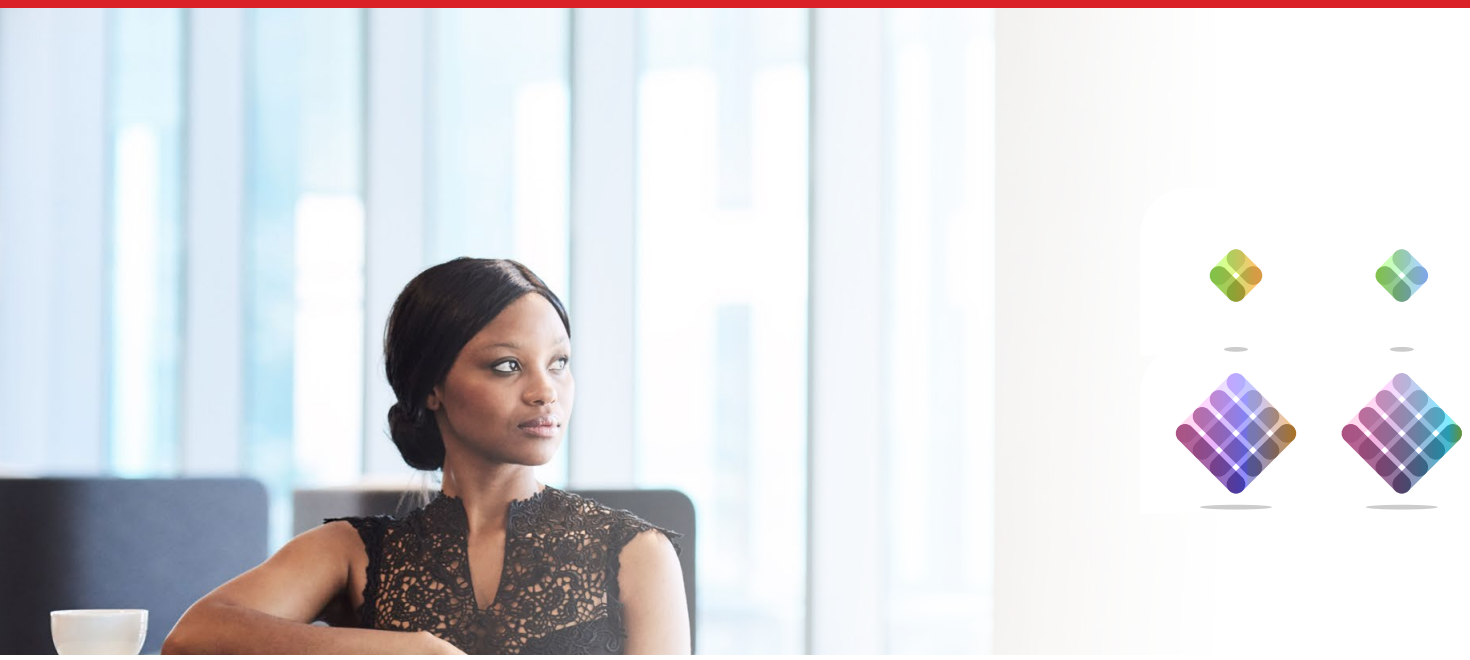
Again, for bigger image classification workloads, M6i instances offered better ResNet50 throughput than their M6a counterparts. Figure 5 compares the relative ResNet50 classification rate that the instances achieved on large, 96vCPU configurations. The M6i instances enabled by 3rd Gen Intel Xeon Scalable processors classified 2.94 times the frames per second on the ResNet50 benchmark as the M6a instances with 3rd Gen AMD EPYC processors. This ability to classify images faster could lead to faster insights in the real world, including quicker answers for medical diagnoses and more reliable self-driving car performance.

Large instance comparison (M6i vs. M6a): Frames per second

ResNet50 workload | Higher is better | Normalized results

Precision: int8 | Batch size: 128

Figure 5: Relative number of frames per second that the large M6i and M6a instances (96 vCPUs) handled using the ResNet50 benchmark. Higher numbers are better. Source: Principled Technologies.



Conclusion

Powerful deep learning solutions that classify your images quickly can deliver insights that let you act sooner. Our testing shows that AWS M6i instances with 3rd Gen Intel Xeon Scalable processors delivered better image classification performance on ResNet50 compared to both previous-gen M5n instances with 2nd Gen Intel Xeon Scalable processors and current-gen M6a instances with 3rd Gen AMD EPYC processors.

Organizations of all sizes that seek to benefit from running deep learning workloads in the cloud could get faster data insights or pay for fewer instances by selecting AWS M6i instances featuring 3rd Gen Intel Xeon Scalable processors.

---

1 Amazon, "Amazon EC2 M6i Instances," accessed December 14, 2021, https://aws.amazon.com/ec2/instance-types/m6i//.

2 Intel, "3rd Gen Intel® Xeon® Scalable Processors," accessed December 14, 2021, https://www.intel.com/content/www/us/en/products/docs/processors/xeon/3rd-gen-xeon-scalable-processors-brief.html.

3 Intel, "3rd Gen Intel® Xeon® Scalable Processors."

Read the science behind this report at https://facts.pt/Mh2McZA ▶

Principled Technologies®

Facts matter.®

This project was commissioned by Intel.