A Principled Technologies report: Hands-on testing. Real-world results.

Classify images up to **6.17x more images per second**

Generate ad recommendations **up to 2.94x as fast**

Speed up deep learning tasks with Amazon Web Services instances featuring 2nd Gen Intel Xeon Scalable processors

Newer M5n instances featuring Intel Xeon Platinum 8272CL processors performed more inference operations per second than M4 instances with older processors

Big data—the massive amounts of information that organizations collect—is useful only if the information is sorted and classified to deliver insights they can act on. By using deep learning networks for quick image classification and prediction, computers can offer insight into business patterns and offer suggestions to consumers in real time. Amazon Web Services (AWS) Elastic Compute Cloud (EC2) offers several cloud instances that can support deep learning models, including general-purpose M5n instances. These newer AWS M5n instances run 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 deep learning inference benchmarks from the Model Zoo for Intel Architecture—ResNet50, which classifies images, and Wide & Deep recommendation system, which generates advertisement recommendations—to compare the inference performance of older M4 instances to newer M5n instances at various instance sizes. We found that for both deep learning frameworks, the upgraded M5n instances offered significantly better inference performance, which shows that M5n instances featuring 2nd Generation Intel Xeon Scalable processors can help organizations make sense of their data faster.

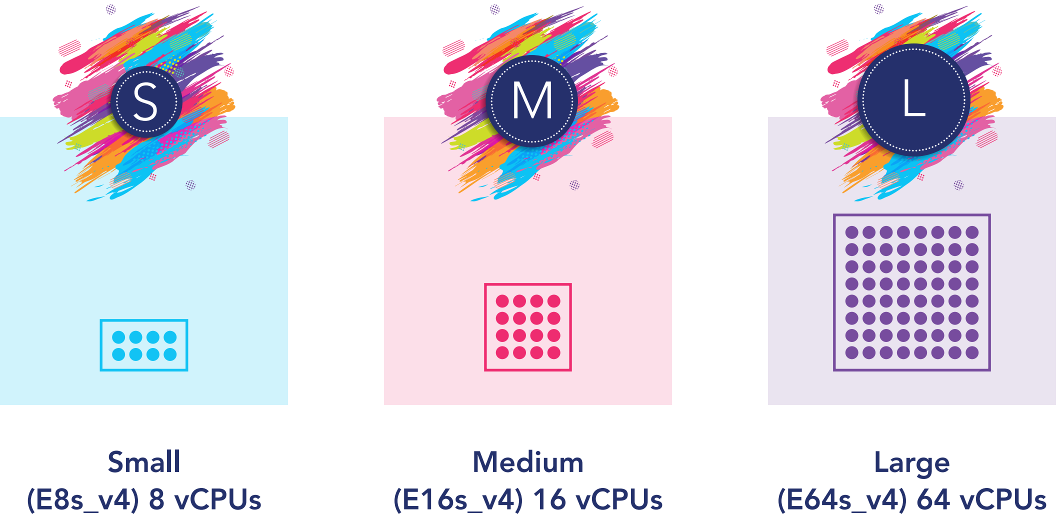
How we tested

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

- Newer M5n instances featuring 2nd Generation Intel Xeon Scalable processors (Cascade Lake)
- Older M4 instances featuring Intel Xeon E5-2686 v4 processors (Broadwell)

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 deep learning demands can benefit from choosing M5n instances, we tested small (8 vCPU), medium (16 vCPU), and large (64 vCPU) VM sizes.



Small
(E8s_v4) 8 vCPUs

Medium
(E16s_v4) 16 vCPUs

Large
(E64s_v4) 64 vCPUs

Figure 1: Key specifications for each instance size 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.

Why choose M5n instances?

Compared to older M4 instances, M5n instances offer:

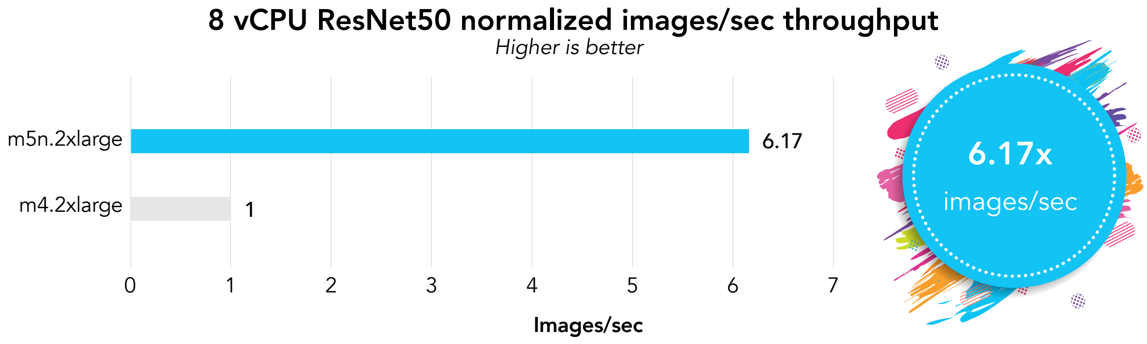
- 2nd Generation Intel Xeon Scalable Processors with a sustained all-core Turbo CPU frequency of 3.1 GHz, maximum single core turbo frequency of 3.5 GHz, and Intel Vector Neural Network support (AVX-512 VNNI)
- Peak bandwidth of 25 Gbps for small instances or 100 Gbps for large instances
- EBC or NVMe™ SSDs physically attached to the host server

Classifying images – ResNet50

From Model Zoo for Intel Architecture, which offers machine learning models, we chose the popular ResNet50 deep learning model for testing. ResNet50 is a convolutional neural network that runs 50 layers deep; organizations use it to recognize and classify images. Deep learning for image classification 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 AWS instance with 8 vCPUs could meet your image classification needs. We found that a newer AWS M5n instance with 8 vCPUs featuring 2nd Gen Intel Xeon Scalable processors classified 6.17 times the number of images per second using the ResNet50 benchmark (with INT8 precision) as the small-sized M4 instance with previous-generation Intel Xeon processors (with FP32 precision).



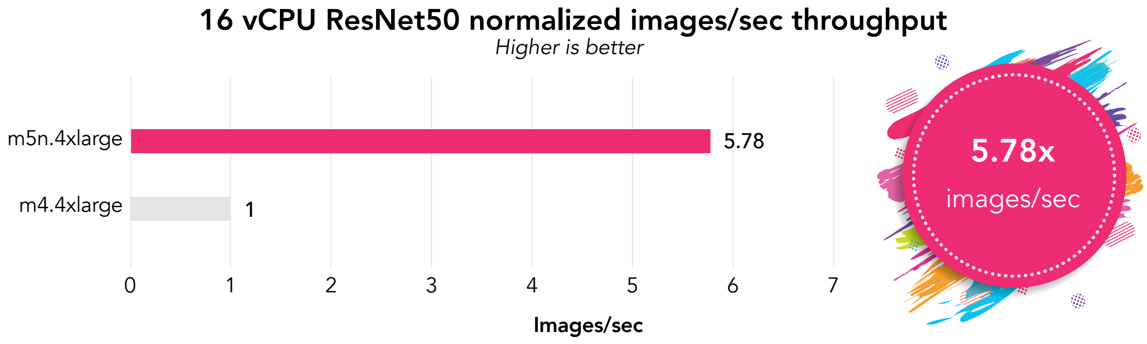
8 vCPU ResNet50 normalized images/sec throughput

Higher is better

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

Medium instances

Medium instances (16 vCPUs), which are better suited for larger models or datasets, also benefit from newer processors. We found that a newer AWS M5n instance with 16 vCPUs featuring 2nd Gen Intel Xeon Scalable processors classified 5.78 times the number of images per second using the ResNet50 benchmark (with INT8 precision) as the medium-sized M4 instance with previous-generation Intel Xeon processors (with FP32 precision).



16 vCPU ResNet50 normalized images/sec throughput

Higher is better

Figure 3: Relative number of images per second that the medium-size instances (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 extract insights or make recommendations from even larger datasets, instances with 64 vCPUs can better tackle your needs. We found that a newer AWS M5n instance with 64 vCPUs featuring 2nd Gen Intel Xeon Scalable processors classified 5.23 times the number of images per second using the ResNet50 benchmark (with INT8 precision) as the large-sized M4 instance with previous-generation Intel Xeon processors (with FP32 precision).

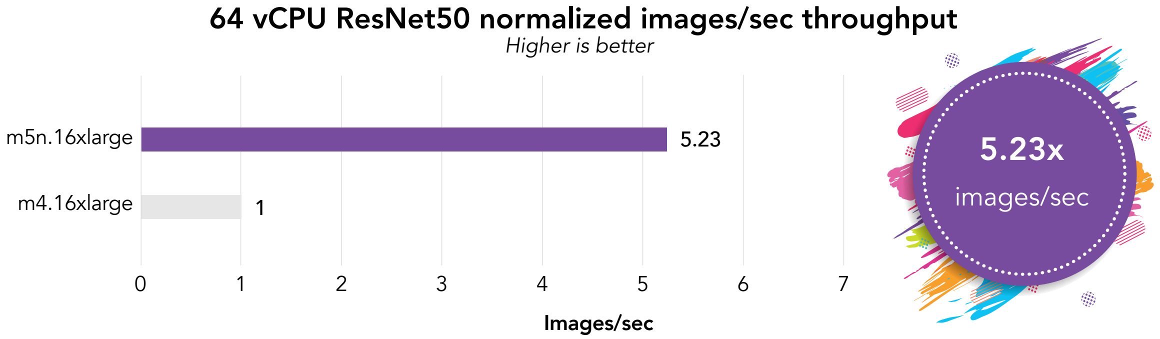


Figure 4: Relative number of images per second that the large-size instances (64 vCPUs) classified using the ResNet50 benchmark. Higher numbers are better. Source: Principled Technologies.

Get more value from your cloud instances

Based on our machine learning test results, newer AWS M5n instances offer up to 6.17 times the ResNet50 performance for as little as 1.19 times the cost. This means that upgraded M5n instances with 2nd Gen Intel Xeon Scalable processors can provide better overall value compared to older M4 instances.

Real-time recommendations based on ad-click historical data – Model Zoo for Intel Architecture Wide & Deep model

We used a TensorFlow-based model from the Model Zoo for Intel Architecture to conduct Wide & Deep testing. Wide & Deep uses wide linear models and deep neural networks to infer meaningful relationships between data and deliver recommendations based on that data.

Small instances

Organizations with smaller model sizes and datasets can run deep learning on instances configured with 8 vCPUs. We found that a newer AWS M5n instance with 8 vCPUs featuring 2nd Gen Intel Xeon Scalable processors (with INT8 precision) handled 2.86 times the number of samples per second using the Wide & Deep benchmark as the small-sized M4 instance with previous-generation processors (with FP32 precision).

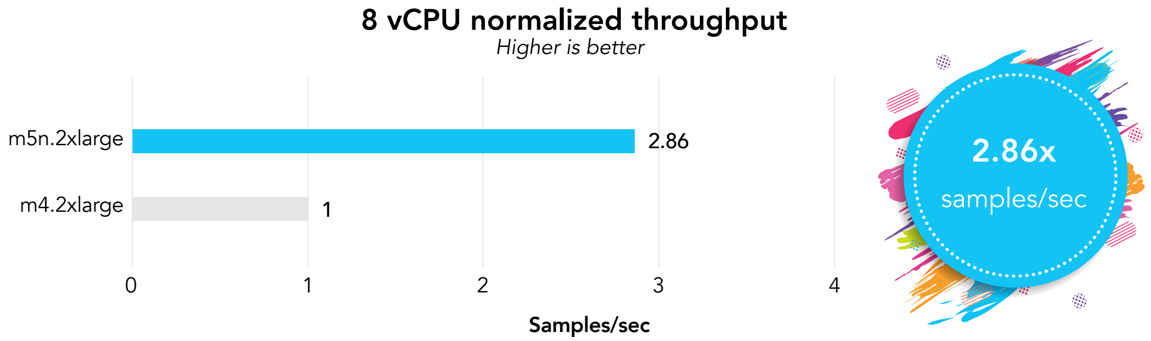


Figure 5: Relative number of samples per second that the small-size instances (8 vCPUs) handled using the Wide & Deep benchmark. Higher numbers are better. Source: Principled Technologies.

Medium instances

For organizations seeking to make recommendations based on mid-sized datasets, 16 vCPU instances may be more appropriate. We found that a newer AWS M5n instance with 16 vCPUs featuring 2nd Gen Intel Xeon Scalable processors (with INT8 precision) handled 2.94 times the number of samples per second using the Wide & Deep benchmark as the medium-sized M4 instance with previous-generation processors (with FP32 precision).

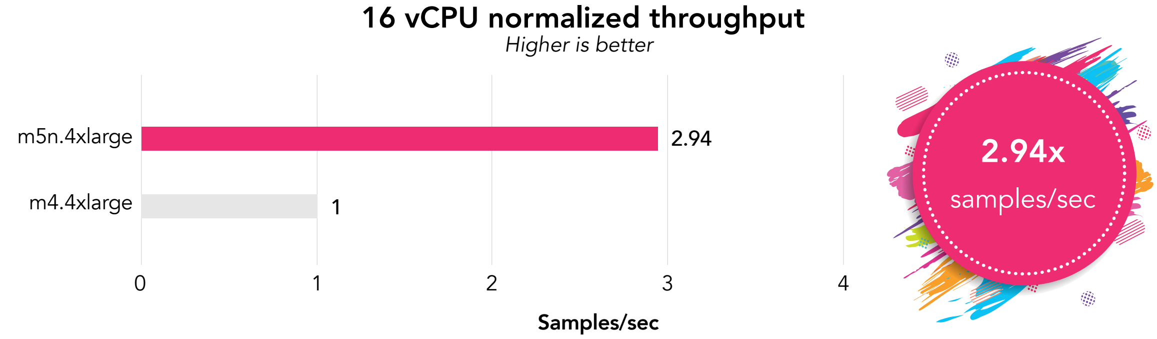


Figure 6: Relative number of samples per second that the medium-size instances (16 vCPUs) handled using the Wide & Deep benchmark. Higher numbers are better. Source: Principled Technologies.

Large instances

Those that need quick recommendations on larger datasets may require virtual machines with 64 vCPUs. We found that a newer AWS M5n instance with 64 vCPUs featuring 2nd Gen Intel Xeon Scalable processors (with INT8 precision) handled 2.67 times the number of samples per second using the Wide & Deep benchmark as the large-sized M4 instance with previous-generation processors (with FP32 precision).

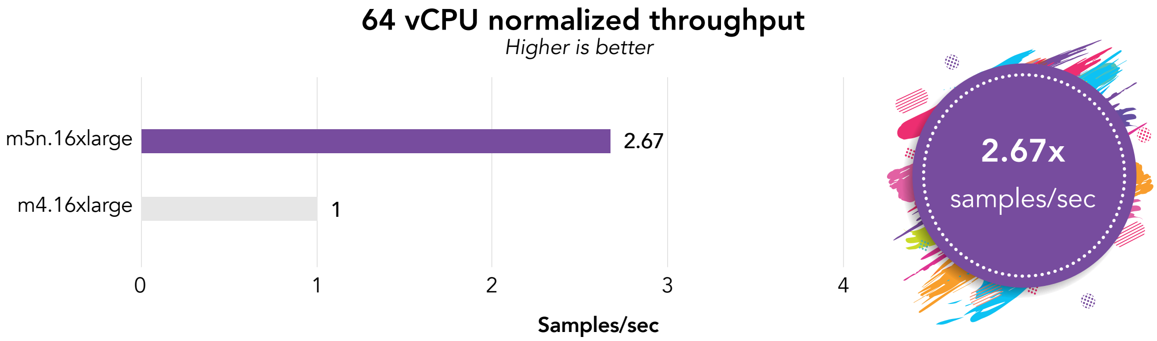


Figure 7: Relative number of samples per second that the large-size instances (64 vCPUs) handled using the Wide & Deep benchmark. Higher numbers are better. Source: Principled Technologies.

Choose AWS M5n instances featuring 2nd Gen Intel Xeon Scalable processors for timely insights from data

Getting faster insights from never-ending data streams can improve business agility and lead to greater success. Our test results show that newer AWS M5n instances featuring 2nd Gen Intel Xeon Scalable processors with Intel Deep Learning Boost sped up deep learning inference performance for image classification and recommendation models over older M4 instances. These performance improvements come at little added cost, which means that M5n instances can offer better value per VM. By doing more deep learning work per instance, your organization could ultimately require fewer instances overall, which can help keep budget concerns in check.

By choosing AWS EC2 M5n instances with 2nd Gen Intel Xeon Scalable processors, your organization can get deep learning insights from data faster than with older M4 instances.

[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/7J92SKA ▶

Principled Technologies®

Facts matter.®

This project was commissioned by Intel.