AWS EC2 M6i instances featuring 3rd Gen Intel Xeon Scalable processors improved Wide & Deep recommender performance

Across different instance sizes, M6i instances performed more inference operations per second than M5n instances with 2nd Gen Intel Xeon Scalable processors and M6a instances with 3rd Gen AMD EPYC processors

While visitors browse your ecommerce site, you collect their data. What they click indicates what they want to purchase, and from this data, you can infer other things they might like based on patterns from other visitors. Deep learning workloads—specifically Wide & Deep recommendation engines—can quickly make sense of this data and make real-time ad or purchase recommendations to fuel sales.

We compared the Wide & Deep inference performance of three Amazon Web Services (AWS) EC2 cloud 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. We found that small-, medium-, and large-sized M6i instances with 3rd Gen Intel Xeon Scalable processors outperformed both their M5n and M6a counterparts, indicating that organizations looking to speed recommendations using Wide & Deep inference workloads could gain insights sooner by selecting M6i instances.

Up to 1.67 times the frames per second

on small instances vs. M6a

Up to 1.35 times the frames per second

on medium instances vs. M6a

Up to 1.75 times the frames per second

on large instances vs. M6a

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), medium (64 vCPUs), and large (96 vCPUs) instance sizes.

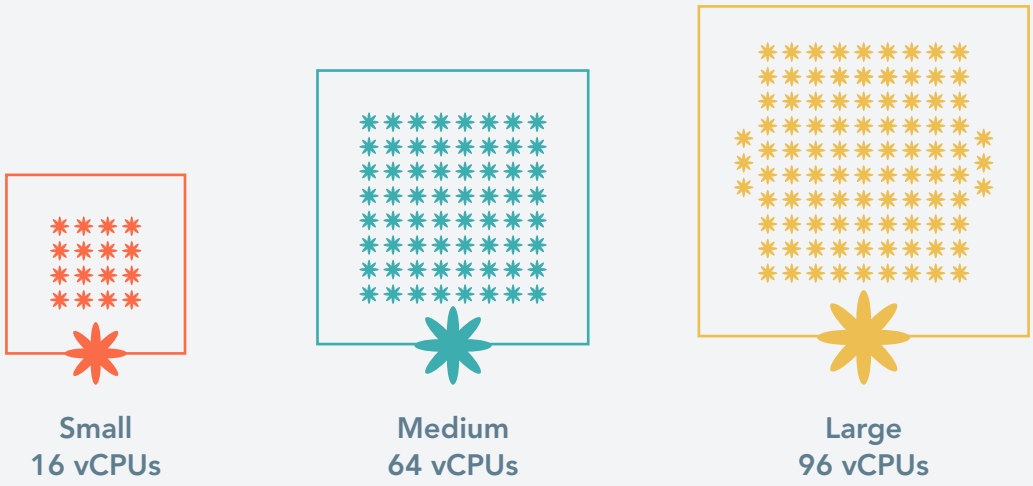

Figure 1: Key specifications for each instance size we tested. Source: Principled Technologies.

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."[1] 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."[2]

For more information, visit http://intel.com/xeonscalable.

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:[3]

- All-core turbo frequency of up to 3.5 GHz
- Always-on memory encryption with Intel Total Memory Encryption (TME)
- Intel Advanced Vector Extensions (AVX-512) instructions with 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

Assessing deep learning performance using Wide & Deep ad recommender

To test Wide & Deep recommendation engine performance, we used the TensorFlow framework. Wide & Deep uses wide linear models and deep neural networks to infer meaningful relationships between data to deliver quick recommendations based on that data. Organizations run Wide & Deep for such purposes as making ad recommendations based on browser clicks or suggesting other purchases that might interest consumers based on their history. In the following sections, we show results of testing using FP32 precision for M6i vs. M6a instances and INT8 precision for M6i vs. M5n instances. At the time of our testing, AMD EPYC processors did not support INT8 precision for Wide & Deep. For comparison, we tested M6i and M5n instances with FP32 as well, and show those results in the science behind the report.

Small instances: M6i vs. M5n

First, we compared the performance of instances with the current generation of processors to that of instances running previous-generation Intel Xeon processors. Upgrading to the latest in Intel Xeon processor technology provided a strong increase in Wide & Deep performance. Figure 2 compares the relative Wide & Deep recommender deep learning performance that the instances achieved on small, 16vCPU configurations. The M6i instances enabled by 3rd Gen Intel Xeon Scalable processors handled 1.22 times the frames per second (FPS) on the Wide & Deep benchmark as the M5n instances with 2nd Gen Intel Xeon Scalable processors.

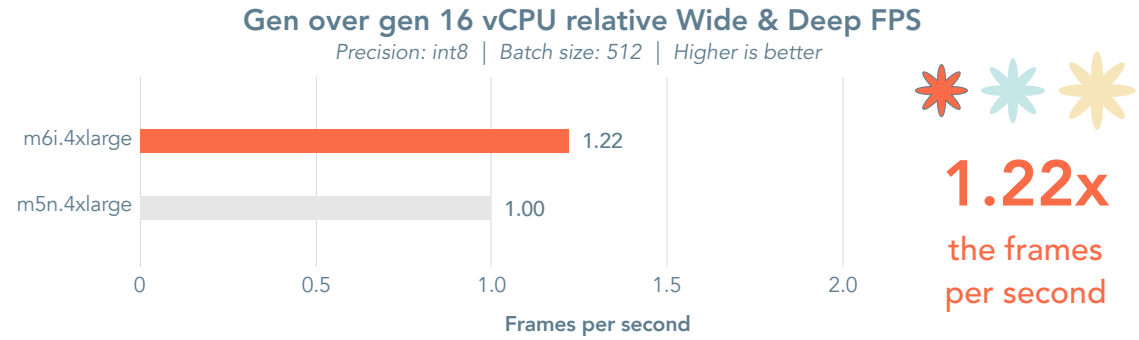

Figure 2: Relative number of frames per second that small M6i and M5n instances (16 vCPUs) handled using the Wide & Deep benchmark. Higher numbers are better. Source: Principled Technologies.

Medium instances: M6i vs. M5n

Testing with medium-sized instances showed the same performance increase from upgrading processors. Figure 3 compares the relative Wide & Deep recommender deep learning performance that the instances achieved on medium, 64vCPU configurations. The M6i instances enabled by 3rd Gen Intel Xeon Scalable processors handled 1.22 times the FPS on the Wide & Deep benchmark as the M5n instances with 2nd Gen Intel Xeon Scalable processors.

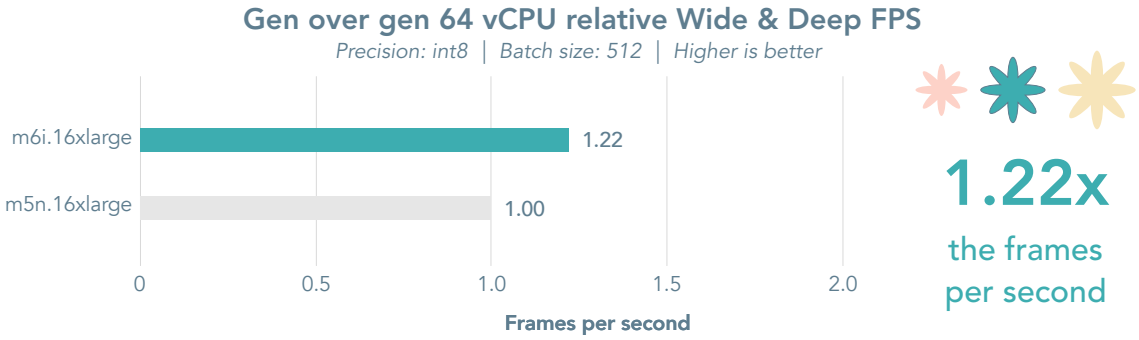

Figure 3: Relative number of frames per second that the medium-size M6i and M5n instances (64vCPUs) handled using the Wide & Deep benchmark. Higher numbers are better. Source: Principled Technologies.

Large instances: M6i vs. M5n

Organizations with larger datasets need instances with more vCPUs, and we found that large instances provided the greatest Wide & Deep performance improvement in our gen-over-gen testing. Figure 4 compares the relative Wide & Deep recommender deep learning performance that the instances achieved on large, 96vCPU configurations. The M6i instances enabled by 3rd Gen Intel Xeon Scalable processors handled 1.33 times the FPS on the Wide & Deep benchmark as the M5n instances with 2nd Gen Intel Xeon Scalable processors.

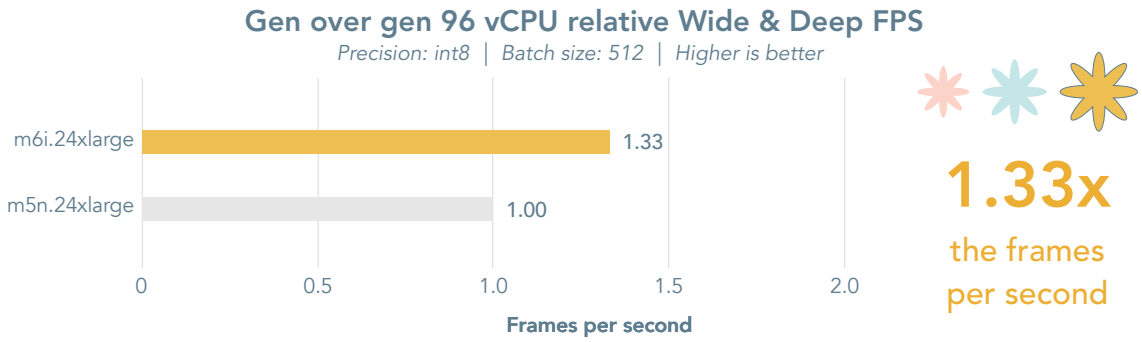

Figure 4: Relative number of frames per second that the large M6i and M5n instances (96 vCPUs) handled using the Wide & Deep benchmark. Higher numbers are better. Source: Principled Technologies.

Small instances: M6i vs. M6a

After assessing gen-over-gen performance of instances enabled by Intel Xeon Scalable processors, we compared the latest Intel Xeon processer-powered instances vs. instances with 3rd Gen AMD EPYC processors, named M6a. Using small instance sizes, M6i instances significantly outperformed their M6a counterparts. Figure 5 compares the relative Wide & Deep recommender deep learning performance that the instances achieved on small, 16vCPU configurations. The M6i instances enabled by 3rd Gen Intel Xeon Scalable processors handled 1.67 times the FPS on the Wide & Deep benchmark as the M6a instances with 3rd Gen AMD EPYC processors.

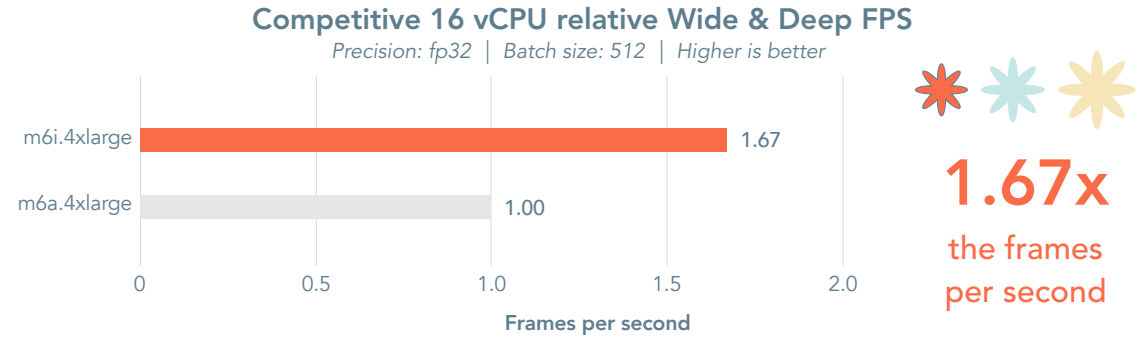

Figure 5: Relative number of frames per second that small M6i and M6a instances (16 vCPUs) handled using the Wide & Deep benchmark. Higher numbers are better. Source: Principled Technologies.

Medium instances: M6i vs. M6a

In testing with medium-sized instances, the performance improvement of M6i instances over M6a instances dropped slightly compared to what we saw in small-instance testing, but M6i instances still provided a significant boost. Figure 6 compares the relative Wide & Deep recommender deep learning performance that the instances achieved on medium, 64vCPU configurations. The M6i instances enabled by 3rd Gen Intel Xeon Scalable processors handled 1.35 times the FPS on the Wide & Deep benchmark as the M6a instances with 3rd Gen AMD EPYC processors.

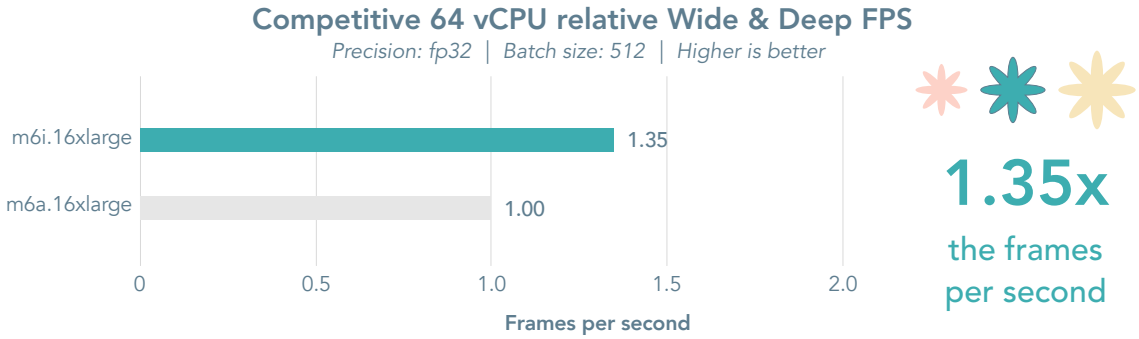

Figure 6: Relative number of frames per second that the medium-size M6i and M6a instances (64 vCPUs) handled using the Wide & Deep benchmark. Higher numbers are better. Source: Principled Technologies.

Large instances: M6i vs. M6a

The greatest performance improvement of M6i over M6a instances came when comparing large instance sizes, indicating that companies with bigger datasets and deep learning workloads can benefit from selecting M6i instances. Figure 7 compares the relative Wide & Deep recommender deep learning performance that the instances achieved on large, 96vCPU configurations. The M6i instances enabled by 3rd Gen Intel Xeon Scalable processors handled 1.75 times the FPS on the Wide & Deep benchmark as the M6a instances with 3rd Gen AMD EPYC processors.

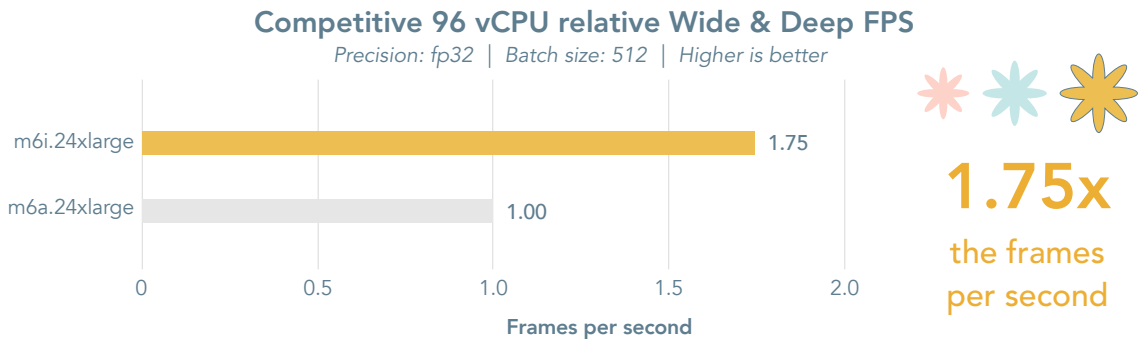

Figure 7: Relative number of frames per second that the large M6i and M6a instances (96 vCPUs) handled using the Wide & Deep benchmark. Higher numbers are better. Source: Principled Technologies.

Conclusion

Our test results show that AWS M6i instances featuring 3rd Gen Intel Xeon Scalable processors with Intel Deep Learning Boost sped up deep learning inference performance for recommendation models over M5n instances with previous-gen processors, with even greater performance improvements over M6a instances with 3rd Gen AMD EPYC processors. This means that by selecting AWS EC2 M6i instances with 3rd Gen Intel Xeon Scalable processors, your organization could get deep learning insights sooner and make real-time recommendations faster.

1. 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.
2. Intel, "3rd Gen Intel® Xeon® Scalable Processors."
3. Amazon, "Amazon EC2 M6i Instances", accessed December 14, 2021, https://aws.amazon.com/ec2/instance-types/m6i//.

Read the science behind this report at https://facts.pt/WItf6C8 ▶

Principled Technologies®

Facts matter.®

This project was commissioned by Intel.