Achieve up to 1.33 Times The Wide & Deep Recommender Performance by Selecting Amazon® M6i Instances Featuring 3rd Gen Intel® Xeon® Scalable Processors

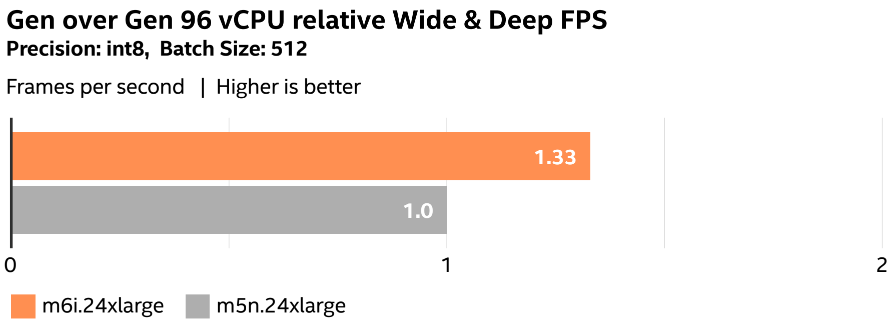
Across Different Instance Sizes, M6i Instances Performed More Inference Operations per Second than M5n Instances with 2nd Gen Intel Xeon Scalable Processors

One application of deep learning inference is generating recommendations for shoppers visiting your website. As they browse, you collect data on the products that interest them. You can use this data, along with patterns from other visitors, to infer other products they might like and recommend them. To perform this data analysis in real time to boost your company's sales, you can use deep learning workloads—specifically Wide & Deep recommendation engines.

We compared the Wide & Deep inference performance of two Amazon Web Services (AWS) EC2 cloud instance types with different processor configurations: M6i instances with 3rd Gen Intel® Xeon® Scalable processors and M5n instances with 2nd Gen Intel Xeon Scalable processors. We found that small-, medium-, and large-sized M6i instances with 3rd Gen Intel Xeon Scalable processors outperformed their M5n counterparts. This means that businesses that want to deliver speedier recommendations with Wide & Deep inference workloads can do so by selecting M6i instances.

Large M6i Instances With 96 vCPUs

To test Wide & Deep recommendation engine performance of the two AWS instance series, we used the TensorFlow framework. As Figure 1 shows, the 96 vCPU m6i.24xlarge instances enabled by 3rd Gen Intel Xeon Scalable processors handled 1.33 times the frames per second (FPS) on the Wide & Deep benchmark as the m5n.24xlarge instances with 2nd Gen Intel Xeon Scalable processors.



Gen over Gen 96 vCPU relative Wide & Deep FPS
Precision: int8, Batch Size: 512

Figure 1. Number of frames per second achieved by an m6i.24xlarge instance cluster with 3rd Gen Intel Xeon Scalable processors and by an m5n.24xlarge cluster with 2nd Gen Intel Xeon Scalable processors. Testing used int8 precision and 512 batch size. Higher is better.

Wide & Deep

Process Up to 1.33 Times the Frames per Second on 96 vCPU m6i.24xlarge Instances Featuring 3rd Gen Intel Xeon Scalable Processors

vs. m5n.24xlarge instances

Process Up to 1.22 Times the Frames per Second on 64 vCPU m6i.16xlarge Instances Featuring 3rd Gen Intel Xeon Scalable Processors

vs. m5n.16xlarge instances

Process Up to 1.22 Times the Frames per Second on 16 vCPU m6i.4xlarge Instances Featuring 3rd Gen Intel Xeon Scalable Processors

vs. m5n.4xlarge instances

Medium-Sized M6i Instances With 64 vCPUs

As Figure 2 shows, the 64 vCPU m6i.16xlarge instances enabled by 3rd Gen Intel® Xeon® Scalable processors handled 1.22 times the FPS as the m5n.16xlarge instances with 2nd Gen Intel Xeon Scalable processors..

Small M6i Instances With 16 vCPUs

As Figure 3 shows, the 16 vCPU m6i.4xlarge instances enabled by 3rd Gen Intel Xeon Scalable processors handled 1.22 times the FPS as the m5n.4xlarge instances with 2nd Gen Intel Xeon Scalable processors.

Conclusion

We tested Wide & Deep recommendation engine performance on two AWS instance series: M6i instances featuring 3rd Gen Intel Xeon Scalable processors and M5n instances featuring 2nd Gen Intel Xeon Scalable processors. At three different sizes, the M6i instances processed more frames per second, as much as 1.33 times as many. To boost your sales, run your Wide & Deep recommendation workloads on Amazon M6i instances with 3rd Gen Intel Xeon Scalable processors.

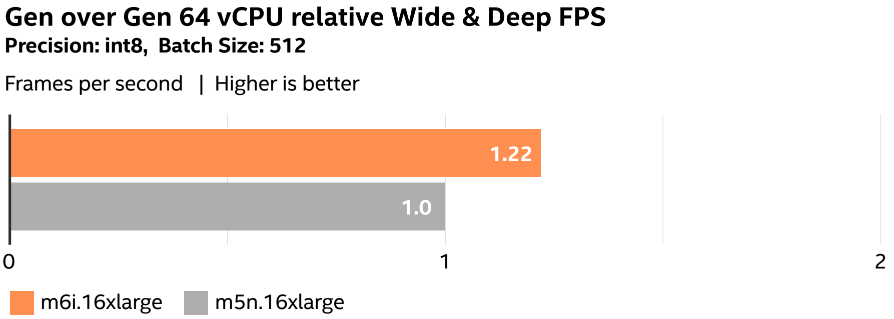

Gen over Gen 64 vCPU relative Wide & Deep FPS
Precision: int8, Batch Size: 512

Figure 2. Number of frames per second achieved by an m6i.16xlarge instance cluster with 3rd Gen Intel Xeon Scalable processors and by an m5n.16xlarge cluster with 2nd Gen Intel Xeon Scalable processors. Testing used int8 precision and 512 batch size. Higher is better.

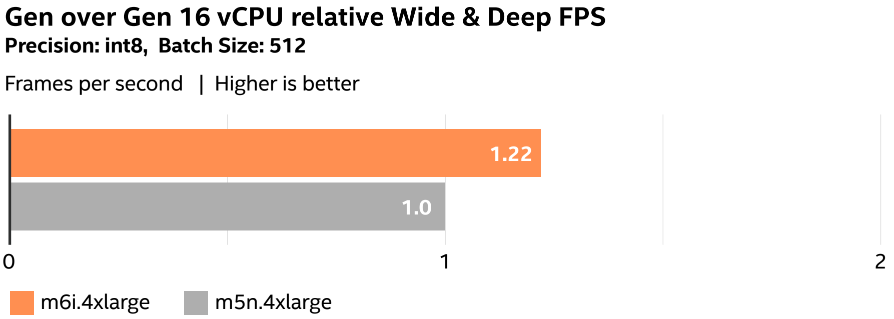

Gen over Gen 16 vCPU relative Wide & Deep FPS
Precision: int8, Batch Size: 512

Figure 3. Number of frames per second achieved by an m6i.4xlarge instance cluster with 3rd Gen Intel Xeon Scalable processors and by an m5n.4xlarge cluster with 2nd Gen Intel Xeon Scalable processors. Testing used int8 precision and 512 batch size. Higher is better.

Learn More

To begin running your Wide & Deep recommendation workloads on Amazon M6i instances with 3rd Gen Intel Xeon Scalable processors, visit https://aws.amazon.com/ec2/instance-types/m6i/.

For complete test details and results showing how these 3rd Gen Intel Xeon Scalable processor-enabled instances fared against instances with 2nd Gen Intel Xeon Scalable processors, read the report at https://facts.pt/ZlqeNXb.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure. Your costs and results may vary.

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

Printed in USA 0322/JO/PT/PDF US001 ♲ Please Recycle