Amazon® M6i Instances Featuring 3rd Gen Intel® Xeon® Scalable Processors Delivered up to 1.75 Times the Wide & Deep Recommender Performance

Wide & Deep

Process Up to 1.75 Times as Many Frames per Second on 96 vCPU m6i.24xlarge Instances Featuring 3rd Gen Intel Xeon Scalable Processors

vs. m6a.24xlarge instances

Process Up to 1.35 Times as Many Frames per Second on 64 vCPU m6i.16xlarge Instances Featuring 3rd Gen Intel Xeon Scalable Processors

vs. m6a.16xlarge instances

Process Up to 1.67 Times as Many Frames per Second on 16 vCPU m6i.4xlarge Instances Featuring 3rd Gen Intel Xeon Scalable Processors

vs. m6a.4xlarge instances

Across Different Instance Sizes, M6i Instances Performed More Inference Operations per Second than M6a Instances with 3rd Gen AMD EPYC processors

If you run an ecommerce site, you might be interested in improving sales with a deep learning workload such as a Wide & Deep recommendation engine. These applications analyze data collected as visitors shop on your site, and generate recommendations of additional products that might interest your customers. By running deep learning applications on cloud instances with powerful underlying hardware, you can deliver these recommendations more quickly.

Testing 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 M6a instances with 3rd Gen AMD EPYC processors. Small, medium-sized, and large M6i instances delivered better performance—measured in frames per second (FPS)—than their M6a counterparts did. By selecting these higher-performing M6i instances for your Wide & Deep inference workloads, your website could deliver speedier recommendations.

Large 96 vCPU Instances

Testing used the TensorFlow framework to evaluate the Wide & Deep recommendation engine performance of the two AWS instance series. As Figure 1 shows, the 96 vCPU m6i.24xlarge instances enabled by 3rd Gen Intel Xeon Scalable processors processed 1.75 times as many FPS on the Wide & Deep benchmark as the m6a.24xlarge instances with 3rd Gen AMD EPYC processors.

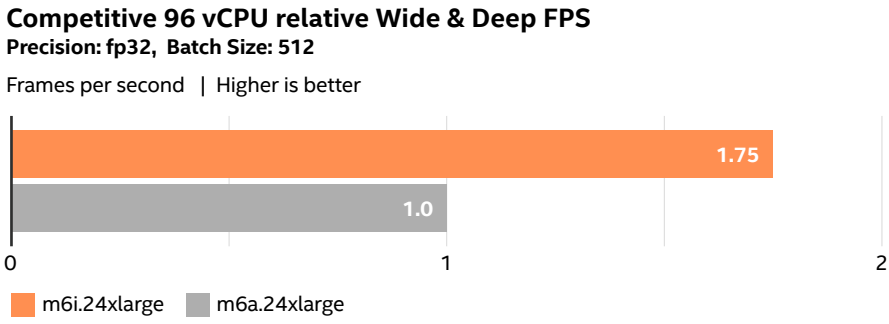

Figure 1. Number of frames per second achieved by an m6i.24xlarge instance cluster with 3rd Gen Intel Xeon Scalable processors and by an m6a.24xlarge cluster with 3rd Gen AMD EPYC processors. Testing used fp32 precision and 512 batch size. Higher is better.

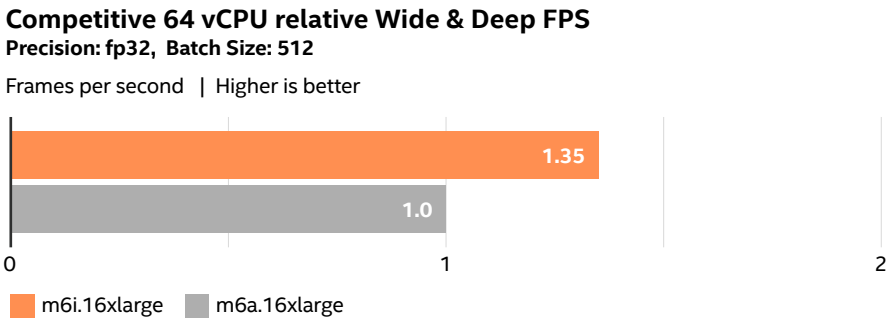

Figure 2. Number of frames per second achieved by an m6i.16xlarge instance cluster with 3rd Gen Intel Xeon Scalable processors and by an m6a.16xlarge cluster with 3rd Gen AMD EPYC processors. Testing used fp32 precision and 512 batch size. Higher is better.

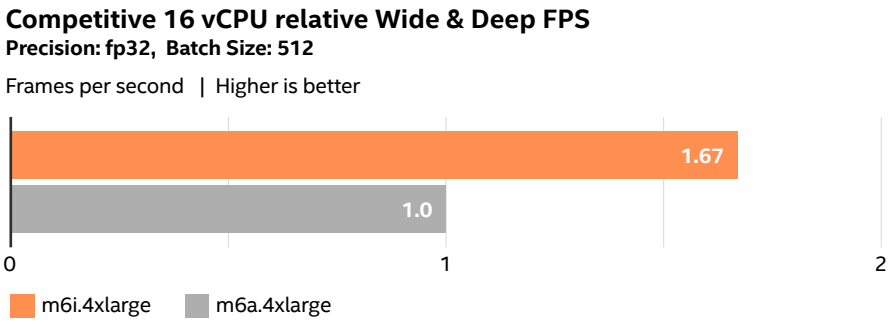

Figure 3. Number of frames per second achieved by an m6i.4xlarge instance cluster with Intel Xeon Scalable processors and by an m6a.4xlarge cluster with 3rd Gen AMD EPYC processors. Testing used fp32 precision and 512 batch size. Higher is better.

Medium-Sized 64 vCPU Instances

As Figure 2 shows, the 64 vCPU m6i.16xlarge instances enabled by 3rd Gen Intel® Xeon® Scalable processors processed 1.35 times as many FPS as the m6a.16xlarge instances with 3rd Gen AMD EPYC processors did.

Small 16 vCPU Instances

As Figure 3 shows, the 16 vCPU m6i.4xlarge instances enabled by 3rd Gen Intel Xeon Scalable processors processed 1.67 times as many FPS as the m6a.4xlarge instances with 3rd Gen AMD EPYC processors did.

Conclusion

Testing of Wide & Deep recommendation engine performance on two AWS instance series revealed that M6i instances featuring 3rd Gen Intel Xeon Scalable processors outperformed M6a instances featuring 3rd Gen AMD EPYC processors. The M6i instances processed up to 1.75 times as many frames per second, which could allow the application to generate customer recommendations more quickly and boost your sales more effectively.

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 3rd Gen AMD EPYC processors, read the report at https://facts.pt/ZlqeNXb.

Printed in USA 0322/JO/PT/PDF US002 ♻ Please Recycle