Achieve up to 64% Better BERT-Large Inference Work Performance by Selecting AWS M6i Instances Featuring 3rd Gen Intel® Xeon® Scalable Processors

BERT-Large

Reap up to 64% better BERT-Large performance on 64-vCPU m6i.16xlarge instances featuring 3rd Gen Intel Xeon Scalable processors

vs. m5n.16xlarge instances

Process up to 40% higher BERT-Large throughput on 32-vCPU m6i.8xlarge instances featuring 3rd Gen Intel Xeon Scalable processors

vs. m5n.8xlarge instances

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

Companies use natural language machine learning inference workloads for a variety of business applications, such as chatbots that analyze text typed by customers and other users. This type of work puts great demands on compute resources, making it very important to select high-performing cloud instances.

BERT is a general-purpose natural language processing (NLP) model we chose to measure the performance of two Amazon Web Services (AWS) EC2 cloud instance types. We tested two sizes of M6i instances with 3rd Gen Intel Xeon Scalable processors and M5n instances with 2nd Gen Intel Xeon Scalable processors. We found that both 32 vCPU and 64 vCPU M6i instances with 3rd Gen Intel Xeon Scalable processors outperformed their M5n counterparts. Based on these findings, businesses can deliver a speedier experience to their users by opting for M6i instances.

M6i Instances With 64 vCPUs

To compare the BERT-Large inference performance of the two AWS instance series, we used the TensorFlow framework. As Figure 1 shows, the 64-vCPU m6i.16xlarge instance enabled by 3rd Gen Intel Xeon Scalable processors delivered 64% higher throughput than the m5n.16xlarge instance with 2nd Gen Intel Xeon Scalable processors.

Relative 64-vCPU BERT-Large INT8 Inference Performance

Speedup | Higher is better

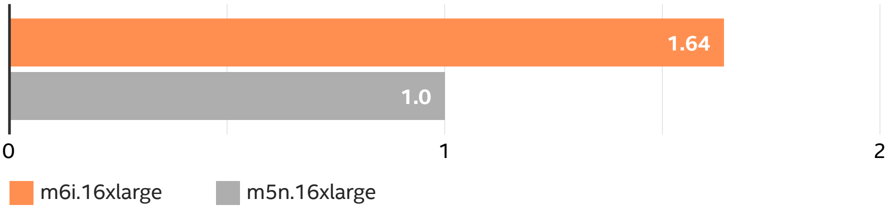

Figure 1. BERT-Large inference performance achieved by an m6i.16xlarge instance with 3rd Gen Intel Xeon Scalable processors and by an m5n.16xlarge instance with 2nd Gen Intel Xeon Scalable processors. Testing used INT8 precision, batch size of 1, and sequence length of 384. Higher is better.

M6i Instances With 32 vCPUs

As Figure 2 shows, the 32-vCPU m6i.8xlarge instance enabled by 3rd Gen Intel® Xeon® Scalable processors delivered 40% higher throughput than the m5n.8xlarge instance with 2nd Gen Intel Xeon Scalable processors.

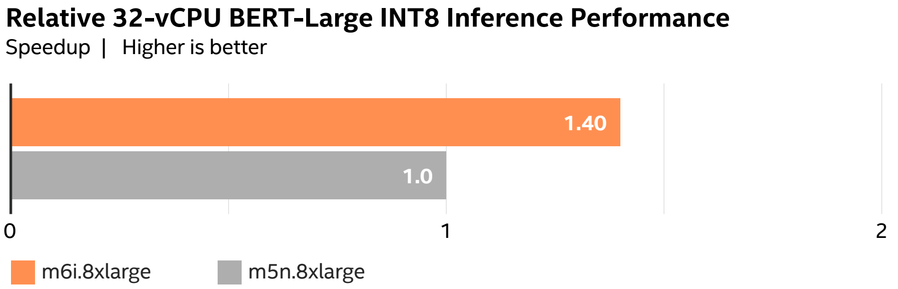

Figure 2. BERT-Large inference performance achieved by an m6i.8xlarge instance with 3rd Gen Intel Xeon Scalable processors and by an m5n.8xlarge instance with 2nd Gen Intel Xeon Scalable processors. Testing used INT8 precision, batch size of 1, and sequence length of 384. Higher is better.

Conclusion

We tested BERT-Large natural language processing inference performance of two AWS instance series: M6i instances featuring 3rd Gen Intel Xeon Scalable processors and M5n instances featuring 2nd Gen Intel Xeon Scalable processors. At two different sizes, the M6i instances outperformed the M5n instances by as much as 64%. To deliver a speedier experience to your customers and other users, run your NLP inference workloads on Amazon M6i instances with 3rd Gen Intel Xeon Scalable processors.

Learn More

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

Single VM tests by Intel on 11/30/2021. All VMs configured with Ubuntu 20.04 LTS, 5.11.0-1022-aws, EBS storage, GCC=8.4.0, Python=3.6.9, tensorflow=2.5.0, Docker=20.10.7, containerd=1.5.5, BERT model, batch size 1, sequence length 384, INT8 precision. Instance details: m6i.8xlarge, 32vcpus, Intel(R) Xeon(R) Platinum 8375C CPU @ 2.90GHz, 128 GB total DDR4 memory; m5n.8xlarge, 32vcpus, Intel(R) Xeon(R) Platinum 8259CL CPU @ 2.50GHz, 128 GB total DDR4 memory; m6i.16xlarge, 64vcpus, Intel® Xeon® Platinum 8375C CPU @ 2.90GHz, 256 GB total DDR4 memory; m5n.16xlarge, 64vcpus, Intel® Xeon® Platinum 8259CL CPU @ 2.50GHz, 256 GB total DDR4 memory

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 securev. Your costs and results may vary.
Intel technologies may require enabled hardware, software or service activation.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

Printed in USA 0522/JO/PT/PDF US001 ♲ Please Recycle