Achieve up to 4.96 Times the BERT-Large Inference Work by Selecting AWS M6i Instances Featuring 3rd Gen Intel® Xeon® Scalable Processors

M6i Instances Performed More Inference Work than M6g Instances Featuring AWS Graviton2 Processors

Natural language machine learning inference workloads underlie chatbots and other business applications. As these workloads analyze text typed by customers and other users, they can put heavy demands on compute resources. This makes it important to choose cloud instances that deliver high performance.

BERT-Large 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 M6g instances with AWS Graviton2 processors. We found that both 32-vCPU and 64-vCPU M6i instances with 3rd Gen Intel Xeon Scalable processors outperformed their M6g counterparts. Our findings illustrate that businesses can deliver a speedier experience to their users by opting for M6i instances. Additionally, at the time of publication, while the M6i series VMs cost 24.6% more than the M6g series VMs, the M6i instances—with performance up to 4.96 times the throughput—offer significantly better performance per dollar.

M6i Instances With 32 vCPUs

To compare the BERT-Large inference performance of the two AWS instance series, we used the TensorFlow framework. We tested two precision levels: FP32, which both series of VMs support, and INT8, which only the M6i series supports with the models we used. As Figure 1 shows, the 32-vCPU m6i.8xlarge instances using INT8 precision delivered 4.96 times the performance of the m6g.8xlarge instances using FP32 precision.

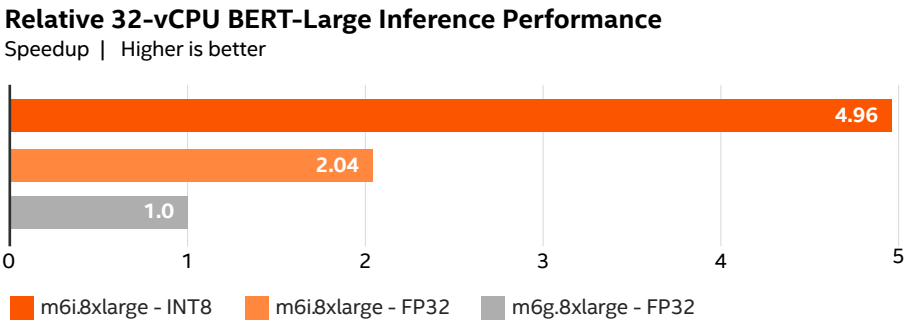

Figure 1. BERT-Large inference performance achieved by an m6i.8xlarge instance cluster with 3rd Gen Intel Xeon Scalable processors and by an m6g.8xlarge instance cluster with AWS Graviton2 processors. Higher is better.

BERT-Large

Get up to 4.96 times the BERT-Large inference work (INT8 precision) with 32-vCPU m6i.8xlarge instances featuring 3rd Gen Intel Xeon Scalable processors

vs. FP32 precision with m6g.8xlarge instances

Get up to 3.07 times the BERT-Large inference work (INT8 precision) with 64-vCPU m6i.16xlarge instances featuring 3rd Gen Intel Xeon Scalable processors

vs. FP32 precision with m6g.16xlarge instances

M6i Instances With 64 vCPUs

As Figure 2 shows, the 64-vCPU m6i.16xlarge instances with 3rd Gen Intel® Xeon® Scalable processors using INT8 precision delivered 3.07 times the performance of the m6g.16xlarge instances with AWS Graviton2 processors using FP32 precision. Note: The BERT-Large model we used for AWS Graviton2 processors does not support INT8 on TensorFlow.

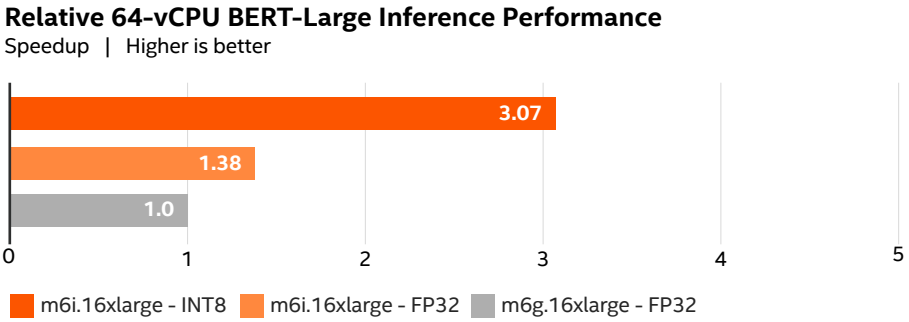

Figure 2. BERT-Large inference performance achieved by an m6i.16xlarge instance cluster with 3rd Gen Intel Xeon Scalable processors and by an m6g.16xlarge instance cluster with AWS Graviton2 processors. Higher is better.

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 M6g instances featuring AWS Graviton2 processors. At two different sizes, the M6i instances outperformed the M6g instances, achieving up to 4.96 times the inference work. To deliver a speedier experience to your customers and other users, run your NLP inference workloads on AWS M6i instances with 3rd Gen Intel Xeon Scalable processors.

Learn More

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

Single VM tests by Intel on 11/10/2021 and 12/01/2021. 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, FP32 and INT8 precision. Instance details: m6i.8xlarge, 32vcpus, Intel® Xeon® Platinum 8375C CPU @ 2.90GHz, 128 GB total DDR4 memory; m6g.8xlarge, 32vcpus, ARM Neovers N1, Arm v8.2 @2.5GHz, 128 GB total DDR4 memory; m6i.16xlarge, 64vcpus, Intel® Xeon® Platinum 8375C CPU @ 2.90GHz, 256 GB total DDR4 memory; m6g.16xlarge, 64vcpus, ARM Neovers N1, Arm v8.2 @2.5GHz, 256 GB total DDR4 memory.

Printed in USA 0722/JO/PT/PDF US002 ♻ Please Recycle