Deliver a Better Customer Support Chatbot Experience with Higher-Value AWS EC2 M7i Instances

RoBERTa

Accelerate Natural Language Processing by up to 10.65x

Reap a Significantly Higher Value

In Natural Language Processing (NLP) Testing, These Instances with 4th Gen Intel® Xeon® Scalable® Processors Not Only Outperformed AWS M7g Instances with AWS Graviton3 Processors, but They Delivered up to 8.62 Times the Performance per Dollar

Many applications, from search result suggestions to text autocorrections to text generation chatbots, use natural language processing (NLP). A great deal of compute power is necessary to power deep learning frameworks for NLP, such as Bidirectional Encoder Representations from Transformers (BERT) and its variations. If your business relies on these workloads and wants to run them in the cloud, selecting the right instance for the job is important because it can lead to speedier text analysis and predictions and can even lead to reduced spending on cloud services.

To help companies shopping for an Amazon Web Services (AWS) Elastic Cloud Compute (EC2) instance for their NLP workloads, Intel commissioned Principled Technologies (PT) to measure the deep learning performance of M7i instances enabled by 4th Gen Intel® Xeon® Scalable processors and M7g instances enabled by AWS Graviton3 processors. Using the Intel® Extension for PyTorch tool, testing compared the instances at different sizes—with 4, 16, and 64 vCPUs—to shed light on the potential performance for businesses with different instance needs. Tests used the RoBERTa model, a variant of BERT, and optimized the model for each processor type. The 4th Gen Intel Xeon Scalable processors featured Intel Advanced Matrix Extensions (Intel AMX) with support for 8-bit integer, bfloat16, and float16. Results of the testing demonstrate that M7i instances featuring 4th Gen Intel Xeon Scalable processors achieved as much as 10.65 times the NLP performance and delivered up to 8.62 times the performance per dollar of the M7g instances with AWS Graviton3 processors.

Intel Workload Proof Series: Natural Language Processing Performance on AWS EC2 M7i vs. M7g

See backup for workloads and configurations. Results may vary.

Spotlight on BF16 Precision

First, PT measured RoBERTa performance at BF16 precision at batch sizes of 1 and 32. At both batch sizes and across vCPU counts, AWS EC2 M7i instances achieved a higher throughput rate, in terms of how many sentences per second each instance analyzed, than M7g instances. As Figure 1 shows, M7i instances delivered up to 10.65 times the throughput.

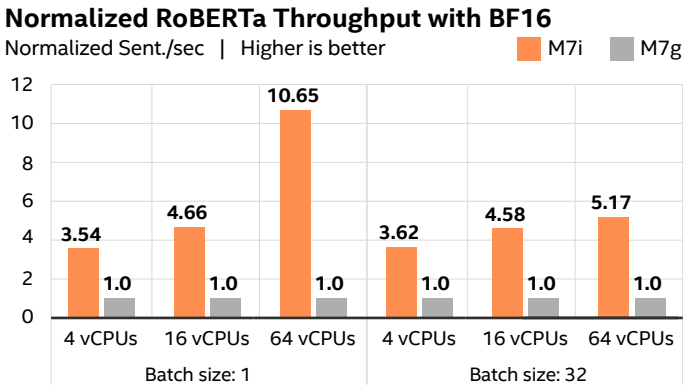

Figure 1. Relative RoBERTa performance of M7i instances, in sentences analyzed per second, compared to M7g instances using BF16 precision and batch sizes 1 and 32. Higher is better. Source: Principled Technologies.

In addition to analyzing sentences at a faster rate, M7i instances also provide better value. When we divide the RoBERTa throughput of each instance by its hourly price, we see that M7i instances with Intel® processors performed more RoBERTa work for every dollar they cost. At BF16 precision, M7i instances delivered up to 8.62 times the throughput per dollar of M7g with AWS Graviton3 processors instances, as Figure 2 shows.

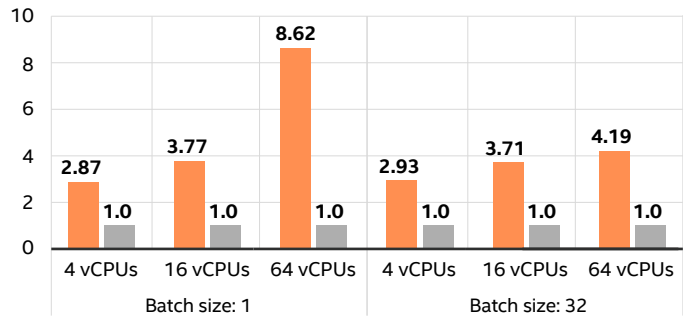

Figure 2. Relative throughput per dollar of M7i instances compared to M7g instances using BF16 precision and batch sizes 1 and 32. Higher is better. Source: Principled Technologies.

Spotlight on FP32 Precision

In the second set of tests, PT measured RoBERTa performance at FP32 precision at two batch sizes, 1 and 32. As we saw with BF16 precision, at both batch sizes and across vCPU counts, AWS EC2 M7i instances with Intel processors outperformed M7g instances with Graviton3 processors. As Figure 3 shows, M7i instances delivered up to 4.25 times the throughput.

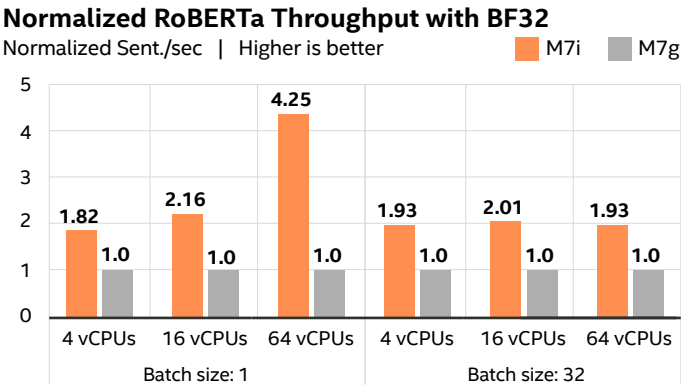

Figure 3. Relative RoBERTa performance of M7i instances, in sentences analyzed per second, compared to M7g instances using FP32 precision and batch sizes 1 and 32. Higher is better. Source: Principled Technologies.

As we saw with BF16 precision, the superior performance of the M7i instances enables them to be more cost-effective. As Figure 4 shows, at FP32 precision, M7i instances delivered up to 3.44 times the throughput per dollar of M7g instances.

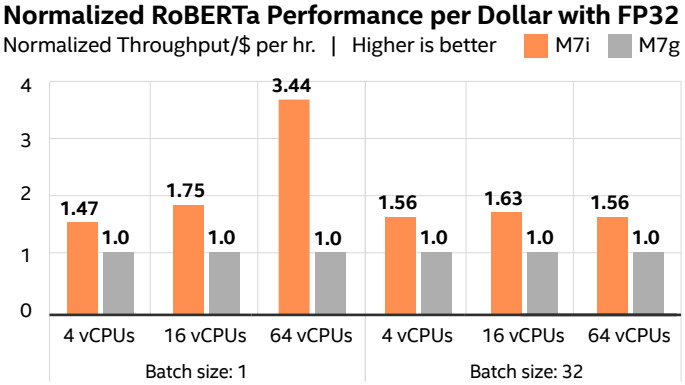

Figure 4. Relative throughput per dollar of M7i instances compared to M7g instances using FP32 precision and batch sizes 1 and 32. Higher is better. Source: Principled Technologies.

Conclusion

If you run NLP workloads to support important work such as customer support chatbots, selecting a high-performing cloud instance could enable your organization to handle more NLP work. In turn, this could save money by doing more work with the same number of instances or using fewer instances to perform a fixed amount of work. In PT testing of the RoBERTa performance of three sizes of AWS EC2 instances, they found that M7i instances with Intel® processors delivered up to 10.65 times the RoBERTa throughput of M7g instances with Graviton processors. This performance advantage combined with the hourly rates for the two instances yields a pricing advantage: M7i Instances with Intel processors delivered as much as 8.62 times the RoBERTa performance per dollar of M7g instances with Graviton processors.

For businesses who want to maximize the return on their cloud investment, AWS EC2 M7i instances are an effective and cost-effective solution.

Learn More

To begin running your Natural Language Processing workloads on AWS EC2 M7i instances, visit https://aws.amazon.com/ec2/instance-types/m7i/.

To learn more about testing and configurations, see https://facts.pt/RfrK3Rr.

Printed in USA 1123/HM/PT/PDF US001 ♻ Please Recycle