4 vCPUs

Up to 5.2x the queries per second

vs. M6a instances

8 vCPUs

Up to 5.1x the queries per second

vs. M6a instances

16 vCPUs

Up to 6.4x the queries per second

vs. M6a instances

AWS EC2 M6i instances featuring 3rd Gen Intel Xeon Scalable processors offered better BERT machine learning performance

vs. M5n instances with 2nd Gen Intel Xeon Scalable processors and M6a instances with 3rd Gen AMD EPYC processors

Many machine learning workloads involve sorting, analyzing, and making relationships between images, but how can organizations quickly make sense of large amounts of text? Bidirectional Encoder Representations from Transformers (BERT) is a machine learning framework for natural language processing (NLP). To analyze text, BERT looks at all the words around a given word to put it in the correct context. This allows applications such as search engines to predict sentences, answer questions, or generate conversational responses.

Using Intel optimization for TensorFlow and ZenDNN integrated with TensorFlow, we compared the BERT machine learning performance of three types of Amazon Web Services (AWS) EC2 series instances: M6i instances with 3rd Gen Intel® Xeon® Scalable processors featuring Intel DL Boost with Vector Neural Network Instructions, M5n instances with 2nd Gen Intel Xeon Scalable processors, and M6a instances with 3rd Gen AMD EPYC™ processors.

In tests at multiple instance sizes, AWS M6i instances offered up to 45 percent better BERT performance on a benchmark from the Intel Model Zoo than the M5n instances with previous-gen processors and up to 6.4 times the BERT performance compared to M6a instances with 3rd Gen AMD EPYC processors. This means that organizations running similar BERT workloads in the cloud could get better performance per instance by choosing M6i instances featuring 3rd Gen Intel Xeon Scalable processors.

How we tested

We purchased three sets of instances from three general-purpose AWS EC2 series:

- M6i instances featuring 3rd Gen Intel Xeon Platinum 8375C processors (Ice Lake)
- M5n instances featuring 2nd Gen Intel Xeon Platinum 8259CL processors (Cascade Lake)
- M6a instances featuring 3rd Gen AMD EPYC 7R13 processors (Milan)

We ran each instance in the US East 1 region.

Figure 1 shows the specifications for the instances that we chose. To show how businesses of various sizes with different machine learning demands can benefit from choosing M6i instances, we tested instances with 4 vCPUs, 8 vCPUs, and 16 vCPUs. To account for different types of datasets, we ran tests using a small batch size of 1 and a large batch size of 32—where batch size is the number of samples that go through the neural network at a time. In this report, we present the comparisons between M6i and M5n instances first, and then present the comparisons between M6i and M6a instances. (Note: For additional test results on even larger instances, see the science behind the report.)

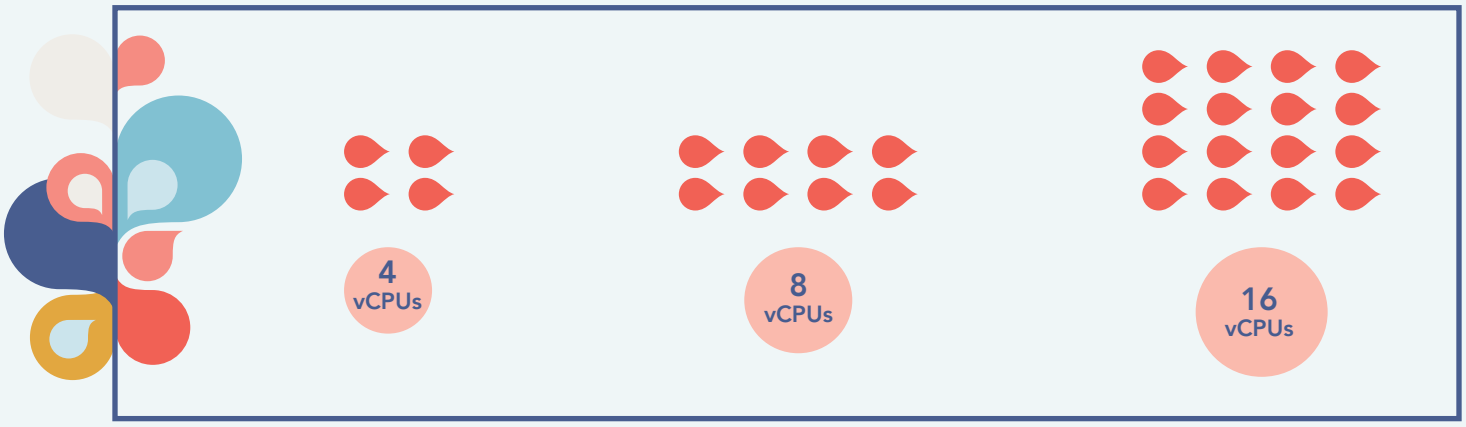


Figure 1: Key specifications for each instance size we tested. Source: Principled Technologies.

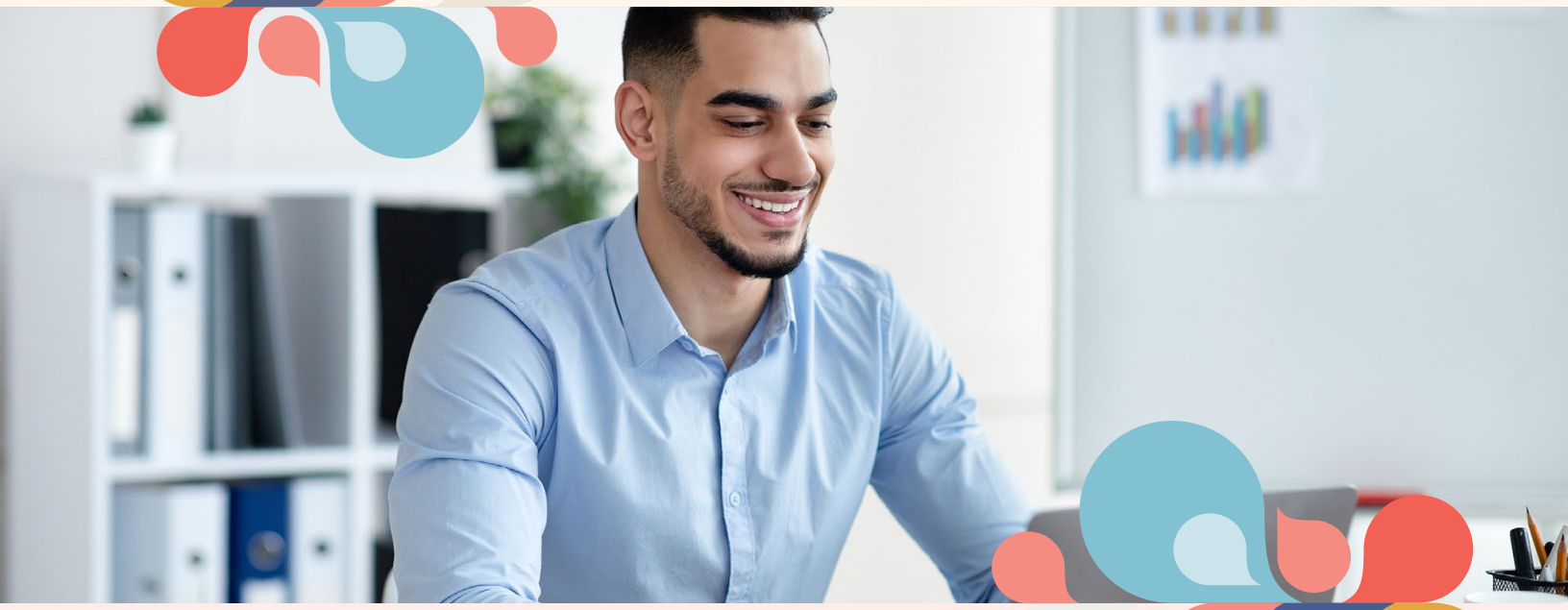
Testing BERT performance in the cloud

The BERT framework, which was trained on text from the English language Wikipedia with over 2.5 million words, works by turning text into numbers to sort, analyze, and make predictions about that text.[1] Depending on the dataset on which an organization needs to run BERT machine leaning, the size of the AWS instances they choose will vary. To account for these different needs, we tested using two batch sizes across three different instance sizes. We used a BERT benchmark from Intel Model Zoo, which offers a range of machine learning models and tools. At the time of our testing, AMD EPYC processors did not support INT8 precision for BERT, so we present FP32 precision results for M6i instances as well for comparison. In all three, the M6i instances enabled by 3rd Gen Intel Xeon Scalable processors outperformed both the previous-gen M5n instances and the current-gen M6a instances.

About 3rd Generation Intel Xeon Scalable processors

According to Intel, 3rd Generation Intel Xeon Scalable processors are "[o]ptimized for cloud, enterprise, HPC, network, security, and IoT workloads with 8 to 40 powerful cores and a wide range of frequency, feature, and power levels."[2] Intel continues to offer many models from the Platinum, Gold, Silver, and Bronze processor lines that they "designed through decades of innovation for the most common workload requirements.[3]

For more information, visit http://intel.com/xeonscalable.



Why choose M6i instances with 3rd Gen Intel Xeon Scalable processors?

New M6i instances with 3rd Gen Intel Xeon Scalable processors offer the following:[4]

- All-core turbo frequency of up to 3.5 GHz
- Always-on memory encryption with Intel Total Memory Encryption (TME)
- Intel DL Boost with Vector Neural Network Instructions (VNNI) that accelerate INT8 performance
- Intel Advanced Vector Extensions 512 (Intel AVX-512) instructions for demanding machine learning workloads
- Support for up to 128 vCPUs and 512 GB of memory per instance
- Up to 50Gbps networking

Instances with 4 vCPUs: M6i vs. M5n

First, we compared BERT performance on smaller instances, looking at the relative amount of text the instance types analyzed on 4vCPU configurations. As Figure 2 shows, M6i instances enabled by 3rd Gen Intel Xeon Scalable processors analyzed up to 18 percent more examples per second than the M5n instances with 2nd Gen Intel Xeon Scalable processors.

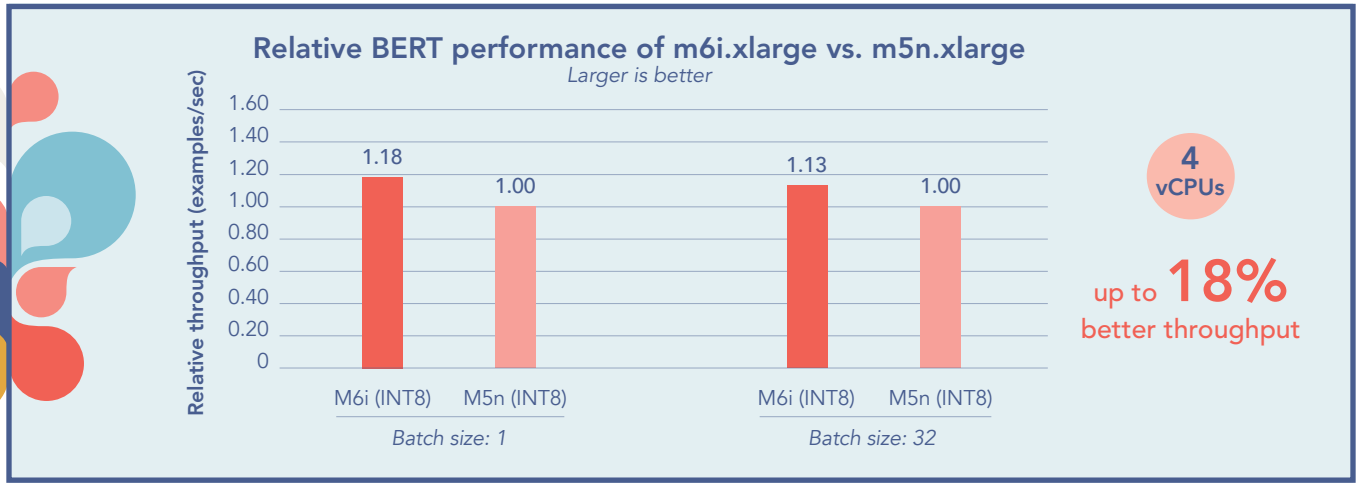


Figure 2: Relative BERT performance for M6i and M5n instances using 4 vCPUs. Higher numbers are better. Source: Principled Technologies.

Instances with 8 vCPUs: M6i vs. M5n

When we doubled the instance size to 8 vCPUs, M6i instances delivered a similar performance increase over previous-gen M5n instances. Figure 3 compares the relative amount of text the instance types analyzed on 8vCPU configurations. The M6i instances enabled by 3rd Gen Intel Xeon Scalable processors analyzed up to 11 percent more examples per second than the M5n instances with 2nd Gen Intel Xeon Scalable processors.

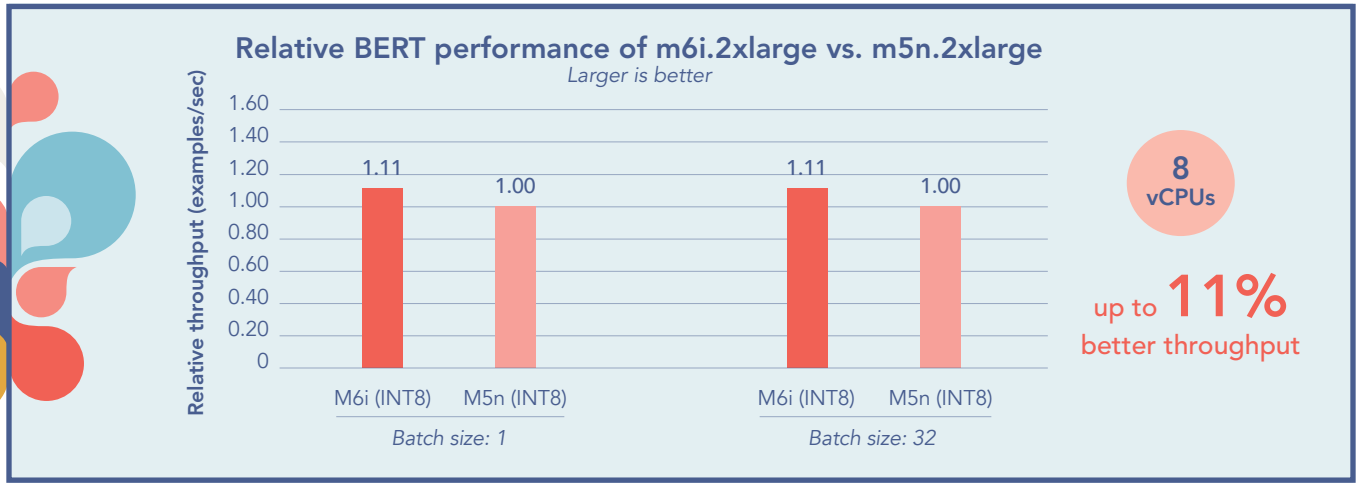


Figure 3: Relative BERT performance for M6i and M5n instances using 8 vCPUs. Higher numbers are better. Source: Principled Technologies.

Instances with 16 vCPUs: M6i vs. M5n

As Figure 4 shows, M6i instances offered the greatest relative BERT performance increase over previous-gen M5n instances using larger 16vCPU configurations. The M6i instances enabled by 3rd Gen Intel Xeon Scalable processors analyzed up to 45 percent more examples per second than the M5n instances with 2nd Gen Intel Xeon Scalable processors. By improving textual data analysis throughput by 45 percent, organizations could reduce the number of instances they need to purchase and manage when they select the M6i instance type.

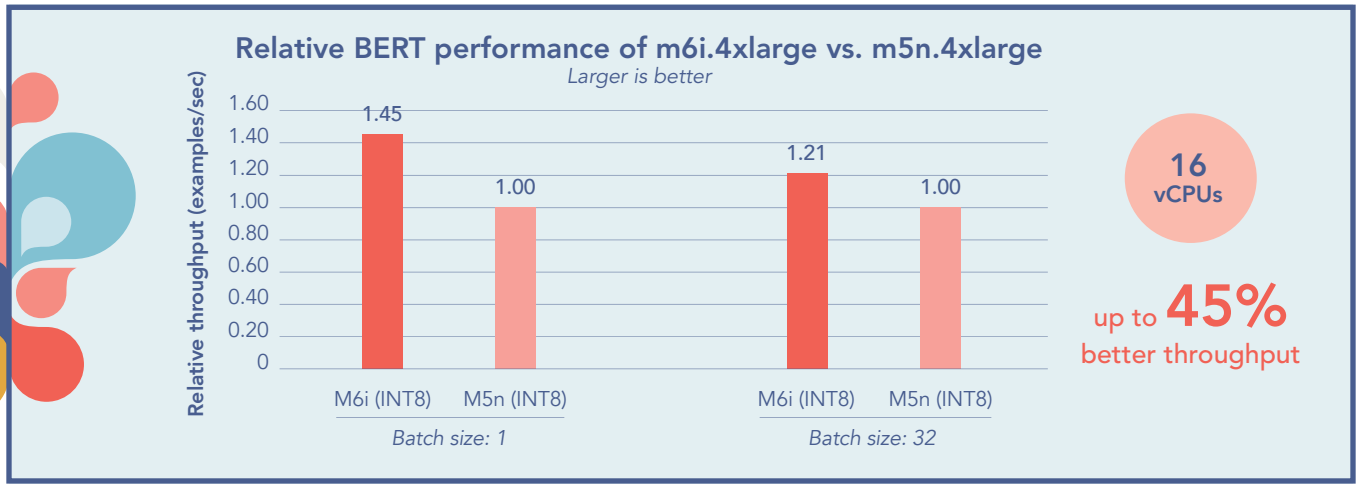


Figure 4: Relative BERT performance for M6i and M5n instances using 16 vCPUs. Higher numbers are better. Source: Principled Technologies.

Instances with 4 vCPUs: M6i vs. M6a

After comparing BERT performance of M6i instances against that of instances based on previous-gen processors, we compared those three sizes of M6i instances against M6a instances with AMD EPYC processors. Figure 5 compares the relative amount of text these instance types analyzed on 4vCPU configurations. The M6i instances enabled by 3rd Gen Intel Xeon Scalable processors with INT8 precision analyzed data 5.29 times as fast as the M6a instances with 3rd Gen AMD EPYC processors using FP32 precision. **Note: At the time of testing, INT8 precision—which can improve performance for these types of machine learning—was not available for BERT workloads on AMD EPYC processors.** Using FP32 precision, M6i instances improved performance over M6a instances by as much as 68 percent.

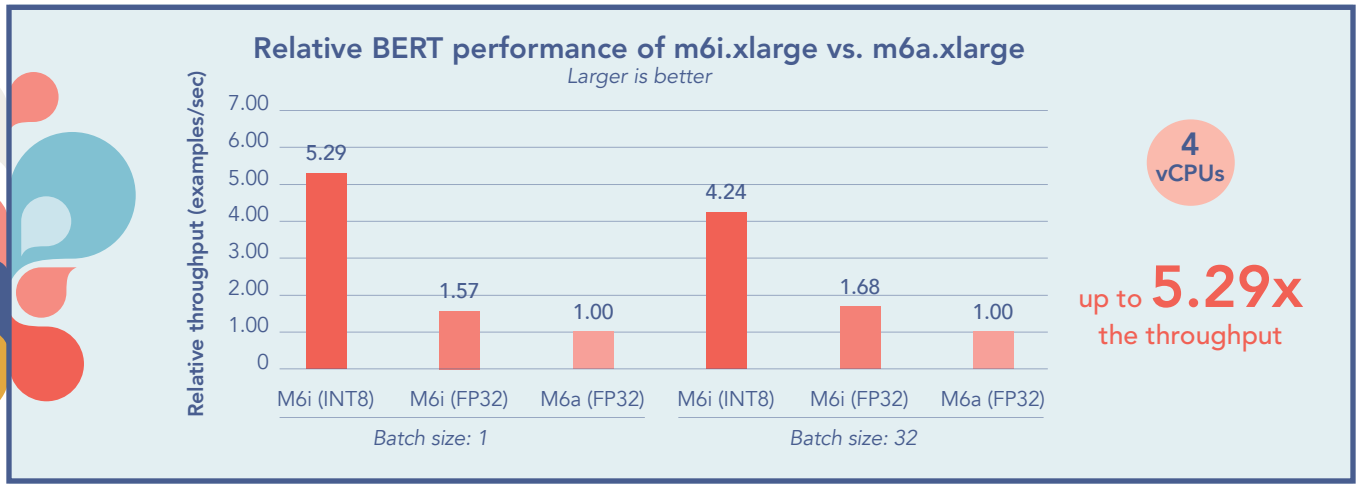


Figure 5: Relative BERT performance for M6i and M6a instances using 4 vCPUs. Higher numbers are better. Source: Principled Technologies.

Instances with 8 vCPUs: M6i vs. M6a

When we increased the instance sizes to 8 vCPUs, performance increases were similar to the 4vCPU configurations. Figure 6 compares the relative amount of text the instance types analyzed on 8vCPU configurations. The M6i instances enabled by 3rd Gen Intel Xeon Scalable processors analyzed data up to 5.10 times as fast as the M6a instances with 3rd Gen AMD EPYC processors.

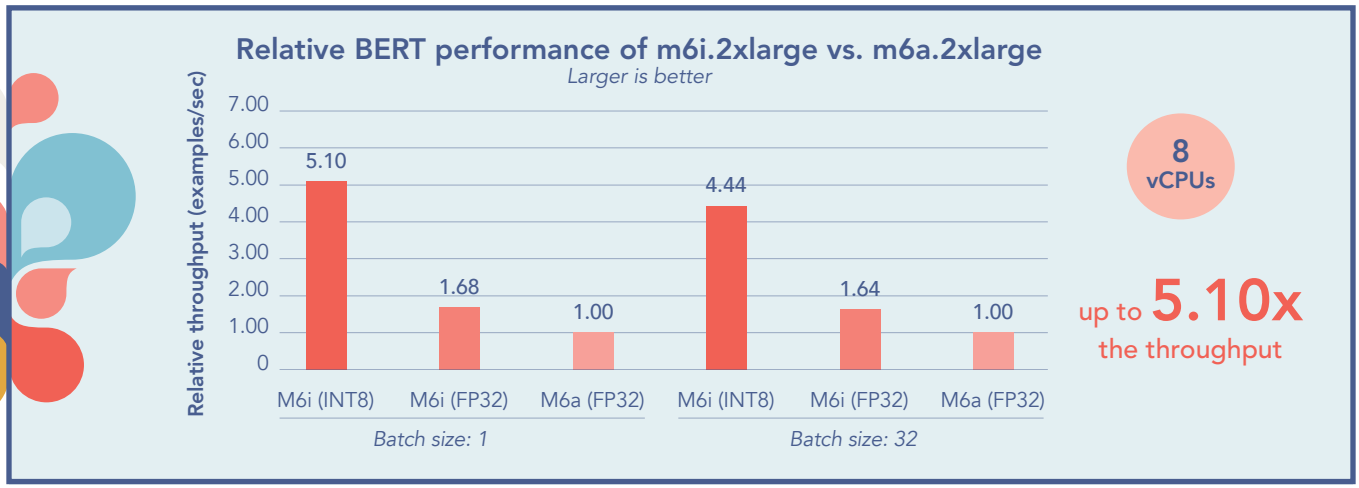


Figure 6: Relative BERT performance for M6i and M6a instances using 8 vCPUs. Higher numbers are better. Source: Principled Technologies.

Instances with 16 vCPUs: M6i vs. M6a

The biggest relative difference in BERT performance occurred in our 16vCPU comparison of M6i and M6a configurations. Figure 7 compares the relative examples per second the instance types analyzed on 16vCPU configurations. The M6i instances enabled by 3rd Gen Intel Xeon Scalable processors analyzed data up to 6.40 times as fast as the M6a instances with 3rd Gen AMD EPYC processors. These results show that for these types of BERT workloads, selecting M6i instances that offer INT8 precision over M6a instances that don't could allow organizations to complete textual analysis workloads using fewer cloud instances.

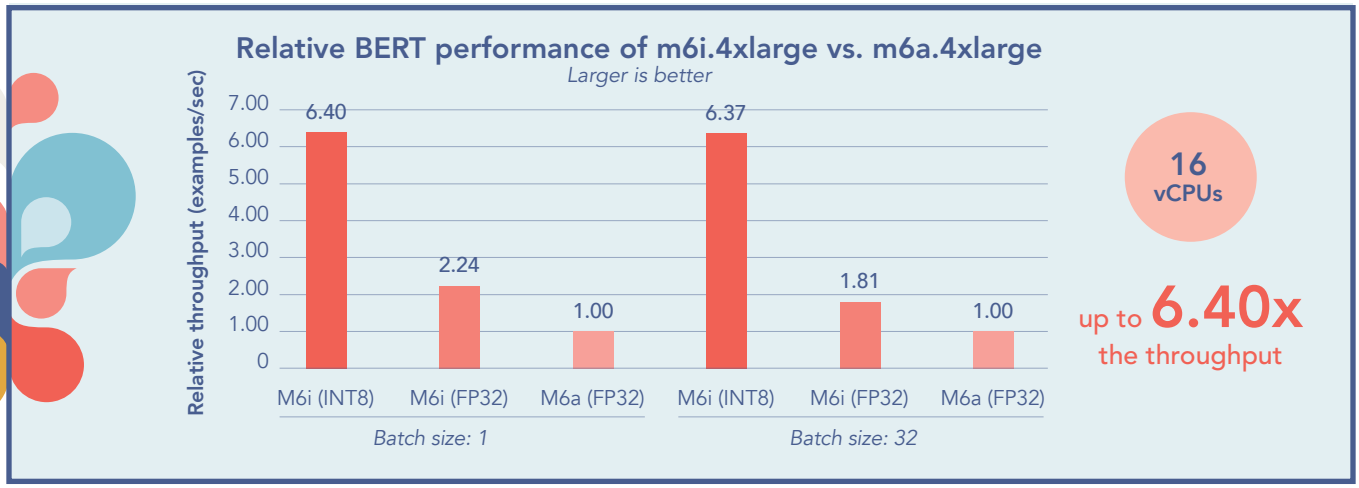


Figure 7: Relative BERT performance for M6i and M6a instances using 16 vCPUs. Higher numbers are better. Source: Principled Technologies.

Scaling BERT workloads

Another consideration for assessing BERT performance is to see how the throughput scales as you increase the size of the instance. Theoretically, performance could double as you double the vCPU count, which would be perfect linear scaling. While resource allocation makes this unlikely in the real world, the closer an instance approaches this ideal, the better.

As Figure 8 shows, using results from our batch size: 1 tests, the M6i instance with 3rd Gen Intel Xeon Scalable processors had better BERT performance scaling from 8 vCPUs to 16 vCPUs compared to the M6a instance with AMD EPYC processors, though slightly worse scaling from 4 vCPUs to 8 vCPUs.

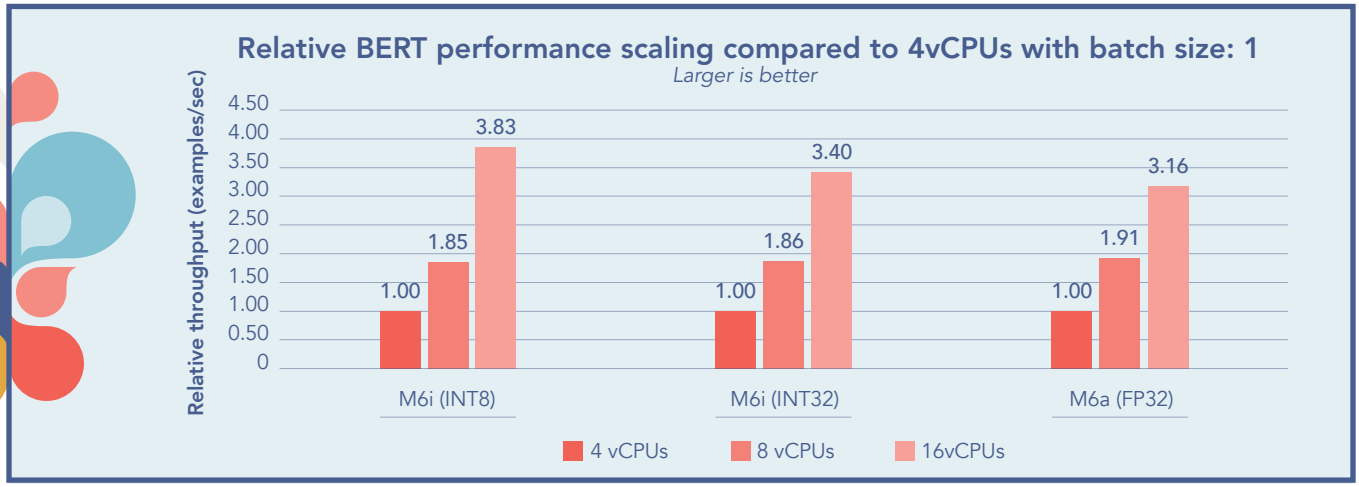


Figure 8: How BERT performance scaled across instance sizes, compared to results from the 4vCPU tests with batch size 1. Higher numbers are better. Source: Principled Technologies.

Figure 9 makes the same comparison, but uses results from our batch size: 32 testing. Again, the M6i instance with 3rd Gen Intel Xeon Scalable processors scaled more linearly from 4 to 16 vCPUs compared to the M6a instance.

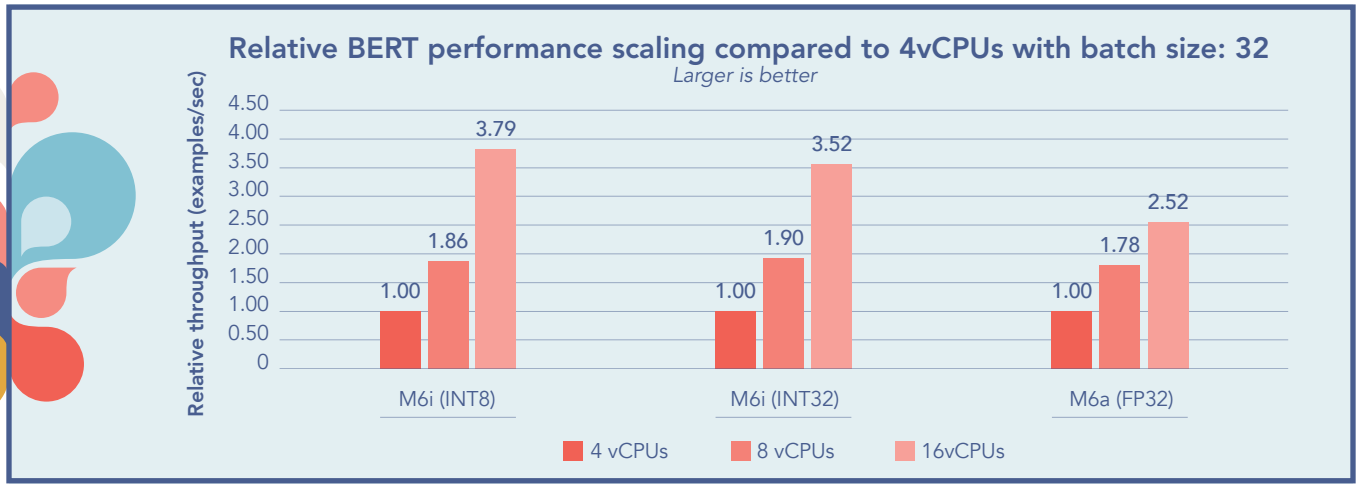


Figure 9: How BERT performance scaled across instance sizes, compared to results from the 4vCPU tests with batch size 32. Higher numbers are better. Source: Principled Technologies.

By selecting M6i instances that offer more linear, predictable performance scaling, organizations could more reliably fix their cloud operating budgets as textual analysis workloads continue to grow.

Conclusion

Organizations analyzing textual data using NLP through the BERT framework must decide which type of instance can deliver the BERT performance they need. In our tests, we found that across instance sizes, AWS M6i instances with 3rd Gen Intel Xeon Scalable processors outperformed both M5n instances with 2nd Gen Intel Xeon Scalable processors and M6a instances with 3rd Gen AMD EPYC processors for BERT machine learning. Plus, the M6i instances offered more predictable scaling at 16vCPUs. These performance increases could help you get quicker insight from textual data to better satisfy consumers and increase revenues.

1. TechTarget, "BERT language model," accessed December 16, 2021, https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model.
2. Intel, "3rd Gen Intel® Xeon® Scalable Processors," accessed December 14, 2021, https://www.intel.com/content/www/us/en/products/docs/processors/xeon/3rd-gen-xeon-scalable-processors-brief.html.
3. Intel, "3rd Gen Intel® Xeon® Scalable Processors."
4. Amazon, Amazon EC2 M6i Instances, accessed December 14, 2021, https://aws.amazon.com/ec2/instance-types/m6i/.

Read the science behind this report at https://facts.pt/ZymlIA3 ▶

Principled Technologies®

Facts matter.®

This project was commissioned by Intel.