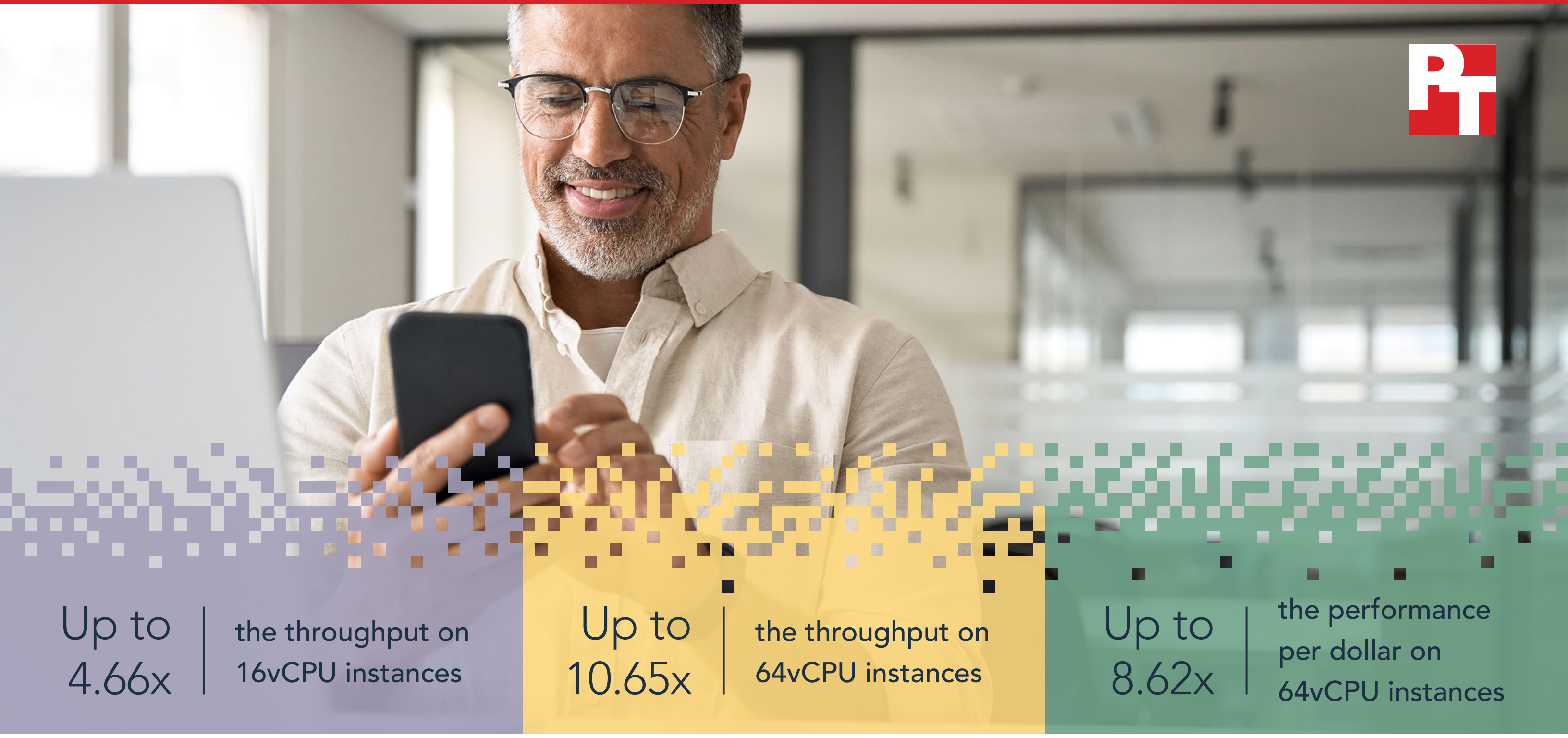
A Principled Technologies report: Hands-on testing. Real-world results.



Up to 4.66x | the throughput on 16vCPU instances

Up to 10.65x | the throughput on 64vCPU instances

Up to 8.62x | the performance per dollar on 64vCPU instances

Accelerate natural language processing with AWS EC2 M7i instances featuring 4th Gen Intel Xeon Scalable processors

Compared to AWS M7g instances with AWS Graviton3 processors, AWS M7i instances with Intel Xeon Scalable processors analyzed up to 10.65 times the sentences per second and delivered up to 8.62 times the performance per dollar in RoBERTa model testing

Natural language processing (NLP) supports many everyday applications, from search result suggestions to text autocorrections to text generation chatbots. Deep learning frameworks for NLP, such as Bidirectional Encoder Representations from Transformers (BERT) and its variations, demand robust compute power to analyze text and make predictions. For businesses running these workloads in the cloud, selecting the right instance can enable faster text analysis.

Using Intel® Extension for PyTorch, we tested the deep learning performance of two types of Amazon Web Services (AWS) Elastic Cloud Compute (EC2) instances: M7i instances enabled by 4th Gen Intel Xeon® Scalable processors and M7g instances enabled by AWS Graviton3 processors. To measure how the instances performed at different sizes, we tested both types with 4, 16, and 64 vCPUs. We ran tests with the RoBERTa model, a variant of BERT, using a benchmark from the Intel Model Zoo. We optimized the model for each processor type.

In our tests, the M7i instances analyzed up to 10.65 times the sentences per second compared to the M7g instances. These results indicate that businesses running similar RoBERTa workloads could analyze textual data faster and complete more work per instance with M7i instances featuring 4th Gen Intel Xeon Scalable processors.

This project was commissioned by Intel.

How we tested

We ran two types of AWS EC2 instances in the US-east-1f region:

- M7i instances featuring 4th Gen Intel Xeon Scalable processors
- M7g instances featuring AWS Graviton3 processors

We optimized the RoBERTa model for each instance type. The 4th Gen Intel Xeon Scalable processors featured Intel Advanced Matrix Extensions (Intel AMX) with 8-bit integer, bfloat16, and float16 support. To determine the performance businesses of various sizes might expect, we ran tests on smaller, medium-size, and larger instances, as Figure 1 shows. To account for different types of datasets businesses may analyze, we ran tests at different batch sizes, which indicate the number of samples that go through the neural network at a time. We chose a small batch size of 1 and a large batch size of 32. Finally, to show performance at different precisions an organization may use, we tested with both BF16 and FP32 precision. While FP32 was "the standard type for neural network computations for a long time," BF16 is a newer precision type that some organizations have used to replace FP16 and FP32 precision.[1]

In addition to measuring the number of sentences per second each instance could analyze at two batch sizes and with two precisions, we also calculated the performance per price of each instance. We based this data on the results of our testing and instance pricing we researched on August 10, 2023. To read more about our tests, configurations, and calculations, see the science behind the report.

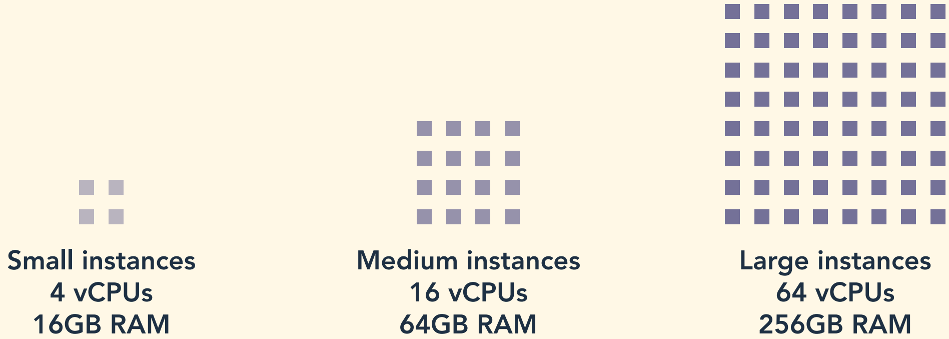


Figure 1: Key specifications for each instance size we tested. Source: Principled Technologies.

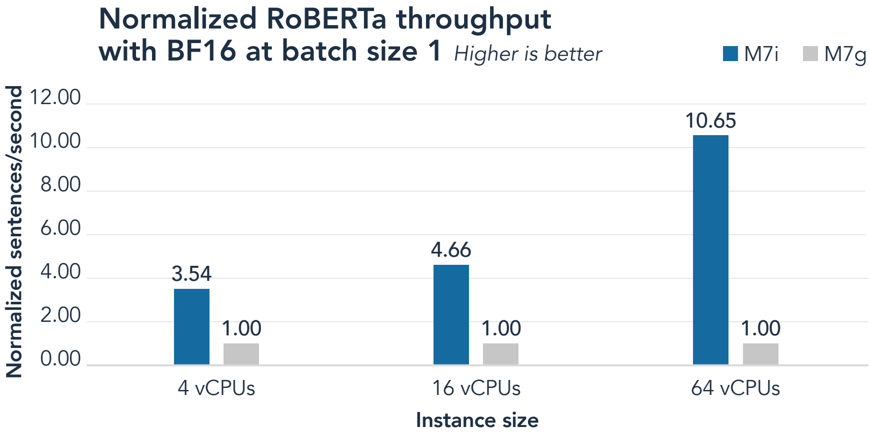
What better RoBERTa performance might mean for you

BERT, pre-trained on the entirety of the English language Wikipedia, currently helps Google interpret user searches.[2] Developed by Facebook AI, the RoBERTa model we used in our testing trained on a larger dataset, including the English language Wikipedia, 63 million news articles, and 38 GB of Reddit content.[3] According to one article, "RoBERTa has been shown to outperform BERT and other state-of-the-art models on a variety of natural language processing tasks, including language translation, text classification, and question answering. It has also been used as a base model for many other successful NLP models and has become a popular choice for research and industry applications."[4] At all three instance sizes we tested, the M7i instances enabled by 4th Gen Intel Xeon Scalable processors handled greater throughput than the M7g instances with Graviton3 processors. And they did so using different batch sizes and precision types. This means that selecting M7i instances could help businesses run RoBERTa workloads faster, which could lead to a more seamless experience for users on RoBERTa-powered apps or the ability to support more users. Additionally, with the better value that the M7i instances delivered, organizations might accomplish more RoBERTa work per instance, improving their bottom lines as they save on instance uptime.

Note: The graphs in this report use two different y-axis scales to keep to a consistent size. Please be mindful of each graph's data range as you compare.

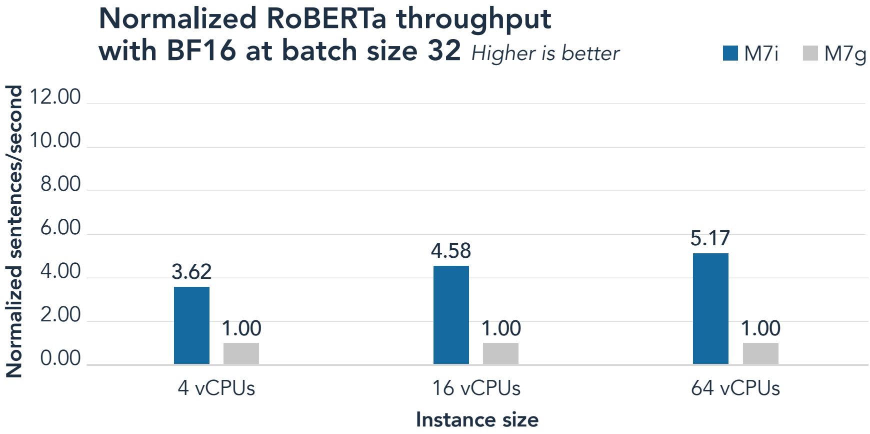
Analyze text faster at BF16 and FP32 precision

We first compared RoBERTa performance at BF16 precision, measuring the sentences per second each instance analyzed. At both batch sizes and across vCPU counts, M7i instances featuring 4th Gen Intel Xeon Scalable processors handled a higher throughput rate than M7g instances with Graviton3 processors. As Figures 2 and 3 show, M7i instances analyzed up to 10.65 times the sentences per second.



Up to 10.65x the throughput on 64vCPU instances

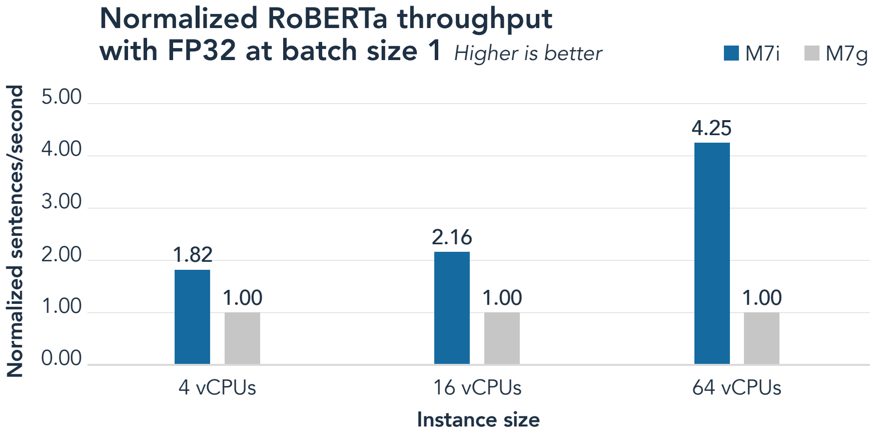
Figure 2: Relative RoBERTa performance of M7i instances, in sentences analyzed per second, compared to M7g instances. Both types of instances used BF16 precision at batch size: 1. Higher is better. Source: Principled Technologies.



Up to 5.17x the throughput on 64vCPU instances

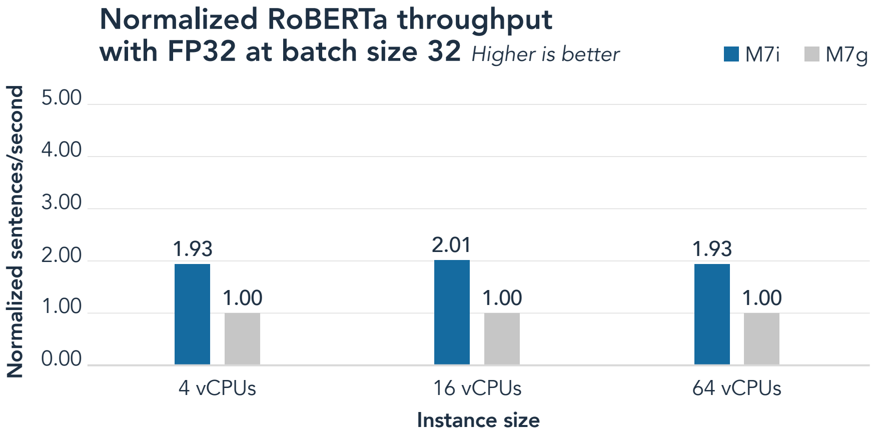
Figure 3: Relative RoBERTa performance of M7i instances, in sentences analyzed per second, compared to M7g instances. Both types of instances used BF16 precision at batch size: 32. Higher is better. Source: Principled Technologies.

When we compared performance using FP32 precision, we again saw performance increases from M7i instances featuring 4th Gen Intel Xeon Scalable processors. For both batch sizes we tested, the M7i instances outperformed the M7g instances by as much as 4.25 times (Figures 4 and 5).



Up to 4.25x the throughput on 64vCPU instances

Figure 4: Relative RoBERTa performance of M7i instances, in sentences analyzed per second, compared to M7g instances. Both types of instances used FP32 precision at batch size: 1. Higher is better. Source: Principled Technologies.



Up to 2.01x the throughput on 16vCPU instances

Figure 5: Relative RoBERTa performance of M7i instances, in sentences analyzed per second, compared to M7g instances. Both types of instances used FP32 precision at batch size: 32. Higher is better. Source: Principled Technologies.

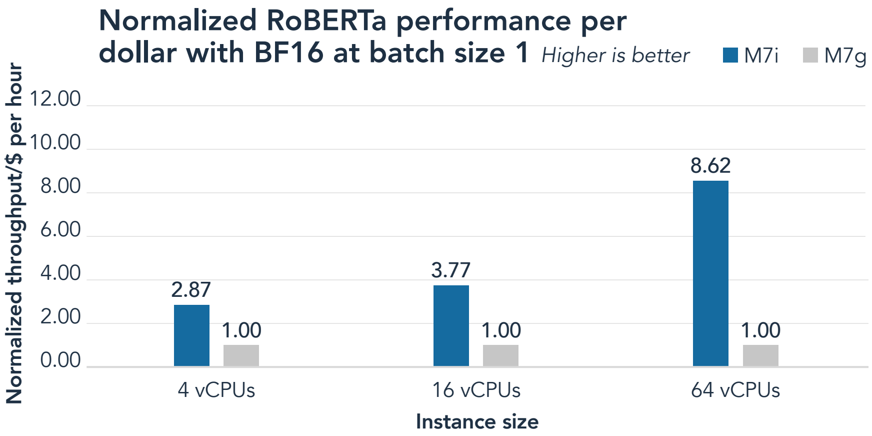
About 4th Gen Intel Xeon Scalable processors

According to Intel, 4th Gen Intel Xeon Scalable processors feature "the most built-in accelerators of any CPU on the market to improve performance in AI, data analytics, networking, storage, and HPC."[5] Along with PCIe Gen5 technology, DDR5 memory, and CXL 1.1 capabilities, 4th Gen Intel Xeon Scalable processors also offer Intel Accelerator Engines for a variety of workloads.[6]

For more information, visit https://www.intel.com/content/www/us/en/products/docs/processors/xeon-accelerated/4th-gen-xeon-scalable-processors.html.

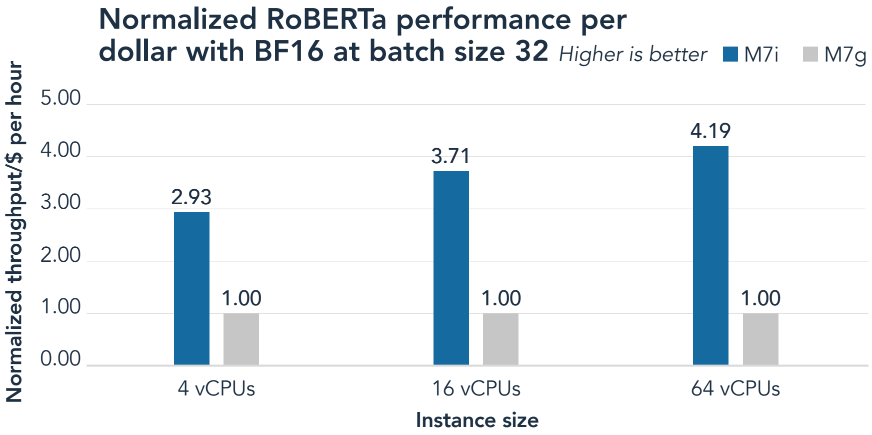
Take advantage of higher-value performance

Not only did M7i instances analyze sentences at a faster rate, but this performance improvement made them a better value. Dividing each instance's RoBERTa throughput by its per-hour price, we found that M7i instances performed more RoBERTa work for every dollar they cost. At BF16 precision, M7i instances delivered up to 8.62 times the throughput per cost of M7g instances, as Figures 6 and 7 show.



Up to 8.62x the performance per dollar on 64vCPU instances

Figure 6: Throughput per instance cost of M7i instances compared to M7g instances. Both instances used BF16 precision at batch size 1. Higher is better. Source: Principled Technologies.



Up to 4.19x the performance per dollar on 64vCPU instances

Figure 7: Throughput per instance cost of M7i instances compared to M7g instances. Both instances used BF16 precision at batch size 32. Higher is better. Source: Principled Technologies.

Using the same calculations as we did for BF16 precision, we saw that with FP32 precision, the M7i instances again offered a better value than the M7g instances we tested. Figures 8 and 9 show that M7i instances handled up to 3.44 times the throughput per cost of the M7g instances using FP32 precision.

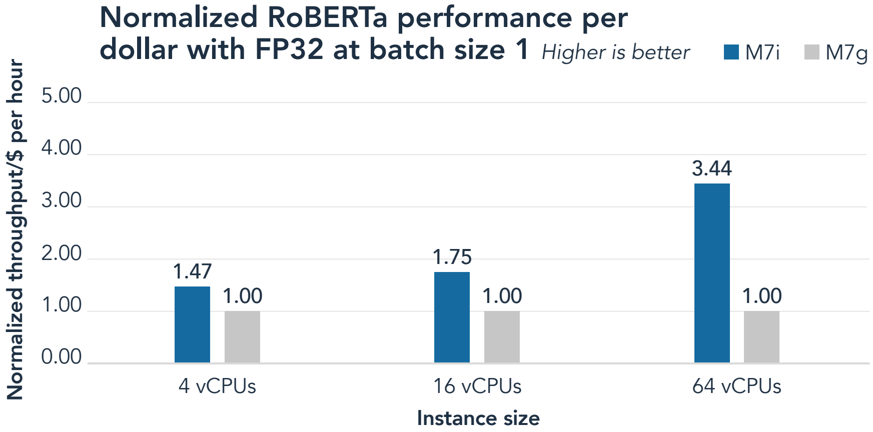


Figure 8: Throughput per instance cost of M7i instances compared to M7g instances. Both instances used FP32 precision at batch size 1. Higher is better. Source: Principled Technologies.

Up to 3.44x the performance per dollar on 64vCPU instances

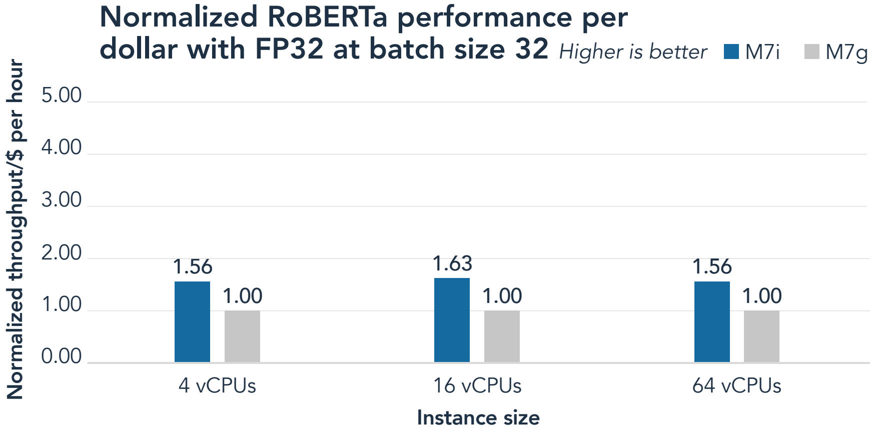


Figure 9: Throughput per instance cost of M7i instances compared to M7g instances. Both instances used FP32 precision at batch size 32. Higher is better. Source: Principled Technologies.

Up to 1.63x the performance per dollar on 16vCPU instances

Conclusion

For natural language processing workloads, the instance type you choose can make an enormous difference in performance—how quickly they can make sense of textual data—and value—how much work they can accomplish at a given cost. Our RoBERTa test results indicate that AWS EC2 M7i instances enabled by 4th Gen Intel Xeon Scalable processors outperformed M7g instances with AWS Graviton3 processors at different vCPU counts, batch sizes, and precisions, processing up to 10.65 times as many sentences per second. These performance gains led to M7i instances delivering a better value, achieving up to 8.62 times the throughput per dollar. With an instance that can analyze text more quickly, your business could offer a smoother experience on RoBERTa-supported apps or potentially condense RoBERTa workloads onto fewer instances.

[1] Grigory Sapunov, "FP64, FP32, FP16, BFLOAT16, TF32, and other members of the ZOO," accessed August 25, 2023, https://moocaholic.medium.com/fp64-fp32-fp16-bfloat16-tf32-and-other-members-of-the-zoo-a1ca7897d407.

[2] Ben Lutkevich, "BERT language model," accessed August 18, 2023, https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model.

[3] GeeksforGeeks, "Overview of ROBERTa model," accessed August 24, 2023, https://www.geeksforgeeks.org/overview-of-roberta-model/.

[4] GeeksforGeeks, "Overview of ROBERTa model."

[5] Intel, "4th Gen Xeon Scalable Processors," accessed August 18, 2023, https://www.intel.com/content/www/us/en/products/docs/processors/xeon-accelerated/4th-gen-xeon-scalable-processors.html/.

[6] Intel, "4th Gen Xeon Scalable Processors."

Read the science behind this report at https://facts.pt/2aqfhRt ▶

This project was commissioned by Intel.