Google Cloud N2 VM instances featuring 3rd Gen Intel Xeon Scalable processors offered better BERT deep learning performance

vs. N2 VM instances with 2nd Gen Intel Xeon Scalable processors and N2D VM instances with 3rd Gen AMD EPYC processors

A deep learning framework for natural language processing (NLP), Bidirectional Encoder Representations from Transformers (BERT) sorts and analyzes textual data to make predictions, answer questions, and even respond to conversation. Organizations running BERT in the cloud can maximize their ability to run these demanding workloads by selecting VMs that can analyze data faster.

From the Principled Technologies data center, using Intel optimization for TensorFlow and AMD ZenDNN integrated with TensorFlow, we compared the BERT deep learning performance of three types of Google Cloud™ VM instances: N2 with 3rd Gen Intel® Xeon® Scalable processors featuring Intel DL Boost with Vector Neural Network Instructions (VNNI) and Intel Advanced Vector Extensions 512 (Intel AVX-512), N2 with 2nd Gen Intel Xeon Scalable processors, and N2D with 3rd Gen AMD EPYC™ processors.

From 4 vCPUs to 16 vCPUs, running a benchmark from the Intel Model Zoo, Google Cloud N2 VM instances with 3rd Gen Intel Xeon Scalable processors offered up to 27 percent better BERT performance than the N2 VM instances with previous-gen processors and up to 5.8 times the BERT performance compared to N2D VM instances with 3rd Gen AMD EPYC processors. Plus, BERT performance scaled more predictably across N2 VM instances. This means that organizations running similar BERT workloads in the cloud could analyze data faster or handle more work per VM by choosing N2 VM instances featuring 3rd Gen Intel Xeon Scalable processors.

Better throughput on 4-vCPU VMs

Up to 5.5x the queries per second vs. N2D VM instances

Better throughput on 8-vCPU VMs

Up to 5.6x the queries per second vs. N2D VM instances

Better throughput on 16-vCPU VMs

Up to 5.8x the queries per second vs. N2D VM instances

How we tested

We purchased three sets of virtual machine instances from three general-purpose Google Cloud series:

- N2 VMs featuring 3rd Gen Intel Xeon Platinum 8373C processors (Ice Lake)
- N2 VMs featuring 2nd Gen Intel Xeon Platinum 6268CL processors (Cascade Lake)
- N2D VMs featuring 3rd Gen AMD EPYC 7B13 processors (Milan)

We ran each instance in the US-Central1-a region.

Figure 1 shows the specifications for the VMs that we chose. To show how businesses of various sizes with different deep learning demands can benefit from choosing N2 VM instances with 3rd Gen Intel Xeon Scalable processors, we tested VMs 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. The Intel Xeon Scalable processors support FP32 precision as well as INT8 precision which can improve performance for types of machine learning. At the time of testing, INT8 support was not available for BERT workloads on AMD EPYC processors, so N2D results reflect FP32 precision. In this report, we first present the comparisons between N2 VM instances with current-gen and previous-gen Intel Xeon Scalable processors, and then present the comparisons between N2 VM instances with 3rd Gen Intel Xeon Scalable processors and N2D VM instances with AMD EPYC processors. (Note: For additional test results on even larger VMs, see the science behind the report.)

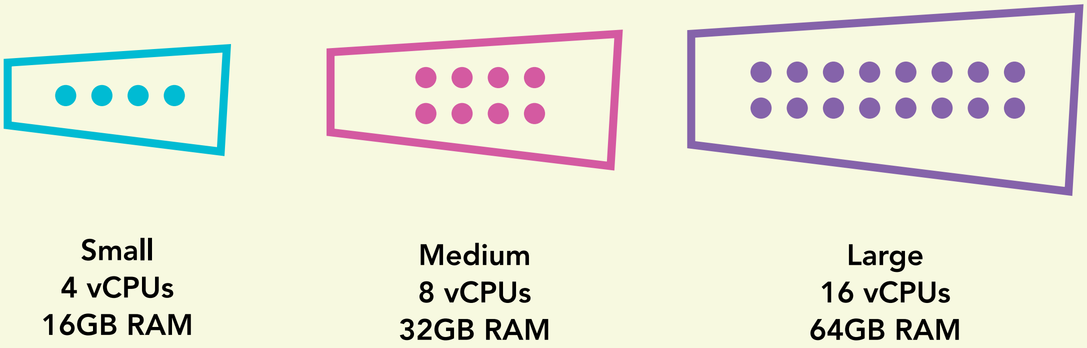
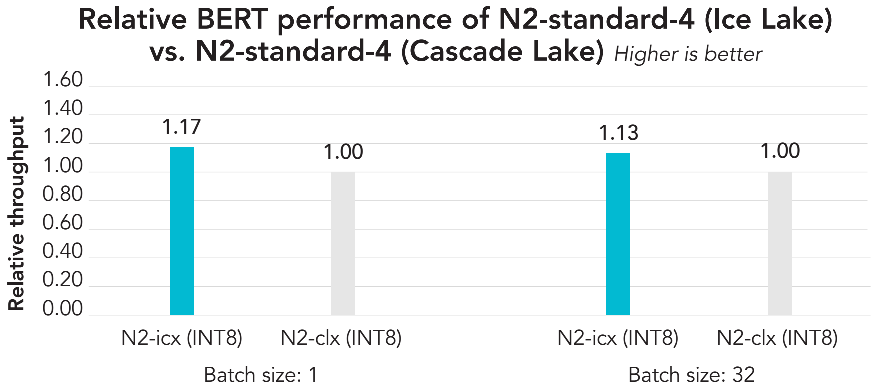


Figure 1: Key specifications for each instance size we tested. Source: Principled Technologies.

Gen-over-gen comparison of N2 VM instances

4 vCPUs

First, we compared BERT performance on smaller sized VMs, looking at the relative amount of text second the VM instance types analyzed on 4vCPU configurations. As Figure 2 shows, N2 VM instances enabled by 3rd Gen Intel Xeon Scalable processors analyzed up to 17 percent more examples per second than the N2 VM instances with 2nd Gen Intel Xeon Scalable processors.



Small instances

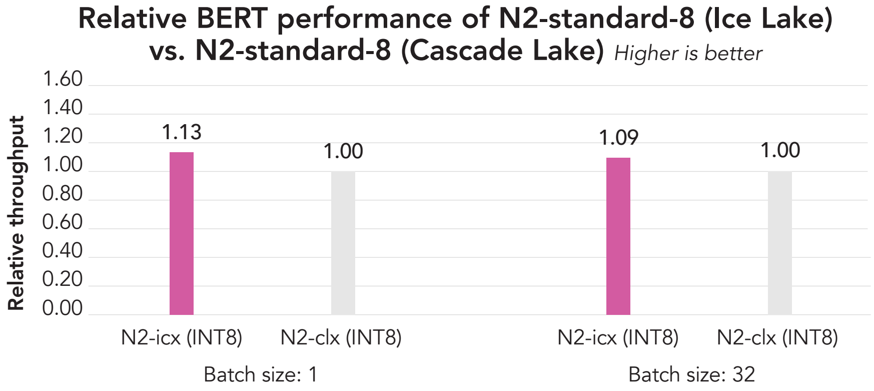
Up to 17%

better throughput

Figure 2: Relative BERT performance between N2 VM instances featuring 3rd Gen Intel Xeon Scalable processors and N2 VM instances with previous-gen Intel Xeon Scalable processors, using 4 vCPUs. Higher numbers are better. Source: Principled Technologies.

8 vCPUs

When we doubled the instance size to 8 vCPUs, current-gen N2 VM instances delivered a similar performance increase over previous-gen N2 VM instances. Figure 3 compares the relative amount of text the instance types analyzed on 8vCPU configurations. The N2 VM instances enabled by 3rd Gen Intel Xeon Scalable processors analyzed up to 13 percent more examples per second than the N2 VM instances with 2nd Gen Intel Xeon Scalable processors.



Medium instances

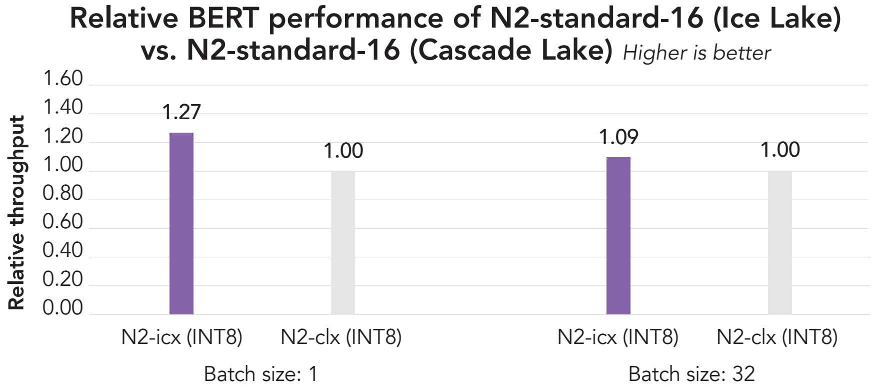
Up to 13%

better throughput

Figure 3: Relative BERT performance between N2 VM instances featuring 3rd Gen Intel Xeon Scalable processors and N2 VM instances with previous-gen Intel Xeon Scalable processors, using 8 vCPUs. Higher numbers are better. Source: Principled Technologies.

16 vCPUs

As Figure 4 shows, N2 VM instances with current-generation processors offered the greatest relative BERT performance increase over previous-gen N2 VM instances using larger 16vCPU configurations. The N2 VM instances enabled by 3rd Gen Intel Xeon Scalable processors analyzed up to 27 percent more examples per second than N2 VM instances with 2nd Gen Intel Xeon Scalable processors. By improving textual data analysis throughput by 27 percent, organizations could reduce the number of VMs they need to purchase and manage.



Large instances

Up to 27%

better throughput

Figure 4: Relative BERT performance between N2 VM instances featuring 3rd Gen Intel Xeon Scalable processors and N2 VM instances with previous-gen Intel Xeon Scalable processors, using 16 vCPUs. Higher numbers are better. Source: Principled Technologies.

Running BERT workloads in the cloud

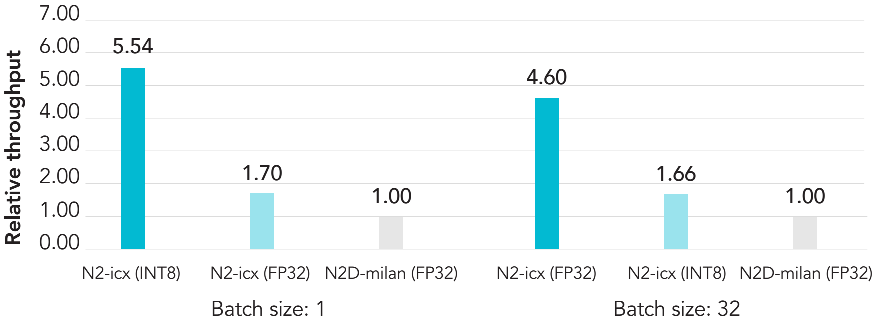
The BERT framework, which trained on text from the English language Wikipedia with over 2.5 million words, turns text into numbers to sort, analyze, and make predictions about that text.[1] Depending on the dataset on which an organization needs to run BERT machine leaning, the size of the Google Cloud VMs they choose will vary. To account for these different needs, we tested using two batch sizes across three different VM 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 N2 VMs as well for comparison. In all three VM sizes, the N2 VM instances enabled by 3rd Gen Intel Xeon Scalable processors outperformed both the previous-gen N2 VM instances and the current-gen N2D VM instances.

N2 VM instances with 3rd Gen Intel Xeon Scalable processors vs. N2D VM instances

4 vCPUs

After comparing BERT performance of N2 VM instances against that of VMs based on previous-gen processors, we compared those three sizes of N2 VM instances against N2D VM instances with AMD EPYC processors. Figure 5 shows the relative amount of text these instance types analyzed on 4vCPU configurations. The N2 VM instances enabled by 3rd Gen Intel Xeon Scalable processors with INT8 precision analyzed data 5.54 times as fast as the N2D VM 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, N2 VMs improved performance over N2D VMs by as much as 70 percent.

Relative BERT performance of N2-standard-4 (Ice Lake) vs. N2D-standard-4 (Milan) *Higher is better*



Small instances

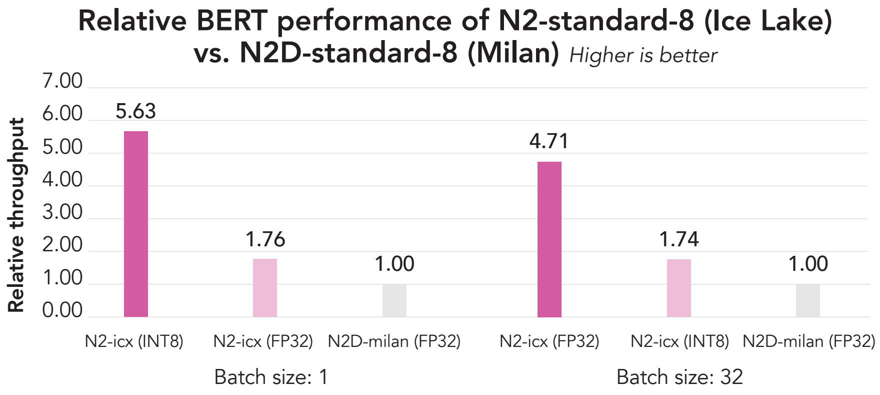
Up to 5.54x

the throughput

Figure 5: Relative BERT performance between N2 VM instances featuring 3rd Gen Intel Xeon Scalable processors and N2D VM instances with 3rd Gen AMD EPYC processors, using 4 vCPUs. Higher numbers are better. Source: Principled Technologies.

8 vCPUs

When we increased the VM sizes to 8 vCPUs, performance increases were similar to the 4vCPU configurations. Figure 6 compares the relative amount of text the VM types analyzed on 8vCPU configurations. The N2 VM instances enabled by 3rd Gen Intel Xeon Scalable processors analyzed data up to 5.63 times as fast as the N2D VM instances with 3rd Gen AMD EPYC processors.



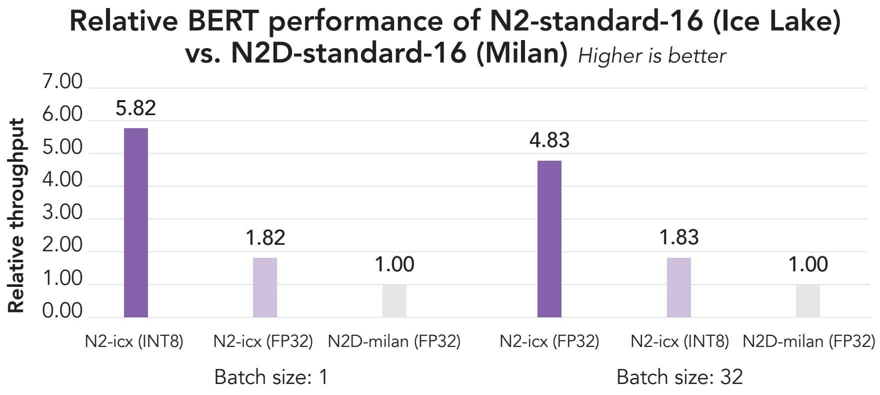
Medium instances

Up to 5.63x the throughput

Figure 6: Relative BERT performance between N2 VM instances featuring 3rd Gen Intel Xeon Scalable processors and N2D VM instances with 3rd Gen AMD EPYC processors, using 8 vCPUs. Higher numbers are better. Source: Principled Technologies.

16 vCPUs

The biggest relative difference in BERT performance occurred in our 16vCPU comparison of N2 and N2D configurations. Figure 7 compares the relative examples per second the instance types analyzed on 16vCPU configurations. The N2 VM instances enabled by 3rd Gen Intel Xeon Scalable processors analyzed data up to 5.82 times as fast as the N2D VM instances with 3rd Gen AMD EPYC processors. These results show that for these types of BERT workloads, selecting N2 VM instances that offer INT8 precision over N2D VM instances that don't could allow organizations to complete textual analysis workloads using fewer cloud VMs.



Large instances

Up to 5.82x the throughput

Figure 7: Relative BERT performance between N2 VM instances featuring 3rd Gen Intel Xeon Scalable processors and N2D VM instances with 3rd Gen AMD EPYC processors, using 16 vCPUs. Higher numbers are better. Source: Principled Technologies.

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.

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 should 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 N2 VM instance with 3rd Gen Intel Xeon Scalable processors had better BERT performance scaling from 4 vCPUs to 8 vCPUs and 8 vCPUs to 16 vCPUs compared to the N2D VM instance with 3rd Gen AMD EPYC processors.

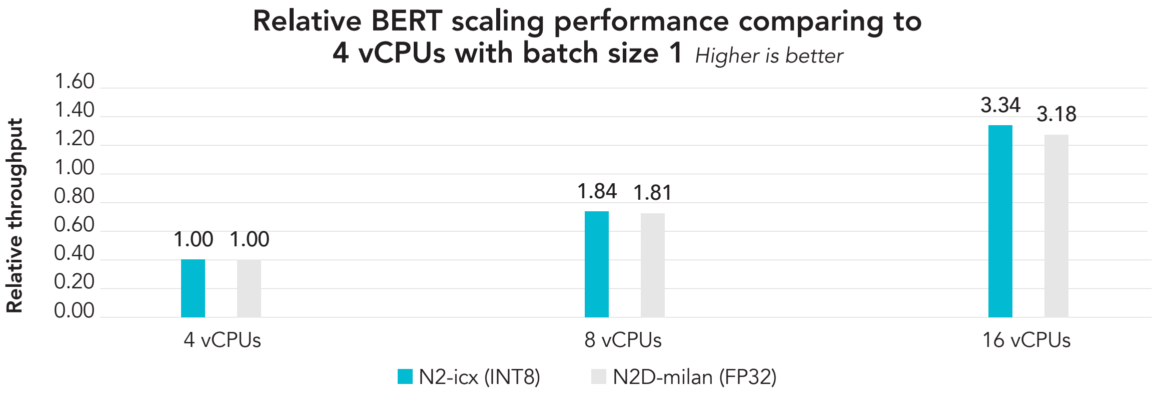


Figure 8: How BERT performance scaled across VM instance sizes, compared to results from the 4 vCPU tests with batch size 1. Higher numbers are better. Source: Principled Technologies.

Figure 9 makes the same comparison, but uses results from our batch size 32 testing. Again, the N2 VM instance with 3rd Gen Intel Xeon Scalable processors scaled more linearly from 4 to 16 vCPUs compared to the N2D VM instance. By selecting N2 VM instances that offer more linear, predictable performance scaling, organizations could more reliably fix their cloud operating budgets as textual analysis workloads continue to grow.

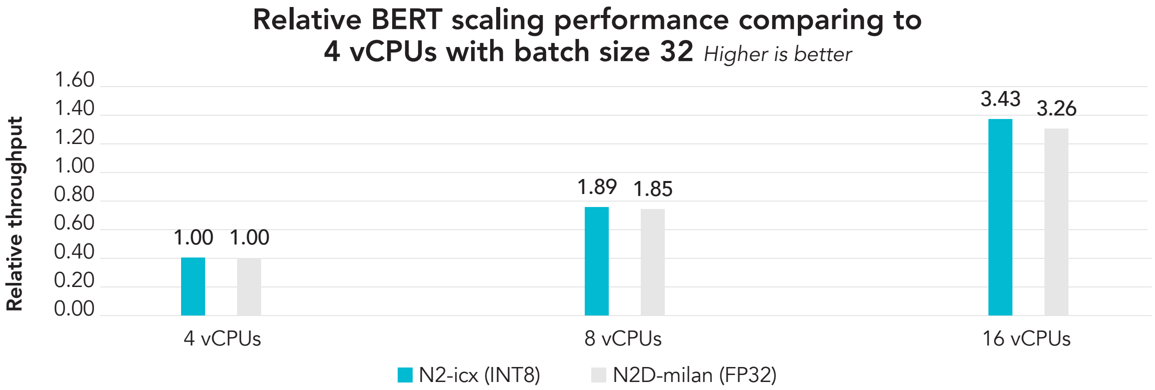


Figure 9: How BERT performance scaled across VM sizes, compared to results from the 4 vCPU tests with batch size 32. Higher numbers are better. Source: Principled Technologies.

Why choose N2 VMs with 3rd Generation Intel Xeon Scalable Processors

New N2 VMs 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

Conclusion

As our tests prove, the VM types you select for running BERT workloads can make a big difference in how quickly you can make sense of textual data. Across VM sizes, Google Cloud N2 VM instances with 3rd Gen Intel Xeon Scalable processors outperformed both N2 VM instances with 2nd Gen Intel Xeon Scalable processors and N2D VM instances with 3rd Gen AMD EPYC processors for BERT machine learning. Plus, the current-gen N2 VM instances offered more predictable scaling from 4 CPUs all the way to 16 vCPUs. 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 July 19, 2022, https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model.

[2] Intel, "3rd Gen Intel® Xeon® Scalable Processors," accessed July 19, 2022, 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] Intel, "3rd Gen Intel® Xeon® Scalable Processors."

Read the science behind this report at https://facts.pt/6pWWUba ▶

Principled Technologies®

Facts matter.®

This project was commissioned by Intel.