Achieve up to 4.3x the BERT AI Deep Learning Performance with Google Cloud C3 High-CPU Virtual Machines with 4th Gen Intel® Xeon® Scalable Processors

BERT Performance

Handle up to 1.48 Times the Sentences per Second at FP32 Precision with C3 High-CPU VMs with 4th Gen Intel Xeon Scalable Processors

vs. N2 Standard VMs with 2nd Gen Processors

Handle up to 4.32 Times the Sentences per Second at INT8 Precision with C3 High-CPU VMs with 4th Gen Intel Xeon Scalable Processors

vs. N2 Standard VMs with 2nd Gen Processors

These VMs Delivered Greater Throughput Than N2 Standard VMs with 2nd Generation Intel Xeon Scalable Processors at Both FP32 and INT8 Precision Levels

Organizations use natural language processing (NLP) frameworks for tasks such as analyzing textual data, making predictions, answering questions, and responding to conversation. Bidirectional Encoder Representations from Transformers (BERT) workloads are a popular NLP framework for analyzing and getting AI insights from text data. When running BERT workloads on Google Cloud VMs, it is important to choose carefully to get the best performance.

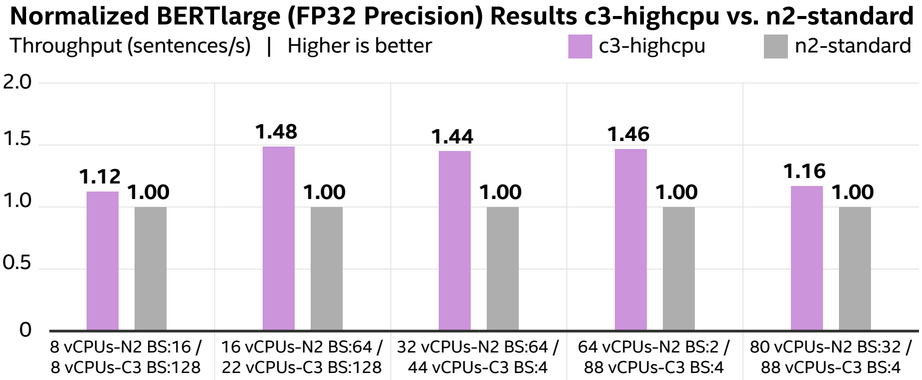
We conducted BERT testing on two types of GCP cloud VMs:

- C3 high-CPU VMs featuring 4th Gen Intel® Xeon® Scalable processors
- N2 standard VMs with 2nd Gen Intel Xeon Scalable processors

We tested at two precision levels: FP32 and INT8. At the time of testing, C3 high-CPU VMs and N2 standard VMs were not available with equal vCPU counts, so we configured our test VMs as comparably as possible. We also provide two points of comparison with the 88vCPU C3 high-CPU VM.

Better Performance with FP32 Precision

As Figure 1 shows, the C3 high-CPU VMs with 4th Gen Intel Xeon Scalable processors performed better at every vCPU count than the N2 VMs with 2nd Gen processors with FP32 Precision. The highest performance gain was 1.48 times the sentences per second at 22 vCPUs and batch size 128 vs. 16 vCPUs and batch size 64.


Normalized BERTlarge (FP32 Precision) Results c3-highcpu vs. n2-standard

Throughput (sentences/s) | Higher is better

Figure 1. Relative number of sentences per second achieved by C3 high-CPU VMs with 4th Gen Intel Xeon Scalable processors vs. N2 standard VMs with 2nd Gen processors using FP32 precision. Higher numbers are better.

See backup for workloads and configurations. Results may vary.

Better Performance with INT8 Precision

As Figure 2 shows, C3 high-CPU VMs with 4th Gen Intel® Xeon® Scalable processors also achieved better performance than the N2 VMs with 2nd Gen Intel Xeon Scalable processors at the INT8 Precision level. The highest performance gain was 4.32 times as many sentences per second at 8 vCPUs (batch size 128 and batch size 16).

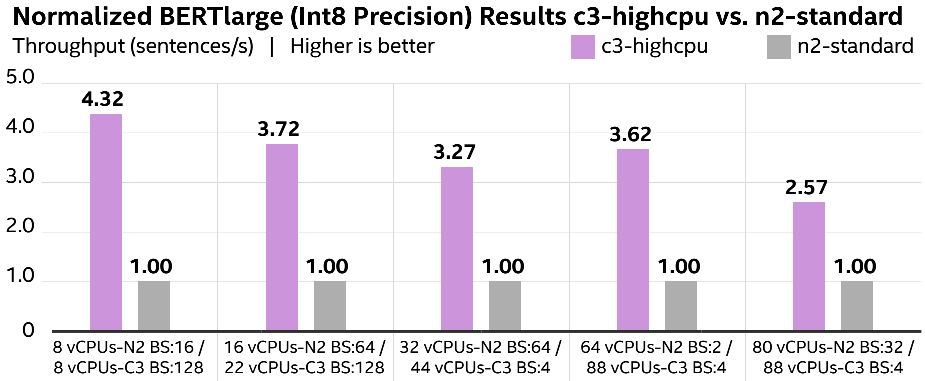


Figure 2. Relative number of sentences per second achieved by C3 high-CPU VMs with 4th Gen Intel Xeon Scalable processors vs. N2 standard VMs with 2nd Gen processors using INT8 precision. Higher numbers are better.

Conclusion

The cloud VM you choose can greatly improve the speed of text analysis on your BERT workloads. Intel testing at various VM and batch sizes shows that C3 high-CPU VMs with 4th Gen Intel Xeon Scalable processors delivered better BERT performance than N2 VMs with 2nd Gen processors did at both FP32 and INT8 precision levels—achieving as much as 4.32 times the throughput in terms of sentences per second.

Learn More

To begin running your AI BERT-like workloads on GCP C3 Virtual Machines with 4th Gen Intel Xeon Scalable processors, visit https://cloud.google.com/compute/docs/general-purpose-machines#c3_series.

Tests by Intel on GCP. Software: Ubuntu 22.04.1 LTS, 5.15.0-1029-gcp, Ubtek TF 2,11,dev202242, Python 3.8, BERT-Large (Inference Mode), FP32 precision, INT8-AMX precision (C3 only), INT8 precision (N2 only), SQuAD 1.1 (seq len=384). C3 instances: Intel Xeon Platinum 8481C @ 2.0GHz, 1x 500GB Direct-att disk, 2300 MB/s Network BW (8-44 vcpu), 3200 MB/s Network BW (88 vcpu), 16-176GB RAM. N2 instances: Intel Xeon Gold 6268CL @ 2.8GHz, 1x 500GB Direct-att disk, 3200 MB/s Network BW, 32-320GB RAM. 8vCPU instances: NUMA nodes: 1; workload specs: Batch Size: FP32/BS128, INT8-AMX/BS32 (C3); FP32/BS16, INT8/BS64 (N2); Step Size:30; Warm Up:10. n2-standard-16: NUMA nodes: 1; workload specs: Batch Size: FP32/BS64, INT8/BS16; Step Size:30; Warm Up:10. c3-highcpu-22: NUMA nodes: 1; workload specs: Batch Size: FP32/BS128, INT8-AMX/BS16; Step Size:30; Warm Up:10. n2-standard-32: NUMA nodes: 2; workload specs: Batch Size: FP32/BS64, INT8/BS16; Step Size:30; Warm Up:10. c3-highcpu-44: NUMA nodes: 1; workload specs: Batch Size: FP32/BS4, INT8-AMX/BS2; Step Size:30; Warm Up:10. n2-standard-64: NUMA nodes:2; workload specs: Batch Size: FP32/BS2, INT8/BS32; Step Size:30; Warm Up:10. n2-standard-80: NUMA nodes:2; workload specs: Batch Size: FP32/BS32, INT8/BS32; Step Size:30; Warm Up:10. c3-highcpu-88: NUMA nodes:2; workload specs: Batch Size: FP32/BS4, INT8-AMX/BS128; Step Size:30; Warm Up:10.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See above for configuration details. No product or component can be absolutely secure. Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Printed in USA 0823/HM/PT/PDF US001 ♲ Please Recycle