Handle up to 1.65x More Inference Work with Hugging Face – BERT Large Using Pytorch on Microsoft Azure Ddsv5 Virtual Machines over Ddsv4 VMs

Enjoy Stronger Performance with New Ddsv5 VMs Featuring 3rd Gen Intel® Xeon® Scalable Processors

Companies increasingly rely on machine learning inference workloads for a range of business activities. Inference is extremely compute-intensive, making it paramount to understand the performance potential of any VMs you are considering when shopping for a cloud solution to host your inference workloads. The latest Microsoft Azure Ddsv5-series VMs enabled by 3rd Gen Intel® Xeon® Scalable processors—available in a range of vCPU counts—can deliver excellent performance. For applications that benefit from high vCPU counts and large amounts of memory, such as inference, these Microsoft Azure Ddsv5-series VMs are a great choice.

We used the Hugging Face - BERT Large inference workload to measure the inference performance of two sizes of Microsoft Azure VMs. We found that new Ddsv5 VMs enabled by 3rd Gen Intel Xeon Scalable processors delivered up to 1.65x more inference work as Ddsv4 VMs with older processors.

Achieve More Inference Work with 32-vCPU VMs

Choosing Microsoft Azure Ddsv5 VMs with newer processors for your inference workloads rather than older VMs can improve performance per VM. In Hugging Face - BERT Large testing of 32-vCPU VMs, Azure Ddsv5 VMs enabled by 3rd Gen Intel Xeon Scalable processors handled up to 1.46x more inference work than a Ddsv4 VM enabled by previous-generation processors (see Figure 1).

Hugging Face - BERT Large

Handle up to 1.46x More Inference Work with 32-vCPU Ddsv5 VMs

vs. Ddsv4 VMs

Handle up to 1.65x More Inference Work with 48-vCPU Ddsv5 VMs

vs. Ddsv4 VMs

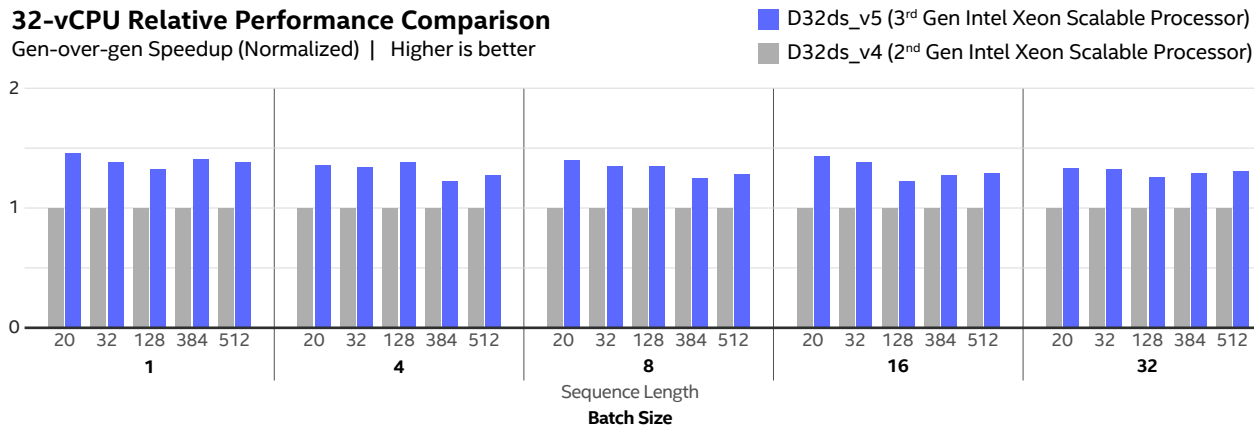

Figure 1. Relative Hugging Face - BERT Large performance of the 32-vCPU Azure Ddsv5 VM and 32-vCPU Azure Ddsv4 VM types.

Achieve More Inference Work with 48-vCPU VMs

In Hugging Face – BERT Large testing of 48-vCPU VMs, Azure Ddsv5 VMs enabled by 3rd Gen Intel® Xeon® Scalable processors handled up to 1.65x more inference work than a Ddsv4 VM enabled by previous-generation processors (see Figure 2).

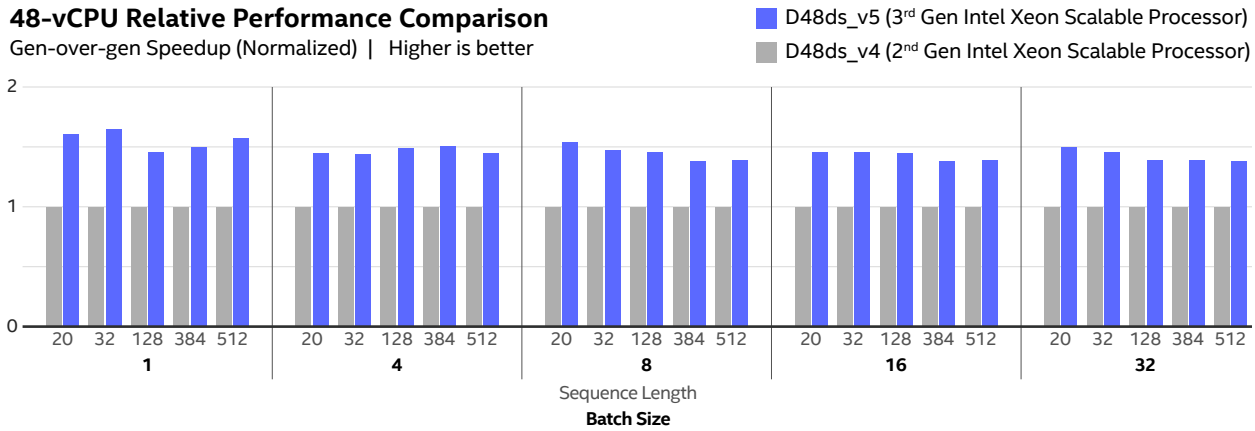

Figure 2. Relative Hugging Face - BERT Large performance of the 48-vCPU Azure Ddsv5 VM and 48-vCPU Azure Ddsv4 VM types.

Learn More

To begin running your DL inference workloads on Microsoft Azure Ddsv5 virtual machines with 3rd Gen Intel Xeon Scalable processors, visit https://intel.com/microsoftazure.

For pricing information, visit https://azure.microsoft.com/en-us/pricing/details/virtual-machines/linux/#pricing.

Testing, conducted by Intel on 8/20/2021, using Ubuntu 20.04.2 LTS, 5.8.0-1039-azure, Direct Attached storage, HF Transformer 4.9.2, oneAPI 2021.3.0.3219, gperftools 2.9.0, Python 3.8.10, Pytorch 1.9.0 + Torchscript.32 vCPU testing compared 1-VM, D32ds_v4 instance with 32x vCPU, Intel Xeon Platinum 8272CL CPU @ 2.60GHz vs. 1-VM, D32ds_v5 instance with 32x vCPU, Intel Xeon Platinum 8370C CPU @ 2.80GHz. Both VMs had 128 GB total DDR4 memory and used 16000 Mbps network. Greatest performance advantage achieved with Hugging Face bert-large-cased model, batch size 1, sequence length 20. 48 vCPU testing compared 1-VM, D48ds_v4 instance with 48x vCPU, Intel Xeon Platinum 8272CL CPU @ 2.60GHz vs. 1-VM, D48ds_v5 instance with 48x vCPU, Intel Xeon Platinum 8370C CPU @ 2.80GHz. Both VMs had 192 GB total DDR4 memory and used 24000 Mbps network. Greatest performance advantage achieved with Hugging Face bert-large-cased model, batch size 1, sequence length 32.

Printed in USA 1221/JO/PT/PDF US001 ♻ Please Recycle