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

Improve AI inference performance with HPE ProLiant DL380 Gen11 servers, powered by 4th Generation Intel Xeon Gold processors

In ResNet-50 image-recognition testing, these servers handled significantly more samples per second than previous-generation HPE ProLiant servers while delivering lower latency

Companies across a wide range of industries are embracing AI inference applications. From healthcare to retail to manufacturing, these applications are solving business problems and providing functionality that would have been hard for most of us to imagine a few decades ago. Because artificial intelligence is computationally demanding, decision makers might assume their only options for running these workloads are engaging costly cloud solutions or, if they want to run applications in their own data centers, investing in expensive servers backed by graphics processing units (GPUs). In fact, many organizations can meet the requirements of their AI applications on site with powerful HPE ProLiant DL380 Gen11 servers featuring 4th Generation Intel® Xeon® Gold processors. These processors include accelerators such as Intel® Advanced Matrix Extensions (AMX) to improve AI performance.

In the PT data center, we conducted a series of ResNet-50 tests to quantify the AI inference performance potential of these servers. First, we tested the Floating Point 32-bit (FP32) inference performance of the HPE ProLiant DL380 Gen11 server and compared it to that of its previous-generation counterpart with older processors. Next, we investigated performance of the new server with lower precision levels, which leverages accelerators in the 4th Generation Intel Gold processor to increase throughput.

With an FP32 precision level, the HPE ProLiant DL380 Gen11 server processed 2.86 times as many images per second as the older server did. At two precision levels that the older server does not accelerate, the HPE ProLiant DL380 Gen11 server also delivered strong performance.

Based on these findings, the HPE ProLiant DL380 Gen11 server is an excellent candidate for running a wide range of image-recognition workloads for many organizations.

Process 2.86x many images per second at FP32 precision levels*

Reduce latency by 30.1% at FP32 precision levels*

Enjoy strong performance at Int8 and bfloat16 precision levels

*HPE ProLiant DL380 Gen11 server featuring Intel Xeon Gold 6430 processors vs. HPE ProLiant DL380 Gen10 server featuring Intel Xeon Gold 6130 processors

Improve AI inference performance with HPE ProLiant DL380 Gen11 servers, powered by 4th Generation Intel Xeon Scalable processors December 2023

The science behind the report:

Improve AI inference performance with HPE ProLiant DL380 Gen11 servers, powered by 4th Generation Intel Xeon Gold processors

This document describes what we tested, how we tested, and what we found. To learn how these facts translate into real-world benefits, read the report **Improve AI inference performance with HPE ProLiant DL380 Gen11 servers, powered by 4th Generation Intel Xeon Gold processors**.

We concluded our hands-on testing on November 28, 2023. During testing, we determined the appropriate hardware and software configurations and applied updates as they became available. The results in this report reflect configurations that we finalized on November 27, 2023 or earlier. Unavoidably, these configurations may not represent the latest versions available when this report appears.

Our results

To learn more about how we have calculated the wins in this report, go to **http://facts.pt/calculating-and-highlighting-wins**. Unless we state otherwise, we have followed the rules and principles we outline in that document.

Table 1: Comparing FP32 precision on the two systems under test. Higher throughput and lower latency are better. Source: Principled Technologies.

	Precision/default batch size	Median throughput (images/second)	Times as many images/second	Median average latency (seconds)	% lower latency
HPE ProLiant DL380 Gen10	FP32/116	306.158		3.026	
HPE ProLiant DL380 Gen11	FP32/116	877.471	2.86x	2.113	30.17%

Table 2: Comparing three precision levels on the HPE ProLiant DL380 Gen11. Higher throughput and lower latency are better. Source: Principled Technologies.

	Precision/default batch size	Median throughput (images/second)	Times as many images/second	Median average latency (seconds)	% lower latency
HPE ProLiant DL380 Gen11	FP32/116	877.471		2.113	
HPE ProLiant DL380 Gen11	bfloat16/80	2,819.42	3.21x	0.454	78.52%
HPE ProLiant DL380 Gen11	Int8/116	5,027.80	5.72x	0.369	82.54%

System configuration information

Table 3: Detailed information on the systems we tested.

System configuration information	HPE ProLiant DL380 Gen10	HPE ProLiant DL380 Gen11
Tested by	Principled Technologies	Principled Technologies
Test date	11/28/2023	11/28/2023
Workload and version	ResNet-50 v1.5 Imagenet	ResNet-50 v1.5 Imagenet
Workload-specific parameters	Cores per instance: 4	Cores per instance: 4
Tensorflow version	2.11.0202242	2.11.0202242
Intel GitHub repository source version	IntelAI/models:origin/master on 11/10/2023	IntelAI/models:origin/master on 11/10/2023
Iterations and result choice	3 runs, median	3 runs, median
Server platform	HPE ProLiant DL380 Gen10	HPE ProLiant DL380 Gen11
BIOS name and version	U30 v2.90	U54 v1.44
Operating system name and version/ build number	Ubuntu 22.04 Kernel 6.2.0-37-generic	Ubuntu 22.04 Kernel 6.2.0-37-generic
Date of last OS updates/patches applied	11/27/23	11/27/23
Processor		
Number of processors	2	2
Vendor and model	Intel® Xeon® Gold 6130	Intel Xeon Gold 6430
Core count (per processor)	16	32
Core frequency (GHz)	2.10	2.10
Family, model, stepping	6, 85, 4	6, 143, 8
SMT	Disabled	Disabled
Turbo	Yes (3.7 GHz)	Yes (3.4 GHz)
Memory module(s)		
Total memory in system (GB)	256	256
Number of memory modules	4	8
DIMM layout	2 x 64GB per CPU (2 of 6 channels used)	4 x 32 GB per CPU (4 of 8 channels used)
Vendor and model	Micron 72ASS8G72LZ-2G3A1	Micron MTC20F2085S1RC48BA1
Size (GB)	64	32
Type	PC4-2666	PC5-4800
Speed (MHz)	2,400	4,800
Speed running in the server (MHz)	2,400	4,400
General hardware		
Storage: NW or Direct Att / Controller	Direct Att Embedded SATA	Direct Att HPE MR416i-p Gen11
OS/data drive		
Number of drives	1	1
Drive size (TB)	1.92	1.92
Drive information (speed, interface, type)	6 Gbps, SATA, SSD	6 Gbps, SATA, SSD

How we tested

Setting up the systems

Configuring BIOS settings

We applied the recommended BIOS adjustments according to Intel guidance and enabled the maximum performance options available on both systems under test (SUTs) (see Table 4).

Table 4: The BIOS settings we applied to both SUTs based on recommendations by Intel.

BIOS setting	Value recommended by Intel
Hyperthreading	Disabled
Turbo Boost	Enabled
Core Prefetchers	Hardware, Adjacent Cache, DCU Streamer, DCU IP
LLC Prefetch	Disabled
CPU Power and Perf Policy	Performance
NUMA-based Cluster	Disabled
Energy Perf Bias	Performance
Energy Efficient Turbo	Disabled
C-States	Enabled

We adjusted additional settings per system, which varied slightly in naming between the different generations of servers (see Table 5).

Table 5: Additional BIOS setting adjustments we made.

BIOS setting	HPE configuration setting
Both SUTs	
Workload Profile	Custom, based on General Peak Frequency Compute
Processor physical addressing	Default
Sub-NUMA Clustering	Disabled
Power Regulator Mode	Static High Performance
Fan and Thermal Options	Enhanced CPU Cooling
HPE ProLiant DL380 Gen11	
Advanced tuning options	
Enhanced Processor Performance Profile	Aggressive
Intel(R) AVX P1	Level 2
IODC Configuration	Auto*
Dead Block Predictor	Disabled*
Snoop Response Hold Off	9*
Snoop Response Hold Off for IOAT Stack	10*

BIOS setting	HPE configuration setting
HPE ProLiant DL380 Gen10	
Processor Jitter Control	Disabled*
Processor Config TDP Level	Level 2
PCI Peer to Peer Serialization	Disabled*
IODC Configuration	Auto*
Posted Interrupt Throttle	Enabled*

*Default option

Configuring Ubuntu 22.04

1. On a default installation of Ubuntu 22.04, log in as the user you defined during setup.
2. Extend the default logical volume and filesystem:

```
sudo lvextend -l +100%FREE /dev/ubuntu-vg/ubuntu-lv
sudo resize2fs /dev/mapper/ubuntu--vg-ubuntu--lv
```

3. Update the OS:

```
sudo apt update && sudo apt full-upgrade
```

4. Reboot.
5. Check the current kernel version, and update the kernel:

```
uname -sr
sudo apt install linux-image-generic-hwe-22.04
```

6. Reboot, and confirm the kernel is updated:

```
sudo reboot
uname -sr
```

7. Set power policy on CPUs to performance:

```
echo performance | sudo tee /sys/devices/system/cpu/cpu*/power/energy_perf_bias
```

8. Install Anaconda with default settings:

```
wget https://repo.anaconda.com/archive/Anaconda3-2023.07-2-Linux-x86_64.sh
bash ~/prereq/Anaconda3-2023.07-2-Linux-x86_64.sh
```

9. Install prerequisites:

```
sudo apt install python3-venv git numactl wget
```

10. Create and activate an Anaconda virtual environment using Python 3.8:

```
:conda create --name intqspy38 python=3.8
conda activate intqspy38
```

11. Create and activate a Python virtual environment in the Anaconda/Python 3.8 environment:

```
pip install virtualenv
virtualenv -p /home/<username>/anaconda3/envs/intqspy38/bin/python venv-tf
source venv-tf/bin/activate
```

12. Install Intel Optimized TensorFlow into the Python environment:

```
pip install intel-tensorflow==2.11.dev202242
pip install keras-nightly==2.11.0.dev2022092907
```

13. Make directories for the ResNet50 data, models, and output logs:

```
mkdir ~/imagenet-tf
mkdir ~/resnet50-log
mkdir ~/models
```

14. Copy the preprocessed data for TensorFlow to `~/imagenet-tf`.

Running TensorFlow ResNet50 v1.5 benchmarks

1. Download the models for each precision:

```
cd ~/models
wget https://zenodo.org/record/2535873/files/resnet50_v1.pb
wget https://storage.googleapis.com/intel-optimized-tensorflow/models/v1_8/resnet50v1_5_int8_
pretrained_model.pb
wget https://storage.googleapis.com/intel-optimized-tensorflow/models/v1_8/resnet50_v1_5_bfloat16.pb
```

2. Pull the Intel® AI Reference Models repository from GitHub:

```
mkdir -p ~/github/intelai
cd ~/github/intelai
git clone https://github.com/IntelAI/models.git
cd ~/github/intelai/models/
```

3. Set the common ResNet-50 environment variables:

```
export DATASET_DIR=~/imagenet-tf
export OUTPUT_DIR=~/resnet50-log
export CORES_PER_INSTANCE=4
export BATCH_SIZE=""
```

Running FP32 precision

```
export PRECISION=fp32
export PRETRAINED_MODEL=~/models/resnet50_v1.pb
./quickstart/image_recognition/tensorflow/resnet50v1_5/inference/cpu/inference_throughput_
multi_instance.sh
```

Running bfloat16 precision

```
export PRECISION=bfloat16
export PRETRAINED_MODEL=~/models/resnet50_v1_5_bfloat16.pb
./quickstart/image_recognition/tensorflow/resnet50v1_5/inference/cpu/inference_throughput_
multi_instance.sh
```

Running Int8 precision

```
export PRECISION=int8
export PRETRAINED_MODEL=~/models/resnet50v1_5_int8_pretrained_model.pb
./quickstart/image_recognition/tensorflow/resnet50v1_5/inference/cpu/inference_throughput_
multi_instance.sh
```

Read the report at https://facts.pt/Jj5UV9r ▶

This project was commissioned by HPE.

Principled Technologies®

Facts matter.®

DISCLAIMER OF WARRANTIES; LIMITATION OF LIABILITY:
Principled Technologies, Inc. has made reasonable efforts to ensure the accuracy and validity of its testing, however, Principled Technologies, Inc. specifically disclaims any warranty, expressed or implied, relating to the test results and analysis, their accuracy, completeness or quality, including any implied warranty of fitness for any particular purpose. All persons or entities relying on the results of any testing do so at their own risk, and agree that Principled Technologies, Inc., its employees and its subcontractors shall have no liability whatsoever from any claim of loss or damage on account of any alleged error or defect in any testing procedure or result.

In no event shall Principled Technologies, Inc. be liable for indirect, special, incidental, or consequential damages in connection with its testing, even if advised of the possibility of such damages. In no event shall Principled Technologies, Inc.'s liability, including for direct damages, exceed the amounts paid in connection with Principled Technologies, Inc.'s testing. Customer's sole and exclusive remedies are as set forth herein.