Champion big decisions and gutsy moves with the new HP Z8 Fury G5 Workstation Desktop PC

vs. an HP Z8 G4 Workstation Desktop PC

For companies with an eye to the future, equipping technical professionals with well-engineered workstations can lead to shorter analysis times for large datasets, giving data scientists the ability to run more iterations as quickly as possible. That's in part because data science is grounded in experimentation. When data scientists are working with big data or large models, the number of times they can try new configurations or explore groundbreaking parameters is limited when they don't have adequate resources to do so in a timely manner.

At Principled Technologies, we ran medical imaging, language processing, and computer vision scenarios from the MLPerf™ Inference Benchmark Suite on the new HP Z8 Fury G5 Workstation and compared its output to that of its predecessor, the HP Z8 G4 Workstation.

These results are relevant to organizations and facilities that want to advance medical research and treatment endeavors, improve customer experiences, and achieve higher levels of productivity.

Accelerate 3D medical image segmentation

More 3D U-Net samples per second

Improve customer experiences

More RNN-T and BERT-99 samples per second

Identify objects and people faster

More ResNet-50 samples per second

How we tested

Before we started testing, we set the G4 workstation power mode to "high performance" and the G5 workstation power mode to "ultimate performance". Other than making and verifying those changes, we used out-of-box OEM performance settings. We tested the best configurations available for each generation:

HP Z8 Fury G5 Workstation
1x 56-core Intel® Xeon® w9-3495X CPU (1.9 - 4.8 GHz)
4x NVIDIA® RTX 6000 Ada-generation GPUs
128GB DDR5-4800 memory
4x 1 TB NVMe® SSDs

HP Z8 G4 Workstation
2x (28-core) Intel Xeon 6258R CPU (2.7 - 4.0 GHz)
2x NVIDIA RTX A6000 GPUs
96GB DDR5-2666 memory
2x 1 TB NVMe SSDs

We ran MLPerf Inference Benchmark Suite machine learning models in the offline scenario, where the workstations can process the data in any order, without latency constraints.[1] We ran each offline model three times and report the median results:

- **3D U-Net**, which measures medical imaging and 3D image segmentation performance.
- **RNN-T**, which measures speech recognition performance.
- **BERT**, which measures natural language processing performance.
- **ResNet**, which measures image classification and detection performance.

The machine learning results we report reflect the specific configurations we tested. Any difference in the configurations you test can affect these results. For a deeper dive into our testing parameters and procedures, see the science behind the report.

About the HP Z8 Fury G5

According to HP, the HP Z8 Fury G5 Workstation contains "transformative" single-socket Intel Xeon w9 processor technology with up to 56 cores, up to 1.5 TB of high-speed memory, up to 56 TB of storage, ISV certification for professional apps, and four NVIDIA RTX 6000 Ada-generation GPUs. This combination enables users to tackle the most complex simulations, virtual production, and high-quality VFX projects.[2]

While we didn't test internal or external security features on this model, the rack-mountable HP Z8 Fury G5 Workstation includes lockable front access carriers, side panel locks with an interlock sensor, and a Kensington lock slot to prevent the physical removal of the workstation. Plus, HP Anyware Remote System Controller allows your designated IT team to remotely manage your workstation fleet from a single interface.[3]

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

Tackle complex AI/ML problems

The final stage of the machine learning process is inference—that's the golden time when your proven ML model has all the data, training, evaluation, and tuning your experts deem necessary for that model to make informed predictions.[4]

At the beginning of 2020, according to Seed Scientific, the amount of data in the world was estimated at 44 zettabytes (a single zettabyte has 21 zeros). This whopping amount of information includes data generated by "social media sites, financial institutions, medical facilities, shopping platforms, automakers, and many other activities online."[5] And, while your business may not want or need to process even a single petabyte of data, your business intelligence analysts, digital transformation specialists, data analysts, and scientists need powerful workstations that can provide usable results as quickly as possible.

The medical imaging, language processing, and computer vision scenarios we ran use trained models to measure how quickly each workstation processed inputs and produced results.[6]

Advance medical research and treatment endeavors

The healthcare sector uses medical imaging (e.g., x-rays, ultrasounds, MRIs, and CT scans) for medical research, disease diagnosis, and drug discovery.[7] The 3D U-Net model we ran "performs volumetric segmentation of dense 3D images for medical use cases."[8] A higher number of samples here equals a cleaner image—but can often result in longer image processing times. As you can see from the significantly shorter wait time (latency), the Z8 Fury G5 Workstation can give medical professionals an advantage when a speedy and correct prognosis can make a big difference.

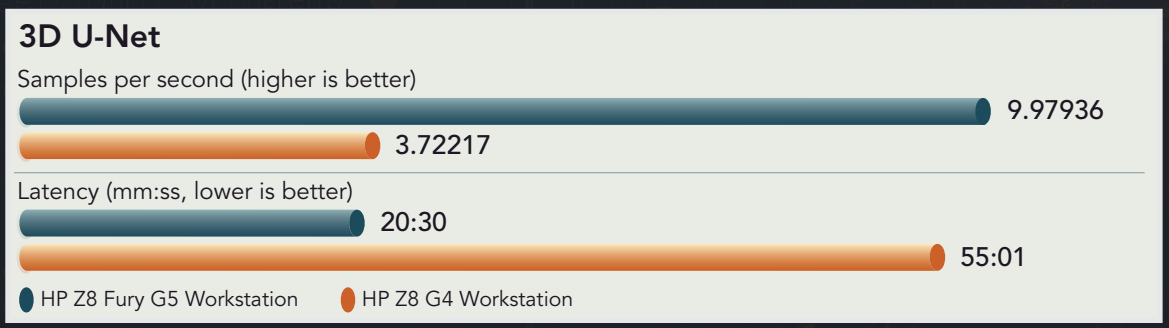

Figure 1: Number of samples per second each workstation classified and latency using the 3D U-Net model in the offline scenario. Higher numbers of samples are better, and lower latency is better. Source: Principled Technologies.

Improve customer experiences

An email spam filter was one of the first natural language processing (NLP) applications. Now, the recurrent neural network (RNN) aspect of NLP, which takes its context from word order and punctuation, helps systems recognize whether an incoming email is categorized as primary, social, promotional, or spam based on its contents.[9] The RNN-T model we ran "recognizes and transcribes audio in real time."[10] Other examples of RNN applications are language translation, stock price prediction, and text mining.[11]

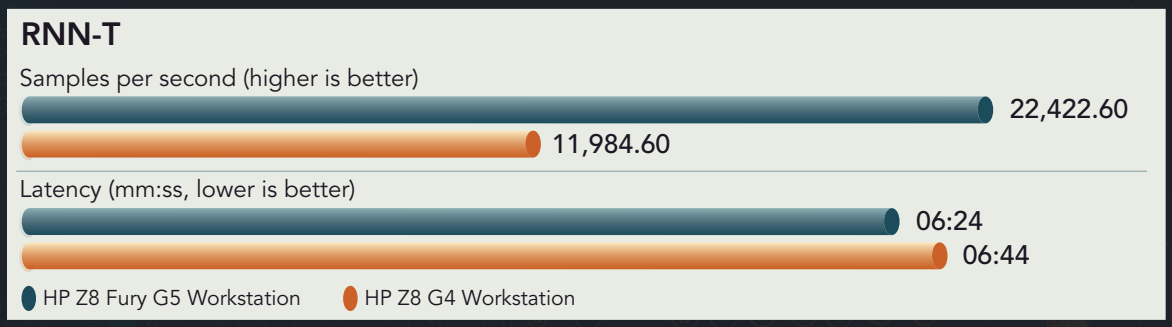

Figure 2: Number of samples per second each workstation classified and mean latency using the RNN-T model in the offline scenario. Higher numbers of samples are better and lower latency is better. Source: Principled Technologies.

Another NLP model is BERT (Bidirectional Encoder Representations from Transformers). BERT, unlike RNN, can "capture the semantic and syntactic features of a text."[12] The BERT model we ran sorts and analyzes text to make accurate language predictions, answer questions correctly, and respond to conversations without errors 99 percent of the time.[13] Real-world examples of BERT applications include sentiment analysis, chatbots, image and video captioning, and virtual assistants (e.g., Alexa, Google Assistant, and Siri).[14]

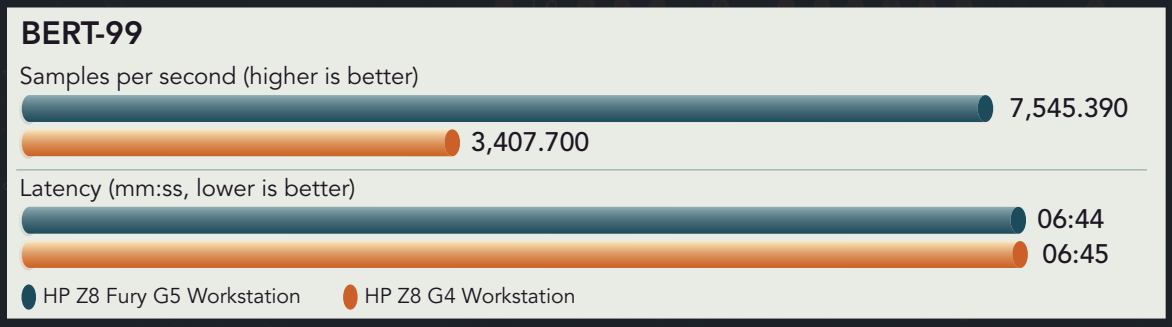

Figure 3: Number of samples per second each workstation classified and latency using the BERT-99 model in the offline scenario. Higher numbers of samples are better and lower latency is better. Source: Principled Technologies.

Identify objects and people faster

The 50-layer ResNet model we ran "[a]ssigns a label from a fixed set of categories to an input image, i.e., applies to computer vision problems."[15] Computer vision enables computers to mimic the way humans use vision to see, identify, and understand both objects and people in images and video. Computer vision application examples include facial recognition, autonomous cars, plant species classification, edge computing, sports performance analysis, and production-line automation.[16]

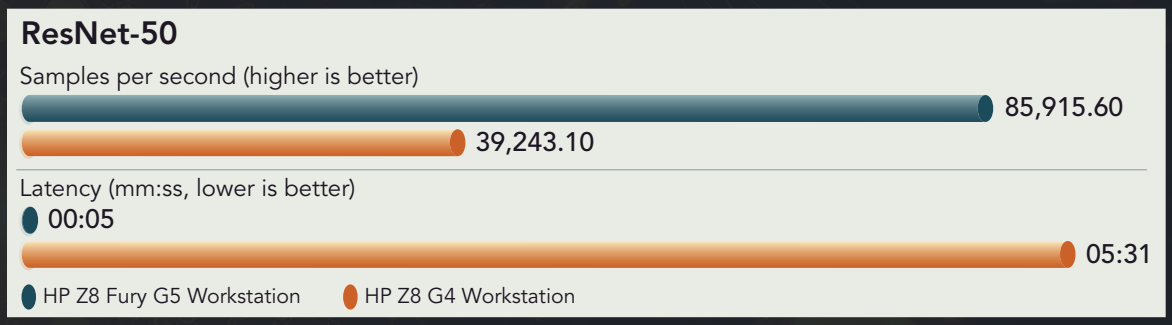

Figure 4: Number of samples per second each workstation classified and mean latency using the ResNet-50 model in the offline scenario. Higher numbers of samples are better and lower latency is better. Source: Principled Technologies.

About the Intel Xeon W-3400 processor architecture

According to Intel, this new line of desktop workstation processors, which includes the Intel Xeon w9-3495X processor we tested, are purpose-built for media and entertainment creatives as well as engineering and data science professionals. With the "breakthrough new compute architecture, faster cores and new embedded multi-die interconnect bridge (EMIB) packaging, the Xeon W-3400 and W-2400 series of processors enable unprecedented scalability for increased performance."[17]

To learn more about the Intel Xeon w9-3495X processor in the HP Z8 Fury G5 Workstation we tested, visit https://www.intel.com/content/www/us/en/products/sku/233483/intel-xeon-w93495x-processor-105m-cache-1-90-ghz/specifications.html.

Conclusion

Our medical imaging, language processing, and computer vision machine learning results show that data scientists, medical personnel, and engineers can process more samples in less time by upgrading to the new HP Z8 Fury G5 Workstation powered by an Intel Xeon w9-3495X CPU and four NVIDIA RTX 6000 Ada-generation GPUs.

1. Sally Ward-Foxton for the EE Times, "Understanding MLPerf Benchmark Scores," accessed September 22, 2023, https://www.eetimes.com/understanding-mlperf-benchmark-scores/.
2. HP, "HP Z8 Fury," accessed August 22, 2023, https://www.hp.com/us-en/workstations/z8-fury.html.
3. HP, "HP Anyware Remote System Controller," accessed August 22, 2023, https://www.hp.com/us-en/solutions/anyware-remote-system-controller.html.
4. Matthew Mayo, "Frameworks for Approaching the Machine Learning Process," accessed September 20, 2023, https://www.kdnuggets.com/2018/05/general-approaches-machine-learning-process.html.
5. Seed Scientific, "How Much Data Is Created Every Day? +27 Staggering Stats,' accessed September 21, 2023, https://seedscientific.com/how-much-data-is-created-every-day/.
6. NVIDIA, "What is MLPerf?" accessed September 22, 2023, https://www.nvidia.com/en-us/data-center/resources/mlperf-benchmarks/.
7. Simplilearn, "Top 25 Deep Learning Application Industries," accessed September 22, 2023, https://www.simplilearn.com/tutorials/deep-learning-tutorial/deep-learning-applications.
8. NVIDIA, "What is MLPerf?" accessed September 22, 2023, https://www.nvidia.com/en-us/data-center/resources/mlperf-benchmarks/.
9. Tableau, "8 Natural Language Processing (NLP) Examples," accessed September 22, 2023, https://www.tableau.com/learn/articles/natural-language-processing-examples#:.
10. NVIDIA, "What is MLPerf?" accessed September 22, 2023, https://www.nvidia.com/en-us/data-center/resources/mlperf-benchmarks/.
11. Great Learning, "What Is Recurrent Neural Network | Introduction of Recurrent Neural Network," accessed September 22, 2023, https://www.mygreatlearning.com/blog/recurrent-neural-network/.
12. LinkedIn, "How do you compare and contrast BERT with other deep learning approaches for sentiment analysis?"Accessed October 16, 2023, https://www.linkedin.com/advice/0/how-do-you-compare-contrast-bert-other-deep-learning#:~:text=BERT.
13. NVIDIA, "What is MLPerf?" accessed September 22, 2023, https://www.nvidia.com/en-us/data-center/resources/mlperf-benchmarks/.
14. Simplilearn, "Top 25 Deep Learning Application Industries," accessed September 22, 2023, https://www.simplilearn.com/tutorials/deep-learning-tutorial/deep-learning-applications.
15. NVIDIA, "What is MLPerf?" accessed September 22, 2023, https://www.nvidia.com/en-us/data-center/resources/mlperf-benchmarks/.
16. Built In, "What Is Computer Vision?" accessed September 22, 2023, https://builtin.com/machine-learning/computer-vision.
17. Intel, "Intel Launches new Xeon Workstation Processors—the Ultimate Solution for Professionals," accessed August 22, 2023, https://www.intel.com/content/www/us/en/newsroom/news/intel-launches-new-xeon-workstation-processors.html#gs.4quj4k.

Read the science behind this report at https://facts.pt/PxCsa38 ▶

Principled Technologies®

Facts matter.®

This project was commissioned by HP.