

## Significant AI inference performance advances with the HPE ProLiant DL380 Gen11 server, powered by 4<sup>th</sup> Generation Intel Xeon Gold processors

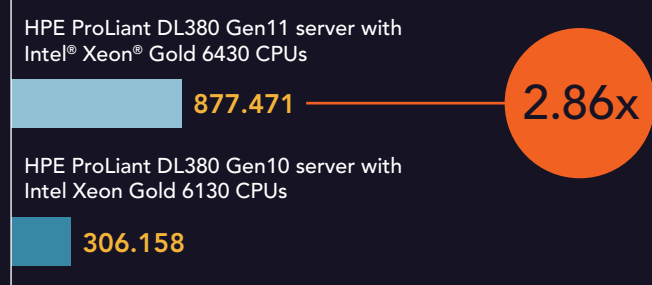
In ResNet-50 image-recognition testing, the Gen11 server handled dramatically more samples per second than previous-generation HPE ProLiant server while delivering lower latency

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

Reduce latency by 30.1% at FP32 precision levels\*

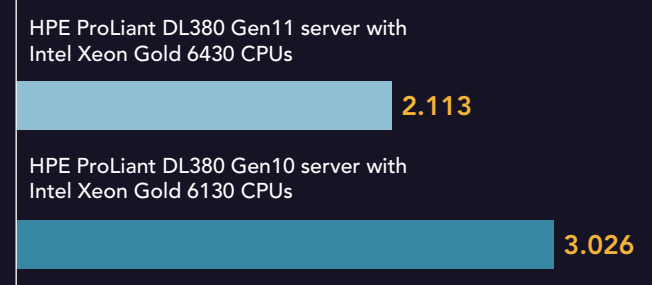
### Images per second

Higher is better



### Latency

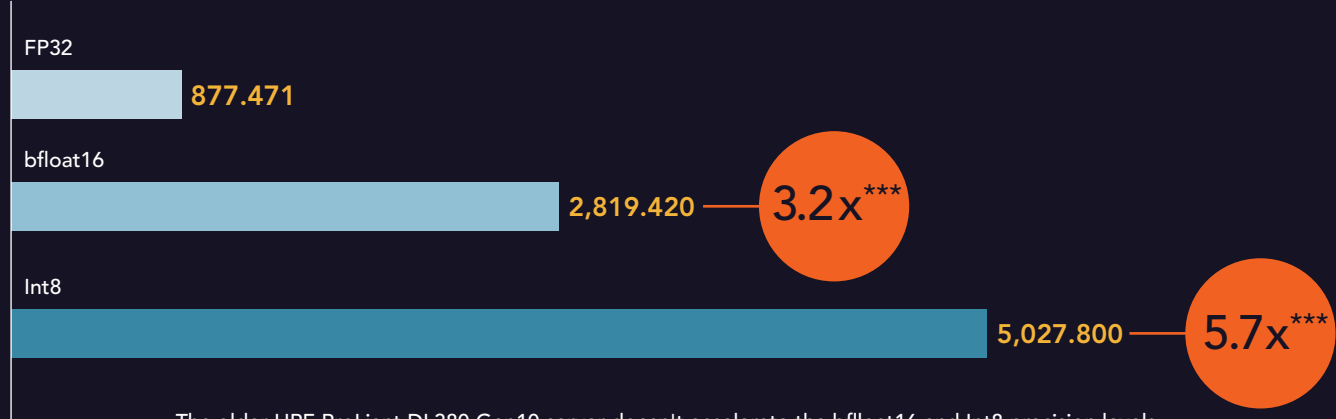
Seconds | Lower is better



Utilize built-in AMX accelerators\*\* in the 4<sup>th</sup> Generation Intel Xeon Gold 6430 processor to increase throughput at lower precision levels

### Images per second the HPE ProLiant DL380 Gen11 server handled at different precision levels

Higher is better



The older HPE ProLiant DL380 Gen10 server doesn't accelerate the bfloat16 and Int8 precision levels.

Learn more at <https://facts.pt/Jj5UV9r>



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

\*\*Intel, "Advanced Matrix Extensions Overview," accessed November 28, 2023, <https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/advanced-matrix-extensions/overview.html>.

\*\*\*times as many images per second as FP32 on the HPE ProLiant DL380 Gen11 server

Copyright 2024 Principled Technologies, Inc. Based on "Improve AI inference performance with the HPE ProLiant DL380 Gen11 server, powered by 4<sup>th</sup> Generation Intel Xeon Gold processors," a Principled Technologies report, January 2024. Principled Technologies® is a registered trademark of Principled Technologies, Inc. All other product names are the trademarks of their respective owners.