

Get Recommendations from Data up to 2.94x Faster with 2nd Gen Intel[®] Xeon[®] Scalable Processor-Based AWS M5n Instances

HiBench



2.86x the samples per second with 8-vCPU M5n instances

vs. M4 instances



2.94x the samples per second with 16-vCPU M5n instances

vs. M4 instances



2.67x more samples per second with 64-vCPU M5n instances

vs. M4 instances

Improve Wide & Deep Inference Performance with AWS M5n instances featuring 2nd Gen Intel[®] Xeon[®] Scalable processors

Collecting data from customers is only useful if you can quickly make relationships between that data to target specific needs and desires to boost sales or increase customer satisfaction. Utilizing wide linear models and deep neural networks to infer relationships between data, Wide & Deep workloads deliver real-time recommendations based on your data. Tests show that choosing AWS M5n Instances enabled by 2nd Gen Intel Xeon Scalable processors over M4 instances with previous-generation processors can improve Wide & Deep recommendation engine performance. The 2nd Gen Intel Xeon Scalable processor family features Intel Deep Learning Boost, which improves machine learning performance. In third-party testing conducted by Principled Technologies, across three different instance sizes, M5n instances featuring Intel Xeon Platinum 8272CL processors handled up to 2.94x the samples per second than M4 instances. With M5n instances, organizations can speed deep learning workloads and make sense of data faster, getting recommendations based on that data in less time.

Improve Deep Learning Performance on Small Instances

The faster your cloud instances can infer meaningful relationships between data, the faster you can act on those recommendations. As Figure 1 shows, 8-vCPU M5n instances enabled by 2nd Gen Intel Xeon Scalable processors outperformed 8-vCPU M4 instances in a deep learning Wide & Deep benchmark test. The 2.86 times increase of samples per second means they can process data and make recommendations faster.

Relative Wide & Deep throughput at 8 vCPU

Samples per second | Higher is better

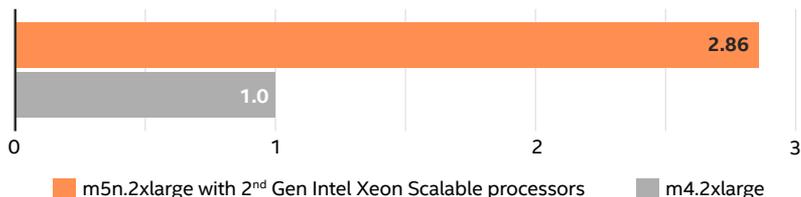


Figure 1. Relative results comparing the Wide & Deep benchmark performance of small (8-vCPU) M5n instances vs. M4 instances (Samples per second).

Improve Deep Learning Performance on Medium Instances

Organizations with mid-sized datasets can also get improved deep learning inference performance by choosing instances with newer processors. As Figure 2 shows, 16-vCPU AWS M5n instances enabled by 2nd Gen Intel® Xeon® Scalable processors handled 2.94 times the samples per second in Wide & Deep tests compared to M4 instances with previous-generation processors.

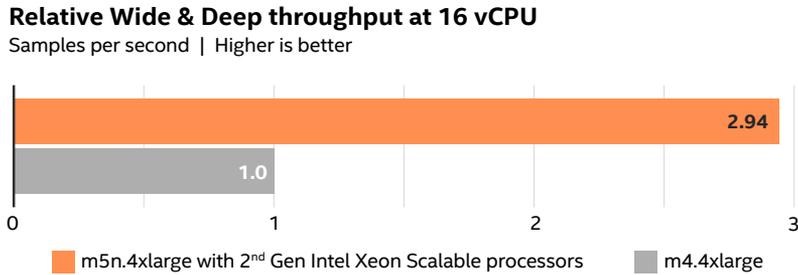


Figure 2. Relative results comparing the Wide & Deep benchmark performance of medium(16-vCPU) M5n instances vs. M4 instances (Samples per second).

Improve Deep Learning Performance on Large Instances

Larger datasets that require larger instances similarly benefit from choosing newer processor architecture for deep learning workloads. In tests, M5n instances featuring 2nd Gen Intel Xeon Scalable processors handled 2.67 times the samples per second using the Wide & Deep benchmark test (see Figure 3).

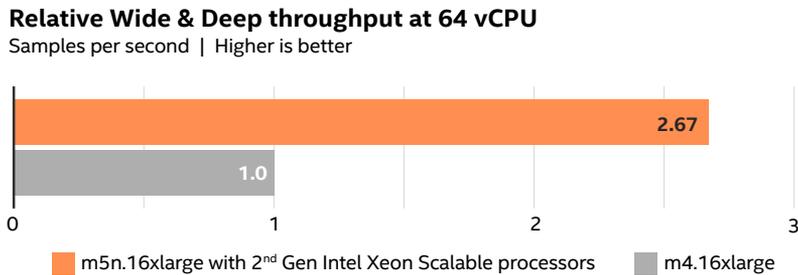


Figure 3. Relative results comparing the Wide & Deep benchmark performance of large (64-vCPU) M5n instances vs. M4 instances (Samples per second).

Whether your datasets are small, large, or somewhere in between, selecting AWS M5n instances with 2nd Gen Intel Xeon Scalable processors instead of M4 instances with older processors can enhance deep learning performance so you can make meaningful relationships from your data faster and make real-time recommendations to teams and consumers..

Learn More

To begin running your Wide & Deep workloads on AWS M5n Instances with 2nd Gen Intel Xeon Scalable processors, visit <http://intel.com/aws>.

For complete testing results, visit <http://facts.pt/oOUDy0E>.



Performance varies by use, configuration and other factors. Learn more at <https://intel.com/benchmarks>.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. Your costs and results may vary.

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

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others

Printed in USA 1021/JO/PT/PDF US001

