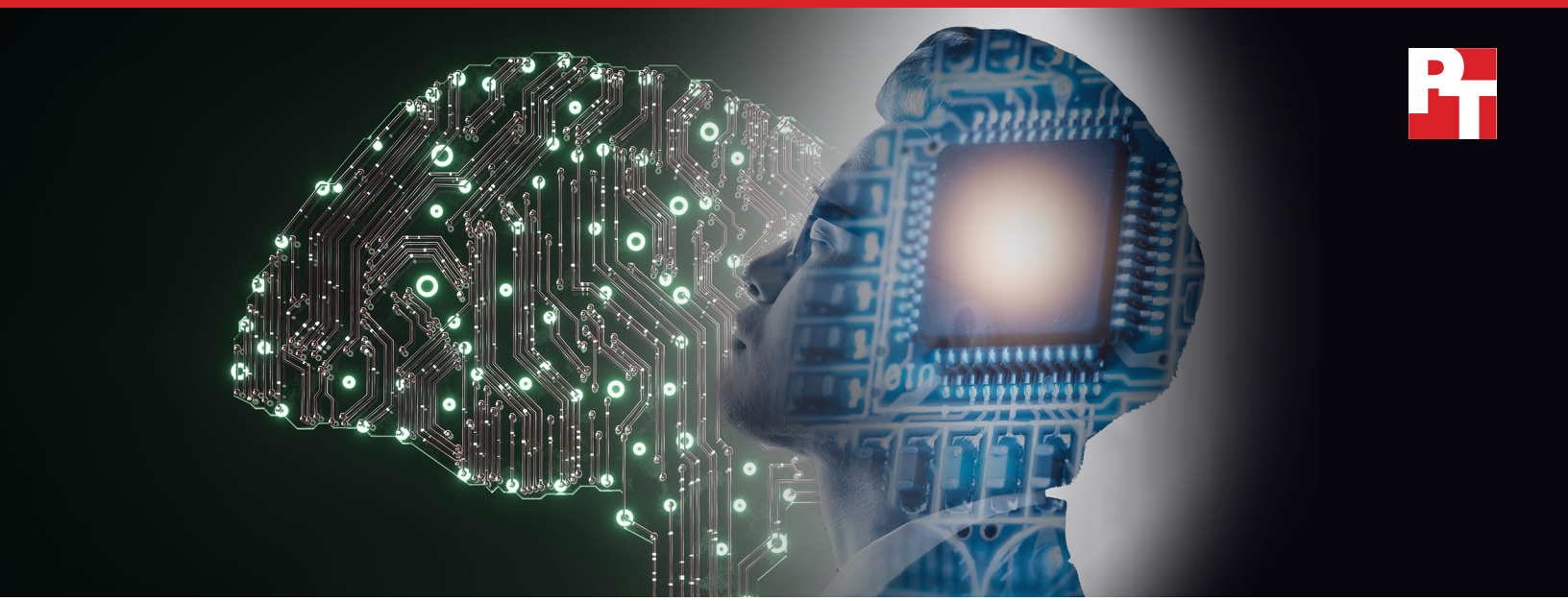


A Dell EMC server with Intel technology delivered more cost-effective performance on three image-classification models than the same server with a GPU

A Dell EMC PowerEdge R740xd with 2nd Generation Intel Xeon Scalable processors achieved comparable training and better inference at a lower hardware cost than the same server equipped with an NVIDIA T4 GPU

As the importance of artificial intelligence and machine learning grows, organizations strive to select the appropriate hardware to run the extremely demanding workloads that AI and ML technologies create. Our study found that in some AI/ML use cases, servers powered by 2nd Generation Intel® Xeon® Scalable processors can deliver stronger and more cost-effective performance on their own than with graphics processing units, or GPUs.

We tested a Dell EMC™ PowerEdge™ R740xd powered by Intel Xeon Gold 6254 processors. We ran an image-classification workload using three different models on this server configured two ways: with only the CPU and with both the CPU and an NVIDIA T4 GPU. Compared to the CPU + GPU configuration, the CPU-only server delivered comparable performance on some training tasks and stronger performance on multiple inference tasks.

Taking hardware costs into account, the CPU-only configuration achieved a better price/performance ratio on training and inference tasks on all three models. This indicates that the Dell EMC PowerEdge R740xd, powered by Intel Xeon Gold 6254 processors and without the addition of GPUs, is a cost-effective option for image-classification training and inference using GoogLeNet, Inception v3, and Inception v4 models.

Better price/ performance on training and inference

up to
20.7% better
on Inception v4

up to
17.3% better
on Inception v3

up to
16.2% better
on GoogLeNet

What we set out to explore

We designed our study around a hypothetical company poised to purchase a Dell EMC PowerEdge R740xd server, which they will use to run a variety of business applications, including machine learning applications such as the image-classification workloads we ran. The IT staff configuring the server has selected the Intel Xeon Gold 6254 processor, part of the 2nd Generation Intel Xeon Scalable processor platform (see sidebar).

To compare the server with and without GPUs, we timed inference and training on the following three image-classification models:

- Inception v3
- Inception v4
- GoogLeNet

We first performed each workload targeting the Intel Xeon Gold 6254 processor on the Dell EMC PowerEdge R740xd. We then repeated the tests, this time targeting the NVIDIA Tesla T4 in the server.

While both the NVIDIA T4 GPU and 2nd Generation Intel Xeon Scalable processors include support for reduced precision INT8 operations, our testing focused on FP32. One of the reasons for this was the lack of support for INT8 in the benchmark package we used, tf_cnn_benchmarks. The package does allow using TensorRT, which automatically applies FP32 to INT8 quantization optimizations for GPUs. However, no equivalent mechanism exists to enable optimizations that make use of Intel AVX-512 Deep Learning (DL) Boost and VNNI instructions.

We considered hand-optimizing model quantization for the CPU-only tests, but were concerned that doing so would make it difficult to assess the fairness of the comparison against TensorRT-optimized models. Altering the quantization of pre-trained models influences the results in subtle ways, chiefly in inferencing accuracy. Because we lacked a good way to compare small percentage accuracy changes to performance changes, and because FP32 still prevails and is the default precision for most platforms, we did not use TensorRT in our GPU tests, nor did we consider INT8 results on either platform.

Both platforms would have performed better with reduced precision INT8 than FP32. However, our goal was to compare absolute performance on the two platforms under test, which we could not accurately assess for reduced precision INT8. Future testing may look more closely at INT8 with highly optimized models if we can determine a way to do so objectively.

The complete details of both server configurations and the testing we performed are available in the science behind this report.

About 2nd Generation Intel Xeon Scalable processors

The latest from Intel, the 2nd Generation Intel Xeon Scalable processor platform features a wide range of processors to support the workloads you run, including Bronze, Silver, Gold, and Platinum. According to Intel, the 2nd Generation Intel Xeon Scalable platform can handle a variety of workloads, including enterprise, cloud, HPC, storage, and communications.[1] This new processor line also supports a new memory and storage technology to further accelerate workloads, Intel Optane™ DC persistent memory.

To learn more about the 2nd Generation Intel Xeon Scalable processor family, visit https://www.intel.com/content/www/us/en/products/docs/processors/xeon/2nd-gen-xeon-scalable-processors-brief.html.

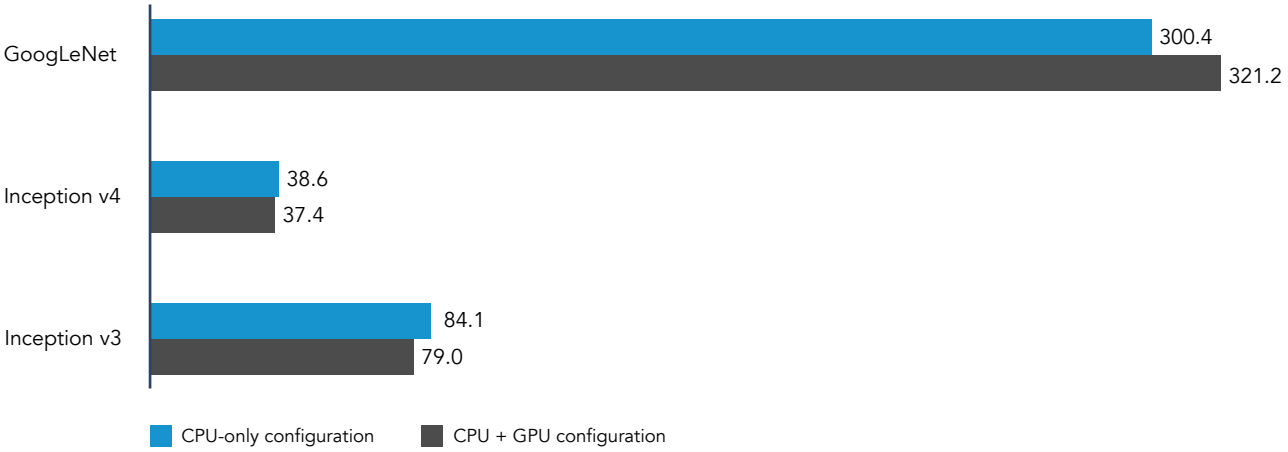
How the two server configurations performed

Generally speaking, machine learning comprises two phases: training and inference. Training prepares a model by feeding it information, and inference uses that training to make predictions. Typically, training is more compute intensive than inference.

In the training phase of our testing, the rate of images per second the Dell EMC PowerEdge R740xd processed was close on all three AI/ML models we tested, regardless of whether the workload targeted the Intel Xeon Gold 6254 Processor or the Tesla T4 GPU. The chart below shows our findings. When the GoogLeNet model targeted the GPU, we saw a modest increase of 20.8 images per second, a 6.4 percent increase over the CPU-only configuration. On the two Inception stacks, targeting the CPU resulted in greater performance than targeting the GPU, with an improvement of 3.3 percent on Inception v4 and 6.5 percent on Inception v3.

Images per second in training testing

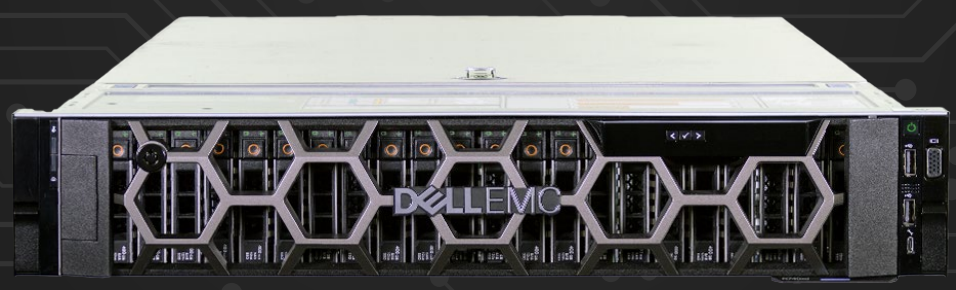
Higher is better



About the Dell EMC PowerEdge R740xd

The Dell EMC PowerEdge R740xd is a 2U, dual-socket platform powered by 2nd Generation Intel Xeon Scalable processors. It features 24 DDR4 DIMM slots and up to 271TB of storage between its front, mid, and rear bays. According to Dell EMC, the PowerEdge R740xd aims to bring scalability and performance to your datacenter.[2]

To learn more about the Dell EMC PowerEdge R740xd, visit https://www.dell.com/en-us/work/shop/povw/poweredge-r740xd.

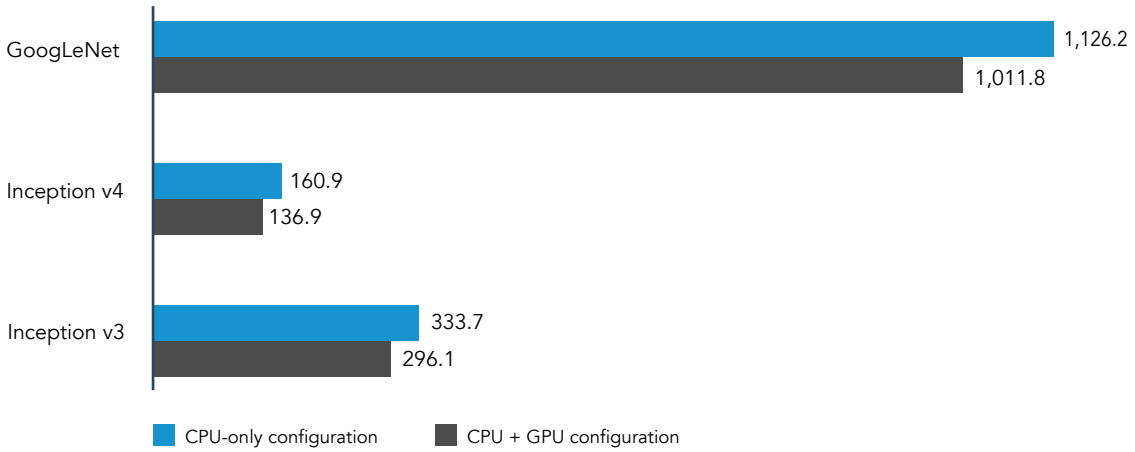


A Dell EMC server with Intel technology delivered more cost-effective performance on three image-classification models than the same server with a GPU

In the inference phase of our testing, the Dell EMC PowerEdge R740xd processed more images per second when the workload targeted the Intel Xeon Gold 6254 processor than when it targeted the Tesla T4 GPU. This was the case on all three AI/ML models, as the chart below shows. We saw the greatest difference on the Inception v4 stack, where the CPU handled 17.5 percent more images each second than the GPU. On the Inception v3 and GoogLeNet stacks, the advantage was 12.7 percent and 11.3 percent respectively.

Images per second in inference testing

Higher is better

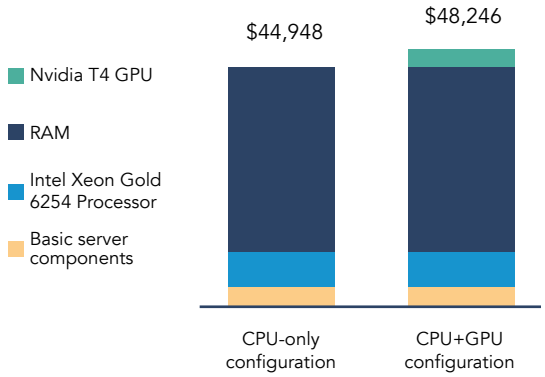


A Dell EMC server with Intel technology delivered more cost-effective performance on three image-classification models than the same server with a GPU

How much the two server configurations cost

The chart below presents list pricing for the two server configurations we tested, which we obtained from the Dell EMC website on September 5, 2019. These prices exclude discounts, sales tax, and shipping fees. (For a breakdown of the costs, see the science behind this report.)

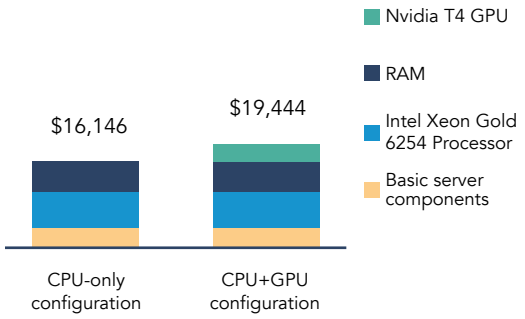
Costs of the configurations we tested
(with extra RAM)



Choosing a CPU-only configuration saves **7.3%**

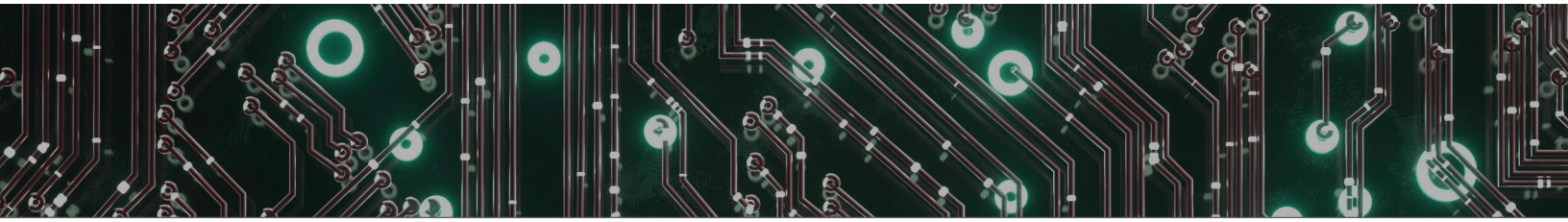
RAM was the most expensive server component, representing upwards of 70 percent the total cost of each configuration. However, we installed more RAM than necessary to ensure that neither server configuration was constrained on memory. In reality, our workloads used less than 15 percent of the physical memory in the server. By removing the cost of the excess RAM, the total cost for each server configuration drops dramatically, as the chart below shows.

Cost of hypothetical configurations
(with adequate RAM)



Choosing a CPU-only configuration saves **20.4%**

In the lower-RAM server configuration, the NVIDIA GPU would represent 16.9 percent of the total server cost—an expense that companies could avoid by choosing a CPU-only solution. Every configuration will vary based on a company's requirements for other general-purpose workloads. However, in both our high-RAM and low-RAM cost scenarios, the price/performance ratio was better on the server using the Intel Xeon processor for the machine learning tasks we tested.



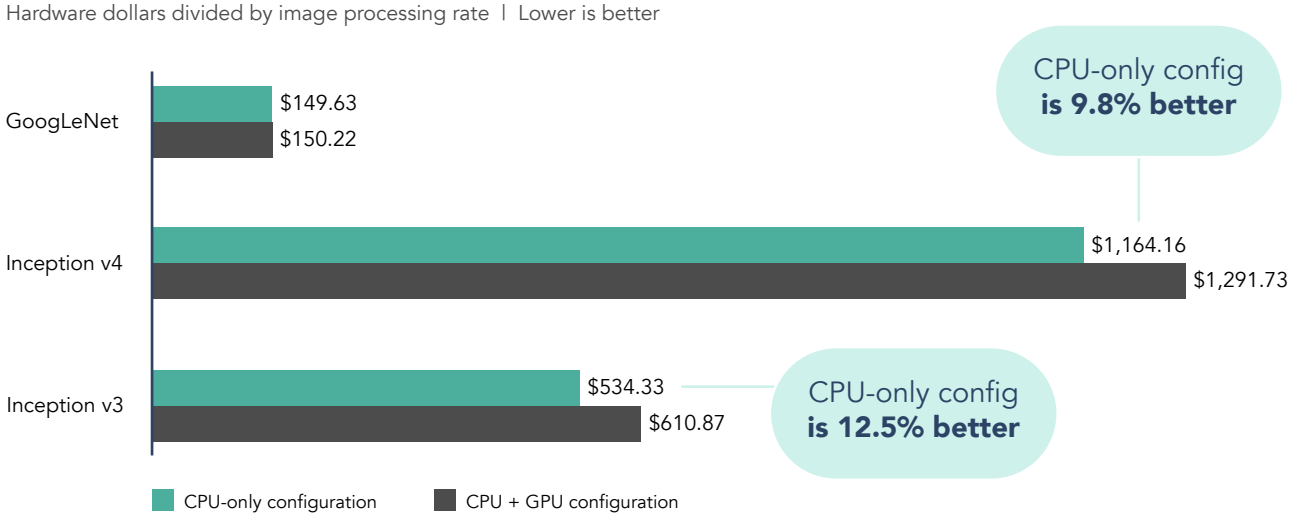
How price and performance determine cost-effectiveness

To determine the relative cost-effectiveness of the Dell EMC PowerEdge R740xd with and without the NVIDIA GPU on the workloads we tested, we divided the total hardware cost for each configuration by the average rate of images per second it achieved.

As the chart below illustrates, when we calculate price/performance using the costs we outlined on page 5 and the results of our training testing, the CPU-only configuration achieved a lower, and therefore better, price/performance ratio on all three models. On the GoogLeNet test, the difference in relative cost was very small (less than 1 percent). On the Inception v3 and v4 tests, the CPU-only configuration completed its work at a cost that was 12.5 percent and 9.8 percent lower than the GPU configuration respectively.

Price relative to performance in training testing (extra RAM config)

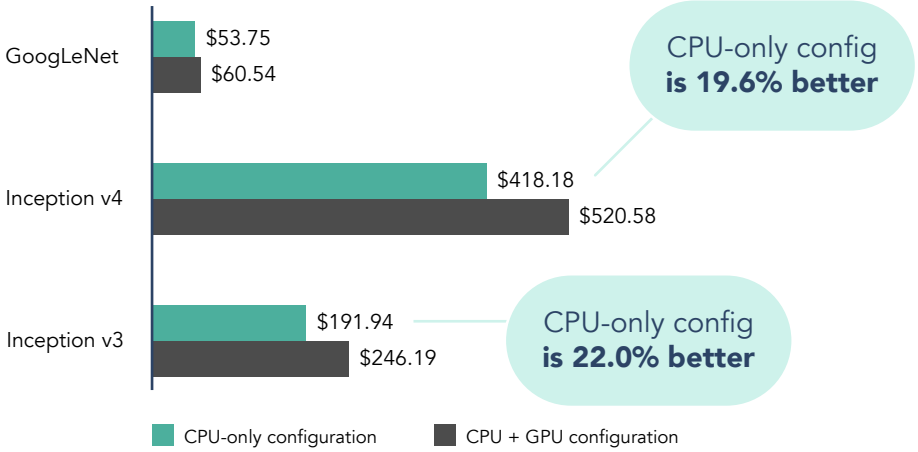
Hardware dollars divided by image processing rate | Lower is better



If we calculate the relative price using the costs we presented in the second chart on page 5, based on the amounts of RAM the workloads actually used, the differences between the two configurations increases. The CPU-only configuration would cost 11.2 percent less relative to its performance than the GPU configuration on the GoogLeNet model, 19.6 percent less on the Inception v4 model, and 22.0 percent less on the Inception v3 model.

Price relative to performance in training testing (hypothetical config)

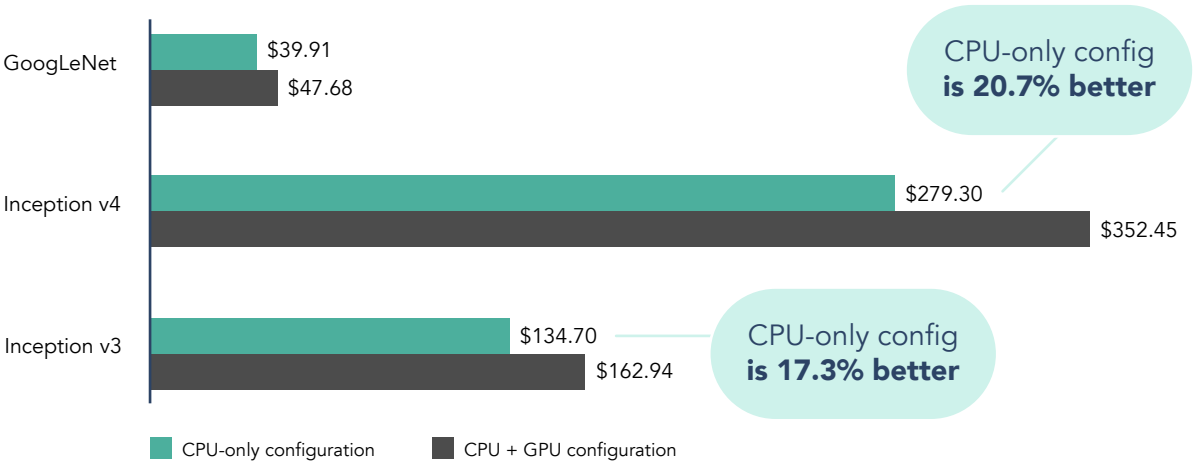
Hardware dollars divided by image processing rate | Lower is better



The chart below illustrates the relative hardware cost for the inference workload. The CPU-only configuration had a lower relative cost for all three models, with the improvements mirroring those we saw with performance. On the Inception v4 stack, the CPU-only configuration had a 20.7 percent lower relative cost than the configuration with the GPU. On the Inception v3 and GoogLeNet tests, the CPU-only configuration carried a 17.3 percent and 16.2 percent lower relative cost than the GPU configuration respectively.

Price relative to performance in training testing (extra RAM config)

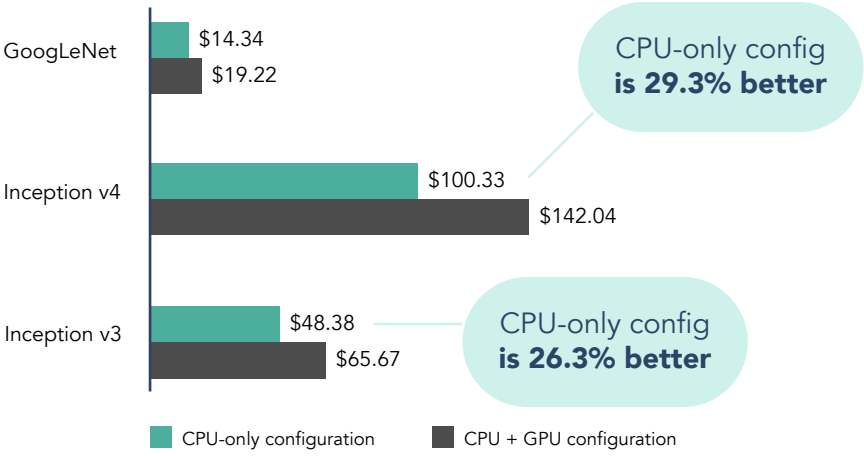
Hardware dollars divided by image processing rate | Lower is better



As we saw with inference, when we calculate the relative hardware price using the hardware costs based on the amounts of RAM the workloads actually used, the differences between the two configurations increase. The CPU-only configuration would have a relative cost that was 29.3 percent lower than the GPU configuration on the Inception v4 model, 26.3 percent lower on the Inception v3 model, and 23.3 percent lower on the GoogLeNet model.

Price relative to performance in training testing (hypothetical config)

Hardware dollars divided by image processing rate | Lower is better



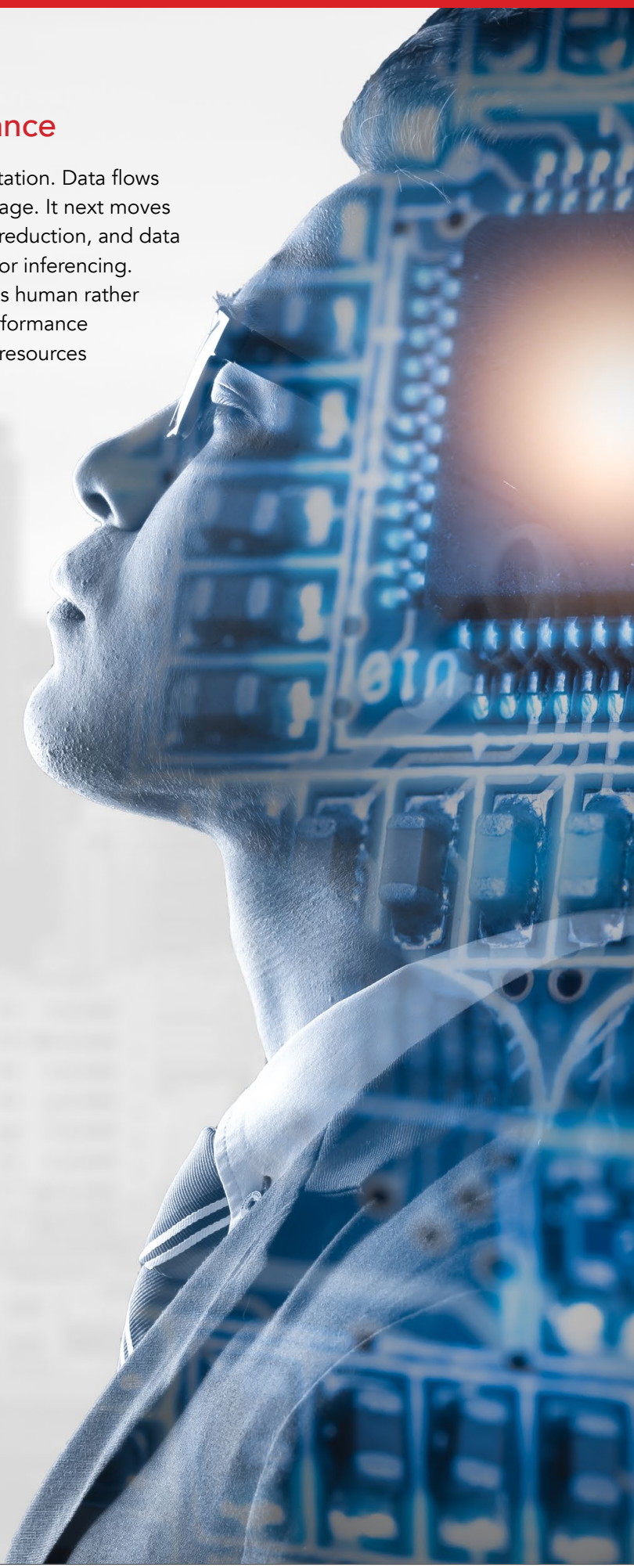
Looking beyond absolute performance

Workflows in machine learning involve more than computation. Data flows through many stages, starting with ingress and initial storage. It next moves through pre-processing, cleaning, transformation and/or reduction, and data engineering before entering neural networks for training or inferencing. At many points along this pipeline, the limiting resource is human rather than machine, and looking at only raw computational performance can paint an incomplete picture. The demand on human resources should be part of any TCO analysis.

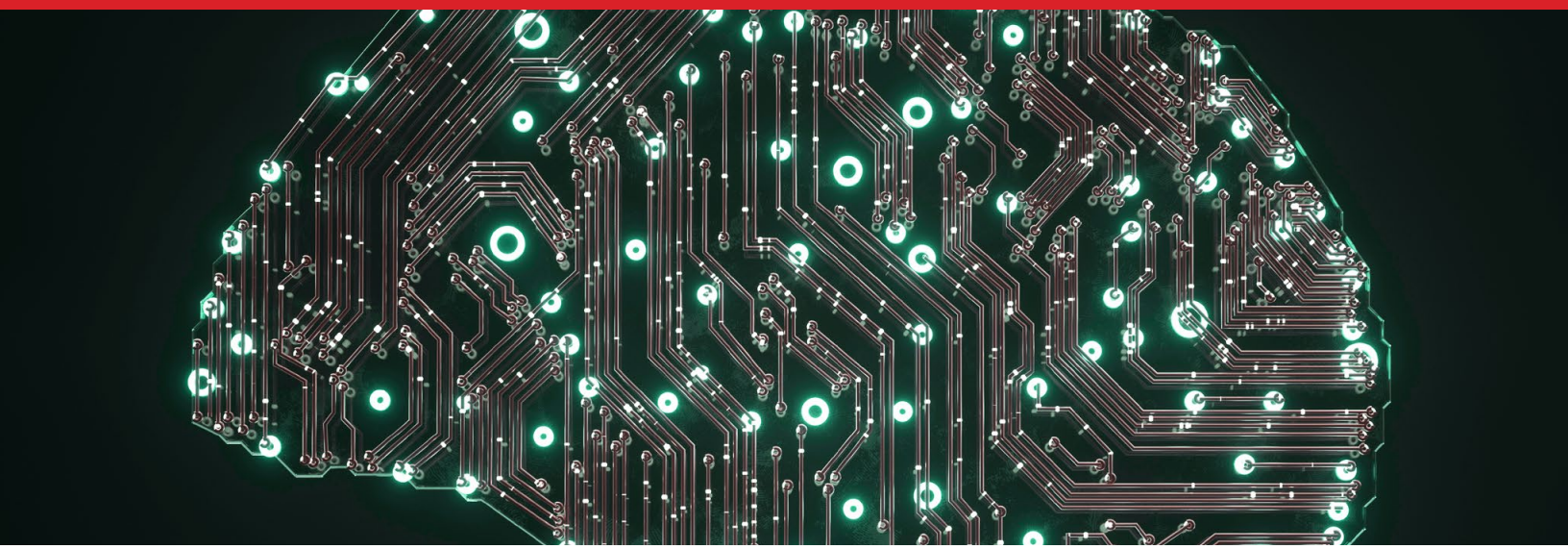
While GPUs have the potential to deliver superior performance, achieving this advantage often requires heavy lifting on the part of data engineers, who must tune problem and data geometry to maximize GPU throughput. In contrast, for many problems at which CPUs excel, strong performance can come without additional engineering.

One example stems from the fact that GPU memory is typically more limited than system RAM. To use GPU computing with models and data that don't fit in GPU memory requires engineers to perform some extra steps, such as breaking up the models into smaller chunks or configuring the ML platform to juggle the data. As long as these models or data fit in system RAM, these steps are unnecessary in a CPU-only approach. (Models and data that won't fit in system RAM require special care with both approaches.) Applications in many disciplines use datasets or models that fall into this size range, which exceeds GPU memory limits but is manageable for servers with a modest memory outlay. Examples include medical imaging, geospatial imaging, meteorology, and astronomy.

Each machine learning problem and application brings unique constraints. When making decisions about hardware architecture, organizations should consider not only the purchase price and raw performance of a solution, but also the demands it will place on their data engineers.



A Dell EMC server with Intel technology delivered more cost-effective performance on three image-classification models than the same server with a GPU



Conclusion

As more companies use artificial intelligence and machine learning to solve business problems, those making decisions about IT purchases need to select hardware that meets the computational demands of these workloads. On an image-recognition workload using the GoogLeNet, Inception v3, and Inception v4 models, a Dell EMC PowerEdge R740xd powered by Intel Xeon Gold 6254 processors delivered stronger inference performance and training performance comparable to that of the same server with an NVIDIA Tesla T4 GPU. Given the additional cost of the GPU, the server on its own delivered better price/performance across the three ML models we tested.

---

1 2nd Gen Intel Xeon Scalable Processors Brief, accessed December 9, 2019, https://www.intel.com/content/www/us/en/products/docs/processors/xeon/2nd-gen-xeon-scalable-processors-brief.html.

2 Dell EMC, "PowerEdge R740xd Rack Server," accessed December 9, 2019, https://www.dell.com/en-us/work/shop/povw/poweredge-r740xd.

Read the science behind this report at http://facts.pt/xu6c2e9 ►

Principled Technologies®

Facts matter.®

This project was commissioned by Dell Technologies.