Run TensorFlow deep learning workloads for less using IBM Cloud bare metal hosting vs. Amazon Web Services

IBM Cloud bare metal servers with GPUs delivered better frames per second (FPS) per dollar on five TensorFlow machine learning models

Organizations are increasingly turning to machine learning, a narrow subset of AI, to help turn the seemingly endless amount of data they acquire into something usable. Machine learning—more specifically deep learning—quickly and effectively solves these problems, but it takes significant computing power. These workloads require powerful GPUs alongside CPUs to get the job done. Public cloud service providers now offer NVIDIA GPU acceleration options—but what level of performance and value do they offer?

We compared the TensorFlow machine learning performance of solutions from two popular cloud hosting services: IBM Cloud®, which offers both virtual machine (VM) and bare metal hosting, and Amazon Web Services™ (AWS), which currently offers only VM hosting for GPU acceleration.

In the five TensorFlow deep learning models we tested, both IBM Cloud bare metal and VM hosting performed comparably to the AWS offering. The IBM Cloud bare metal solution we tested, however, delivered a significant improvement in performance per dollar over the AWS solution, providing the better value for machine learning of the solutions we compared.

If your organization seeks cloud hosting for TensorFlow deep learning workloads, IBM Cloud could help you get the most out of the money you spend.

Get more value for your money

Up to

17.3%

better performance per dollar

compared to Amazon Web Services

The rise of big data and using machine learning to analyze it

Research firm Gartner defines big data as "...high-volume, high-variety information assets that demand innovative forms of information processing that enable enhanced insight, decision making, and process automation."[1] Analyzing this amalgamation of data is not something humans do well—there's simply too much there and too many connections to make.

Artificial intelligence, however, can learn to make connections efficiently to give data meaning and allow businesses to glean real business insights from the information they store. It is becoming so common, in fact, that AI, Intelligent Apps and Analytics, and Intelligent Things made the top three of Gartner's Top 10 Strategic Technology Trends for 2018 list.[2] Machine learning is a specific type of AI, and deep learning is a subset of machine learning that uses multi-layer neural networks to enable the system to learn. Graphics processing units, or GPUs, excel at giving deep learning workloads the computing power they need to complete these types of operations.

With public cloud hosting also on the rise, today's cloud providers offer solutions with GPU acceleration options that make machine learning possible in the cloud. IBM Cloud is a complete cloud computing service that offers infrastructure as a service, cloud migration services, cloud application development, cloud strategy services, and more. To learn more about what IBM Cloud offers, visit www.ibm.com/cloud.

The benefits of bare metal

While virtualization surely has its advantages, bare metal has some key benefits that may fit the needs of certain organizations, including those who:

- Want greater control over customization
- Don't want to share a server, avoiding the possibility that neighboring workloads could periodically hog resources and thus hurt performance
- Have strict service-level agreements that restrict access for data protection

Using IBM bare metal could give you similar performance as AWS while enabling you to trim your cloud services budget and use those resources elsewhere.

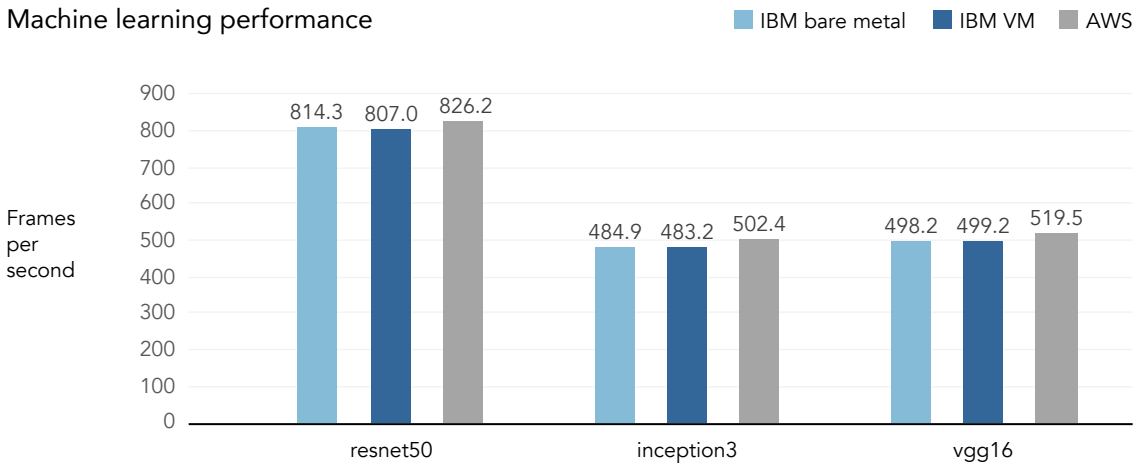
About TensorFlow

In our tests, we used TensorFlow, an open-source library for machine learning models. We chose five of the most popular publicly available deep learning models: resnet50, inception3, vgg16, alexnet, and googlenet. TensorFlow reported frames per second (FPS) that the solutions achieved using these models, with higher scores indicating better performance at these types of machine learning. For more information about TensorFlow, visit https://www.tensorflow.org/.

How the deep-learning solutions compared

If you sort through a stack of pictures, it is likely easy for you to tell which images contain dogs and which do not. That's because you've learned over time what a dog is, and what makes it different from a cat or a horse or a chicken. Computers can also learn to tell dogs from chickens, and that's the kinds of learning the TensorFlow models helped our solutions to do. When we report frames per second, we mean the number of frames, or images, the solution was able to analyze per second. For our comparison, we tested for speed alone and did not check for learning accuracy.

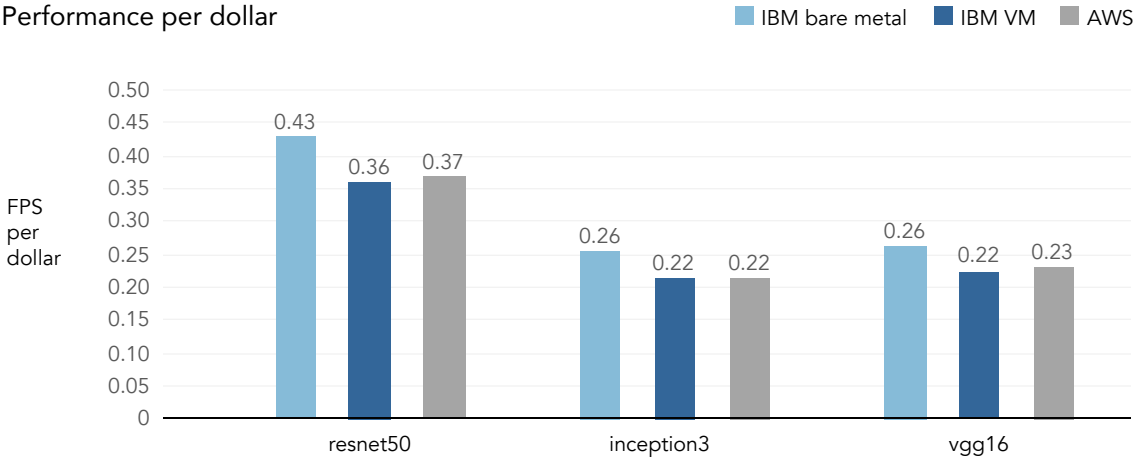
Comparable performance within 5% for each model



Machine learning models require specific tuning to achieve top performance outcomes. With this in mind, we tuned each solution multiple ways and report the best results for each solution. On all five deep learning models, the solutions performed comparably (with 5 percent or less variance), with AWS slightly edging out the IBM solutions in each test. Above, we present results from the three most popular deep learning models: resnet50, inception3, and vgg16. For alexnet and googlenet results, please see the science behind the report.

In our tests, the bare metal solution from IBM Cloud offered the highest performance per dollar, increasing FPS/dollar by 17.3 percent[3] on the resnet50 model. This means that using IBM bare metal could give you similar performance as AWS while enabling you to trim your cloud services budget and use those resources elsewhere.

Up to 17.3% better performance per dollar



The table below shows the solutions we compared at a glance. Comparing exact equivalent solutions can be difficult when using cloud providers, because options are limited. Because we were comparing GPU-intensive workloads, we chose the lowest, baseline performance options available for the processor, RAM, and drive to get a fair price comparison. Our tunings throughout discovery testing showed that these specifications had minimal effect on our GPU-intensive deep-learning workloads. The IBM Cloud bare metal hosting offering as we priced it cost 16.0 percent less than the AWS virtualized offering. For more detailed information, see the science behind the report.

Price comparison			
	IBM bare metal	**IBM VM**	**AWS**
Processor	Intel® Xeon® processor E5-2690 v4	Intel Xeon processor E5-2690 v4	Intel Xeon processor E5-2686 v4
# of cores	28	8	8
Processor speed (GHz)	2.60	2.60	2.30
RAM (GB)	64	60	61
OS drive	1TB SATA HDD	100GB SAN	100GB EBS SSD
GPU	Tesla V100-PCIE-16GB	Tesla V100-PCIE-16GB	Tesla V100-SXM2-16GB
Price/month	$1,889.00	$2,244.09	$2,249.92

IBM Cloud handled demanding TensorFlow deep-learning workloads for a better price

Organizations want to get insights from data through machine learning, but they don't want to overspend to do so. IBM Cloud offers two GPU acceleration solutions to meet this demand: one bare metal, and one virtualized. We found that the IBM Cloud bare metal offering delivered the best TensorFlow performance/dollar of the three solutions we tested, making it a smart deep-learning option for those administrators and CEOs who like to get more performance for less.

---

1 Gartner IT Glossary, accessed March 12, 2019, https://www.gartner.com/it-glossary/big-data/.

2 Kasey Panetta, "Garner Top 10 Strategic Technology Trends for 2018," accessed March 12, 2019, https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2018/.

3 We have rounded to two decimals here, but base our calculations on the numbers TensorFlow provided. Our calculations are based off those more exact numbers, which are in the science behind the report.

Read the science behind this report at http://facts.pt/3nk9q5f ▶

Principled Technologies®

Facts matter.®

This project was commissioned by IBM.