Cisco UCS C480 ML M5 Rack Server Performance Characterization

The Cisco UCS C480 ML M5 Rack Server platform is designed for artificial intelligence and machine-learning workloads.

© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Contents

© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Executive summary

This document summarizes the performance characteristics of the Cisco UCS® C480 ML M5 Rack Server Artificial Intelligence (AI) platform using the NVIDIA Tesla V100 SXM2 graphics processing unit (GPU) with 32 GB of memory and a 7-TB of Non-Volatile Memory Express (NVMe) drive. The goal of this document is to help you to understanding the performance of C480 ML MR rack server for Deep-learning (DL) workloads. Performance data was obtained using the TensorFlow Deep Learning Framework.

This document contains performance validation information for a scalable AI platform. We validated the operation and performance of this system using industry-standard benchmark tools. As reported by the validation testing results, the platform delivers excellent training performance for Deep Learning models.

Scope

The goal of this document is to characterize the performance of the Cisco UCS C480 ML M5 AI platform for Deep Learning workloads using NVIDIA Tesla V100 SXM2 GPUs with NVIDIA GPU Cloud optimized containers. The performance tests described here were run using eight Tesla V100 SXM2 GPUs, Intel® Xeon® 8180 Scalable CPUs, 768 GB of physical memory, and 7.2-TB NVMe drives.

Solution summary

Cisco® machine-learning computing solutions ease the challenges faced by IT organizations and data scientists, supporting the needs of machine-learning workloads while making them part of the enterprise data center. With Cisco solutions, you can power AI workloads at scale and help extract more intelligence from data to make better decisions.

AI projects span all industries

AI is real, and interest in it spans all industries. From finance to healthcare to security to manufacturing—and across all other vertical markets—AI is a top priority for enterprise IT. Organizations are receiving data from many sources and are challenged to extract more intelligence from it and to use it to augment human capabilities to make better decisions. Cisco is uniquely positioned to address these needs with our fabric- and infrastructure-based differentiated solutions.

New powerful additions to the Cisco UCS portfolio can power AI initiatives across a wide range of industries (Figure 1).

Figure 1. Support AI projects across industries

Finance	Healthcare	Media Entertainment	Security and Defense	Retail	Manufacturing
Fraud Detection	Cancer Cell Detection	Video Captioning	Face Recognition	Theft Detection	Reduce Product Defects
Cryptocurrencies	Drug Discovery	Content Based Search	Crowd Analytics	Auto Checkout	Increase Production Speed
Algorithmic Trading	Medical Research	NLP, VR and AR	Cyber Security	Targeted Marketing	Shorten Downtime

© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

A new approach for deep-learning workloads

The new Cisco UCS C480 AI platform accelerates deep learning: a computation-intensive form of machine learning that uses neural networks and large data sets to train computers for complex tasks. Using powerful NVIDIA GPUs, it is designed to accelerate many of today's best-known machine-learning software stacks. Data scientists and developers can experiment with machine learning on a laptop. But deep learning at scale demands much greater computing capability. It requires an IT architecture that can ingest vast sets of data and tools that can make sense of this data and use it to learn. That is why Cisco is working with its technology partners to validate many of today's most popular machine-learning tools: to help simplify deployments and accelerate time to insight.

With the addition of the Cisco UCS C480 ML M5 Rack Server for machine learning, we now offer a complete array of computing options sized to each element of the AI lifecycle: data collection and analysis near the edge, data preparation and training in the data center core, and real-time inference at the heart of AI. Our cloud-based management makes it easy to extend accelerated computing to the right locations across an increasingly distributed IT landscape. You gain these benefits:

- Gain performance and capacity. The Cisco UCS C480 ML M5 offers flexible options for CPU, memory, networking, and storage while providing outstanding GPU acceleration.
- Demystify your machine-learning software ecosystem with validated solutions.
- Simplify operations with the Cisco Intersight™ platform to extend accelerated computing to the locations where it is needed.

Platform summary

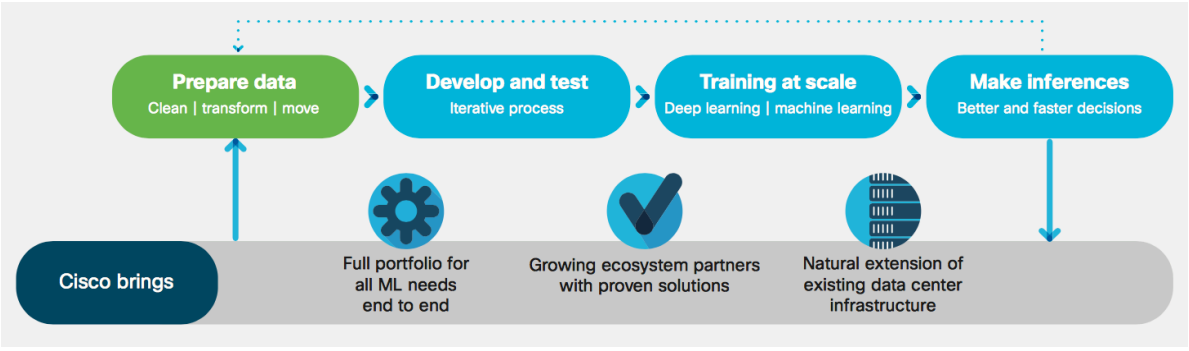
Cisco offers a new server portfolio for artificial intelligence and machine-learning workloads.

Cisco UCS C480 ML M5 AI platform

The Cisco UCS C480 ML M5 Rack Server is the latest addition to the Cisco Unified Computing System™ (Cisco UCS) server portfolio and its first server built from the start for AI and machine-learning workloads. With this addition to the Cisco UCS portfolio, you have a complete range of computing options designed for each stage of the AI and machine-learning lifecycles, enabling you to extract more intelligence from your data and use it to make better, faster decisions (Figure 2).

In a four-rack-unit (4RU) form factor, the Cisco UCS C480 ML M5 server is specifically built for deep learning. It is storage and I/O optimized to deliver industry-leading performance for training models. It is designed for the most computation-intensive phase of the AI and machine-learning lifecycles: deep learning. This server integrates GPUs and high-speed interconnect technology with a large storage capacity and up to 100-Gbps network connectivity.

Figure 2. Support your data scientists with a complete portfolio of AI and machine-learning servers



© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

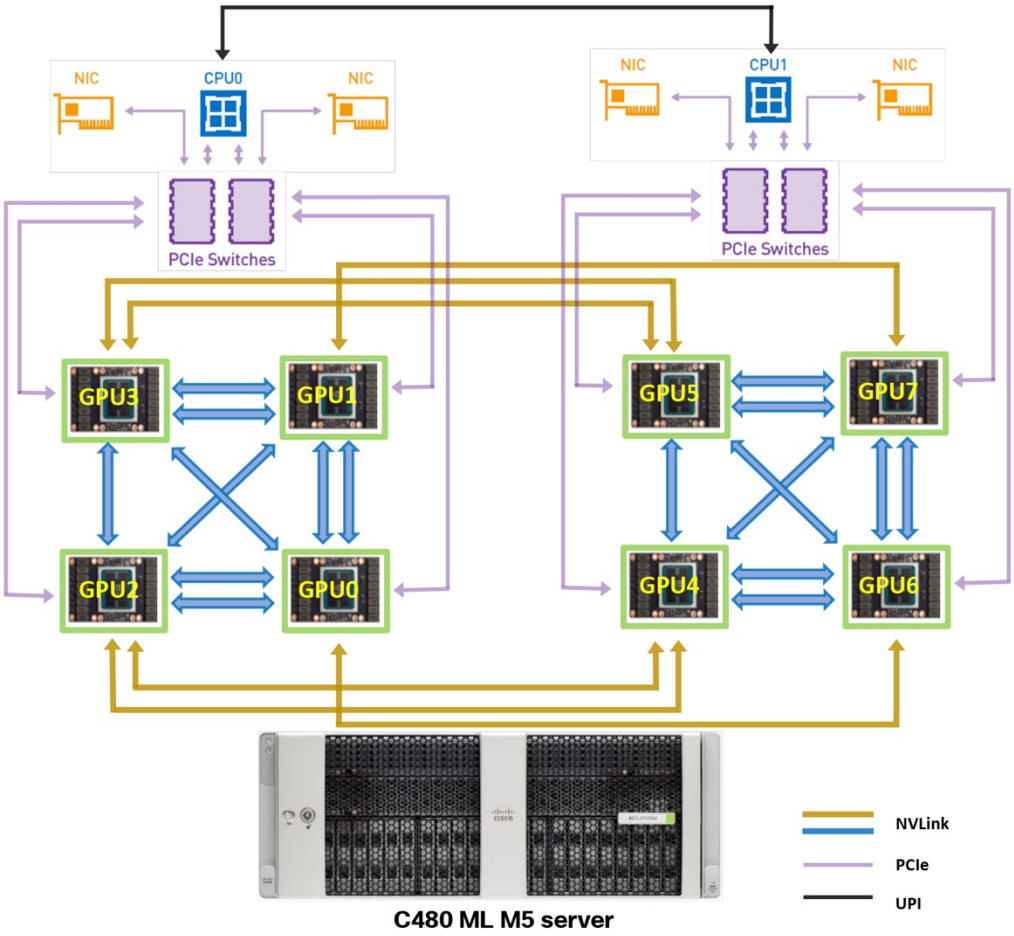
The Cisco UCS C480 ML M5 offers these features and benefits:

- **GPU acceleration:** Eight NVIDIA Tesla V100 SXM2 32-GB modules are interconnected with NVIDIA NVLink technology for fast communication across GPUs to accelerate computing. NVIDIA specifies TensorFlow performance of up to 125 teraflops per module, for a total of up to 1 petaflop of processing capability per server.
- **Internal NVLink topology:** NVLink is a high-speed GPU interconnect. Eight GPUs are connected through an NVLink cube mesh. Each NVLink is capable of 25 GBps of send and receive processing, for a total bandwidth of about 300 GBps among the eight GPUs.

Note that each GPU has six NVLinks. Thus, not all GPUs are directly connected through NVLink. Therefore, performance may be negatively affected if a particular training model has a heavy load of GPU-to-GPU communication.

The eight-GPU hybrid cube-mesh NVLink topology provides the highest bandwidth for multiple collective communication primitives, including broadcast, gather, all-reduce, and all-gather primitives, which are important to deep learning. Figure 3 shows the NVLink topology.

Figure 3. Modoc NVLink topology

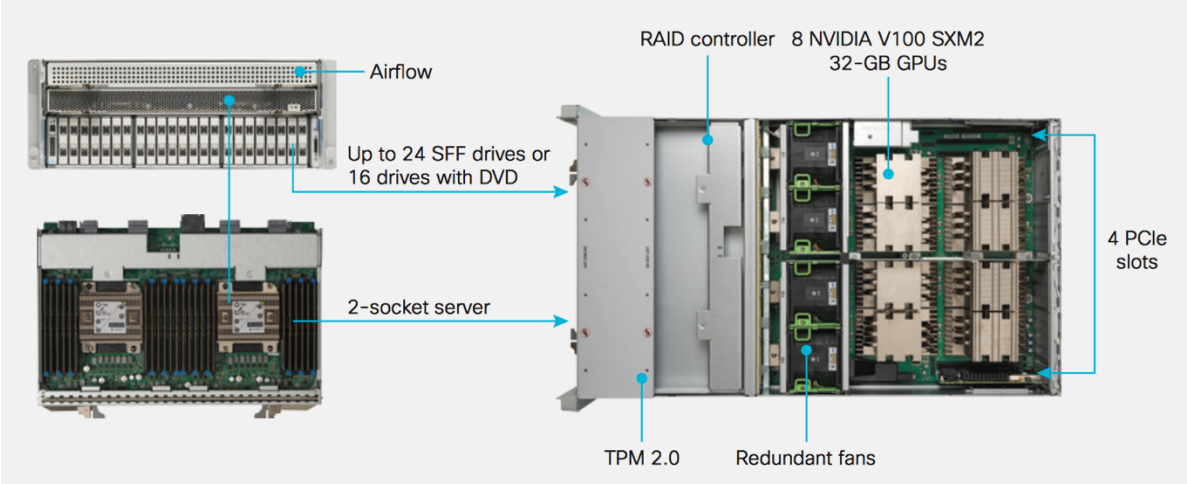


© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

- The latest Intel Xeon Scalable CPUs: Two CPUs with up to 28 cores each manage the machine-learning process and send calculations to the GPUs.
- Storage capacity and performance: Data locality can be important for deep-learning applications. Up to 24 hard-disk drives (HDDs) or solid-state disks (SSDs) store data close to where it is used and are accessed through a midplane-resident RAID controller. Up to six disk-drive slots can be used for NVMe drives, providing best-in-class performance.
- Up to 3 TB of main memory: The system uses fast 2666-MHz DDR4 DIMMs.
- High-speed networking: Two built-in 10 Gigabit Ethernet interfaces accelerate the flow of data to and from the server.
- PCI Express (PCIe) expandability: Four PCIe switches feed four x16 PCIe slots for high-performance networking. Options include Cisco UCS virtual interface cards (VICs) and third-party network interface cards (NICs), for up to 100-Gbps connectivity.
- Unified management: By expanding the Cisco UCS portfolio with the new Cisco UCS C480 ML M5 server, we continue to support any workload without adding management complexity.

Figure 4 shows the physical design of the server.

Figure 4. Cisco UCS C480 ML M5 Rack Server physical design



Cisco UCS differentiators and main features

The Cisco UCS C480 ML M5 delivers outstanding storage expandability and performance options for both standalone systems and Cisco UCS managed environments and for Cisco Intersight cloud-based environments.

The Cisco UCS C480 ML M5 can be managed with Cisco Intersight, which is a new cloud-based management platform that uses analytics to deliver proactive automation and support. By combining intelligence with automated actions, you can reduce costs dramatically and resolve issues more quickly. Cisco Intersight enables transparency and management in all of an organization's data centers from a central source. Insights are based on analytics and machine learning, which enables a new level of systems management. Its machine-learning capabilities also offer proactive guidance to secure and optimize the data center and identify possible points of failure so that they can be corrected before problems occur.

© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

The Cisco UCS C480 ML M5 can be deployed as a standalone server or within a Cisco UCS managed environment. When used in combination with Cisco UCS Manager, the C480 ML M5 brings the power and automation of unified computing to enterprise applications, including Cisco SingleConnect technology, drastically reducing switching and cabling requirements. Cisco UCS Manager uses service profiles, templates, and policy-based management to enable rapid deployment and help ensure deployment consistency. It also enables end-to-end server visibility, management, and control in both virtualized and bare-metal environments. The Cisco Integrated Management Controller (IMC) delivers comprehensive out-of-band server management with support for many industry standards, including:

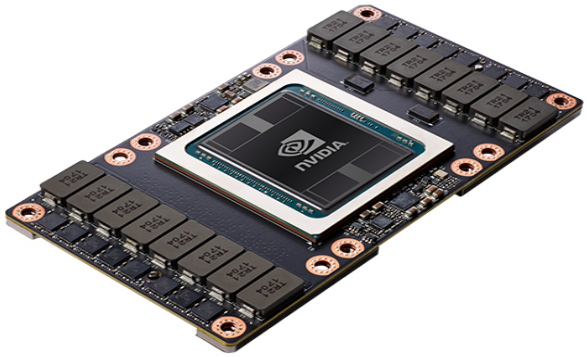
- Redfish Version 1.01 (v1.01)
- Intelligent Platform Management Interface (IPMI) v2.0
- Simple Network Management Protocol (SNMP) v2 and v3
- Syslog
- Simple Mail Transfer Protocol (SMTP)
- HTML5 GUI
- HTML5 virtual keyboard, video, and mouse (vKVM)
- Command-line interface (CLI)
- XML APIs

Management software development kits (SDKs) and DevOps integrations exist for Python, Microsoft PowerShell, Ansible, Puppet, Chef, and other technologies.

NVIDIA Tesla V100 SXM2 with 32 GB of memory

The NVIDIA Tesla V100 (Figure 5) is the world's most advanced data center GPU ever built to accelerate AI, high-performance computing (HPC), and graphics processing. Powered by NVIDIA Volta, the latest GPU architecture, the Tesla V100 offers the performance of up to 100 CPUs in a single GPU—enabling data scientists, researchers, and engineers to address challenges that were once thought impossible.

Figure 5. NVIDIA Tesla V100 SXM2 GPU



© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

- **Volta architecture:** By pairing CUDA cores and Tensor cores within a unified architecture, a single server with Tesla V100 GPUs can outperform hundreds of commodity CPU servers for certain deep learning applications.
- **Tensor core:** Equipped with 640 Tensor cores, the Tesla V100 delivers 125 teraflops of deep-learning performance. Thus, Tensor offers 12 times more floating-point operations per second (FLOPS) for deep-learning training and 6 times more FLOPS for deep-learning inference than NVIDIA Pascal GPUs.
- **Next-generation NVIDIA NVLink:** NVLink in the Tesla V100 delivers twice the throughput of the previous generation of technology. Up to eight Tesla V100 accelerators can be interconnected at up to 300 GBps to achieve the highest application performance possible on a single server.
- **Maximum-efficiency mode:** The new maximum-efficiency mode allows data centers to achieve up to 40 percent greater computing capacity per rack within the existing power budget. In this mode, the Tesla V100 runs at peak processing efficiency, providing up to 80 percent of the performance at half the power consumption.
- **Programmability:** The Tesla V100 is designed from the foundation to simplify programmability. Its new independent thread scheduling enables synchronization and improves GPU utilization by sharing resources among small jobs.

Every major deep-learning framework is optimized for NVIDIA GPUs, enabling data scientists and researchers to use AI for their work. When running deep-learning training and inference frameworks, a data center with Tesla V100 GPUs can save over 90 percent in server and infrastructure acquisition costs.

The Tesla V100 platform for deep-learning training offers these main features:

- Caffe, TensorFlow, and the Microsoft Cognitive Toolkit (previously called CNTK) are up to three times faster with the Tesla V100 than with the NVIDIA P100 GPU.
- 100 percent of the top deep-learning frameworks are GPU accelerated.
- The platform offers up to 125 teraflops of TensorFlow operations per GPU.
- The platform offers up to 32 GB of memory capacity per GPU.
- The platform offers memory bandwidth of up to 900 GBps per GPU.

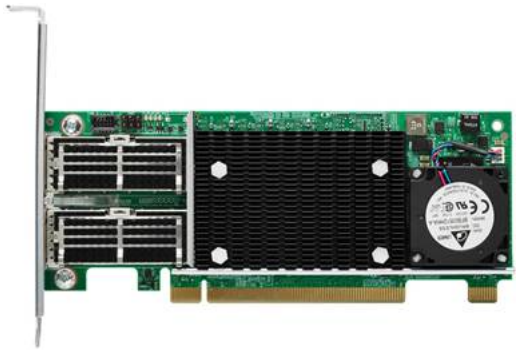
Cisco UCS Virtual Interface Card 1385 network adapter

The Cisco UCS VIC 1385 (Figure 6) is a Cisco innovation. It provides a policy-based, stateless, agile server infrastructure for your data center. This dual-port Enhanced Quad Small Form-Factor Pluggable (QSFP) half-height PCIe card is designed exclusively for Cisco UCS C-Series Rack Servers.

The card supports 40 Gigabit Ethernet and Fibre Channel over Ethernet (FCoE). It incorporates Cisco's next-generation converged network adapter (CNA) technology and offers a comprehensive feature set, providing investment protection for future feature software releases. The card can present more than 256 PCIe standards-compliant interfaces to the host, and these can be dynamically configured as either NICs or host bus adapters (HBAs). In addition, the VIC supports Cisco Data Center Virtual Machine Fabric Extender (VM-FEX) technology. This technology extends the Cisco UCS fabric interconnect ports to virtual machines, simplifying server virtualization deployment.

© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Figure 6. Cisco UCS VIC 1385 network adapter



The Cisco UCS VIC 1385 provides the following features and benefits:

- Stateless and agile platform: The personality of the card is determined dynamically at boot time using the service profile associated with the server. The number, type (NIC or HBA), identity (MAC address and World Wide Name [WWN]), failover policy, bandwidth, and quality-of-service (QoS) policies of the PCIe interfaces are all determined using the service profile. The capability to define, create, and use interfaces on demand provides a stateless and agile server infrastructure
- Network interface virtualization: Each PCIe interface created on the VIC is associated with an interface on the Cisco UCS fabric interconnect, providing complete network separation for each virtual cable between a PCIe device on the VIC and the interface on the fabric interconnect

Software components overview

This section describes the software components of the Cisco AI solution with the Cisco UCS C480 ML M5 server.

Red Hat Enterprise Linux

Red Hat Enterprise Linux (RHEL) is a high-performing operating system that has delivered outstanding value to IT environments for more than a decade. More than 90 percent of Fortune Global 500 companies use Red Hat products and solutions, including RHEL. As the world's most trusted IT platform, RHEL has been deployed in mission-critical applications at global stock exchanges, financial institutions, leading telecommunications companies, and animation studios. It also powers the websites of some of the most recognizable global retail brands.

RHEL offers these main features:

- Delivers high performance, reliability, and security
- Is certified by the leading hardware and software vendors
- Scales from workstations, to servers, to mainframes
- Provides a consistent application environment across physical, virtual, and cloud deployments

Designed to help organizations make a seamless transition to emerging data center models that include virtualization and cloud computing, RHEL includes support for major hardware architectures, hypervisors, and cloud providers, making deployments across different physical and virtual environments predictable and secure. Enhanced tools and new capabilities in the current release enable administrators to tailor the application environment to efficiently monitor and manage computing resources and security.

© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

NVIDIA deep-learning framework and tools

The NVIDIA deep-learning SDK accelerates widely used deep learning frameworks such as NVCaffe, Caffe2, Microsoft Cognitive Toolkit, MXNet, TensorFlow, PyTorch, Torch, and TensorRT. NVIDIA GPU Cloud (NGC) provides containerized versions of these frameworks optimized for the Cisco AI server platform. These frameworks, including all necessary dependencies, are prebuilt, tested, and ready to run. For users who need more flexibility to build custom deep-learning solutions, each framework container image also includes the framework source code to enable custom modifications and enhancements, along with the complete software development stack.

Most deep-learning frameworks have begun to merge support for half-precision training techniques that exploit Tensor core calculations in Volta. Some frameworks include support for FP16 storage and Tensor core math. To achieve optimum performance, you can train a model using Tensor core math and FP16 mode on some frameworks.

ImageNet

Deep learning attempts to model data through multiple processing layers containing nonlinearities. It has proven to be very efficient in classifying images, as shown by the impressive results of deep neural networks in the ImageNet competition for example. However, training these models requires very large data sets and is time consuming.

ImageNet is a large visual database designed for use in visual object recognition software research. It offers a data set, from Stanford and MIT, that contains more than 14 million images and more than 10 thousand categories (labels). It is the data set most commonly used by major training models (ResNet, Inception, VGG16, etc.).

On the surface, many new architectures appear to be very different, but most of them are reapplying well-established training principles. ImageNet provides a common foundation for comparing architectures. Networks trained on ImageNet are often starting points for other visioning tasks. Architectures that perform well on ImageNet have been successful in other domains.

TensorFlow

In this guide, the TensorFlow deep-learning framework is used to test popular convolutional neural network (CNN) training models such as ResNet, Inception, VGG, AlexNet, and GoogleNet.

TensorFlow Serving is a flexible, high-performance serving system for machine-learning models. It is designed for production environments. TensorFlow Serving lets you easily deploy new algorithms and experiments while keeping the same server architecture and APIs. TensorFlow Serving offers ready-to-use integration with TensorFlow models, but it can easily be extended to serve other types of models and data.

TensorFlow is a software library, developed by the Google Brain Team within the Google Machine Learning Intelligence research organization, for the purpose of conducting machine-learning and deep neural network research.

The main features of TensorFlow include the following:

- Capability to define, optimize, and efficiently calculate mathematical expressions involving multidimensional arrays (tensors)
- Programming support for deep neural networks and machine-learning techniques
- Transparent use of GPU computing, automating management and optimization of the same memory and the data used; you can write the same code and run it on either CPUs or GPUs
- High scalability of computation across machines and huge data sets

© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

TensorFlow supports single- and multiple-GPU processing. NVIDIA offers the following TensorFlow optimizations:

- Integration with NVIDIA CUDA Deep Neural Network Library (cuDNN) v7.0 and with CUDA 9.0
- Integration with the latest version of the NVIDIA Collective Communications Library (NCCL) with NVLink for improved multiple-GPU scaling
- Support for FP16 storage and Tensor core math
- Support for the ImageNet preprocessing script

Convolutional neural network training models

The TensorFlow CNN benchmarks include benchmarks for several convolutional neural networks. The TensorFlow CNN benchmarks contain implementations of several popular convolutional models and are designed to be as fast as possible. The benchmarks can be run on a single machine or in distributed mode across multiple hosts.

In general, each deep-learning neural network training workload runs in the following manner:

- First, a given neural network architecture is replicated on each GPU. Then the neural network is trained by processing an image data set sequentially in batches or iterations.
- For each batch, images are divided among available GPUs for data parallelism. For the training, each GPU processes its images, resulting in a series of model activations and floating-point operations, resulting in distinct values for each GPU's copy of the model's parameters.
- At the end of each iteration, all-reduce operations help ensure that each GPU's model has an identical copy of the model's parameters.

The following popular CNN architecture training models were tested and validated for this study:

- **ResNet:** The Deep Residual Learning Network (ResNet) introduced the concept of a residual block. Each block consists of two convolution layers along with a connection adding the output of the second block to the input of the first. Residual blocks are designed to allow the training of substantially deeper models than had been trained previously. By adding the input of the block to its output, ResNet allows the residual block to learn the residual function. It then forwards the activations to deeper layers than was possible with earlier technologies. One advantage of ResNet is that it can improve the accuracy of the model while avoiding parameter explosion: that is, the ResNet blocks increase the depth (and inner layers) of the network instead of the width.
- **Inception:** Inception v3 is trained for the ImageNet Large Visual Recognition Challenge (ILSVRC) using the data from 2012. This is a standard task in computer visioning, in which models try to classify entire images into 1000 classes,
- **VGG16:** VGG is a CNN model proposed by K. Simonyan and A. Zisserman from the University of Oxford in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition." The model achieves 92.7 percent top-five test accuracy in ImageNet, using a data set of over 14 million images belonging to 1000 classes.
- **AlexNet:** AlexNet has five convolution layers, three pooling layers, and two fully connected layers. This CNN architecture requires about 1.4 million activations per image and 60 million parameters. AlexNet has performed well on ILSVRC 2012.
- **GoogleNet:** GoogleNet has two convolution layers, two pooling layers, and nine inception layers. Each inception layer consists of six convolution layers and one pooling layer. The goal of the inception layer is to cover larger area of images while maintaining fine resolution for smaller information in the images. GoogleNet performs significantly better than AlexNet on ImageNet and recent ILSVRC challenge data sets. This CNN architecture has about 5.5 million parameters and 10.8 million activations per image.

© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Choice of model

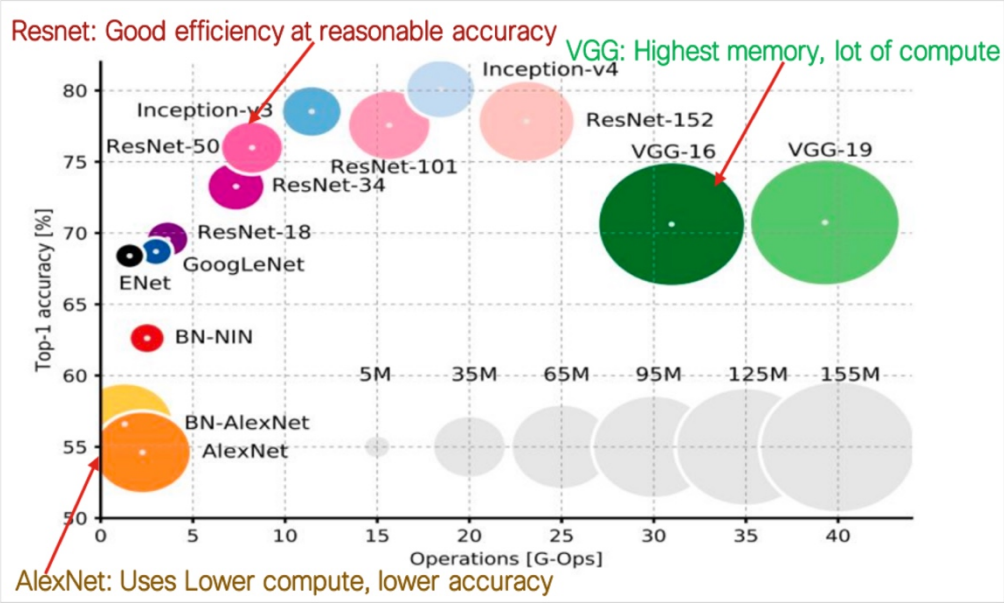
The first step is to choose which model to choose. ILSVRC requires models to classify 1000 categories of images, and some suggested models cannot show this type of super performance. However, if you choose to classify 10 to 100 categories of images, the models can fit the architecture discussed here.

CNN models have evolved, and some of them have complicated architecture. If you want to modify certain parts of an entire layer, or if you want to troubleshoot to find the part that is the bottleneck in a problem, you must understand how models works behind the scenes.

For example, ResNet training models are extremely popular. They demonstrate reasonable accuracy at a relatively low number of giga operations per second (GOPS). ResNet-152 has moderate memory and computing requirements with reasonable accuracy. The VGG16 model requires the greatest amount of memory and high computing requirements with moderate accuracy. The AlexNet model has lower computing requirements and produces lower accuracy.

Figure 7 shows how the models have evolved over time.

Figure 7. Operations versus accuracy among the training models



© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Platform requirements

Table 1 lists the hardware specifications, and Table 2 lists the software specifications for the C480 ML M5 platform.

Table 1. Hardware specifications

Hardware model	Quantity
Cisco UCS C480 ML M5 AI platform	1
AI server components	
NVIDIA SXM2 V100 32G Memory	8
Intel Xeon Platinum 8180 Scalable processor	2
32-GB 2666-MHz DIMMs	24
7-TB NVMe drives	1
Cisco UCS VIC 1385 PCIe adapter	1

Table 2. Software versions

Software	Version
RHEL	Release 7.5
NVIDIA driver	Release 396.44
CUDA toolkit	Release 9.2.148
Docker container	docker-ce-18.09.0-3
NVIDIA Docker container	nvcr.io/NVIDIA/TensorFlow:18.08-py3
Machine-learning framework	TensorFlow Version 1.9

Performance validation

This section describes the testing performed to validate the performance of Deep Learning workloads on the Cisco UCS C480 ML M5 platform. All tests described in this section were conducted using the hardware and software listed in Table 1 and Table 2.

Validation test plan

The validation testing was performed using industry-standard deep learning workloads with TensorFlow on a number of computing configurations to demonstrate the performance of the Cisco UCS C480 ML M5 AI platform. The ImageNet data set was hosted on a single 7-TB NVMe drive on a Cisco UCS C480 ML M5 system.

The TensorFlow deep-learning framework was downloaded from NGC using containers optimized for NVIDIA GPUs. The containers from NGC integrate the framework, necessary drivers, and libraries and are optimized across the stack by NVIDIA for the highest GPU-accelerated performance. NGC containers incorporate the NVIDIA CUDA toolkit, which provides the NVIDIA CUDA driver, the NVIDIA cuDNN, and much more.

The NGC containers also include the NCCL for multiple-GPU and multinode collective communication primitives, enabling topology awareness for deep-learning training. NCCL can be deployed in single-process or multiprocess applications, handling required interprocess communication transparently. When the containers are deployed, the TensorFlow deep-learning framework automatically uses this version of NCCL when run on multiple GPUs.

© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

TensorFlow was used as the deep-learning framework for all models that were tested. Computing and GPU (NVLink) performance metrics were captured for each test case.

The following CNN models, with different degrees of computing and storage complexity, were used to demonstrate training rates:

- ResNet-50 delivers better accuracy with faster processing time.
- Inception v3 is another common TensorFlow model.
- VGG16 produces the most inter-GPU communication.
- ResNet-152 is generally considered to be the most accurate training model.
- AlexNet has had a large impact on the field of machine learning: specifically, in the application of deep learning to machine vision.
- The GoogleNet deep network is 22 layers deep, the quality of which is assessed in the context of classification and detection.

For all workloads, ILSVRC 2012 ImageNet was used. It has 1000 object classes and 1.43 million annotated images, each with a size of 256 x 256 pixels.

Each of these models was tested with various software configurations to study the effects of each option on performance:

- We tested each model with both synthetic data and the read data from the ImageNet reference data set.
- We used ImageNet data with distortion disabled to reduce the overhead of CPU processing before copying the data into GPU memory.
- We tested each model by using Tensor cores and CUDA cores to demonstrate the performance improvements that the Tensor cores provide.
- For each neural network model, we varied a number of TensorFlow parameters to find the settings for each model that provided the best training performance, measured as images per second. TensorFlow has many other configurable options as well (Table 3).

Table 3. TensorFlow parameters

	ResNet-50	Inception v3	VGG16	ResNet-152	AlexNet	GoogleNet
Batch size	256	256	256	128	256	256

- Increasing the batch size has several effects on the system that ultimately result in higher overall training rates, lower inter-GPU communication requirements, and higher storage bandwidth requirements.
- All performance metrics were gathered from runs of several hours, and stable performance results were observed during the training. Each test was run three times, and the mean values for the performance metrics that were observed are reported.

Test results

This section presents highlights of the computing (CPU, GPU, PCIe, and NVLink) performance data that was collected during the tests using the TensorFlow deep-learning framework.

Note the following details about the data that is presented here:

© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

- Model training performance is measured in images per second.
- Storage performance is measured using throughput (MBps) and latency (microseconds).
- NVLink utilization and bandwidth are measured using the NVIDIA system management interface: NVIDIA-smi.
- CPU utilization is measured using dstat system resource statistics on the RHEL host operating system.

Overall training throughput

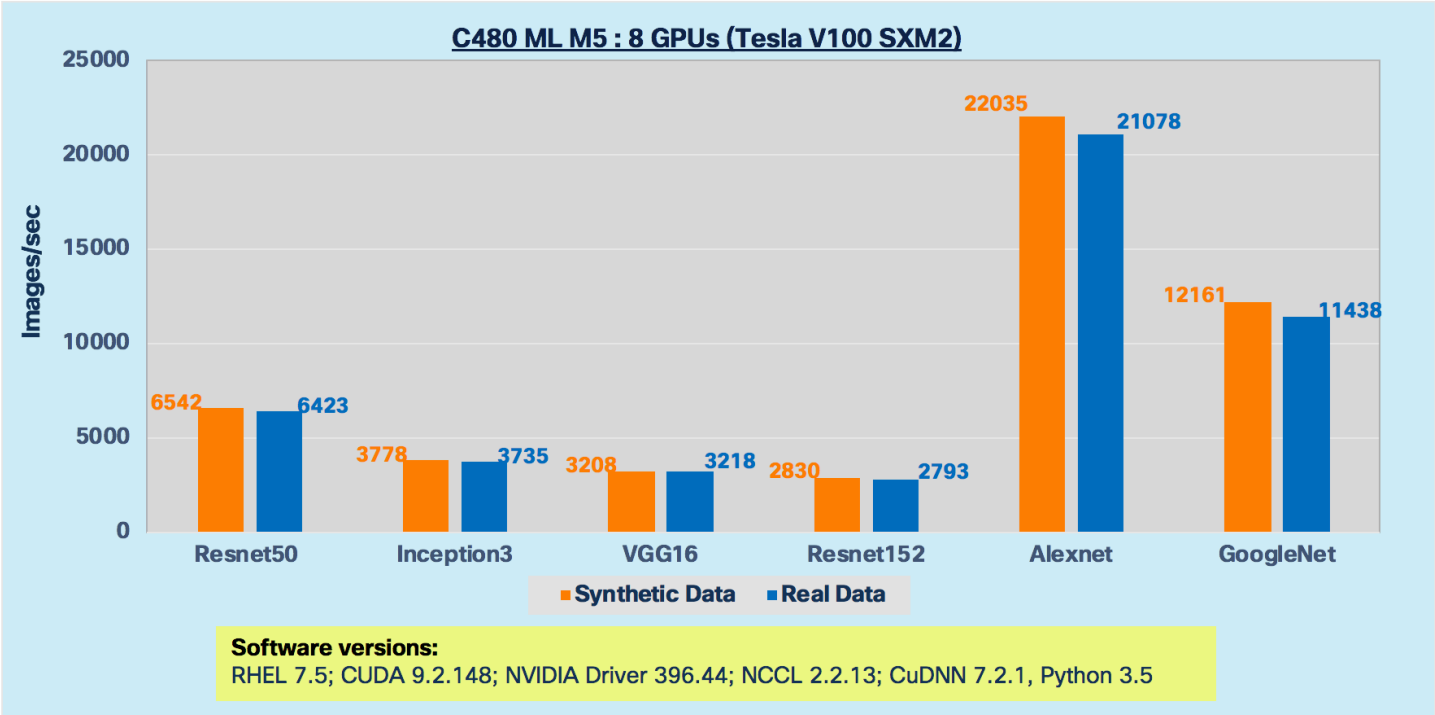
The charts that follow show the maximum number of training images per second achieved with each of the models that were tested using the TensorFlow framework for maximum performance.

The CNN benchmarks were run with synthetic data and actual data. Synthetic data can be considered the theoretical maximum because it does not exercise any underlying I/O subsystem or even hardware (including CPU) for data preparation. The very small differences observed indicate that the system is well balanced, and no significant bottlenecks were observed while fetching and preparing data for the real data tests.

Figure 8 demonstrates the training throughput for the TensorFlow machine-learning workload that was achieved by using ImageNet real data and synthetic data for a baseline comparison. It also shows the theoretical maximum that is achievable, in which all GPUs train synthetic data independently without updating parameters with each other.

As shown in Figure 8, the achieved throughput for ImageNet data is very close to the throughput achieved for synthetic data.

Figure 8. Training throughput for all models tested using TensorFlow



© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Cisco UCS C480 ML M5 system performance charts for TensorFlow DL workloads

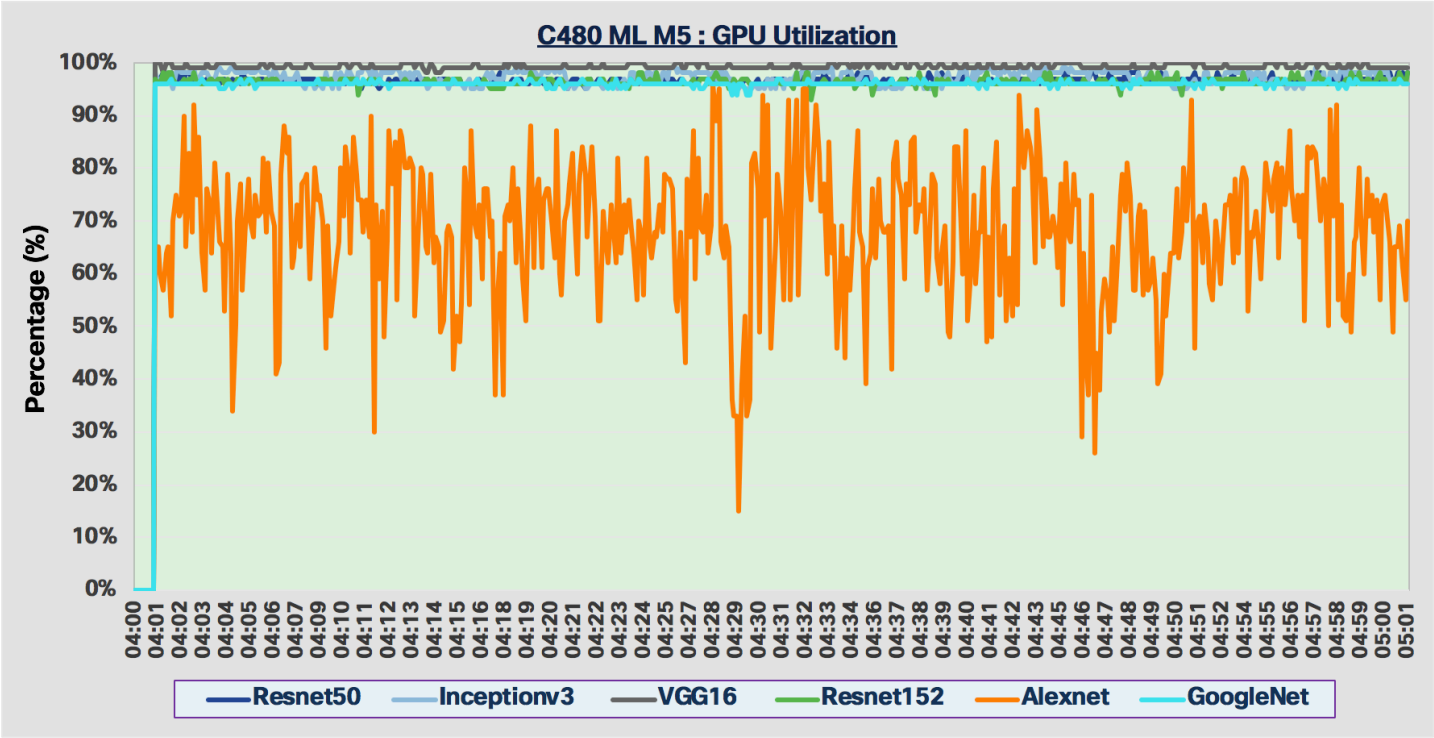
The following performance charts demonstrate the capabilities of the Cisco UCS C480 ML M5 Rack Server using the NVIDIA Tesla V100 GPU, Intel Xeon processor 8180, NVLink technology, and a 7-TB NVMe drive under a full load.

GPU utilization results

Figure 9 shows the GPU utilization for the Cisco UCS C480 ML M5 that occurred when running each TensorFlow model using eight V100 GPUs.

The C480 ML M5 with V100 GPUs begin processing data almost immediately, and GPU utilization remained consistent throughout the test run. This graph shows the results for ResNet50, Inception v3, VGG16, ResNet152, and GoogleNet, which produced the highest level of GPU utilization (approximately 96 percent) in the testing reported here. One model, AlexNet, produced moderate GPU utilization (approximately 65 percent) in the testing.

Figure 9. GPU utilization for Cisco UCS C480 ML M5 based on TensorFlow benchmark



NVIDIA-smi output below shows that the GPU utilization for AlexNet is moderate compared to other models:

```
root@C480-ML ~]# NVIDIA-smi
Thu Nov 27 11:12:25 2018
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 396.44                 Driver Version: 396.44                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
```

© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

```
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  Off  | 00000000:1B:00.0 Off |                    0 |
| N/A  55C    P0   256W / 300W |      31644MiB / 32510MiB |      68%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  Off  | 00000000:1C:00.0 Off |                    0 |
| N/A  60C    P0    243W / 300W |      31640MiB / 32510MiB |      70%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla V100-SXM2...  Off  | 00000000:42:00.0 Off |                    0 |
| N/A  58C    P0    244W / 300W |      317201MiB / 32510MiB |     65%      Default |
+-------------------------------+----------------------+----------------------+
|   3  Tesla V100-SXM2...  Off  | 00000000:43:00.0 Off |                    0 |
| N/A   57C    P0    243W / 300W |     31684MiB / 32510MiB |      59%      Default |
+-------------------------------+----------------------+----------------------+
|   4  Tesla V100-SXM2...  Off  | 00000000:89:00.0 Off |                    0 |
| N/A   61C    P0    244W / 300W |     31732MiB / 32510MiB |      62%      Default |
+-------------------------------+----------------------+----------------------+
|   5  Tesla V100-SXM2...  Off  | 00000000:8A:00.0 Off |                    0 |
| N/A   59C    P0    247W / 300W |     31689MiB / 32510MiB |      68%      Default |
+-------------------------------+----------------------+----------------------+
|   6  Tesla V100-SXM2...  Off  | 00000000:B2:00.0 Off |                    0 |
| N/A   58C    P0    248W / 300W |     31841MiB / 32510MiB |      66%      Default |
+-------------------------------+----------------------+----------------------+
|   7  Tesla V100-SXM2...  Off  | 00000000:B3:00.0 Off |                    0 |
| N/A   60C    P0    255W / 300W |      31789MiB / 32510MiB |      64%      Default |
+-------------------------------+----------------------+----------------------+
```

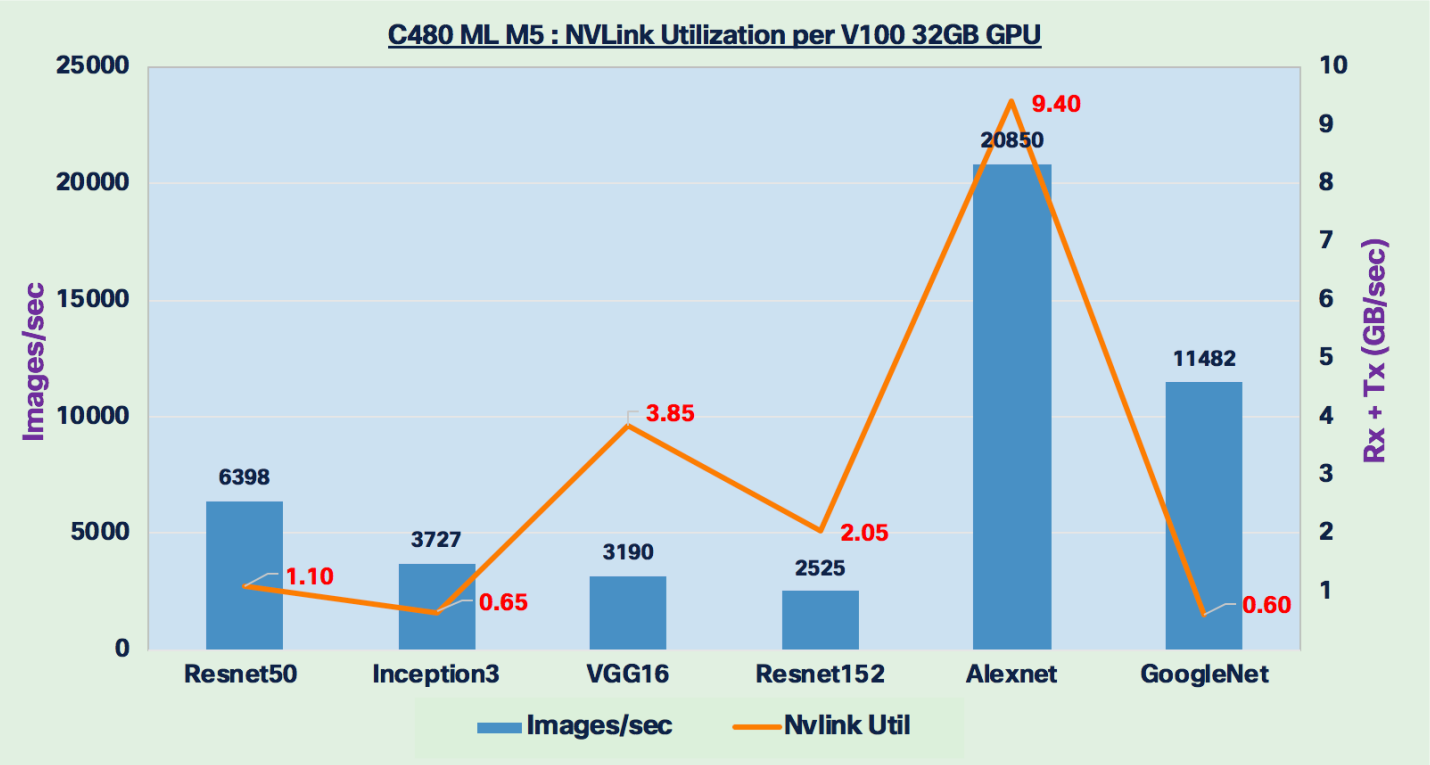
NVLink utilization results

Figure 10 shows the NVLink bandwidth utilization for the Cisco UCS C480 ML M5 that occurred when running each TensorFlow model using eight NVIDIA Tesla V100 GPUs.

Each NVIDIA Tesla V100 GPU has six NVLink connections, each capable of 50 GBps of bidirectional bandwidth, for an aggregate of up to 300 GBps of bidirectional bandwidth. The C480 ML M5 with eight NVIDIA Tesla V100 GPUs interconnected using NVLink technology achieved the highest application performance possible on a single server. The graph shows the results for each of the TensorFlow deep-learning models measured in images per second and the NVLink utilization. The tests showed that AlexNet had the most NVLink traffic, VGG16 had moderate traffic, and rest of the models had very low NVLink traffic.

© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Figure 10. NVLink utilization per NVIDIA Tesla V100 GPU based on TensorFlow



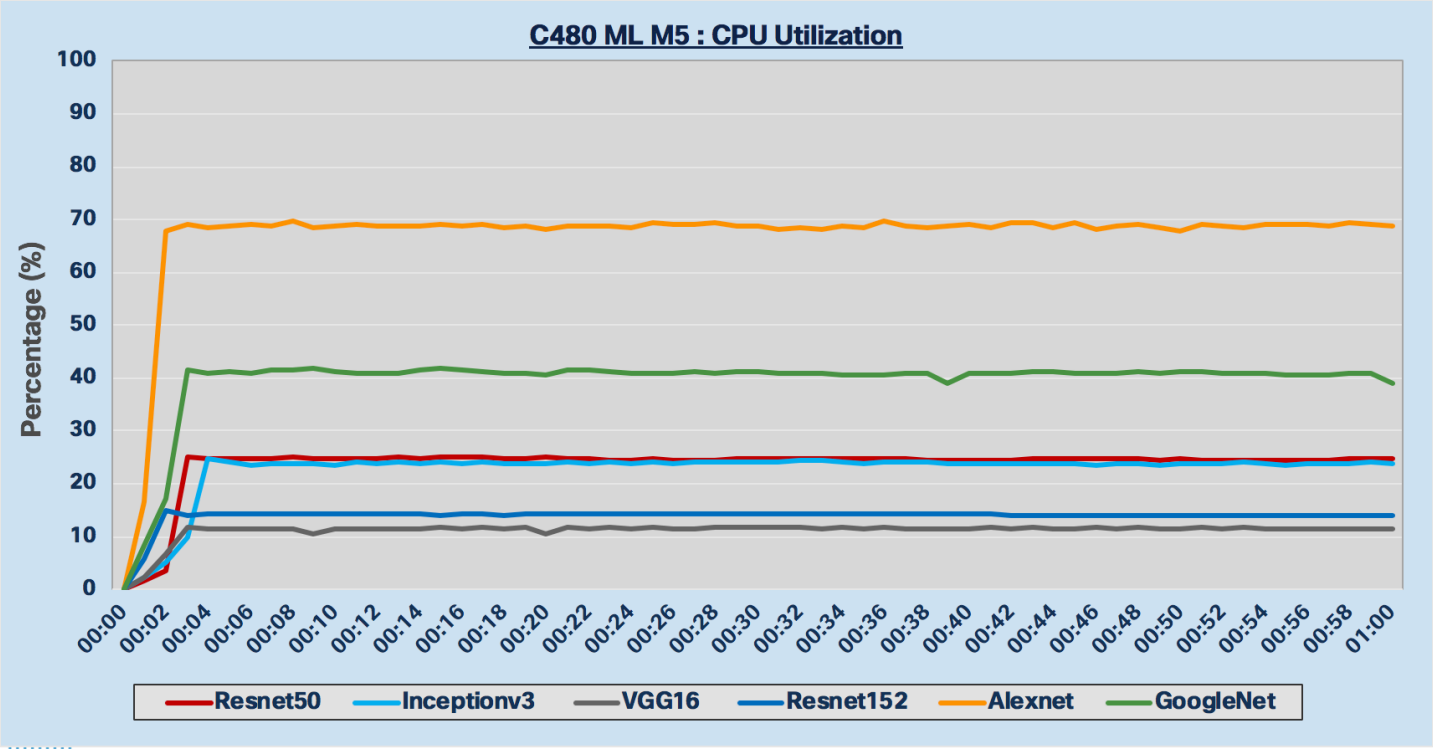
© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

CPU utilization results

Figure 11 shows the CPU utilization for the Cisco UCS C480 ML M5 that occurred when running each model using eight V100 GPUs.

The C480 ML M5 with dual Intel Xeon processor 8180 CPUs with the highest core density provided better computing performance for deep-learning workloads. The CPU utilization was captured to demonstrate the deep-learning performance for each of the tested models. The graph shows the CPU utilization for each TensorFlow model for the image training (in images per second). It clearly shows that the lowest CPU utilization was between 10 and 20 percent, for the ResNet50, Inception v3, VGG16, and ResNet152 models; the highest CPU utilization was between 40 and 70 percent, for the GoogleNet and AlexNet models.

Figure 11. CPU utilization for Cisco UCS C480 ML M5 based on TensorFlow benchmark



© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Figure 12 shows the Cisco UCS C480 ML M5 eight-GPU topology visualized using NVIDIA-smi tools. It also shows how the GPUs will communicate with each other and the CPUs to which they are connected.

The topology illustrates that GPU intercommunication from GPU0 to GPU1 uses two NVLinks (100 GBps), and intercommunication from GPU0 to GPU3 uses one NVLink (50 GBps). Most important, traffic from GPU0 to GPU4, GPU5, and GPU7 uses the PCIe bandwidth between CPU1 to the CPU2 Intel Ultra Path Interconnect (UPI).

Note: On the basis of Figure 12 and Figure 13, you can also see that the AlexNet training model has the highest CPU utilization as a result of the NVLink topology. Figure 9 shows that AlexNet model GPU utilization is low compared to other models.

Figure 12. Cisco UCS C480 ML M5 eight-GPU topology

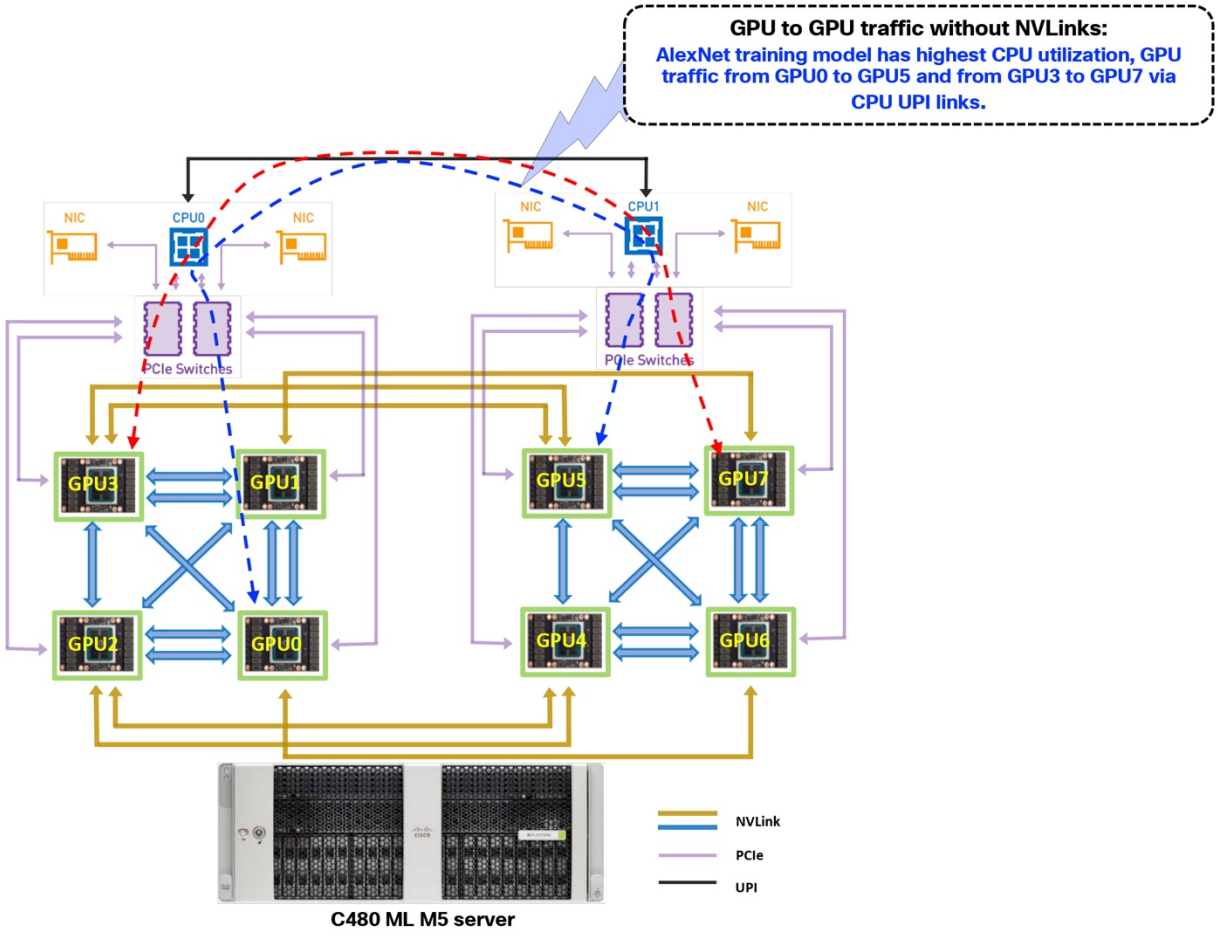
```
# nvidia-smi topo -mq
        GPU0    GPU1    GPU2    GPU3    GPU4    GPU5    GPU6    GPU7    CPU Affinity
GPU0     X      NV2     NV2     NV1     SYS     SYS     NV1     SYS     0-11,24-35
GPU1    NV2      X      NV1     NV2     SYS     SYS     SYS     NV1     0-11,24-35
GPU2    NV2     NV1      X      NV1     NV2     SYS     SYS     SYS     0-11,24-35
GPU3    NV1     NV2     NV1      X      SYS     NV2     SYS     SYS     0-11,24-35
GPU4    SYS     SYS     NV2     SYS      X      NV1     NV2     NV1     12-23,36-47
GPU5    SYS     SYS     SYS     NV2     NV1      X      NV1     NV2     12-23,36-47
GPU6    NV1     SYS     SYS     SYS     NV2     NV1      X      NV2     12-23,36-47
GPU7    SYS     NV1     SYS     SYS     NV1     NV2     NV2      X      12-23,36-47

Legend:

  X    = Self
  SYS  = Connection traversing PCIe as well as the SMP interconnect between NUMA nodes (e.g., QPI/UPI)
  NODE = Connection traversing PCIe as well as the interconnect between PCIe Host Bridges within a NUMA node
  PHB  = Connection traversing PCIe as well as a PCIe Host Bridge (typically the CPU)
  PXB  = Connection traversing multiple PCIe switches (without traversing the PCIe Host Bridge)
  PIX  = Connection traversing a single PCIe switch
  NV#  = Connection traversing a bonded set of # NVLinks
```

© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Figure 13. Highest CPU utilization for AlexNet training model



© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.

Conclusion

Experts and data scientists today believe that new industry leaders will be enterprises that invest in AI and turn their data into intelligence. Engineers at NVIDIA and Cisco partnered to design an affordable and simple yet powerful infrastructure that delivers high performance specifically for AI and machine-learning workloads.

This integrated deep-learning framework from NGC, combined with NVIDIA Tesla V100 and NVLink technology, helps ensure that the Cisco UCS C480 ML M5 AI platform outperforms similar ready-to-use systems.

Authors

Tushar Patel, Principal Engineer, UCS Performance and AI/ML Solutions

Vijay Durairaj, Technical Marketing Engineer, UCS Performance and Solutions

For more information

For more information about Cisco UCS C480 ML M5 AI servers, see https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/datasheet-c78-741211.html.

For more information about deep-learning benchmarks and data sets, see:

- TensorFlow CNN benchmark: https://github.com/TensorFlow/benchmarks/tree/master/scripts/tf_cnn_benchmarks
- ImageNet data sets: http://www.image-net.org/

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: www.cisco.com/go/trademarks. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)

Printed in USA C87-741689-01 1/19

© 2019 Cisco and/or its affiliates. All rights reserved. This document is Cisco Public Information.