WHITE PAPER

FlashStack for AI: Scale-Out Infrastructure for Deep Learning

POWERED BY NVIDIA GPUS

March 2019

TABLE OF CONTENTS

EXECUTIVE SUMMARY

FlashStack is an excellent platform for delivering mission-critical workloads and empowering users by building on extensible, continuously available, Cisco® validated architecture. With FlashStack, IT organizations can upgrade capacity without downtime and deploy cloud-like capabilities to help ensure constant access to data across the organization. Now IT administrators can avoid the rip-and-replace cycle of traditional infrastructure ownership and use familiar, integrated tools to manage the entire data center stack without the need for additional manual configuration and tuning. FlashStack solution-based data centers free IT staff to innovate by reducing the amount of time spent on infrastructure, automating common tasks, getting systems up and running in minutes, and keeping them running to support a business's vital applications.

Our new solution, which uses FlashStack for artificial intelligence (AI), combines the storage scalability and simplicity of Pure Storage FlashBlade servers and the performance of the new Cisco UCS® C480 ML M5 Rack Server platform to power AI workloads at scale and help extract more intelligence from data to help organizations make better decisions.This document summarizes the architecture and performance characteristics of the FlashStack AI system. It validates the operation and performance of the FlashStack AI system using industry-standard benchmark tools and demonstrates that the platform delivers excellent training performance and scalability for deep-learning training. In fact, the benchmark testing highlights near- linear scaling (within 2%) of deep learning training performance from one GPU to 32 GPUs on 4 x C480 ML: Proving the scale-out performance of a thoroughly engineered system with compute, network, storage and GPU accelerators. It also presents suggestions for integrating the FlashStack AI system into your team's the overall data science and data management ecosystem.

SOLUTION OVERVIEW

This section presents the FlashStack artificial intelligence (AI) solution and its implementation.

Introduction

Advances in deep neural networks have prompted a profusion of new applications for AI. Powerful new tools and techniques have enabled breakthroughs in applications as diverse as self-driving cars, natural-language translation, and predictive healthcare. As a result, investment in AI initiatives has increased dramatically. Spending on AI and machine learning is estimated to grow from US$12 billion in 2017 to US$57.6 billion by 2021 (*IDC*).

Designing and configuring an infrastructure to enable large-scale deep learning requires a significant investment of time and resources to avoid unforeseen delays, bottlenecks, or downtime. AI teams are overloaded with information, and decisions at the infrastructure architecture stage can make or break AI projects.

Cisco and Pure Storage have joined forces to develop the FlashStack AI solution, a fully

integrated platform that accelerates deep learning.

The FlashStack AI solution enables seamless scaling for both purpose-built graphics processing unit (GPU) base Cisco UCS C480 ML M5 Rack Servers and Pure Storage systems. The FlashStack AI solution is built for organizations that want a faster time to insight and that cannot afford to let any single infrastructure component artificially limit their pipeline throughput. As computing demands grow, additional C480 ML M5 servers can be provisioned in the high-performance fabric and instantly gain access to all available data sets. Similarly, as storage capacity or performance demands grow, additional blades can be added to the FlashBlade system with no downtime or reconfiguration.

FlashStack for AI: A critical component in your overall data strategy

Data is fragmented. Some of it is stored in data warehouses, and some are lost in data lakes. And when data is not unified, the velocity of data is greatly diminished. Why is it so hard for storage systems to unify data on a single platform? The problem is that each application has different requirements for its data, leading to a proliferation of data silos. Storage needs to be rethought.

A data hub is a data-centric architecture for storage that powers analytics and AI. It enables enterprises to consolidate data silos and share data in today's rapidly evolving, data-first world. A data hub integrates the main strengths of each silo into a single unified platform that includes four essential qualities: high-throughput files and objects, native scale-out architecture, multidimensional performance, and massively parallel architecture (Figure 2).

Even within a single AI project, the various phases of development have different infrastructure requirements. Each of these requirements is easily met with Pure Storage FlashBlade as the storage data hub and the Cisco Unified Computing System™ (Cisco UCS) family of products for the various computing activities (Figure 3).

FIGURE 1. FLASHSTACK AI SYSTEM

Deep learning at scale 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. With the addition of FlashStack AI for deep learning, Cisco and Pure Storage together now offer a complete array of joint solutions for every 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 and analytics.

FlashStack for AI helps data science teams:

- Demystify machine-learning infrastructure with validated solutions for all components of their data pipeline
- Flexible and seamlessly increase performance and capacity

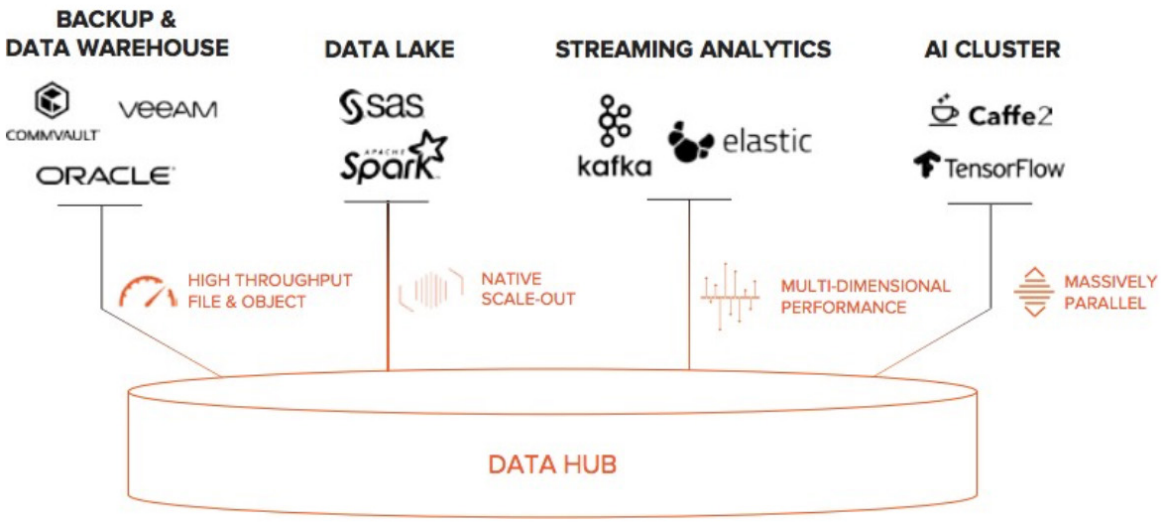
Implementation overview

This solution architecture is built with two main considerations in mind:

- Build a dynamic and agile infrastructure to support the iterative nature of this project
- Protect both computing and storage investments as organizations' AI-based solutions and capabilities mature

The solution presented here demonstrates scale-out capability from 1 GPU to 32 GPUs (four C480

FIGURE 2. ESSENTIAL FEATURES OF A DATA HUB



ML M5 servers) while running TensorFlow-based training with synthetic data and an ImageNet data set.

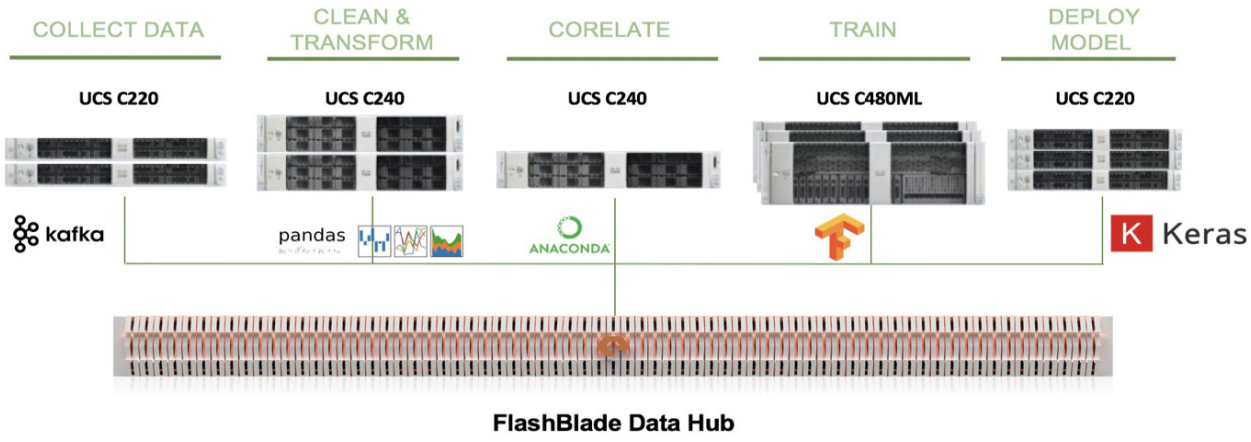
The FlashStack AI architecture is designed for scale-out deep-learning workloads and is not restricted to this size. As data sets and workload requirements scale, additional C480 ML servers can be provisioned and instantly gain access to all available data. Similarly, as storage capacity or performance demands grow, blades can be added to the FlashBlade system with no downtime or reconfiguration.

Additionally, FlashBlade's "tuned for everything" architecture makes it consistently high performing across a wide range of analytics workloads. This capability enables other analytics teams within the organization to use this same central data hub, reducing data duplication and data management.

Audience

The audience for this document includes sales engineers, field consultants, professional services, IT managers, partner engineers, IT architects, and customers who want to take advantage of an infrastructure that is built to deliver IT efficiency and enable IT innovation. The reader of this document is expected to have the necessary training and background to install and configure Red Hat Enterprise Linux (RHEL), the Cisco UCS platform, and Cisco Nexus® switches and to understand AI and machine-learning workloads. External references are provided where applicable, and familiarity with these additional documents is highly recommended.

FIGURE 3. PURE STORAGE FLASHBLADE AND CISCO UCS DATA HUB



TECHNOLOGY OVERVIEW

This section provides a brief introduction to the hardware and software components used in this solution.

Cisco Unified Computing System

Cisco UCS is a state-of-the-art data center platform that unites computing, network, storage access, and virtualization into a single cohesive system.

These are the main components of Cisco UCS:

Computing: The system is based on an entirely new class of computing system that incorporates rack-mount and blade servers based on Intel® Xeon® processor E5 and E7 CPUs. The Cisco UCS servers offer the patented Cisco Extended Memory Technology, which supports applications with large data sets and allows more virtual machines per server.

Network: The system is integrated onto a low-latency, lossless, 40/100Gbps unified network fabric. This network foundation consolidates LANs, SANs, and high-performance computing networks, which are separate networks today. The unified fabric lowers costs by reducing the number of network adapters, switches, and cables and by decreasing power and cooling requirements.

Storage access: The system provides consolidated access to both SAN storage and network-attached storage (NAS) over the unified fabric. By unifying storage access, Cisco UCS can access storage over Ethernet (NFS or Small Computer System Interface over IP [iSCSI]), Fibre Channel, and Fibre Channel over Ethernet (FCoE). This capability provides customers with choices for storage access and investment protection. In addition, server administrators can preassign storage-access policies for system connection to storage resources, simplifying storage connectivity and management for increased productivity.

Cisco UCS is designed to deliver:

- Lower total cost of ownership (TCO) and increased business agility
- Increased IT staff productivity through just-in-time provisioning and mobility support
- A cohesive, integrated system that unifies the technology in the data center
- Industry standards supported by a partner ecosystem of industry leaders

Cisco UCS C480 ML M5 server AI platform

The Cisco UCS C480 ML M5 Rack Server is the latest addition to the Cisco UCS server portfolio. It is a Cisco UCS server built 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 4).

The four-rack-unit (4RU) Cisco UCS C480 ML M5 server is specifically built for deep learning. It is optimized for storage and I/O 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 large storage capacity and up to 100-Gbps network connectivity.

The Cisco UCS C480 ML M5 (Figure 5) offers these features and benefits:

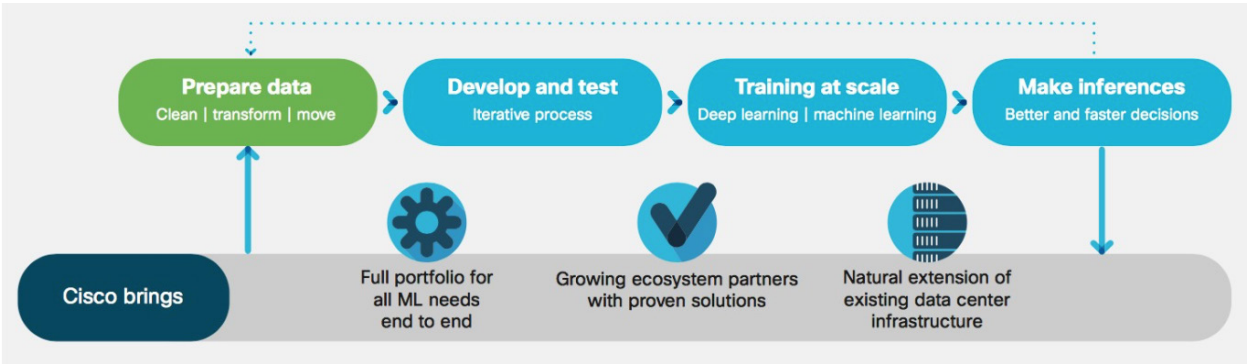
- ***GPU acceleration***: Eight NVIDIA Tesla V100 SXM2 32-GB modules are interconnected with NVIDIA NVLink technology for fast communication across GPUs to accelerate computing. NVIDIA specifies TensorFlow performance of up to 125 teraflops per module, for a total of up to 1 petaflop of processing capability 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 NVLink interconnects. 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 for deep learning.

- ***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

FIGURE 4. SUPPORT YOUR DATA SCIENTISTS WITH A COMPLETE PORTFOLIO OF AI AND MACHINE-LEARNING SERVERS



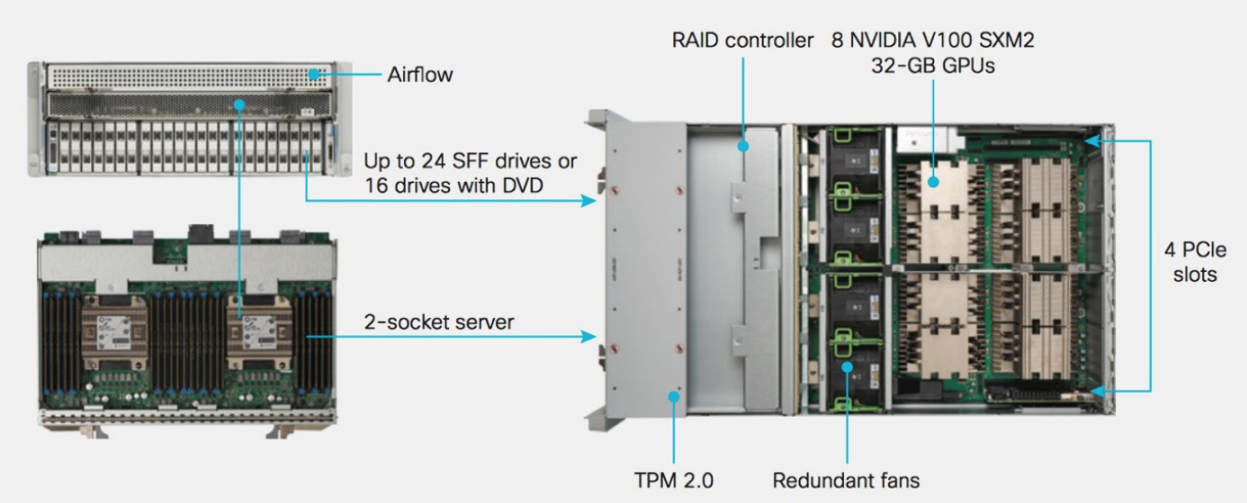
slots can be used for Non-Volatile Memory Express (NVMe) drives, providing best-in-class performance.

- ***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.

NVIDIA Tesla V100 SMX2

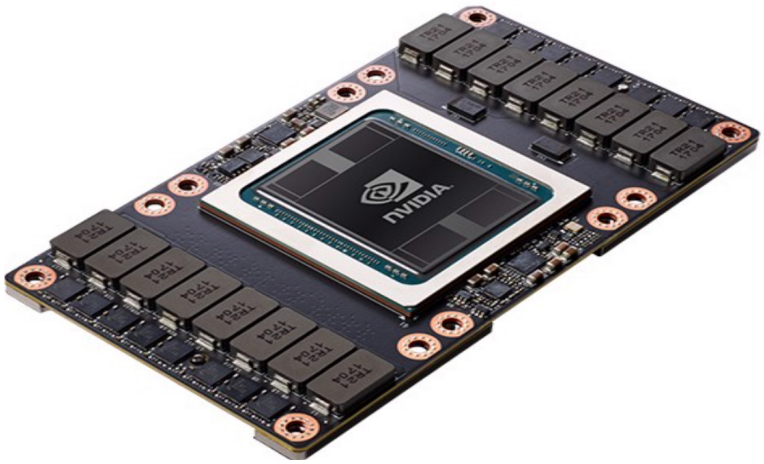
The NVIDIA Tesla V100 (Figure 6) is the world's most advanced data center GPU ever built to accelerate AI, high-performance computing (HPC), and graphics processing. Powered by NVIDIA Volta, the latest GPU architecture, the Tesla V100 offers the performance of up to 100 CPUs in a single GPU, enabling data scientists, researchers, and engineers to address challenges that were once thought impossible.

FIGURE 5. CISCO UCS C480 ML M5 RACK SERVER PHYSICAL DESIGN



- ***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.

FIGURE 6. NVIDIA TESLA V100 SXM2 GPU



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.

Mellanox ConnectX-5 EN card

ConnectX-5 EN (Figure 7) supports 100 Gigabit Ethernet connectivity and delivers extremely high message rates and PCIe switch and NVMe over fabric offloads. ConnectX-5 provides extremely high performance and an extremely flexible solution for the most demanding applications: machine learning, data analytics, and more.

ConnectX-5 EN uses remote direct memory access (RDMA) over converged Ethernet (RoCE) technology, delivering low-latency and high performance. ConnectX-5 enhances RDMA network capabilities by completing the switch adaptive routing capabilities and supporting data delivered out of order, while maintaining ordered completion semantics. It provides

multipath reliability and efficient support for all network topologies, including DragonFly.

ConnectX-5 supports dual ports, each with 100-Gbps Ethernet connectivity, latency of less than 700 nanoseconds, and a very high message rate. It also supports PCIe switch and NVMe-over-fabric offloads, providing a high-performance and extremely flexible solution for the most demanding applications and markets.

Pure Storage FlashBlade

Data is the most valuable asset in an organization today. However, slow and complex traditional storage systems often prevent data from being put to use. FlashBlade (Figure 8) is the industry's most advanced file and object storage platform: a data hub built to consolidate data silos such as backup appliances and data lakes to accelerate tomorrow's discoveries and insights.

Deep learning requires more than fast computing and high-bandwidth interconnects. When designing a full-stack platform for large-scale deep learning, the system architect's goal is to provision as many GPUs as possible while helping ensure linearity of performance as the environment is scaled and keeping the GPUs fed with data. Keeping the GPUs fed requires a high- performance data pipeline all the way from the storage system to the GPUs. When defining storage for deep-learning systems, architects must consider three requirements:

- ***Diverse performance***: Deep learning often requires I/O rates of multiple gigabytes per second, but it isn't restricted to a single data type or I/O size. Training deep neural network models for applications as diverse as machine vision, natural- language processing, and anomaly detection requires a variety of data types and data set sizes. Storage systems that fail to deliver the performance required during neural network training will starve the GPU tier for data and prolong the length of the run, inhibiting developer productivity and efficiency. Providing consistency of performance at various I/O sizes and profiles at capacity scale helps ensure success.
- ***Scalable capacity***: Successful machine-learning projects often have ongoing data acquisition and continuous training requirements, resulting in continued growth of data over

FIGURE 7. CONNECTX-5 EN SINGLE-PORT ADAPTER SUPPORTS

time. Furthermore, enterprises that succeed with one AI project find ways to apply these powerful techniques to new application areas, resulting in further data expansion to support multiple use cases. Storage platforms with inflexible capacity limits result in high administrative overhead to federate disparate pools.

- ***Strong resiliency***: As the value of AI grows in an organization, so does the value of the infrastructure that supports its delivery. Storage systems that result in excessive downtime or require extensive administrative outages can cause costly project delays and service disruptions.

Existing storage systems fail to meet one or more of these requirements or force architects and administrators to experience excessive deployment and management complexity.

Initial deployments for deep learning often start with direct-attached storage (DAS), resulting in hard capacity limits and challenges in sharing data sets across multiple compute units. Collecting multiple DAS servers into a shared file system with the Hadoop Distributed File System (HDFS) can alleviate capacity concerns, but the performance cost is high for the small, random I/O patterns common in many deep-learning use cases. Furthermore, burdening the CPUs in a GPU server with storage management can lead to bottlenecks in the overall pipeline and poor resource utilization.

Parallel file systems such as the General Parallel File System (GPFS) and Lustre, designed specifically for the needs of high- performance computing (HPC), can be tuned by expert-level administrators to meet the requirements of a particular workload. However, a new data set or training model inevitably requires a new configuration and tuning process, resulting in project delays and potential stranded capacity. Traditional NAS offerings can provide strong resilience and scalable capacity but often fail to deliver the performance required across a range of I/O patterns and large-scale computing clusters.

Pure Storage FlashBlade, with its scale-out, all-flash architecture and

FIGURE 8. PURE STORAGE FLASHBLADE

a distributed file system purpose-built for massive concurrency across all data types, is the only storage system that delivers on all of these characteristics while keeping the required configuration and management complexity to a minimum. In addition, with multichassis support, FlashBlade seamlessly scales from terabytes to petabytes in one name space.

Cisco Nexus 93180LC-EX Switch

The Cisco Nexus 9000 Series Switches include both modular and fixed-port switches that are designed to overcome these challenges with a flexible, agile, low-cost, application-centric infrastructure.

The Cisco Nexus 93180LC-EX Switch (Figure 9) is the industry's first 50-Gbps 1RU switch that provides flexible line-rate Layer 2 and 3 feature sets. Designed with Cisco Cloud Scale technology, it supports flexible migration options. It is well suited for highly scalable cloud architectures and enterprise data centers and operates in ACI mode.

Cisco provides two modes of operation for the Cisco Nexus 9000 Series. Organizations can use Cisco NX-OS Software to deploy the Cisco Nexus 9000 Series in standard Cisco Nexus switch environments. Organizations also can use a hardware infrastructure that is ready to support Cisco Application Centric Infrastructure (Cisco ACI,Ñ¢) to take full advantage of an automated, policy- based, systems management approach.

In ACI mode, the Cisco Nexus 93180LC-EX Switch has twenty-four 40- and 50-Gbps Quad Enhanced Small Form-Factor Pluggable Plus (QSFP+) ports and six 40- and 100-Gbps QSFP28 uplink ports. This 1RU switch supports 3.6 Tbps of bandwidth and over 2.6 billion packets per second (bpps) across twenty-four fixed 40 and 50-Gbps QSFP+ ports. The switch provides these main features:

Architectural flexibility

- Deploy top-of-rack, fabric extender aggregation, or middle-of-row fiber-based server access connectivity for traditional and leaf-and-spine architectures.
- Increase scale and simplify management through support for Cisco Nexus 2000 Series Fabric Extenders.

Comprehensive feature set

- Enhanced Cisco NX-OS Software is designed for performance, resiliency, scalability, manageability, and programmability.
- Virtual extensible LAN (VXLAN) routing provides network services.

FIGURE 9. CISCO NEXUS 93180LC-EX SWITCH

- Real-time buffer utilization per port and per queue enables monitoring of traffic microbursts and application traffic patterns.

Real-time visibility and telemetry

- Cisco Tetration Analytics™ platform support with built-in hardware sensors enable comprehensive traffic flow telemetry and line-rate data collection.
- Cisco Nexus Data Broker support enables network traffic monitoring and analysis.

Highly available and efficient design

- Deploy a high-performance, nonblocking architecture.
- Easily deploy the switch in either a hot-aisle or cold-aisle configuration.
- The switch supports redundant, hot-swappable power supplies and fan trays.

Simplified operations

- Automate and configure switches with DevOps tools such as Puppet, Chef, and Ansible.
- Python scripting gives programmatic access to the switch command-line interface (CLI).
- The switch supports hot and cold patching and online diagnostics.

Deep-learning neural framework and tools

The NVIDIA deep-learning software development kit (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 and 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 increased training throughput, 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 nonlinear data. It has proven to be efficient at classifying images, as shown by the impressive results of deep neural networks in the ImageNet competition, for example. However, training these models requires large data sets and is time consuming.

ImageNet is a large visual database designed for use in visual object recognition software research, developed by Stanford and MIT. It is the data set most commonly used by major training models (ResNet, Inception, VGG16, etc.) for performance benchmarking. ImageNet provides a common

foundation for comparing architectures. Networks trained on ImageNet can be starting points for other computer vision tasks.

TensorFlow

In this document, 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 is a software library, developed by the Google Brain Team within the Google Machine Learning Intelligence research organization.

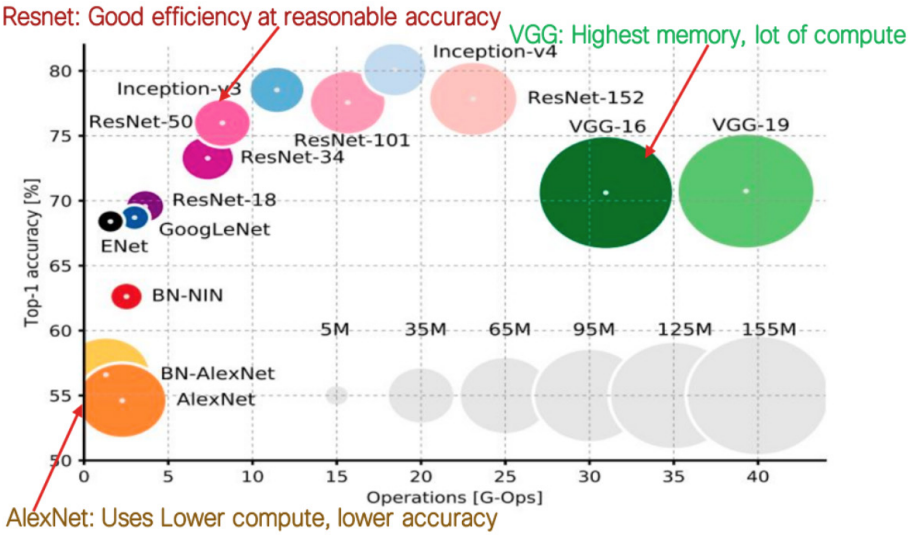
The main features of TensorFlow include the following:

- 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

The TensorFlow CNN benchmarks contain implementations of several popular convolutional models and can be run on a single machine or in distributed mode across multiple hosts.

FIGURE 10. OPERATIONS VERSUS ACCURACY AMONG THE TRAINING MODELS

Refer: https://arxiv.org/pdf/1605.07678.pdf



Choice of model

The first step is to choose the model that you want to use. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) requires models to classify 1000 categories of images, and some suggested models cannot provide 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 the models work behind the scenes.

CNN models have evolved to support a range of trade-offs between time to accuracy and number of operations. For example, ResNet50 requires relatively fewer operations to achieve relatively high accuracy. Visual Geometry Group (VGG) models require the greatest amount of memory and high computing requirements with moderate accuracy for a single forward pass (Figure 10).

DISTRIBUTED TENSORFLOW

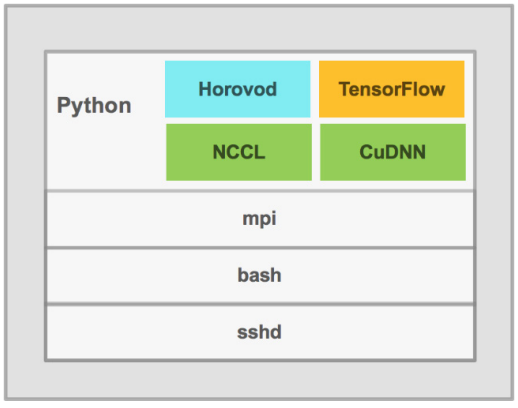
This section presents benchmark results from training deep neural networks on the FlashStack AI platform. Training performance was measured for a variety of convolutional networks using the popular ImageNet data set, which consists of 1.28 million images, with a total size of 143 GB. All tests described in this section were conducted using the hardware and software components described in the "*Solution design*" section of this document.

Methodology

To orchestrate tests across distributed storage and GPUs, several widely used public benchmarking tools were used for individual layers of an AI pipeline, shown in Figure 11.

Each Cisco UCS C480 ML M5 server runs a Docker container with Python libraries for optimized distributed training.

FIGURE 11. SOFTWARE STACK USED IN DISTRIBUTED TRAINING BENCHMARK TESTS



TENSORFLOW AND MODEL TUNING

TensorFlow's standard *tf_cnn_benchmarks* were used for the primary test orchestration for the toolkit.

In TensorFlow, the stage in which data is read from storage is called the input pipeline. Testing revealed that, of the many configurable options for TensorFlow, the options that adjust aspects of the input pipeline affect performance the most when training using real-world data. Following common best practices for TensorFlow performance optimization, the input parameters were tuned for four models supported by *tf_cnn_benchmarks*, solving for maximum training performance.

For the full list of tunables, see GitHub: *Command for TensorFlow performance comparison*.

Horovod

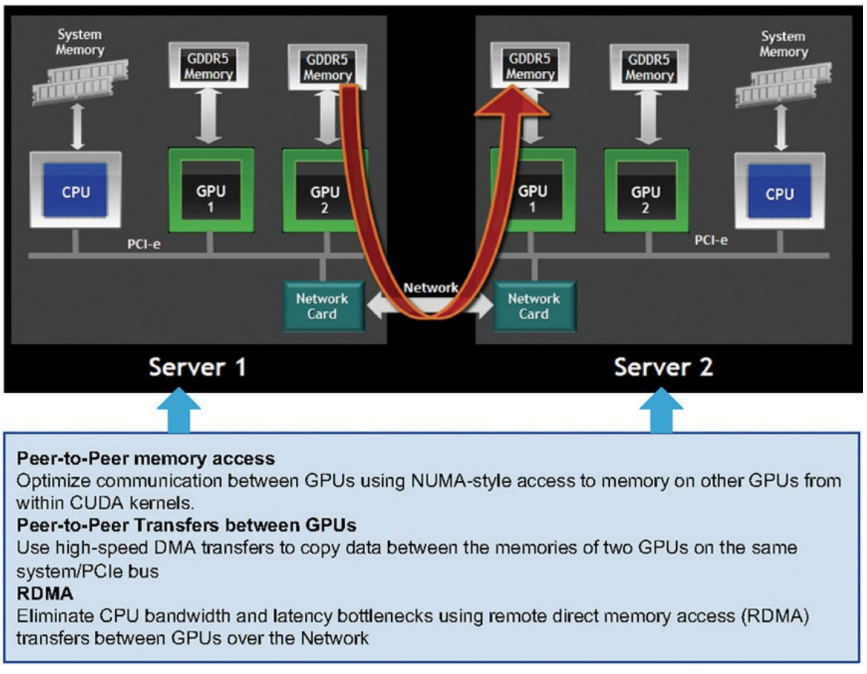
Horovod is a library that optimizes inter-process communication (using the message passing interface [*MPI*]) during training on distributed GPUs. Scaling efficiency is *double* that of standard distributed TensorFlow. All data exchanges-gradient updates from neural-network training-are performed by the NVIDIA Collective Communications Library (NCCL). NCCL includes significant optimizations for communication over NVLink within each NVIDIA DGX-1 server, as well as communication over RDMA between DGX-1 servers.

Horovod is integrated as part of the *tf_cnn_benchmarks*.

NVIDIA GPUDirect

NVIDIA GPUDirect RDMA is a technology that enables a direct path for data exchange between the GPU and third-party peer devices using standard PCIe features. Enabled on Tesla GPUs, GPUDirect RDMA relies on the ability of NVIDIA GPUs to expose portions of device memory in a PCIe base address register region, as shown in Figure 12.

FIGURE 12. NVIDIA GPUDIRECT PROCESSING FLOW



SOLUTION DESIGN

This section provides an overview of the hardware and software components used in this solution, as well as the design factors you should consider to make the system work as a single, highly available solution.

Physical topology

Figure 13 shows the physical topology of the FlashStack AI solution.

Hardware and software stack definitions

Tables 1 and 2 list the hardware and software versions used to achieve the results described in this document.

FIGURE 13. FLASHSTACK AI PHYSICAL TOPOLOGY

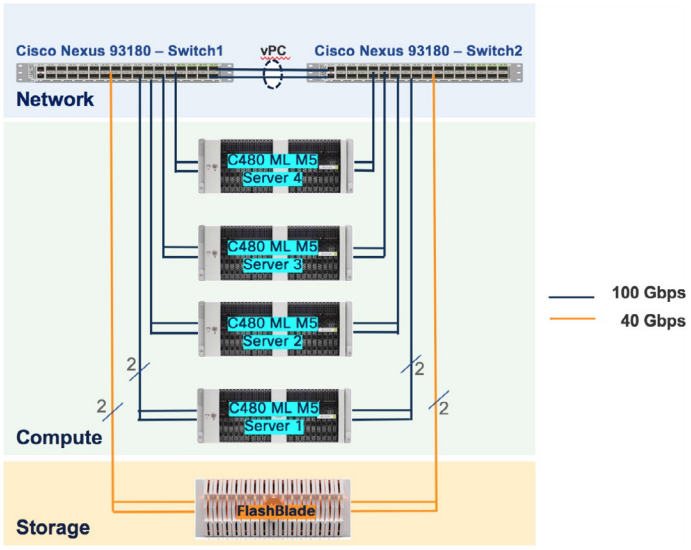


TABLE 1.HARDWARE STACK

COMPONENT	MODEL	QUANTITY
Storage	Pure Storage FlashBlade: 7 x 17 TB	1
Computing	Cisco UCS C480 ML M5 Rack Server	4
GPU per server	NVIDIA SXM2 V100 with 32 GB of memory	8
CPU per server	Intel Xeon Platinum 8180 Scalable processor	2
Memory per server	32-GB 2666-MHz DIMMs	24
Local disks per server	7-TB NVMe drives	1
Storage controller per server	Cisco 12-Gbps SAS Modular RAID Controller	1
Ethernet adapter per server	Mellanox CX-515A EN 100 Gigabit Ethernet	4
Network	Cisco Nexus 93180LC-EX Switch	2

TABLE 2. SOFTWARE VERSIONS

SOFTWARE	VERSION
Cisco UCS C480 ML M5	Cisco Integrated Management Controller (IMC): Release 4.0.2c BIOS: C480 M5.4.0.2a
Pure Storage FlashBlade	Purity//FB 2.2.12
Cisco Nexus 93180LC-EX	Cisco NX-OS Release 7.0(3)I7(1)
Red Hat Enterprise Linux	Release 7.5
NVIDIA driver	Release 410.72
NVCR Docker image	TensorFlow 18.10-py2 • TensorFlow: 1.10.0 • Horovod: 0.13.10 • OpenMPI: 3.0.0 • CUDA: 10.0.130

SOLUTION SCALE BENCHMARK RESULTS

The server was tested with end-to-end deep-learning training workloads that resemble our customers' real-world workloads. Training throughput when reading data sets from external storage into C480 ML servers was optimized, helping ensure that the platform provided linearly scaling performance.

Benchmark testing consisted of two main types of tests:

- Single-server scale tests
- Multiple-server scale tests

Single-server scale tests

Benchmarking training performance with data read from an external storage system exercises both the TensorFlow input pipeline and the TensorFlow core training pipeline. In contrast, testing with synthetic data generated inside the GPUs exercises only the core training pipeline.

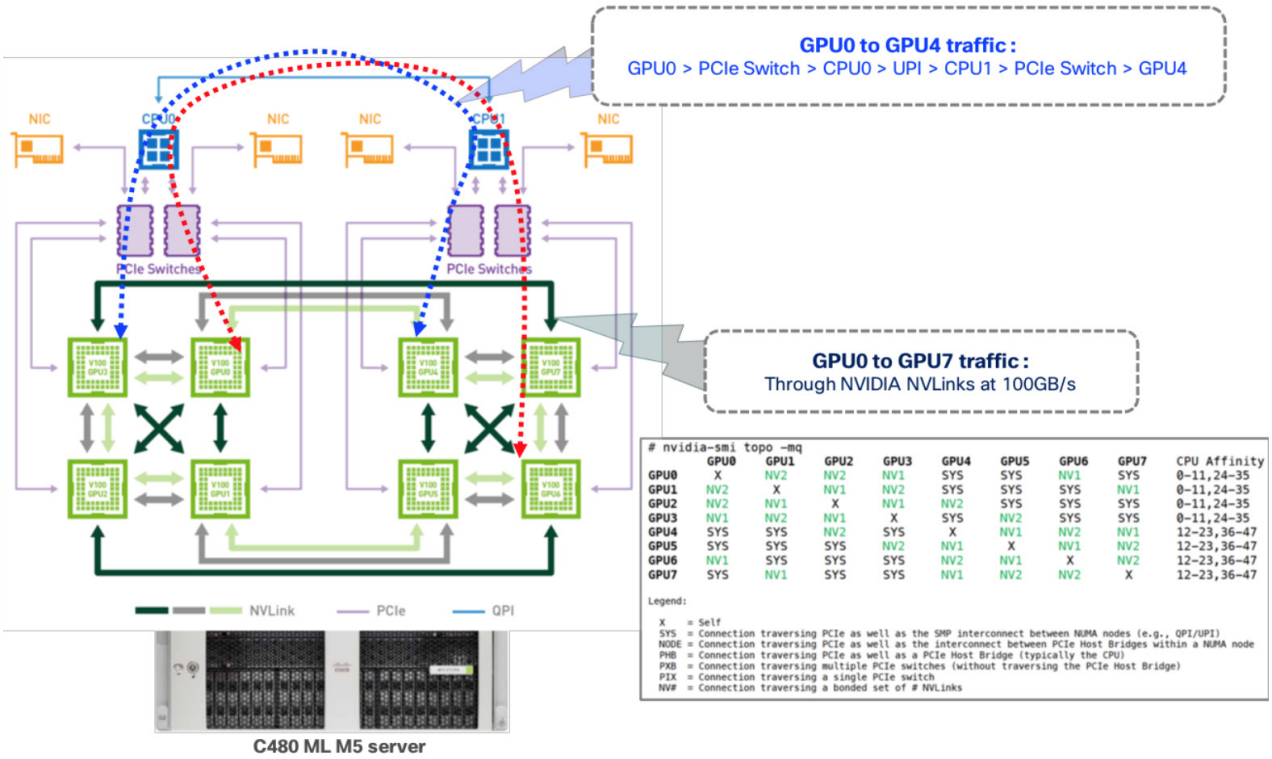
Optimally, when reading real-world data from storage, training performance should be as close to this synthetic baseline as possible. When the results from synthetic data and real data are nearly equal, the input pipeline is essentially invisible, allowing GPU throughput close to the hypothetical maximum.

Figure 14 shows the traffic flow and Figure 15 shows the CPU

utilization for the Cisco C480 ML M5 that occurred when running each model using eight V100 GPUs. Figure 15 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. The Cisco C480 ML M5 server provides flexibility in the choice of CPU model, which enables organizations to tune their CPU:GPU computing ratio if desired.

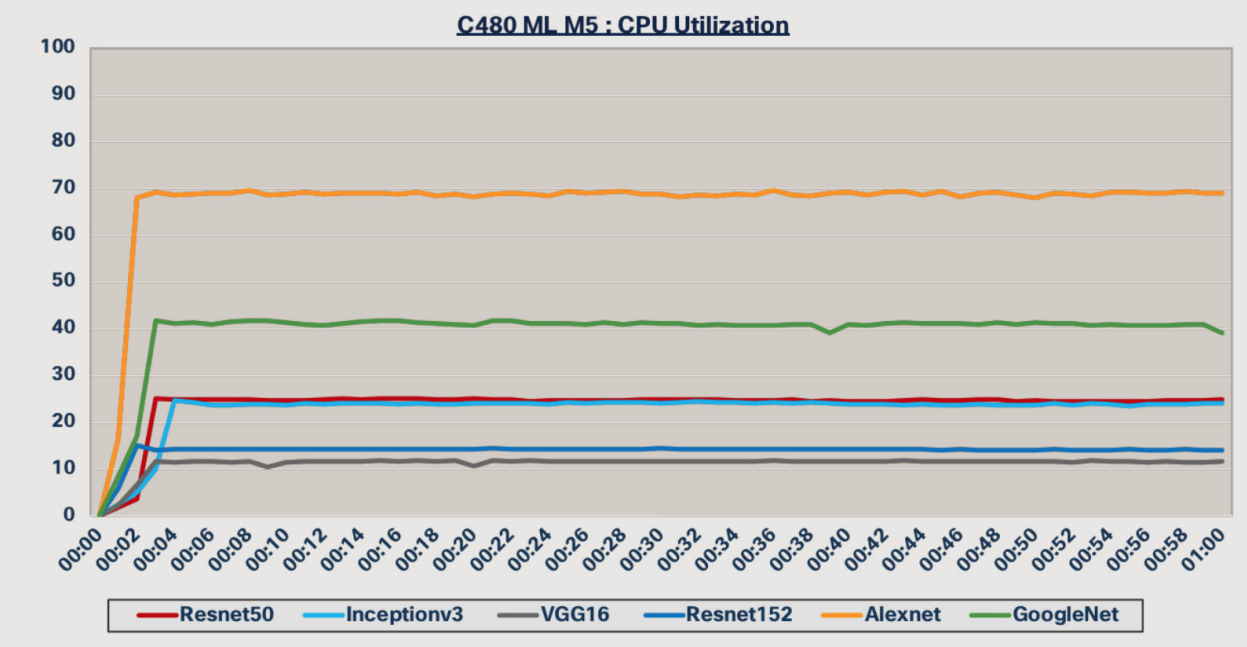
In addition to overall training throughput, you should consider CPU utilization during deep-learning training jobs. Many data processing steps are run on the CPU, and the CPU can quickly become the bottleneck, leaving GPUs starved for data.

FIGURE 14. GPU-TO-GPU TRAFFIC USING NVIDIA NVLINK TOPOLOGY



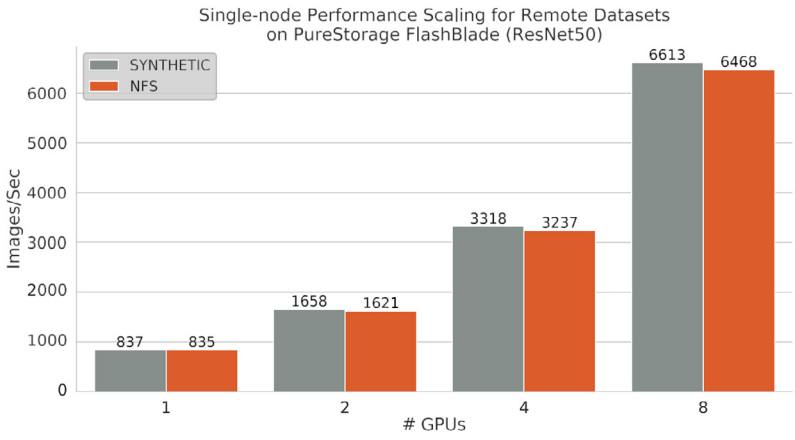

```
# nvidia-smi topo -mq
        GPU0    GPU1    GPU2    GPU3    GPU4    GPU5    GPU6    GPU7    CPU Affinity
GPU0     X      NV2     NV2     NV1     SYS     SYS     NV1     SYS     0-11,24-35
GPU1    NV2      X      NV1     NV2     SYS     SYS     SYS     NV1     0-11,24-35
GPU2    NV2     NV1    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

FIGURE 15. CPU UTILIZATION FOR CISCO UCS C480 ML M5 BASED ON TENSORFLOW BENCHMARK



The tests compared the performance of ImageNet data and synthetic input data for four neural network architectures (models) across varying numbers of GPUs. Across all the models, FlashStack AI throughput is within 5 percent of the results obtained with synthetic data. These results were obtained after tuning the input pipeline in TensorFlow. Figure 16 shows the results.

FIGURE 16. SINGLE-SERVER PERFORMANCE RESULT



Multiple-server GPU scale tests

Data science teams should be able to depend on infrastructure scalability as their data sets, team sizes, and number of projects grow. For optimal team productivity, infrastructure must provide near-linear performance scaling as computing and storage nodes are added to the system.

The FlashStack AI platform can deliver linear throughput scaling with low scale-out networking cost as C480 ML M5 servers are added.

Test runs and Observations:

For scale test, all GPUs are communicating among each other within the same C480 ML server and also across the servers and we observed 3 kinds of GPU traffic patterns.

1. For single server, GPUs connected directly via NVIDIA NVlink, the GPU traffic flows via NVlinks (100GB/s).
2. GPUs to GPUs traffic travesing via CPU UPI interconnect links within the same server. In this case, GPUs are not connected directly via NVLinks. In other words, the traffic does not flow via NVlink hops.
3. For Mutli server scale with GPU direct, the GPUs to GPU traffic flows via RoCE v2 over Nexus switches.

During the multiple-node tests, full GPU utilization can be achieved, making use of the valuable computing resources (Figure 17).

Figure 18 shows multiple-node networking traffic paths.

FIGURE 17. FULL GPU UTILIZATION ACHIEVED DURING MULTIPLE-NODE TESTS

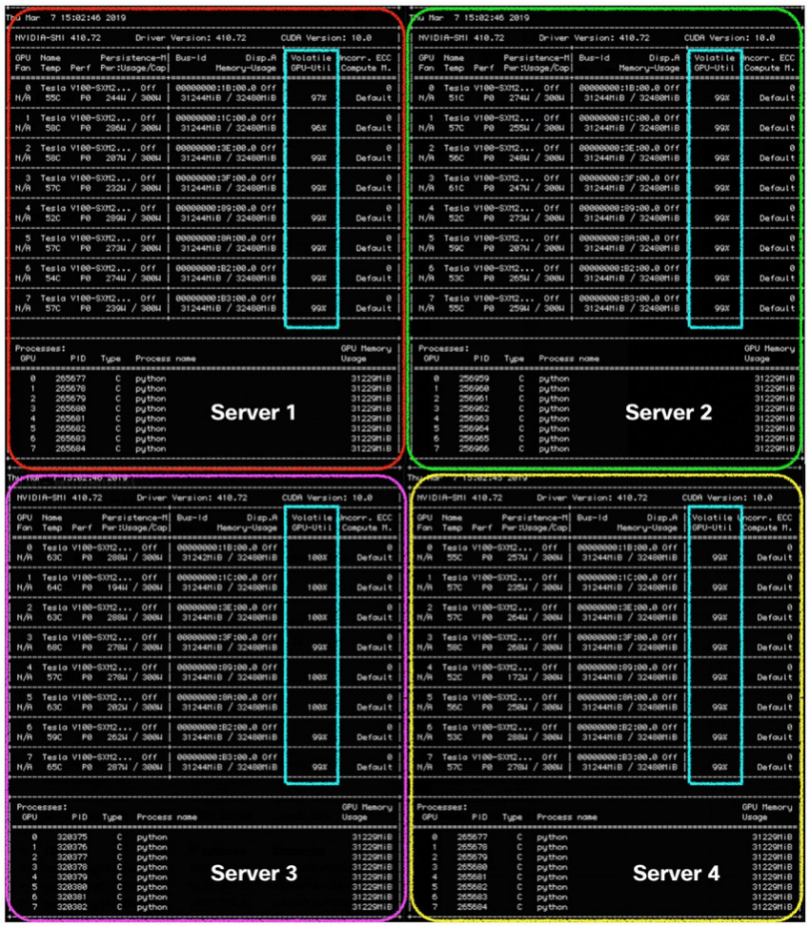


FIGURE 18. ROCEV2 TRAFFIC BETWEEN C480 ML M5 SERVERS THROUGH CISCO NEXUS 93180 SWITCHES

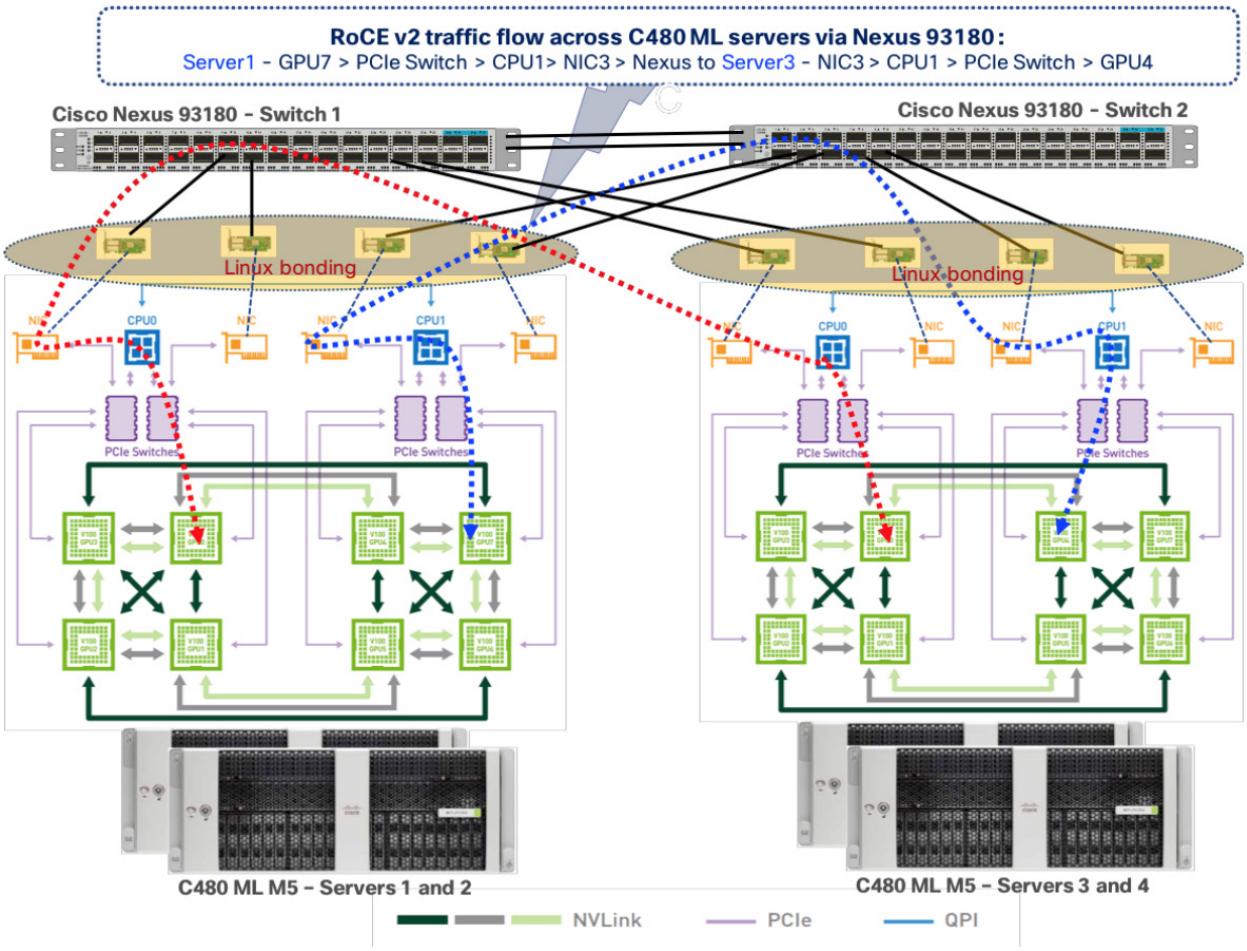
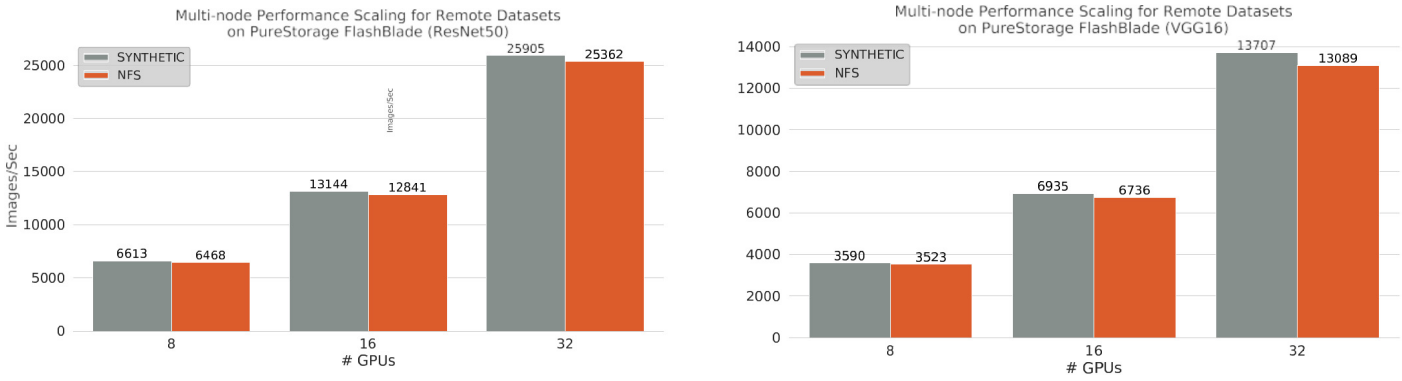


FIGURE 19. DISTRIBUTED TRAINING PERFORMANCE RESULTS FOR RESNET50 (BATCH SIZE 256) AND VGG16 (BATCH SIZE 64): BOTH MODELS DEMONSTRATE LINEAR PERFORMANCE SCALING ACROSS GPU COUNT



Results from both the relatively small ResNet50 model and the relatively large VGG16 model are shown in Figure 19.

Observations

The following observations were made from the test results:

- Throughput is nearly identical between training on remote data sets saved to FlashBlade and training on synthetic data sets (generated directly on the GPUs), meaning that the networking pipeline from storage to CPU to GPU is nearly invisible, even at multiple-node scale.
- The end-to-end FlashStack AI system can provide linearly scaling performance. As C480 ML M5 computing nodes are added, minimal scale-out networking overhead is added for both small models (such as ResNet50) and larger models (such as VGG16). AI teams can grow their environments as needed and depend on scalable networking performance that stays simple to manage.

CONCLUSION

For today's enterprises, data is the fuel that drives insight and innovation. Deep-learning projects can bring new value to businesses, and FlashStack for AI delivers the massive performance and scalability they require.

Deploying an end-to-end AI platform doesn't have to be a complex process. FlashStack for AI offers seamless integration and is built to accelerate the entire data pipeline, from data ingestion and extract, transform, and load (ETL) processing to training and model deployment.

THIS APPENDIX SHOWS A SAMPLE CISCO NEXUS CONFIGURATION FOR ROCE V2 TRAFFIC.

```
# Create Policy Map and QoS group for RoCE traffic

policy-map type network-qos ROCE-NQ-policy
    class type network-qos c-8q-nq3
        pause pfc-cos 3
        mtu 9216
    class type network-qos c-8q-nq7
        mtu 1500
    class type network-qos c-8q-nq-default
        mtu 9216
    class-map type qos match-any ROCE-class
        match cos 3
        match dscp 24-31
    class-map type qos match-any control-class
        match cos 5-7
        match dscp 40-63
    policy-map type qos marking-policy
        class control-class
            set qos-group 7
        class ROCE-class
            set qos-group 3
        class class-default
            set qos-group 0
    system qos
        service-policy type network-qos ROCE-NQ-policy

# Configure Priority Flow Control (PFC) and QoS policy on Ethernet ports

interface Ethernet1/1
    switchport mode trunk
    switchport trunk allowed vlan 1,239-240
    priority-flow-control mode on
    spanning-tree port type edge trunk
    speed 100000
```

```
    mtu 9216
    no negotiate auto
service-policy type qos input marking-policy

interface Ethernet1/5
    switchport mode trunk
    switchport trunk allowed vlan 1,239-240
    priority-flow-control mode on
    spanning-tree port type edge trunk
    speed 100000
    mtu 9216
    no negotiate auto
    service-policy type qos input marking-policy

interface Ethernet1/7
    switchport mode trunk
    switchport trunk allowed vlan 1,239-240
    priority-flow-control mode on
    spanning-tree port type edge trunk
    speed 100000
    mtu 9216
    no negotiate auto
    service-policy type qos input marking-policy

interface Ethernet1/9
    switchport mode trunk
    switchport trunk allowed vlan 1,239-240
    priority-flow-control mode on
    spanning-tree port type edge trunk
    speed 100000
    mtu 9216
    no negotiate auto
    service-policy type qos input marking-policy
```

```
interface Ethernet1/11
    switchport mode trunk
    switchport trunk allowed vlan 1,239-240
    priority-flow-control mode on
    spanning-tree port type edge trunk
    speed 100000
    mtu 9216
    no negotiate auto
    service-policy type qos input marking-policy

interface Ethernet1/13
    switchport mode trunk
    switchport trunk allowed vlan 1,239-240
    priority-flow-control mode on
    spanning-tree port type edge trunk
    speed 100000
    mtu 9216
    no negotiate auto
    service-policy type qos input marking-policy

interface Ethernet1/15
    switchport mode trunk
    switchport trunk allowed vlan 1,239-240
    priority-flow-control mode on
    spanning-tree port type edge trunk
    speed 100000
    mtu 9216
    no negotiate auto
    service-policy type qos input marking-policy
```

APPENDIX B: MELLANOX CONNECTX-5 CONFIGURATION

Convert InfiniBand mode to Ethernet (IP) mode.

1. Install the latest MLNX_OFED release (Release 4.0 or later).

   root@c480-1:~# ofed_info -s
   MLNX_OFED_LINUX-4.0-1.0.1.0:

2. Check that the adapters are recognized by running the ***lspci*** command:

   root@c480-1:~# lspci | grep Mellanox
   1a:00.0 Ethernet controller: Mellanox Technologies MT27800 Family [ConnectX-5]
   3d:00.0 Ethernet controller: Mellanox Technologies MT27800 Family [ConnectX-5]
   88:00.0 Ethernet controller: Mellanox Technologies MT27800 Family [ConnectX-5]
   b1:00.0 Ethernet controller: Mellanox Technologies MT27800 Family [ConnectX-5]
   oot@c480-1:~#
   Note: In ConnectX-5, each port is identified by a unique number.

3. Change the link protocol to Ethernet using the MFT mlxconfig tool.
   Note: The default link protocol for ConnectX-5 is InfiniBand.

4. Start MFT.

   root@c480-1:~# mst start
   Starting MST (Mellanox Software Tools) driver set
   Loading MST PCI module - Success
   Loading MST PCI configuration module - Success
   Create devices
   Unloading MST PCI module (unused) - Success
   root@c480-1:~#

5. Extract the vendor_part_id parameter.
   Note: The ConnectX-5 ID is 4119.

```
root@c480-1:~# ibv_devinfo     |     grep vendor_part_id
vendor_part_id:                      4119
vendor_part_id:                      4119
vendor_part_id:                      4119
vendor_part_id:                      4119
```

6. Query the host about ConnectX-5 adapters.

```
root@c480-1:~#
root@c480-1:~# mlxconfig -d /dev/mst/mt4119_pciconf0 q
Device #1:
----------
Device type:  ConnectX5
PCI device:    /dev/mst/mt4119_pciconf0
Configurations: Current
...
LINK_TYPE_P1 1
LINK_TYPE_P2 1
....
```

Note: LINK_TYPE_P1 and LINK_TYPE_P2 equal 1 (InfiniBand) by default.

7. Change the port type to Ethernet (LINK_TYPE = 2).

```
#mlxconfig -d /dev/mst/mt4119_pciconf0 set LINK_TYPE_P1=2 LINK_TYPE_P2=2
Device #1:
----------
Device type: ConnectX5
PCI device: /dev/mst/mt4119_pciconf0 Configurations:
Current New
LINK_TYPE_P1 1 2
LINK_TYPE_P2 1 2
Apply new Configuration? ? (y/n) [n] : y
Applying... Done!
-I- Please reboot machine to load new configurations
```

8. Reboot the server. After the reboot, make sure that Mellanox Interfaces appears under ifconfig.

FOR MORE INFORMATION

For additional information, see the following resources:

- ***FlashStack and AI solutions:***
  https://www.cisco.com/c/en/us/solutions/data-center-virtualization/flashstack/index.html
- ***Pure Storage FlashBlade:***
  https://www.purestorage.com/solutions/applications/artificial-intelligence.html
- ***Cisco UCS C480 ML M5 Servers:***
  https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/datasheet-c78-741211.html
- ***Deep-learning benchmarks and data sets:***
  TensorFlow CNN benchmark: *https://github.com/TensorFlow/benchmarks/tree/master/scripts/tf_cnn_benchmarks*
  ImageNet data sets: *http://www.image-net.org/*
  Horovod: *https://github.com/horovod/horovod*

Pure Storage, Inc. 650 Castro Street, Mountain View, CA 94041

PS-FS-WP-FSMini-0417-0006v1-LE59701.pdf

FLASHSTACK@PURESTORAGE.COM

WWW.CISCO.COM/GO/FLASHSTACK