IBM SpectrumAI

WITH NVIDIA

IBM SpectrumAI with NVIDIA® – Converged Infrastructure Solutions for AI workloads

The engine to power your AI data pipeline

Introduction:

Artificial intelligence (AI) – including deep learning (DL) and machine learning (ML) – is the engine rapidly powering innovation across industries from healthcare to autonomous vehicles and agriculture. By 2020, IBM® projects that the world's volume of digital data will exceed 44 zettabytes.[1] Organizations that recognize the value of their data for decisions and actions are turning to DL systems that can rapidly ingest, accurately interpret, and quickly provide key data insights from the volumes of new data generated now and in the future.

Enterprises are increasing investment in AI research and innovation, with related patents growing more than 30% and academic papers by 13% during the last decade.[2] Arguably, only one kind of enterprise will survive and thrive in the future - the data-driven enterprise.

Highly performant and scalable DL systems must excel at data ingestion, data preparation, data training, and verification, and deliver inferences and classifications while handling the growing demands of DL in the organization. The DL algorithm accuracy may improve by increasing the neural network size, and data quantity and quality used for the model training step. However, this accuracy improvement comes with a significant increase in computational complexity and increased demand on DL compute, storage, and network resources.

NVIDIA® has led the AI computing revolution, leveraging the power of the modern GPU with its massive processor core count and parallel architecture, uniquely suited to the massively parallelized operations that are core to DL, and which exceed the limitations of traditional CPU based architectures. IBM Storage delivers industry-leading, high-performance, low-latency, and cost effective all-flash storage and software defined clustering with proven scalability that enables massively parallel processing of different data types such as image, audio, video, text, or time series data.

Together NVIDIA and IBM Storage provide an integrated, individually scalable compute and storage solution with end-to-end parallel throughput from flash to GPU for accelerated DL training and inference. This paper is intended for enterprise leaders, solution architects, and other readers interested in learning how the IBM SpectrumAI with NVIDIA DGX™ solution simplifies and accelerates AI. The scalable infrastructure solution integrates the NVIDIA DGX-1™ server with IBM Spectrum Scale™ which powers the IBM Elastic Storage Server (ESS) and the upcoming NVMe all flash appliance. The solution provides storage linear performance while scaling from one to nine DGX-1 servers with both synthetic workloads and ImageNet data.

[1] https://www.ibm.com/blogs/systems/ibm-and-nvidia-further-collaboration-to-advance-open-source-gpu-acceleration/

[2] https://web.luxresearchinc.com/hubfs/18%20for%202018/Lux%20Research%2018%20for%202018.pdf

Table of contents

AI data pipeline

The productivity of the data science team depends on the ready availability of the latest development frameworks; ample compute power and data accessibility. While performance is important, it is not the only consideration. Data preparation and ingest can consume most of the AI development timeline. For each project, data must be extracted from other sources and properly organized so that it can be used for model training. Once a model is developed, the data must be retained for traceability. The value of data grows with diverse use across multiple users, systems, and models. Data scientist productivity depends on the efficacy of the overall data pipeline as well as the performance of the infrastructure used for running AI workloads. Moreover, the underlying storage, and network technologies play a crucial role in both these workflow aspects.

As organizations move from prototyping AI to deploying it in a production environment, the first challenge is to embed AI into the existing data pipeline or build a data pipeline that can leverage existing data repositories.

A typical data pipeline for AI with storage requirements for each stage in the pipeline is represented below.

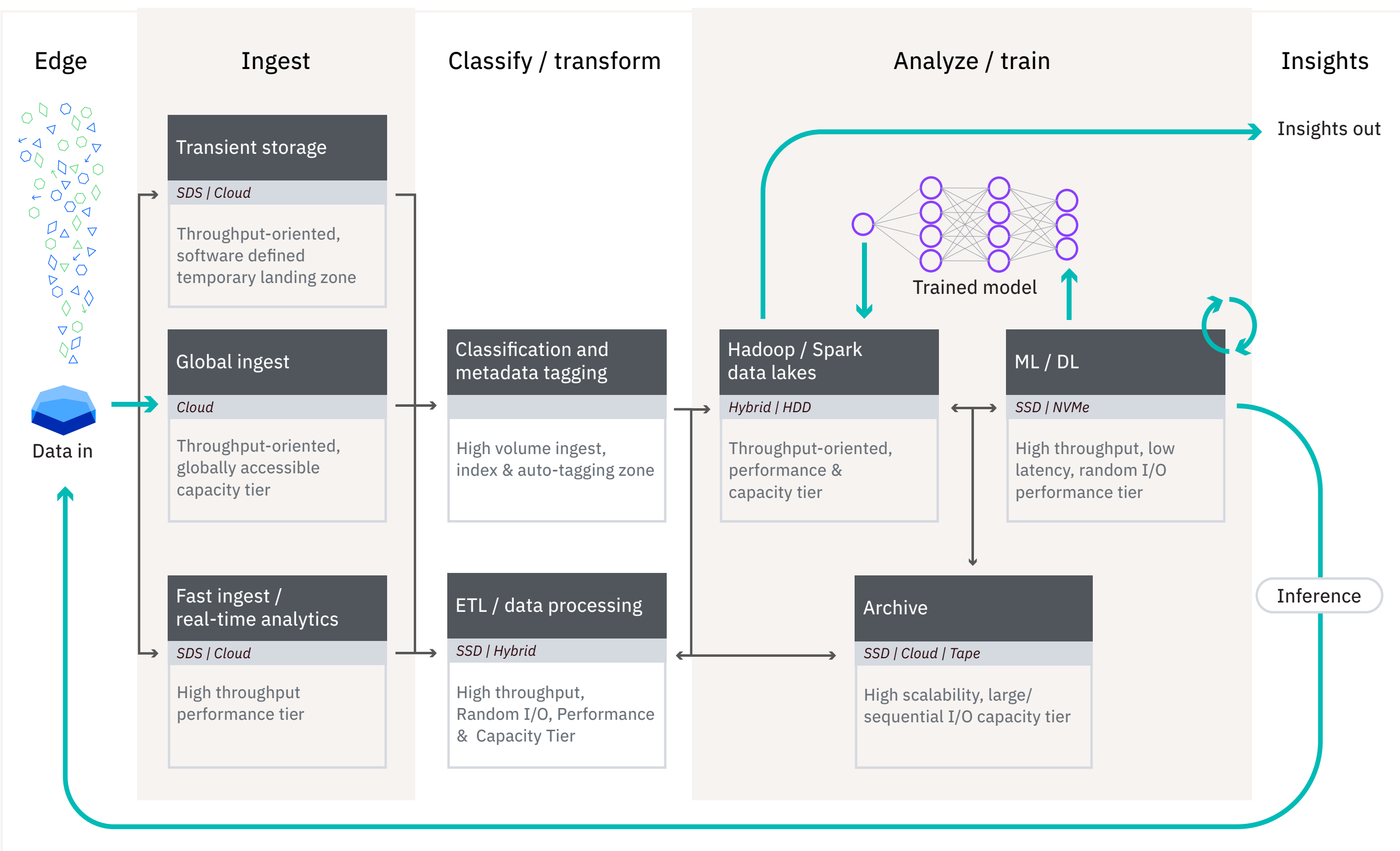


Figure 1: AI data pipeline with storage requirements

IBM delivers a comprehensive portfolio of software defined storage products (**Figure 2**) that enable customers to build their data pipelines with the right performance and cost characteristics for each stage. This includes IBM Cloud Object Storage for global ingest, geographically dispersed repositories and data preparation, as well as a variety of different Elastic Storage Server models and NVMe appliance (2019) powered by IBM Spectrum Scale for high performance file storage and scalable data lakes. Additionally, IBM Spectrum Archive enables direct file access to data stored on tape for inexpensive archives.

A new addition to this portfolio is IBM Spectrum Discover.[3] IBM Spectrum Discover is metadata management software that provides data insight for unstructured storage. IBM Spectrum Discover easily connects to IBM Cloud Object Storage and IBM Spectrum Scale to rapidly ingest, consolidate, and index metadata across petabytes of storage with billions of files and objects. IBM Spectrum Discover fulfills a leading role in the data classification phase of the overall data pipeline, but also provides capabilities that support data governance requirements and enable storage optimization along the pipeline.

Organizations who take this end-to-end data pipeline view when choosing storage technologies can benefit from improved data governance. Data science teams benefit from rapid time-to-insight with a minimum number of data copies. Infrastructure teams benefits from simplified management, scalability and improved TCO.

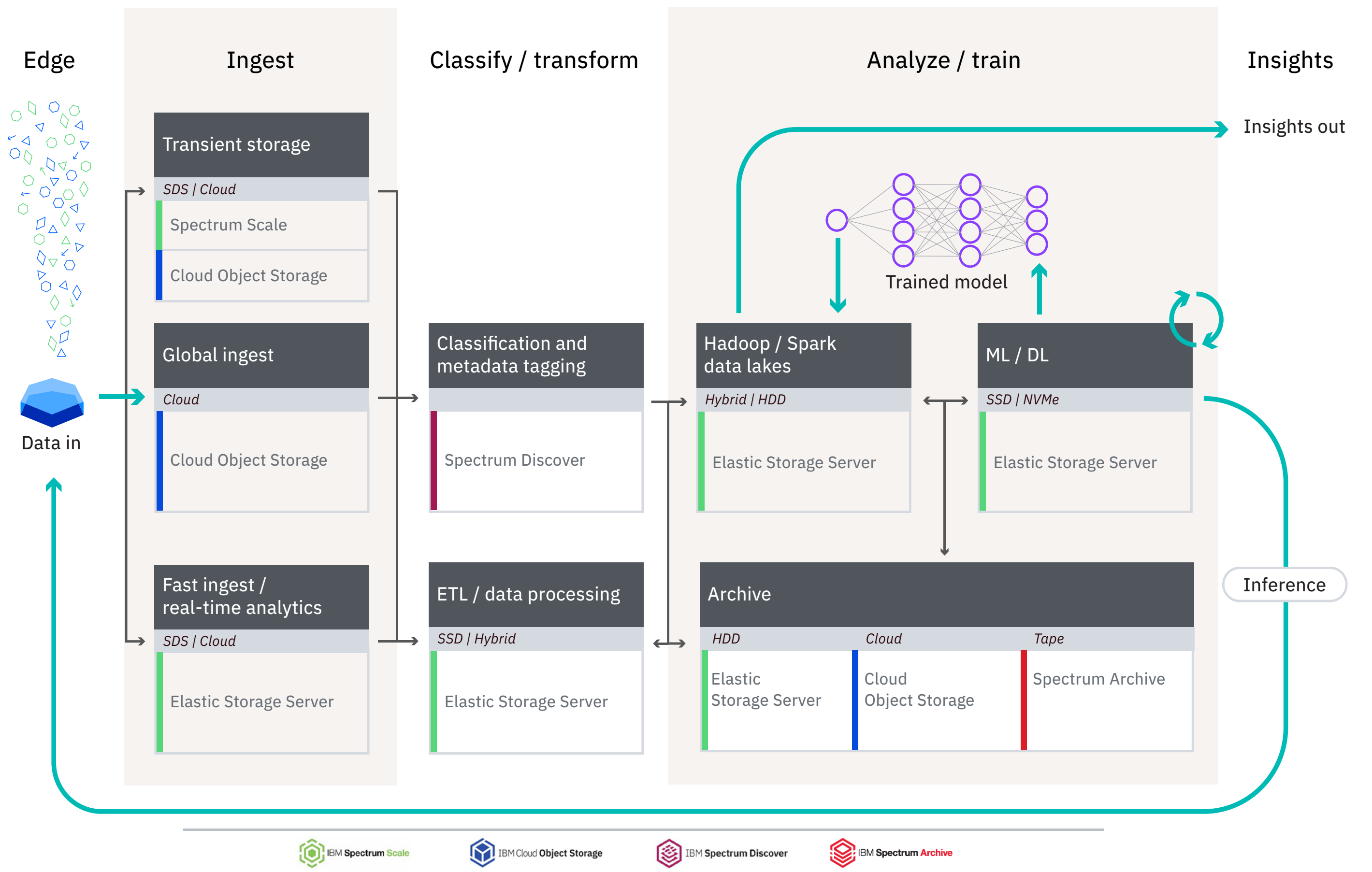


Figure 2: IBM solutions within AI data pipeline

[3] https://www.ibm.com/us-en/marketplace/spectrum-discover

If we zoom into this pipeline, represented by figure 3 below, to focus more on the AI (ANALYZE / TRAIN) stages, this phase is an iterative process itself. Once a trained neural network model is developed, it needs to be tested and retrained continuously to keep it current and to improve its accuracy. As projects grow beyond the first test systems, appropriate storage and networking infrastructures are needed so the AI systems can sustain the growth and eventually deliver the required insights to make business decisions.

The design goal for the IBM SpectrumAI with NVIDIA DGX converged solution is to reduce the complexity and time required to plan and execute growth for these AI stages (data prep, model training, and inference).

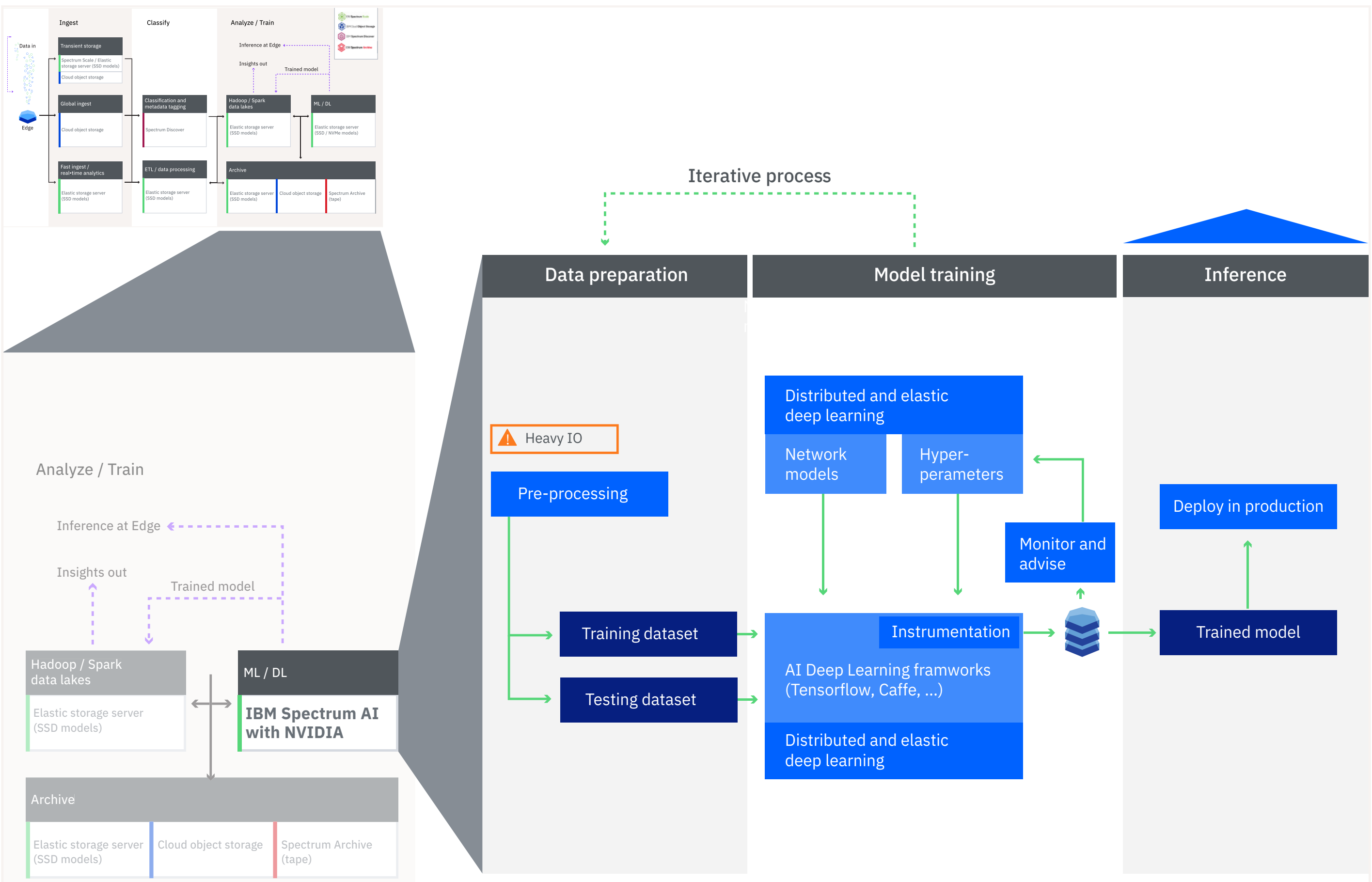


Figure 3: IBM SpectrumAI with NVIDIA ML/DL ANALYZE/TRAIN pipeline stages

Components of IBM SpectrumAI with NVIDIA converged solution

Figure 4: NVIDIA DGX-1 server

NVIDIA DGX-1 Server

The NVIDIA DGX family of supercomputers is a purpose-built fully integrated hardware and software solution for AI applications. The NVIDIA DGX POD™ implements data center design best practices for DL, which have been incorporated within this reference architecture.

Each DGX-1 server (**Figure 4**) integrates eight NVIDIA Tesla™ V100 Tensor Core GPUs configured in a hybrid cube mesh topology that uses NVIDIA NVLink™ technology. This fabric provides low-latency, high-bandwidth GPU-to-GPU communication that enables scalable multi-GPU training while eliminating the PCIe-based interconnect bottleneck found in traditional architectures that result in non-linearity of performance as more GPUs are used for training. The DGX-1 server includes four Mellanox VPI cards enabling EDR InfiniBand or 100 GbE network ports for multi-node clustering with high speed RDMA capability for exceptional storage-to-DGX server data rates.

The DGX-1 server is powered by the DGX software stack which includes the server operating system (DGX server OS) and optimized DL containers engineered by NVIDIA for maximized GPU-accelerated performance. This DGX software stack facilitates rapid development and deployment on a single DGX-1 server (or multiple DGX-1 servers), and multi-GPU and multi-system scale-up of applications on the DGX server platform saving time and developer effort. The software components offered with the NVIDIA DGX POD include cluster management and orchestration tools, and workload scheduling which can be leveraged for management of the architecture described in this paper.

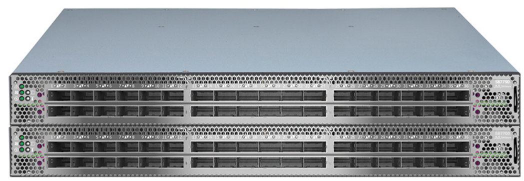


Figure 5: Mellanox SB7700 Series

Mellanox InfiniBand Network

For this reference architecture, the IBM Spectrum Scale on NVMe storage is attached to the DGX-1 servers by a Mellanox® EDR InfiniBand network to provide scalability of the GPU workloads and datasets beyond a single DGX-1 server while providing the inter-node communication between DGX-1 servers. Mellanox is an industry-leading supplier of switches, cables, and network adapters for Ethernet and InfiniBand. Delivering a complete, end-to-end interconnect solutions for high performance GPU clusters used for AI workloads, leveraging advanced technologies such as RDMA, GPUDirect® technology to accelerate GPU-to-GPU communications and SHARP™ technology to accelerate machine learning algorithms.

IBM SpectrumAI all-flash storage

The Non-Volatile Memory Express (NVMe) flash storage component of IBM SpectrumAI with NVIDIA DGX tested during this benchmark is running IBM's software-defined file storage offering – IBM Spectrum Scale. It available as software only, on the public cloud and in multiple IBM solutions, including the NVMe all-flash appliance (2019) and IBM Elastic Storage Server (ESS). The IBM Spectrum Scale ESS is a family of pre-integrated storage hardware offerings designed for highly available, high-performance storage. The IBM SpectrumAI with NVIDIA DGX converged solution can be composed using any of the IBM ESS all-flash models or the upcoming NVMe all-flash appliance (**Figure 6**).

Figure 6: IBM Spectrum Scale NVMe all-flash appliance

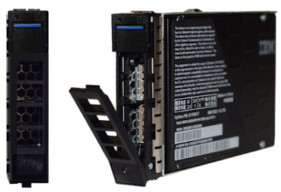


Figure 7: IBM NVMe FlashCore Module

IBM Spectrum Scale NVMe all-flash appliance

IBM Spectrum Scale NVMe all-flash appliance combines the performance of flash and NVMe end-to-end with the reliability and innovation of IBM FlashCore® technology (**Figure 7**) and the rich features of IBM Spectrum Scale, along with several high-speed attachment options such as 100 Gb/s IB— all in a powerful 2U storage system. IBM Spectrum Scale on NVMe is designed to be the market leader in all-flash performance, and scalability with a bandwidth of around 40 GB/s per NVMe all-flash appliance, that can scale-out with practically linear performance and100 microseconds latency. Providing data-driven multicloud storage capacity, the NVMe all-flash appliance is deeply integrated with the software defined capabilities of IBM Spectrum Storage™ allowing you to plug it into your AI data pipeline easily.

NVMe all-flash appliance supports a variety of connectivity options as well as the IBM NVMe FlashCore Module (FCM) and other industry standard NVMe flash drive options. The NVMe all-flash appliance is well designed to address the full range of AI application workloads and business use cases.

IBM Spectrum Scale and IBM Spectrum Scale RAID

IBM Spectrum Scale is an industry leader in high performance parallel file system software thousands of installations globally in practically every industry It also supports the world's #1 and #2 fastest super-computers, named Summit and Sierra. "Summit will have the capacity of 30B files and 30B directories and will be able to create files at a rate of over 2.6 million I/O file operations per second. That is opening every book in the US Library of Congress in 10 seconds."[4]

A key ability that IBM Spectrum Scale provides is a single namespace (or data plane) so that each data source can add data to the repository using NFS, SMB, Object, or a POSIX interface. This single data plane allows the data prep tools to access the data in place – no copying required. Training tools can access the data in place also, as can the inference applications, all with no copying, all through industry standard interfaces.

Another key strength is Spectrum Scale enables data to be tiered automatically and transparently to and from more cost-effective storage, including hard disk drive (HDD), tape, and cloud. As shown previously in figures 1-3, IBM provides full AI pipeline support.

IBM Spectrum Scale RAID is a software implementation of storage erasure code technologies within IBM Spectrum Scale that provides sophisticated data placement and error-correction algorithms to deliver high levels of storage performance, availability, and reliability. Shared file systems are created from the Network Shared Disks (NSD) defined with IBM Spectrum Scale RAID. This file system can be accessed concurrently by all compute nodes in the configuration to efficiently meet the capacity and performance requirements of modern scale out applications such as AI.

Along with the most proven scalable capabilities to handle enterprise workloads, IBM Spectrum Scale RAID delivers enhanced data protection and support to run GPU processor-intensive AI workloads at the performance levels required for multi-system GPU clusters such as the DGX POD.

[4] https://www.ibm.com/blogs/systems/fastest-storage-fastest-system-summit

Scaling with GPUs

The IBM SpectrumAI with NVIDIA DGX architecture provides scalability for capacity and performance. Deployments can start with a single IBM Spectrum Scale NVMe all-flash appliance and a single DGX-1 server or multiples of each as workloads demand, all while providing a cost-effective solution.

Figure 8 shows some of the potential IBM Spectrum Scale NVMe all-flash appliance configurations in DGX POD racks. Each configuration delivers balanced performnce, capacity and scalability for expansion of existing production workflows and additional DL applications. IBM Spectrum Scale NVMe all-flash appliance is power efficient to allow maximum flexibility when designing rack space and addressing power requirements.

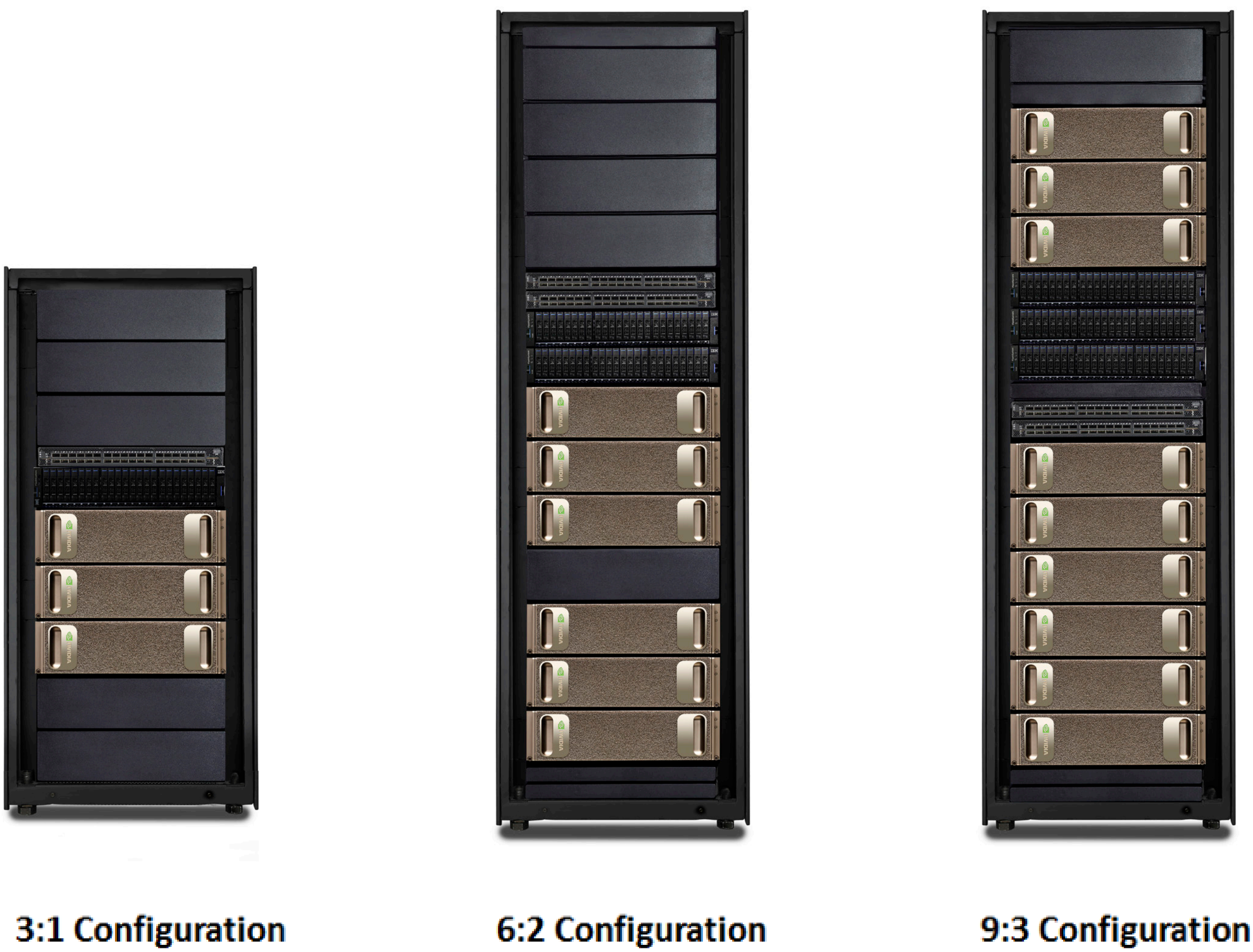


Figure 8: Rack elevations for scaling of IBM Secrum Scale NVMe all-flash appliance with NVIDIA DGX-1 servers (3:1, 6:2, 9:3 configurations – illustrative only)

1 NVIDIA DGX-1 server

2 Mellanox EDR IB network switch

3 IBM Spectrum Scale NVMe all-flash appliance

Figure 9 illustrates one DGX-1 server to one Spectrum Scale NVME all-flash appliance configuration example.

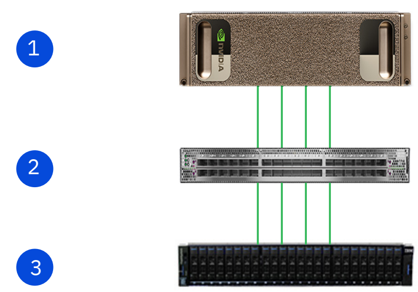


Figure 9: IBM SpectrumAI with NVIDIA in a 1 DGX-1 Server + 1 Spectrum Scale NVMe all-flash appliance configuration

Figure 10 illustrates four DGX-1 servers in a IBM two Spectrum Scale NVMe all-flash appliance unit configuration example to show another potential architecture option with either two switches as shown, or one switch if desired. Note this is a flat network topology example showing two switch to switch links but could require up to eight links. The sizing of the number of links required between switchs is dependent on the anticipated bandwidth the AI workload will drive through each network port and should be carefully considered in the overall network design. Other network topologies such as fat tree should also be considered when sizing new or scaling up existing infrastructure performance beyond what a flat network topology provides.

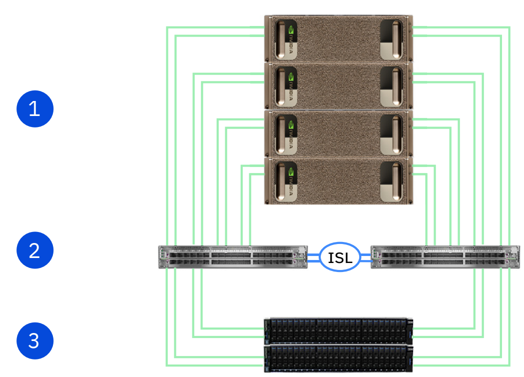


Figure 10: IBM SpectrumAI with NVIDIA in a 4 DGX-1 Server + 2 Spectrum Scale NVMe all-flash appliance configuration

Figure 11 depicts a full DGX POD configuration with three Spectrum Scale NVMe all-flash appliance 2U units. Here is the full reference architecture used in the IBM Spectrum Scale NVMe all-flash appliance benchmark, however only two IBM Spectrum Scale NVMe all-flash appliance units were necessary to obtain the model images/sec results reported in this paper.

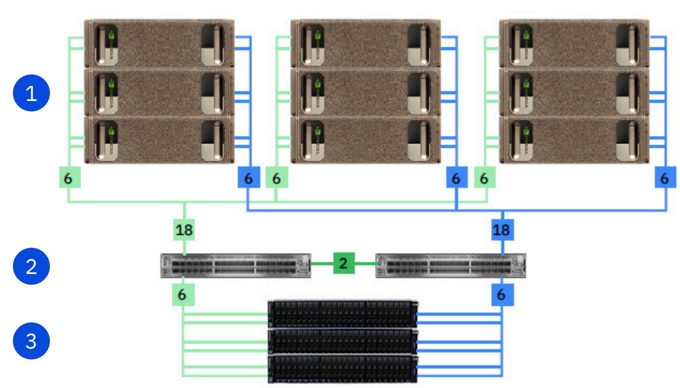
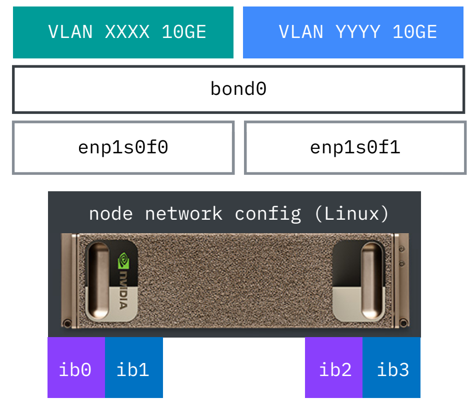


Figure 11: IBM SpectrumAI with NVIDIA in a 9 DGX-1 Server + 3 Spectrum Scale NVMe all-flash appliance configuration

IBM SpectrumAI with NVIDIA DGX solution reference architecture

Figure 12 illustrates the end-to-end DL reference architecture of IBM SpectrumAI with NVIDIA DGX-1 server solution environment.

In this cutting-edge test environment, IBM Spectrum Scale RAID v5 is installed on the NVME all-flash appliance base Linux OS. The IBM Spectrum Scale RAID software is generally available as part of the Spectrum Scale software stack for IBM Elastic Storage Server (ESS) deployments. As configured, each Spectrum Scale NVME all-flash appliance provides a pair of fully redundant NSD servers within the Spectrum Scale cluster. The Spectrum Scale NVMe all-flash appliance is connected over EDR InfiniBand with eight links to a Mellanox SB7800 fabric interconnect switch. The DGX-1 servers also connect with four links to the EDR InfiniBand switch. In addition to the high speed EDR InfiniBand fabric, the Spectrum Scale NVMe all-flash appliance, and DGX-1 servers are connected by a 10GbE management network to Ethernet switches.



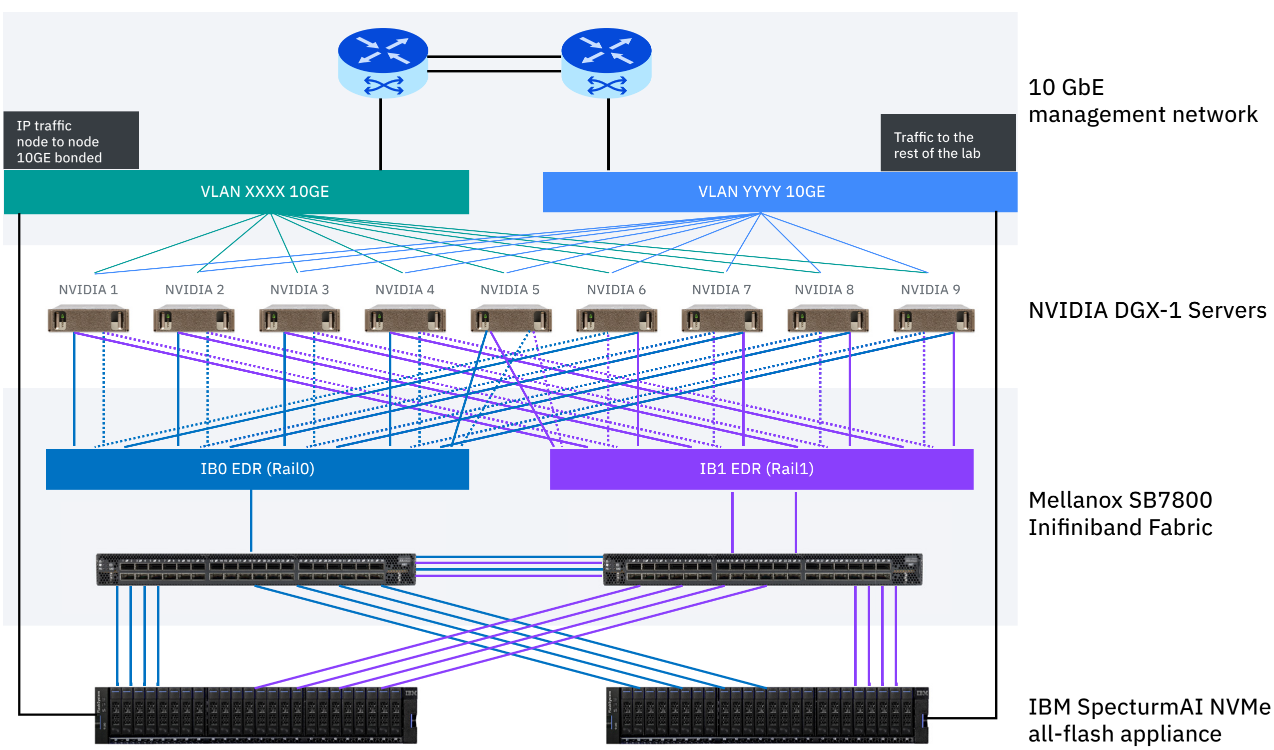


Figure 12: Solution architecture diagram

Architcture

Sotware

- IBM Spectrum Scale RAID (v5.3)
- Ubuntu 16.04
- Docker 18.09-py3
- NVIDIA AI OS
- NVIDIA GPU Cloud (container registry)
- TensorFlow 1.10
- Precision: FP16

Hardware

- IBM NVMe all-flash appliance
- Mellanox SB7800 IB Switch
- NVIDIA DGX-1 Server

Network

- EDR Infiniband
- 10 GB Ethernet (management)

Storage and network performance testing

Our testing was primarily focused on providing system performance capabilities to help DL development teams plan their infrastructure effectively. We ran benchmarks tests on DL model training performance and DL model inference performance, as wel as the incremental and total throughput capabilities when scaling up from one DGX-1 server to a full DGX POD. The IBM Spectrum Scale NVMe all-flash appliance performance was tested with synthetic throughput test applications such as IOR and fio and with the DL framework TensorFlow using several models such as ResNet-50, ResNet-152, Inception-v3, AlexNet, and other networks with the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) dataset.

The testing utilized NVIDIA NGC containers. Each AI container has the NVIDIA GPU Cloud Software Stack, a pre-integrated set of GPU-accelerated software. The stack includes the chosen application or framework, NVIDIA CUDA Toolkit, NVIDIA DL libraries, and a Linux OS — all tested and tuned to work together with no additional setup.

For developers, the NVIDIA Deep Learning SDK offers powerful tools and libraries for the development of DL frameworks such as Caffe2, Cognitive toolkit, MXNet, PyTorch, TensorFlow, and others. These frameworks rely on GPU-accelerated libraries such as cuDNN and NCCL to deliver high-performance multi-GPU accelerated training.

Developers, researchers, and data scientists can get easy access to NVIDIA optimized DL framework containers, performance tuned and tested for NVIDIA GPUs. This eliminates the need to manage packages and dependencies or build DL frameworks from source. Visit [NVIDIA GPU Cloud (NGC)] to learn more and get started.

TensorFlow benchmarks were performed using scripts made available by TensorFlow on GitHub.[3]

System throughput results

The total system throughput performance for one to nine DGX-1 servers with one to three IBM Spectrum Scale NVMe all-flash appliances shows the NVMe appliance performance scales as we add IBM Stpectrum Scale appliances. The total data throughput easily meets the benchmark testing demands to keep all DGX-1 server GPUs saturated. The results of the 4k random read tests using fio demonstrate that the IBM Spectrum Scale on NVMe solution maximizes the potential throughput of the data infrastructure scaling linearly from around 40 GB/s read performance for one Spectrum Scale NVMe all-flash appliance 2U unit, and scaling linearly with multiple Spectrum Scale NVMe all-flash appliance units.

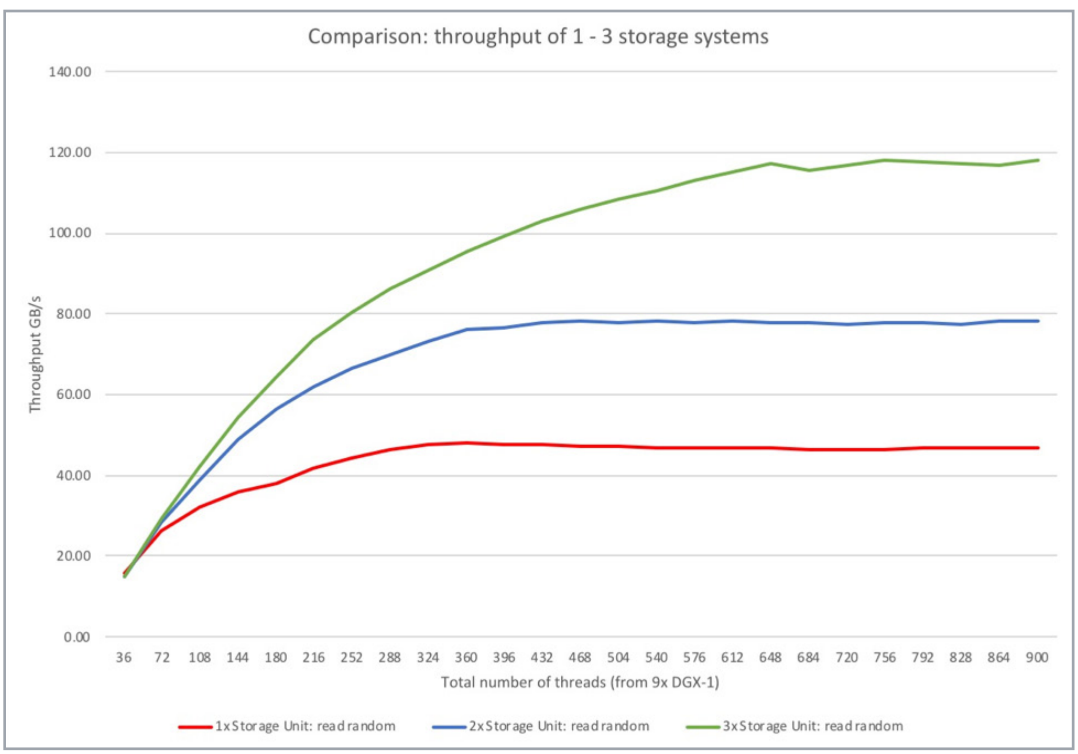


Figure 13: System scalable throughput using fio benchmark

[3]https://github.com/tensorflow/benchmarks/tree/master/scripts/tf_cnn_benchmarks

Placing the single srver GPU utilization and the storage throughput on the same graph illustrates the NVMe flash storage can efficiently feed the DGX POD servers (**Figure 14**). When running ResNet-50 GPU utilization is nearly 100% average for the DL workload while the overall bandwidth demands for a single DGX-1 server are not taxing the Spectrum Scale NVMe appliances. Extrapolating storage throughput capabilities and requirements from both charts below show that a single storage appliance unit as tested can handle the model training workloads presented in this paper.

To demonstrate the flexibility of the NVMe all-flash storage solution additional throughput tests were run for sequential versus random IO access patterns (**Figure 15**). Sequential read performance versus random read performance with the same file sizes shows some prefetch advantage which fades when the number of job threads increases. A fully utilized rack of NVIDIA DGX-1 servers will have many workloads running across the 72 GPUs. Therefore, sustained random read performance of the storage can be more important than peak sequential data bandwidth in evaluating storage peformance for DL. The NVMe all-flash appliance shows

Overall, the IBM Spectrum Scale NVMe all-flash storage appliance with DGX-1 servers throughput results show that IBM Spectrum Scale on NVMe storage performs well with one to nine DGX-1 servers making full use of the GPUs. This performance capability provides development teams the ability to add compute resources when needed knowing the storage system performance can accommodate their increasing workload demands and, if needed, by seamlessly adding additional NVMe all-flash appliance building blocks into the AI storage cluster.

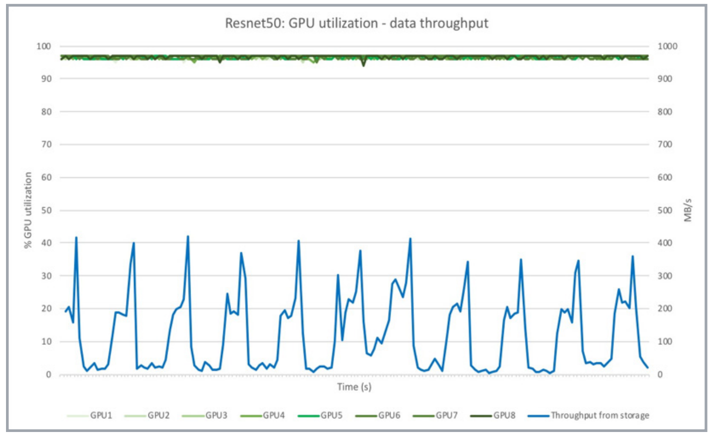


Figure 14: DGX-1 Server GPU utilization versus IO Bandwidth

Figure 15: Sequential vs Random Read throughput with fio benchmark

Training results – single DGX-1 server

The figure below shows the images per second training throughput with common models, which we tested with the publicly available ImageNet reference dataset. In this testing, we used AlexNet, ResNet-50, ResNet-152, Inception-v3, LeNet, Inception-v4, and GoogLeNet models across different numbers of GPUs on a single DGX-1 server with a separate container per GPU and comparing training runs between the IBM Spectrum Scale filesystem and local ramdisk only runs.

Model training results for the single DGX-1 server demonstrate minimal to no penalty when comparing IBM Spectrum Scale file system performance to internal system ramdisk performance for DL models with GPUs. Shared storage eliminates the need to copy the data to the local storage, decreasing the overall time to insights. As shown in **Figure 14**, a single IBM Spectrum Scale NVMe all-flash appliance effectively feeds the DGX-1 server GPUs, keeping them saturated with data to optimize results with shared storage.

As shown, some models scale up with linearity as the number of GPUs increase while others present a consistent non-linear scale up pattern whether using the IBM Spectrum Scale on NVMe storage or local ramdisk. This indicates that the scalability in these cases is not storage IO constrained whether local or shared storage, but rather a pattern of the DL model scalability within the compute infrastructure itself.

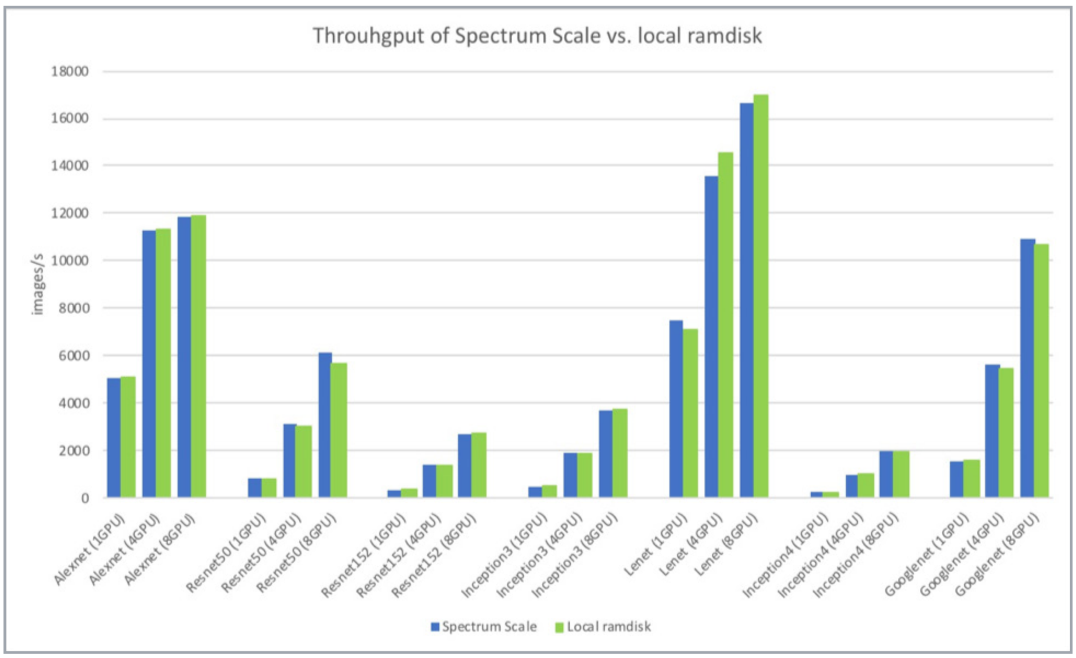


Figure 16: Single DGX-1 Server model to GPU performance and Spectrum Scale filesystem vs. ramdisk performance

Training results — multiple DGX-1 systems

For multiple DGX-1 servers with separate containers, IBM Spectrum Scale on NVMe architecture demonstrates linear scale up to full saturation of all DGX-1 server GPUs simultaneously running from one to nine DGX-1 servers for a total of 72 GPUs providing the following performance results of aggregate images/sec using ImageNet datasets for the models shown in **Figure 17** below.

The multiple DGX-1 server image processing rates shown above demonstrate IBM Spectrum Scale on NVMe scalability for training application performance with Inception-v4, ResNet-152, VGG-16, Inception-v3, ResNet-50, GoogLeNet, and AlexNet models using eight GPUs on each DGX-1 server. The Spectrum Scale NVMe all-flash appliance solution shows linear scale up when adding additional DGX-1 servers. In each case, testing begins with one DGX-1 server with eight GPUs ramping up to nine DGX-1 servers with consistent full saturation of GPUs for all 72 GPUs tested.

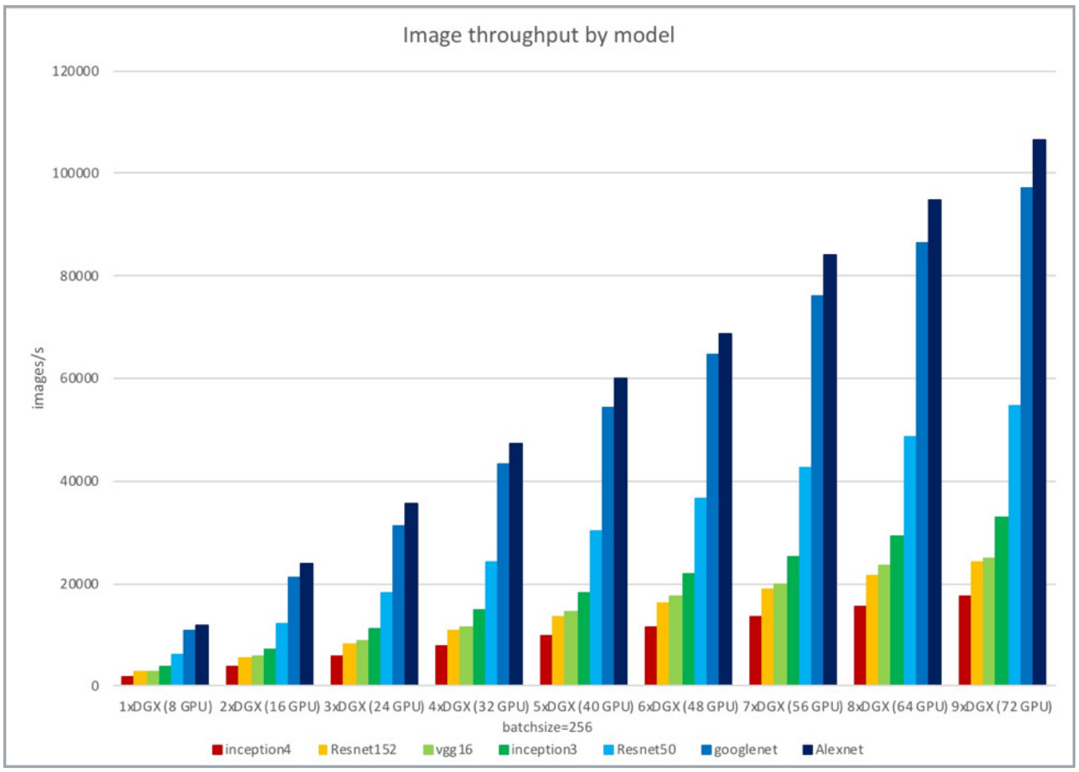


Figure 17: : Multiple DGX-1 Server training rates with TensorFlow models

Inference results – multiple DGX-1 systems

For multiple DGX-1 servers with separate containers, IBM Spectrum Scale on NVMe architecture demonstrates linear scale up to saturation of all the DGX-1 server GPUs simultaneously running from one to nine DGX-1 servers for a total of 72 GPUs. The following performance results of aggregate images/sec inference using ImageNet datasets for the models shown in **Figure 18**.

As tested, inference image processing rates are between 1.5x to almost 4x the training rates of the corresponding TensorFlow models. The DGX-1 server with Spectrum Scale on NVMe solution provides data scientists the ability to run in mixed training and inference mode on a single DGX-1 server as needed, dedicating one or two GPUs to inference and the remaining GPUs in the DGX-1 server to training jobs.

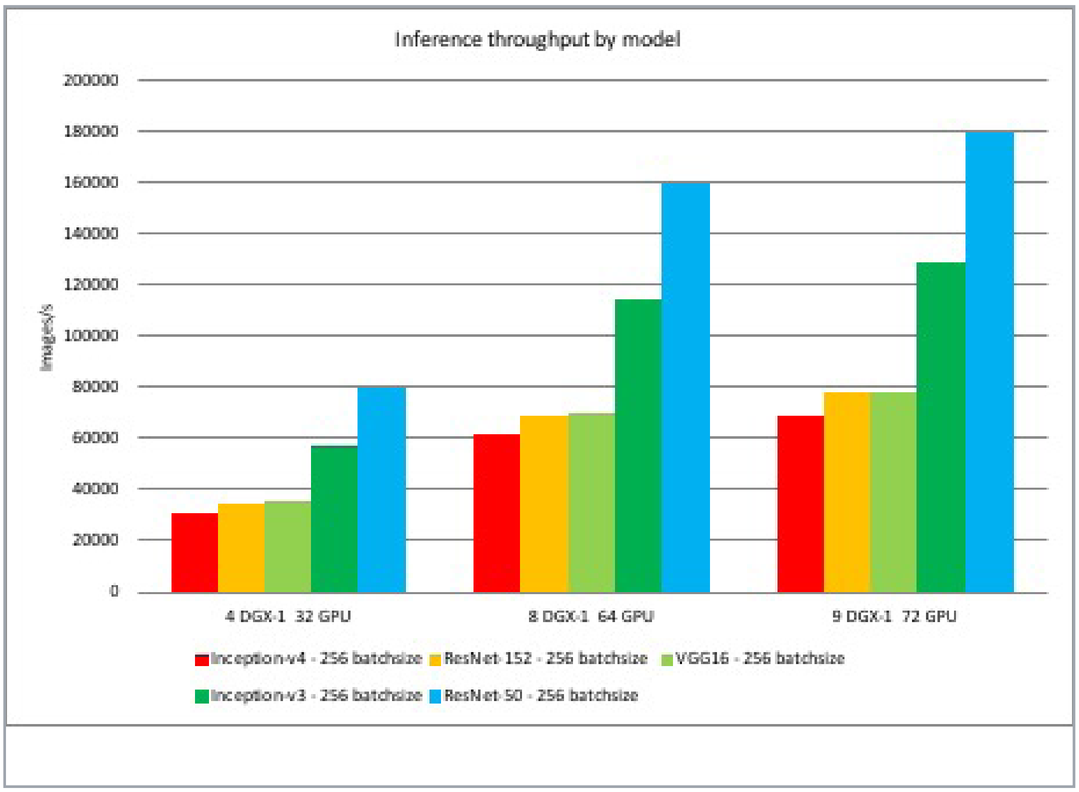


Figure 18: Multiple DGX-1 Server inference rates with TensorFlow models

Conclusion

The IBM SpectrumAI with NVIDIA DGX as tested delivers test results, data throughput and scaling utilizgin Spectrum Scale NVMe all-flash appliance architecture over a Mellanox EDR InfiniBand fabric. The convereged solution demonstrated capabilities to run multiple DL workload training and inference workloads. Proper sizing of the solution and network eabled high-performance parallel processing on multiple GPUs with high bandwidth and low latency storage. The result was full utilization of GPUs when running on multiple DGX-1 servers.

IBM SpectrumAI with NVIDIA is a tested convereged solution of the NVIDIA DGX-1 server integrated with the NVIDIA GPU Cloud Software Stack in combination with the IBM Spectrum family of software defined storage solutions. Because IBM Spectrum Scale is capable of scaling and tiering across multiple physical storage arrays and is extensible with IBM Cloud Object Storage, IBM Specrtrum Discover, and IBM Spectrum Archive, organizations can confidently start with a basic configurtaion and grow as needed. IBM SpectrumAI with NVIDIA DGX is the infrastructure to support workload consolidation, data preparation and management, and process automation to streamline the AI data pipeline, speed AI development and ease integration into existing infrastructure.

Additional resources

- Introduction Guide to The IBM Elastic Storage Server