Looking at IBM Spectrum Storage for AI with NVIDIA DGX performance

Abstract

This document covers the results of AI and throughput testing of IBM Spectrum Scale with next gen NVMe Arrays.

Recently IBM® participated in the development of a reference architecture with NVIDIA®. As part of the program, we benchmarked a number of common AI workloads as well as the performance of the reference architecture, IBM Spectrum Storage for AI with NVIDIA DGX. 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 array in 2019. This benchmark tested the IBM Spectrum Scale NVMe all-flash appliance storage near linear performance while scaling from one to nine DGX-1 servers with both synthetic workloads and TensorFlow models using ImageNet data.

Benchmark Tests

We tested deep learning (DL) model training performance as well as DL model inference performance and the incremental and total throughput capabilities when scaling up from one DGX-1 server to a full DGX POD with IBM Spectrum Scale NVMe all-flash. We ran performance evaluations with synthetic throughput test applications such as IOR and fio and then with the DL framework TensorFlow using several models such as ResNet-50, ResNet-152, Inception-v3, AlexNet, and other networks. All tests used the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) dataset. The benchmark we ran utilizes the NVIDIA GPU Cloud (NGC) containers. Each AI container has the NVIDIA GPU Cloud Software Stack, a pre-integrated set of GPU-accelerated software. The stack included the chosen application or framework, NVIDIA CUDA Toolkit, NVIDIA DL libraries, and a Linux OS — all tested and tuned to work together immediately with no additional setup.

For IT professionals tasked with supporting their growing data science teams, providing sufficient performance for model training is critical. Although several groups are attempting to codify scalable ML/DL testing suites, such as ML Perf, there isn't a benchmark suite that has found common acceptance to demonstrate the limits of storage for AI. SpecSFS benchmarks are not a direct comparison to AI workloads run today but can provide some indication of the top-end performance of maximized storage configurations. Comparisons are difficult because they lack relative comparison information such as price/performance or performance per rack unit. However, IBM Spectrum Scale has been used in the majority of SpecSFS benchmarks by IBM, our partners, and start-ups who want to showcase the highest performing lowest latency shared file systems.

AI data needs are different. It is generally accepted that the majority of data flow for storage during the model training phase is random read as the model loads across multiple GPUs. However, other patterns are emerging in IBM research and with our clients in production. Depending upon the data structures, metadata performance for traversing directories of small files can be critical. In another example, such as our client's genomics research, the AI training is actually in two phases. The first phase uses AI to develop a preliminary model, then writes out large data files to train a complementary model that uses the data classification of the first model to improve the model accuracy.

What is most important is the configuration built for training be performant end-to-end to feed the GPU accelerated servers. IBM tested both sequential and random read data throughput as part of our testing using FIO for both sequential and random reads.

What's the IBM Spectrum Storage for AI with NVIDIA DGX solution?

Reference architecture under test
IBM testing was on a fully populated IBM Spectrum Storage for AI with NVIDIA DGX reference architecture. We had 9 NVIDIA DGX-1 servers connected by Mellanox EDR Infiniband to three IBM next-generation NVMe arrays running IBM Spectrum Scale. Each DGX-1 server includes eight NVIDIA Tesla™ V100 Tensor Core GPUs. For more information on the configuration, see IBM Spectrum Storage for AI with NVIDIA DGX Reference Architecture.

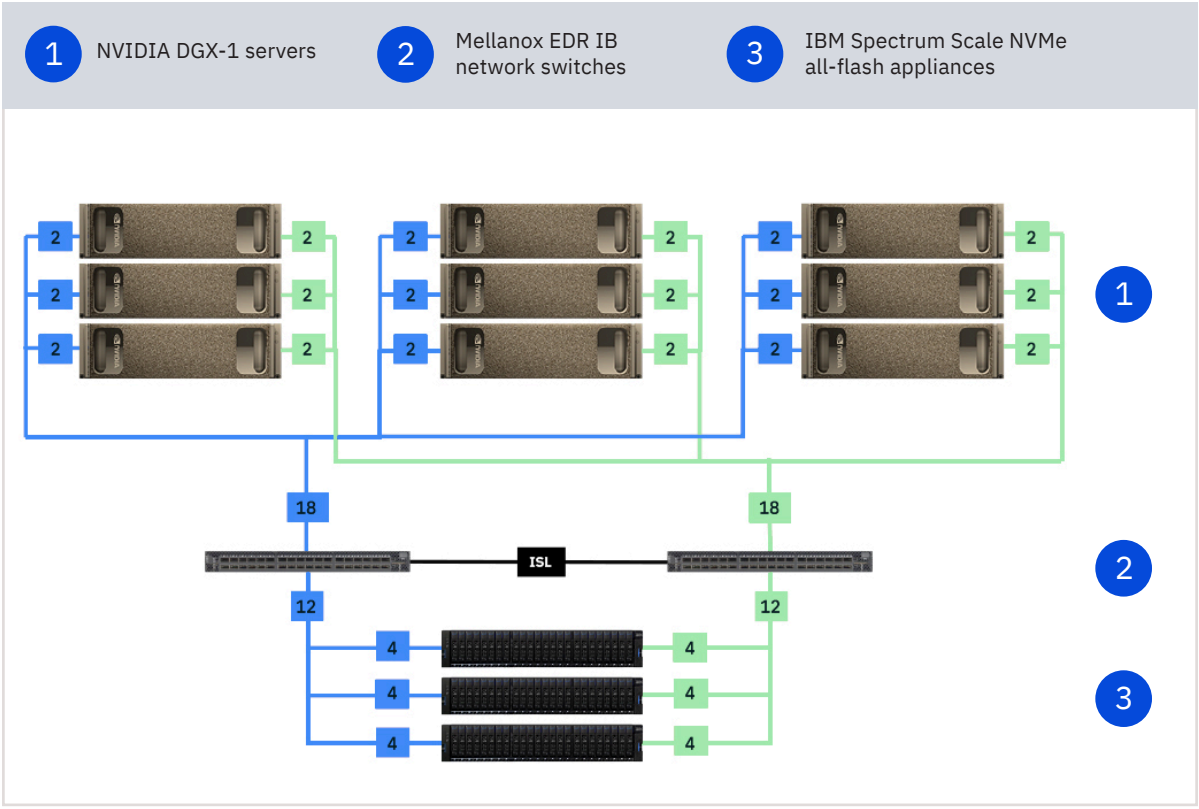

Figure 1: IBM Spectrum Storage for AI with NVIDIA in a 9 DGX-1 Server + 3 Spectrum Scale NVMe all-flash appliance configuration

IBM Spectrum Scale NVMe all-flash appliance deployment close up
In this cutting-edge test environment, we used IBM Spectrum Scale version 5. IBM Spectrum Scale RAID was installed on the NVMe all-flash array with the Linux OS. The IBM Spectrum Scale RAID software is generally available as part of the IBM Spectrum Scale software stack for IBM Elastic Storage Server (ESS) deployments. As configured, each IBM Spectrum Scale NVMe all-flash provided a pair of fully redundant Network Shared Disks (NSD) servers within the IBM Spectrum Scale cluster over the EDR IB network.

Storage and network performance results

We ran a series of tests designed to measure the total system throughput for one to nine DGX-1 servers with one to three IBM Spectrum Scale NVMe all-flash arrays for synthetic data and ImageNet data for DL image training and DL image inference. To qualify the storage performance characteristics, we tested random IO versus sequential IO access patterns. We also measured the performance impact of using IBM Spectrum Scale filesystem versus ramdisk on a single NVIDIA DGX-1 server.

System throughput results

We tested the total maximum read throughput while increasing the number of fio threads across 9 DGX-1 servers. **Figure 2** shows the NVMe performance scaled almost linearly from around 40 GB/s read performance for one IBM NVMe all-flash 2 rack unit array to around 120 GB/s with three NVMe all-flash arrays. In this configuration, the IBM Spectrum Scale solution demonstrated 4.5x more data throughput for a rack of DGX-1 systems than comparable solutions. The IBM NVMe all-flash solution demonstrated the ability to keep all DGX-1 server GPUs saturated for the DL benchmark tests.

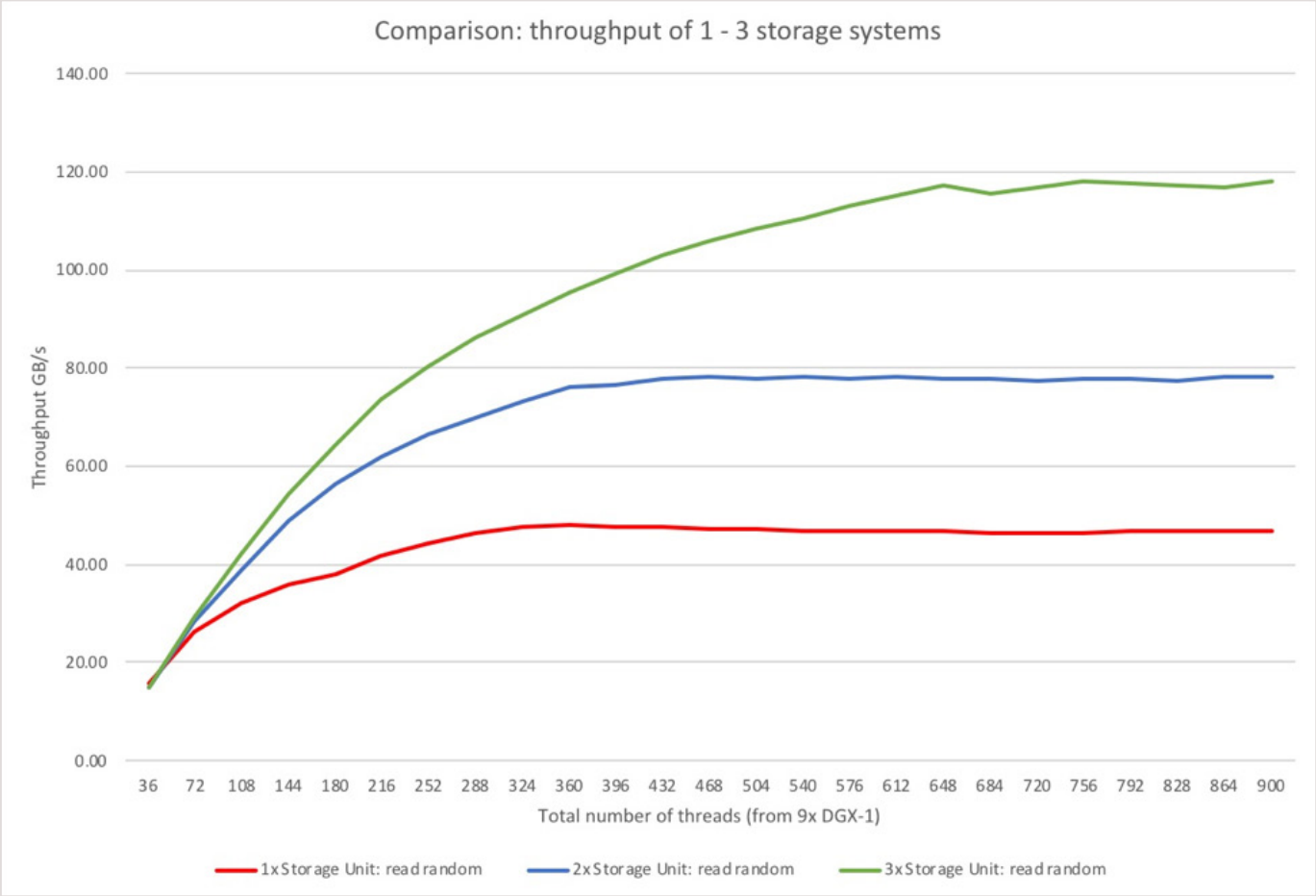

Figure 2: System scalable throughput using fio benchmark

We also ran IO pattern throughput tests to demonstrate the flexibility of the NVMe all-flash storage solution. Sequential read performance versus random read performance showed some prefetch advantage which faded as the number of job threads increased. The NVMe all-flash solution showed robust throughput capabilities regardless of the IO type.

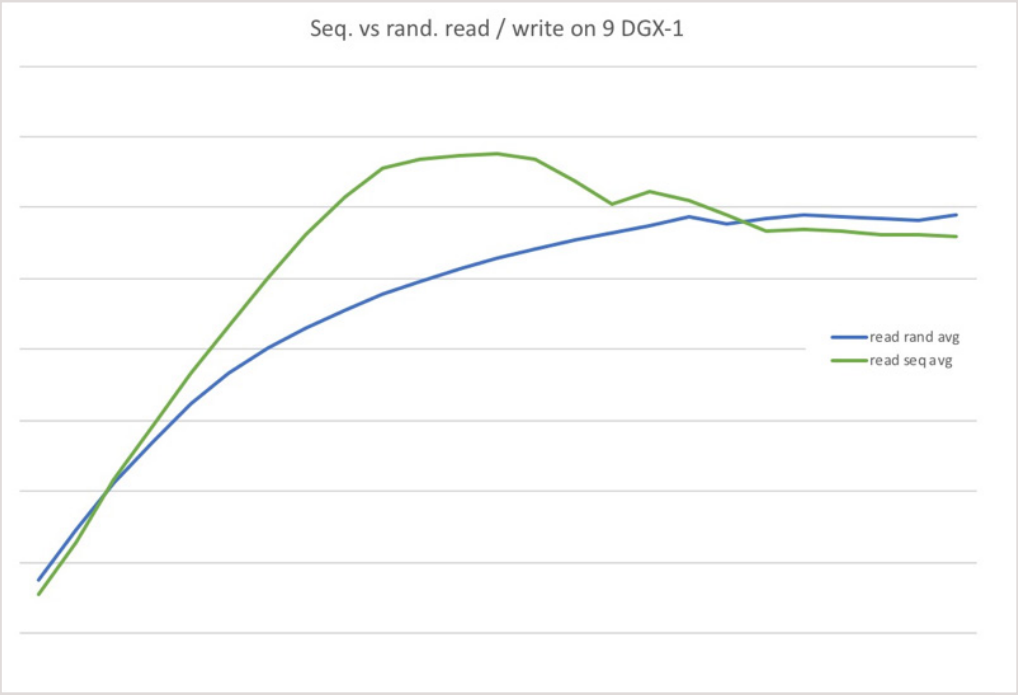

Figure 3: DGX-1 Server GPU utilization versus IO Bandwidth

Figure 3 shows near 100% GPU average utilization measured for the DL workloads compared to the overall bandwidth demands for a single DGX-1. A single server with 8 GPUs does not tax IBM Spectrum Scale on NVMe during the benchmark.

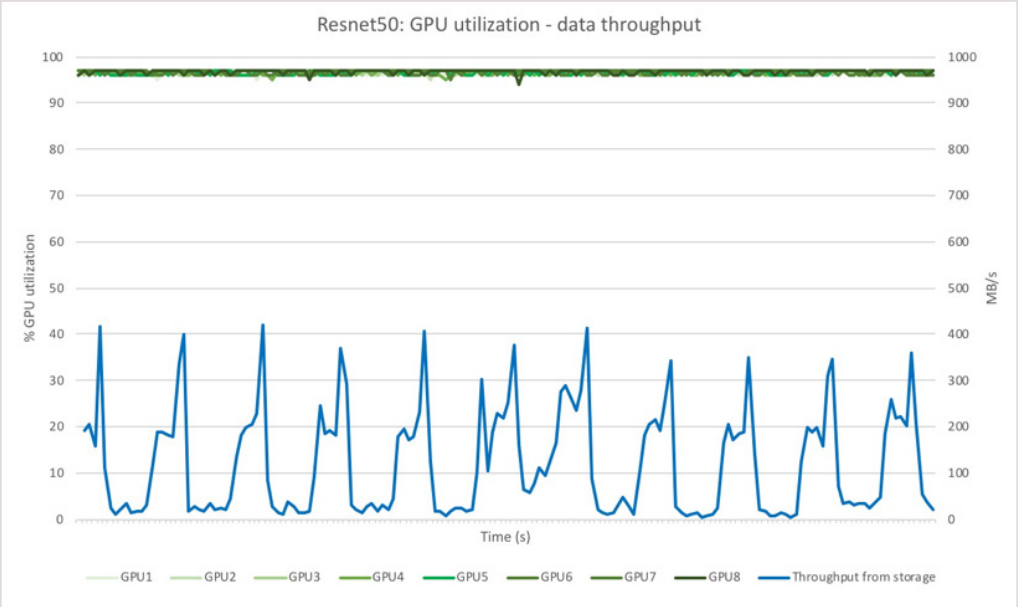

Figure 4: Sequential vs Random Read throughput with fio benchmark

ImageNet training results for a single DGX-1 server
We tested the images per second training throughput with different models using different numbers of GPUs on a single NVIDIA DGX-1 server with IBM Spectrum Scale and then with local ramdisk only. Comparing the performance of local, ramdisk storage to the IBM Spectrum Scale filesystem network storage, the performance of the applications tested was similar.

Note: The network storage was unmounted and remounted on each client to clear the cache before each test run. IBM Spectrum Scale Pagepool was set lower to 16 GB so it could only partially contain the ~140 GB test data sets. We also compared the NVMe device throughput to the NVIDIA DGX-1 server throughput to validate the cache had cleared.

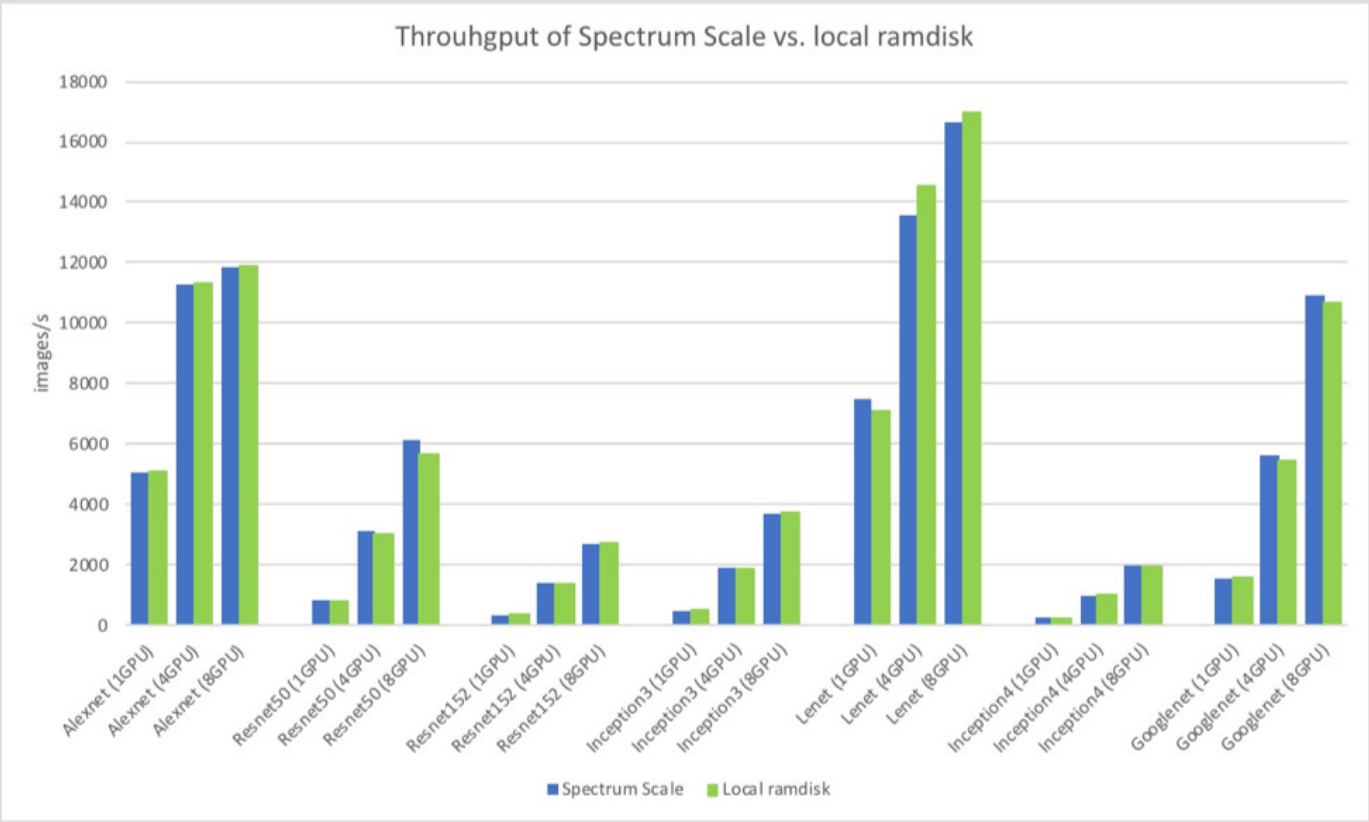

Figure 5: Single DGX-1 Server model to GPU performance and Spectrum Scale filesystem vs. ramdisk performance

Some models scaled up with near linearity as the number of GPUs increased while others presented a consistent non-linear scale up whether using the IBM Spectrum Scale on NVMe storage or local ramdisk. Indicating the scalability is not storage IO constrained whether local or shared storage, but rather a pattern of the DL model scalability within the compute infrastructure itself.

ImageNet Training Results for Multiple NVIDIA DGX-1 Servers
For multiple NVIDIA DGX-1 servers, IBM Spectrum Scale NVMe all-flash demonstrated near 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 using IBM Spectrum Scale.

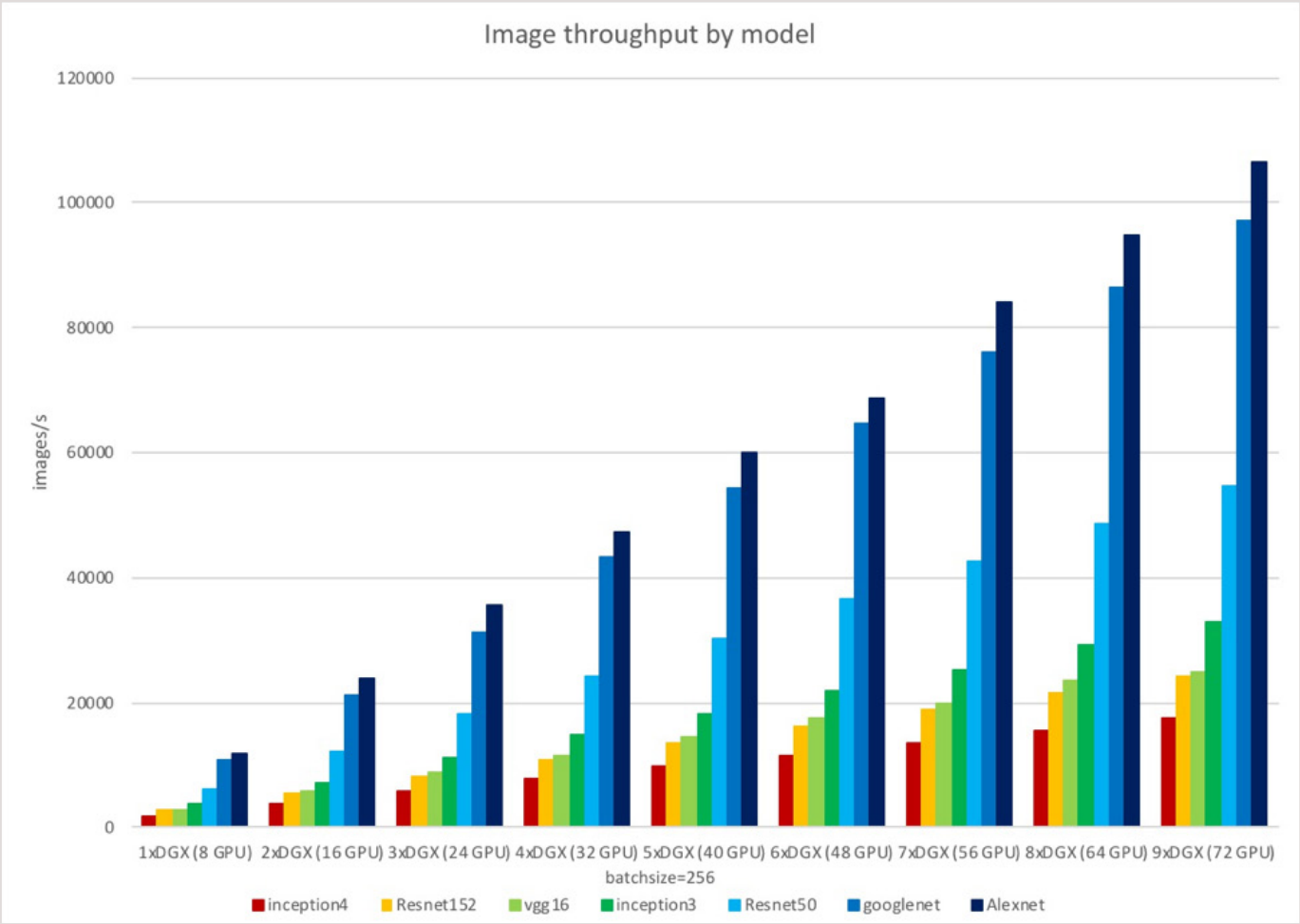

Figure 6 : Multiple DGX-1 Server training rates with TensorFlow models

ImageNet Inference Results for Multiple DGX-1 Servers

Due to benchmark time constraints, we tested the following DGX-1 server number of GPUs and model combinations as shown. Once again IBM Spectrum Scale on NVMe architecture demonstrated near linear scale up to full saturation of all the DGX-1 server GPUs simultaneously running up to a total of 72 GPUs.

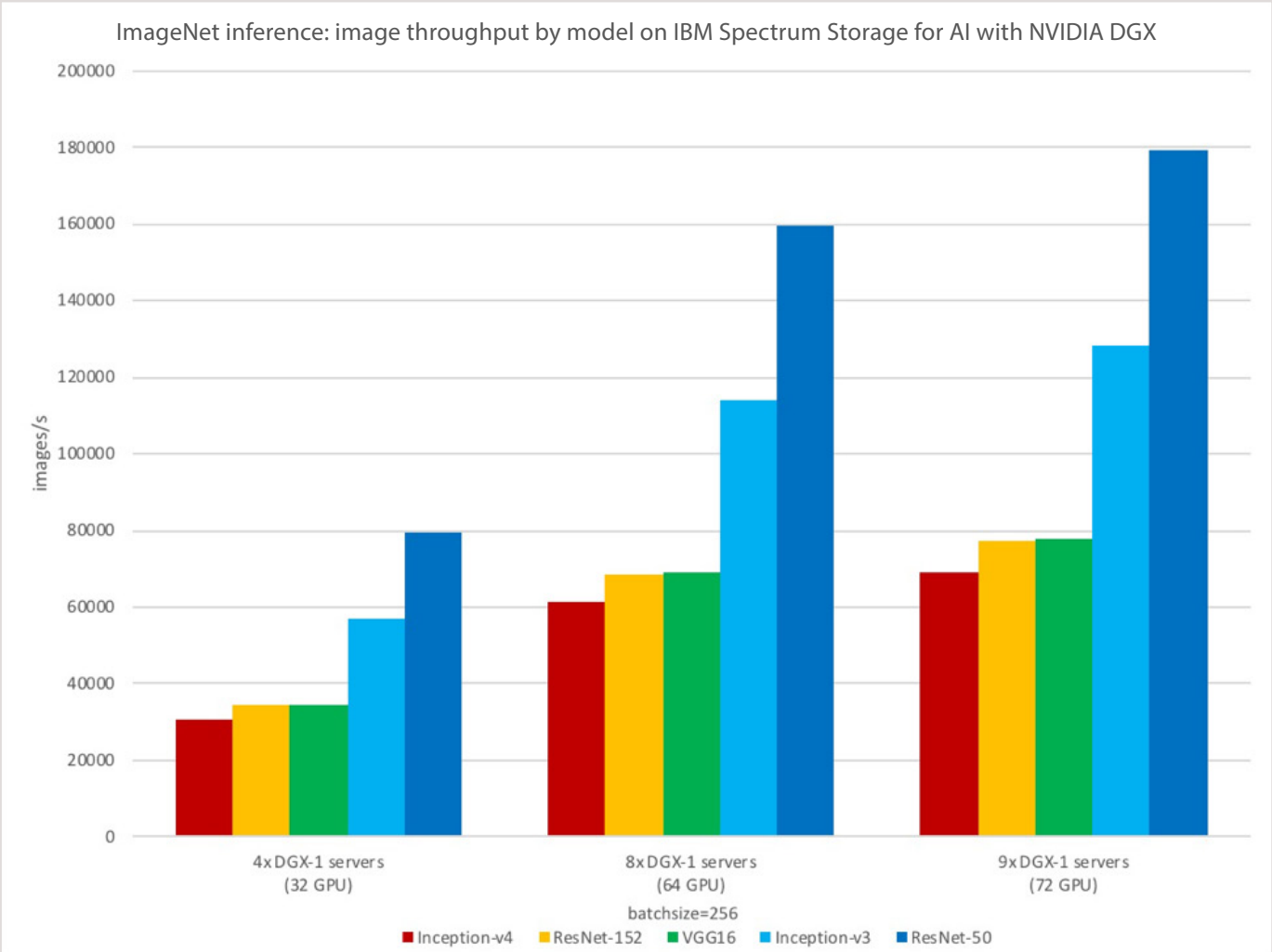

Figure 6: Multiple DGX-1 Server inference rates with TensorFlow models

As tested, inference image processing rates are between 1.5x to almost 4x the training rates of the corresponding TensorFlow models. Demonstrating that a DGX-1 server with IBM 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.

Conclusion

IBM storage offers best-in-class performance for AI solutions. There are multiple options for shared file storage in the Elastic Storage Server (ESS) line and the next generation NVMe platform with excellent performance for AI workloads. IBM Spectrum Scale software and solutions can deliver scalable storage throughput as needed for future data growth or compute needs. With performance to spare, IBM Spectrum Storage for AI with NVIDIA DGX is ready to support the NVIDIA ecosystem and the AI data pipeline that drives development productivity.

The IBM Spectrum Scale NVMe all-flash architecture over a Mellanox EDR InfiniBand fabric is the choice in providing leading-edge DL workload training and inference results while servicing high-performance parallel processing at high bandwidth and low latency on any GPU accelerated systems. The testing of NVIDIA DGX-1 servers with various AI training and inference models showed full utilization of GPUs.

Additional resources

For more information about IBM Storage for AI and IBM Spectrum Storage for AI with NVIDIA DGX, please visit: www.ibm.com/it-infrastructure/storage/ai-infrastructure

For more information about other IBM Systems solutions for AI, including IBM Storage, IBM AC922, IBM PowerAI Enterprise, and IBM Spectrum Computing, please visit:

- IBM AI Infrastructure Solutions
- IBM Systems AI Infrastructure Reference Architecture
- IBM Spectrum Computing for AI