Simplifying Deep Learning Infrastructure with Dell EMC

Boni Bruno, CISSP, CISM, CGEIT
Chief Solutions Architect

TRADITIONAL CLUSTER

Workload Profile:

Financial Services Company

192TB data set

16 years, 68 quarters

34.7 Million Customers

1.85 Billion performance records

XGBoost training set: 50 features

300 Servers | $3M | 180 kW

Dell EMC Solution

1/6 the Cost

1/14 the Space

1/9 the Power

Integrated Deep Learning Platforms with Isilon

The Buy Options

Ready Solution for AI:
Deep Learning with Intel

- Intel Nauta Platform
- Evolves into a fully integrated software environment for optimizing the complete range of Intel based processors and acceleration hardware

Ready Solution for AI:
Deep Learning with NVIDIA

- Dell EMC Compute with embedded GPUs, GPU-aware cluster management and wide set of deep learning container libraries
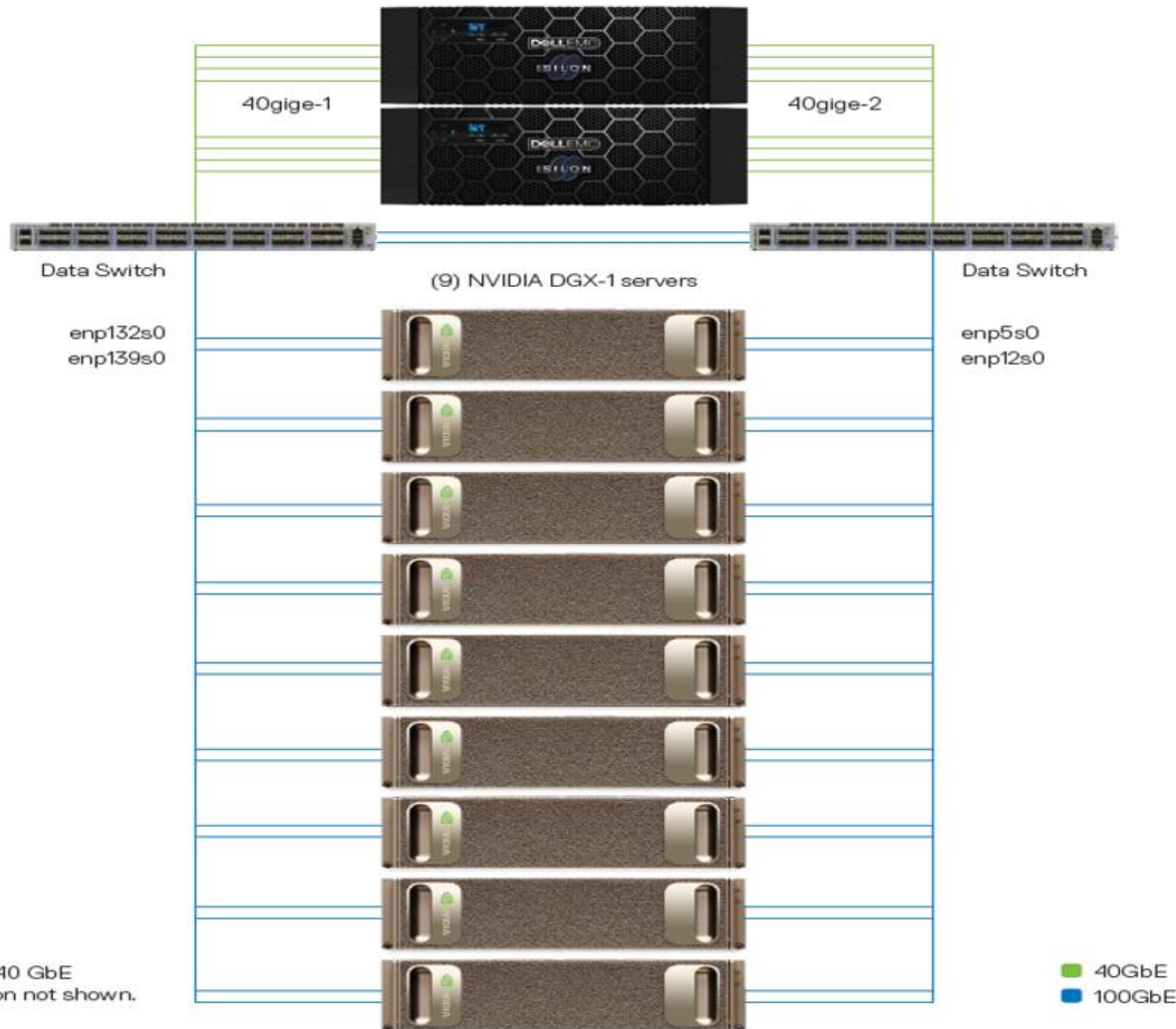- Specifically for deep learning model training.

Reference architecture
Isilon with NVIDIA DGX-1

- DGX-1 appliance form factor with NGC container libraries
- Specifically for workloads that need 8-way GPUs with high-speed interconnect (ex. High dimensionality data sets)

Isilon is the foundation for AI platforms

DELL EMC

(8) Isilon F800 nodes in (2) Isilon Chassis

40gige-1

40gige-2

Data Switch

(9) NVIDIA DGX-1 servers

Data Switch

enp132s0
enp139s0

enp5s0
enp12s0

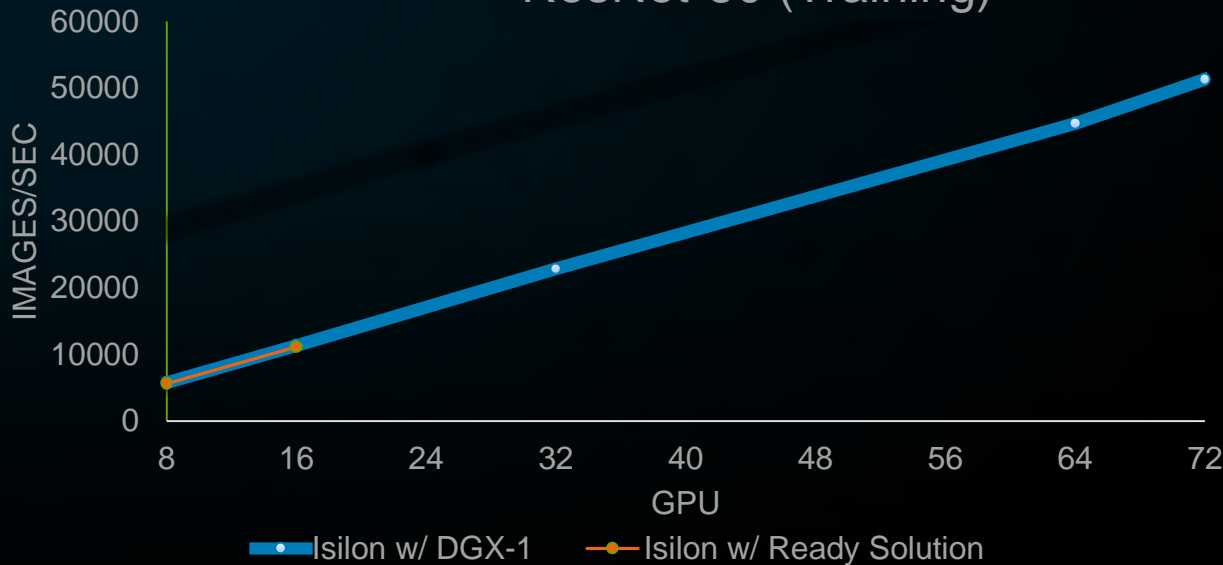**Note:** Backend 40 GbE switches for Isilon not shown.

■ 40GbE
■ 100GbE

Reference Architecture

DELLEMC

Isilon with NVIDIA GPUs Benchmark

Image detection with Tensorflow and 22 TB dataset
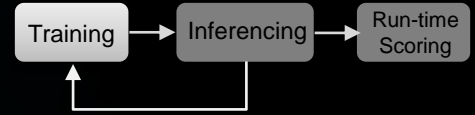
ResNet-50 (Training)

Highlights

- **Record performance:** 96% or more of theoretical max

- **Linear Scaling:** From 8 to 72 GPUs

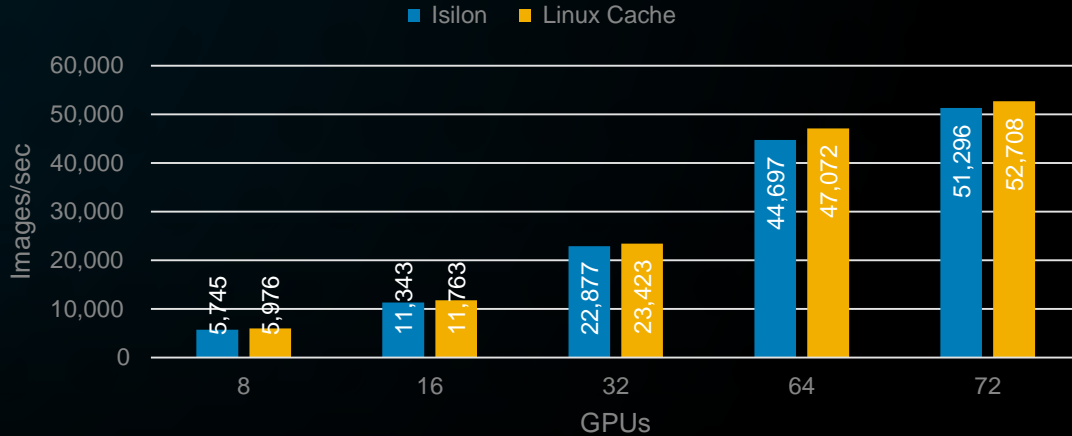- Only vender to publish above 32 GPUs

Not STAC Benchmarks

DELLEMC

Isilon with DGX-1: Benchmark Results

Training: Image Classification with TensorFlow and ImageNet Data Set

Training → Inferencing → Run-time Scoring

ResNet-50

■ Isilon ■ Linux Cache

GPUs	Isilon	Linux Cache
8	5,745	5,976
16	11,343	11,763
32	22,877	23,423
64	44,697	47,072
72	51,296	52,708

Images/sec vs GPUs

Training

- 97% GPU utilization or higher
- 96% of local memory throughput with Isilon
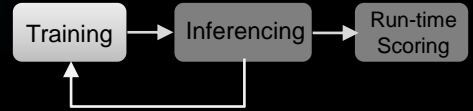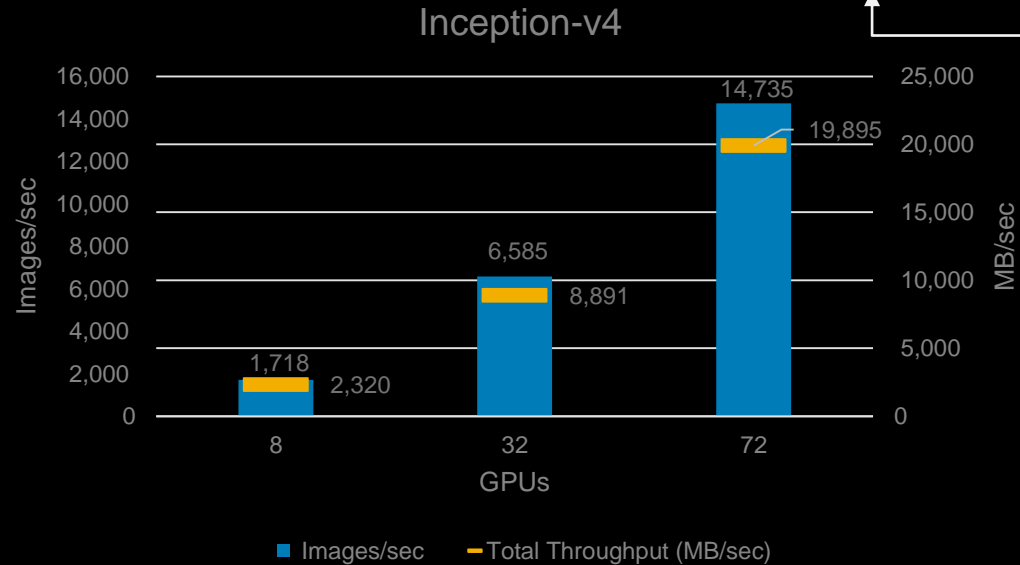- Linear Scaling from 8 to 32 to 72 GPUs

Not STAC Benchmarks

DELLEMC

Isilon with DGX-1: Benchmark Results

Training: Large Image Classification with TensorFlow and ImageNet

Training → Inferencing → Run-time Scoring

Inception-v4

Chart — Images/sec (left axis) and MB/sec (right axis) vs GPUs:

GPUs	Images/sec	Total Throughput (MB/sec)
8	1,718	2,320
32	6,585	8,891
72	14,735	19,895

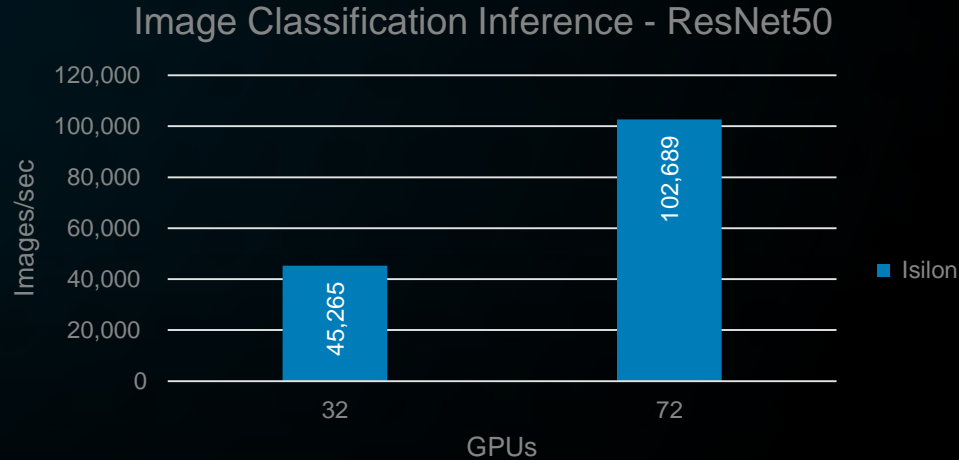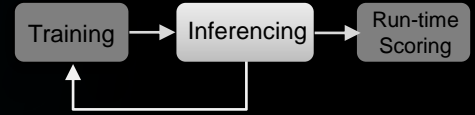■ Images/sec — Total Throughput (MB/sec)

Training

- 98% GPU utilization
- Linear Scaling from 8 to 32 to 72 GPUs
- 19.9 GB/s throughput with 72 GPUs

Not STAC Benchmarks

DELL EMC

Isilon with DGX-1: Benchmark results

Inferencing: Image Classification with TensorFlow and ImageNet Data Set

Training → Inferencing → Run-time Scoring

Image Classification Inference - ResNet50

Images/sec vs GPUs

- 32 GPUs: 45,265
- 72 GPUs: 102,689

■ Isilon

Inferencing

- 100% of local memory throughput with Isilon
- Linear Scaling from 32 to 72 GPUs

Not STAC Benchmarks

DELL EMC

Isilon eliminates the I/O bottleneck for AI at any scale

All Flash

Ultra Dense
72TBs up to 924TBs in 4U

Seamless Scale to Over 58PBs
In a Single File System

No Special Storage Drivers Needed
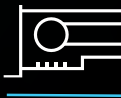Simply use NFS for data access

Faster training and validation of AI models

Higher model accuracy

Improve data science productivity

Maximize ROI of compute investments

DELL EMC

Additional Reading Material…

- [Tick Data Analytics – Scaling Concurrency and I/O Performance](#)

- [Deep Learning with Dell EMC Isilon](#)

- [Digital Banking](#)

DELLEMC

Thank You

DELLEMC