# S91016 AI Growing Pains: Platform considerations for moving from POCs to Large Scale Deployments


Sai Devulapalli

Global Head, Data Analytics Platform Portfolio

Unstructured Data Solutions

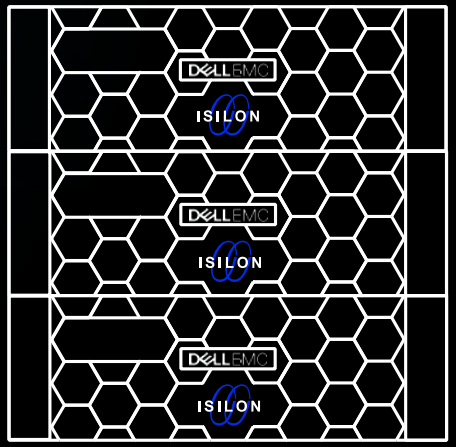
www.linkedin.com/in/saidevulapalli

Claudio Fahey

Principal Architect, Data Analytics and AI Platforms

Unstructured Data Solutions

www.linkedin.com/in/claudiofahey


DELL EMC

# It’s all about extracting value from your data

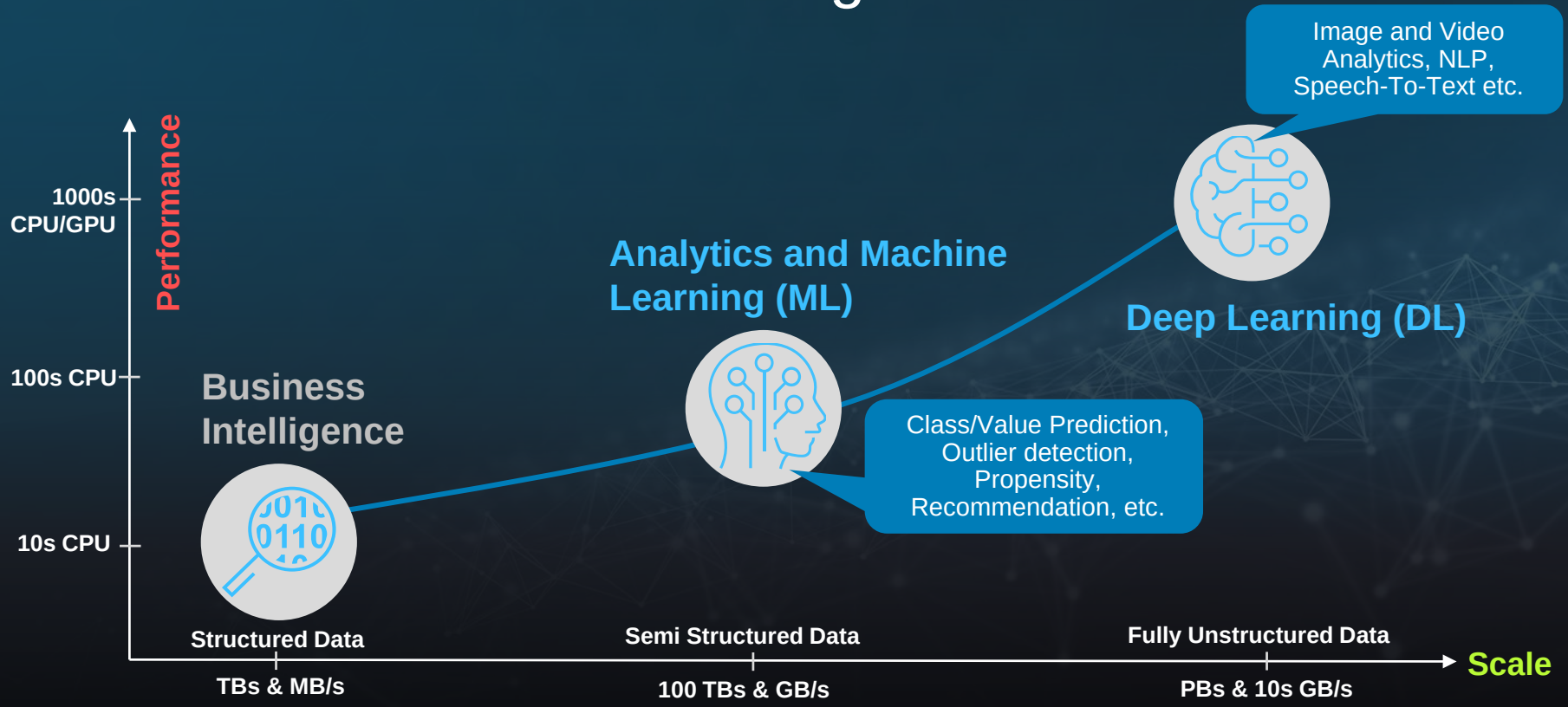
## TRADITIONAL ASSETS

- Human Capital
- Intellectual Property
- Operations
- Infrastructure

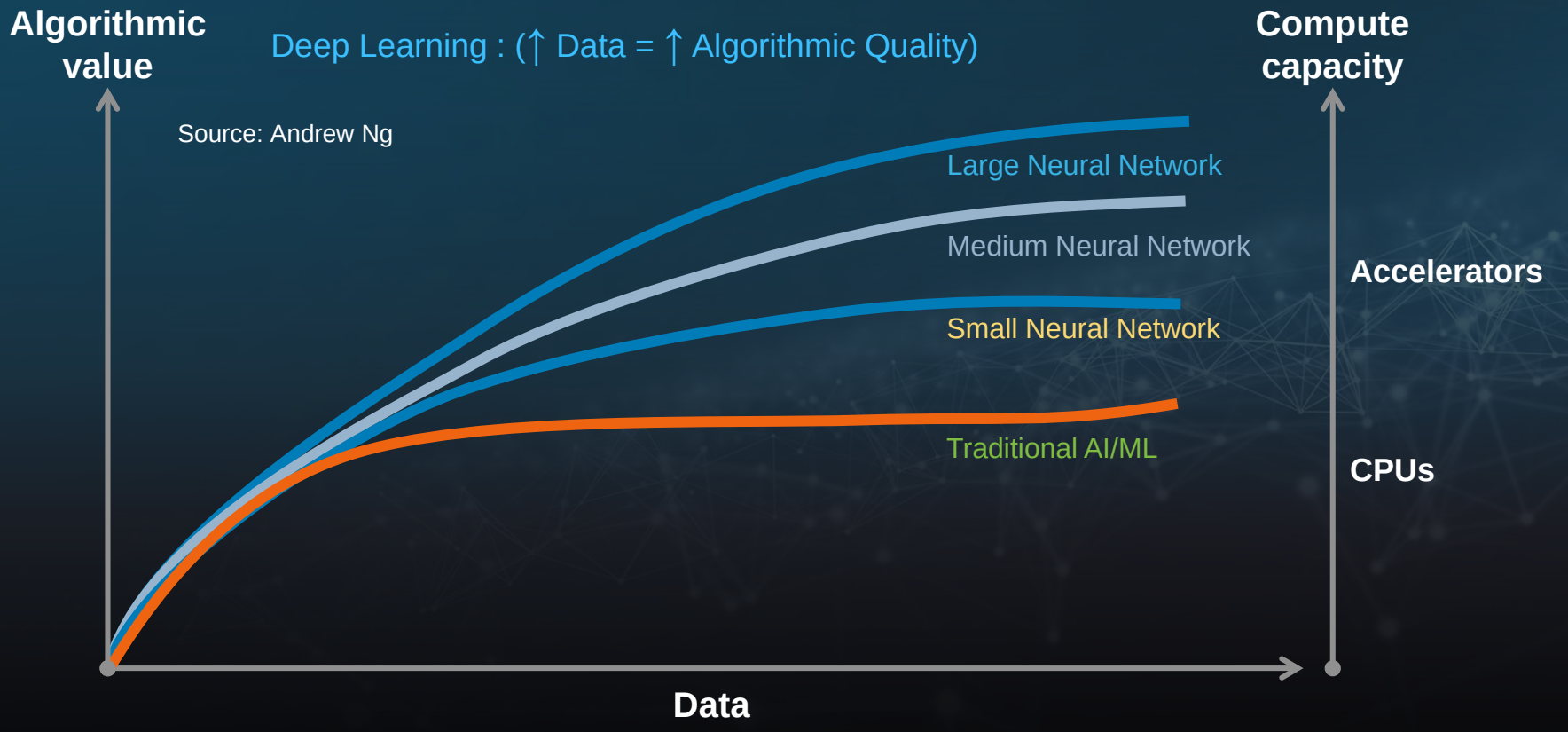
## DATA CAPITAL

Data is now an organizations’ most differentiating asset

# AI accelerates infrastructure growth



# The data-compute-algorithm eureka



# Sample deep learning use cases

AUTOMOTIVE
Road to Autonomous Driving

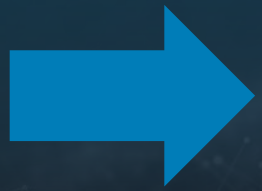
LIFE SCIENCES
Precision Medicine

SMART CITIES
Traffic Analytics, Green Cities

OIL & GAS
Drilling exploration sensor analysis

MEDIA/ENTERTAINMENT
Content enrichment with Metadata

# Deep Learning: From POCs to production

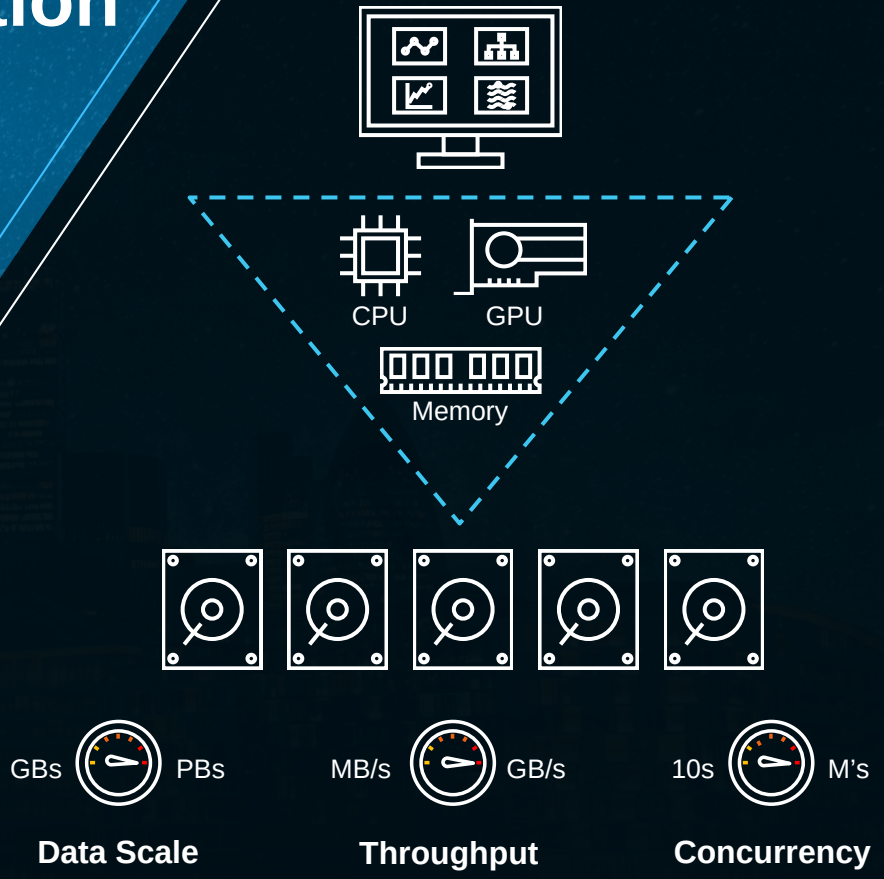


# Data gravity

# I/O bottlenecks AI innovation

## I/O constraints impacts AI

- **Lengthens** model development cycles
- **Difficult to** capture the full value of GPU
- **Limits** analytic accuracy
- **Hard to scale** to large-scale production



# Economic scaling

# Production BI robustness in deep learning

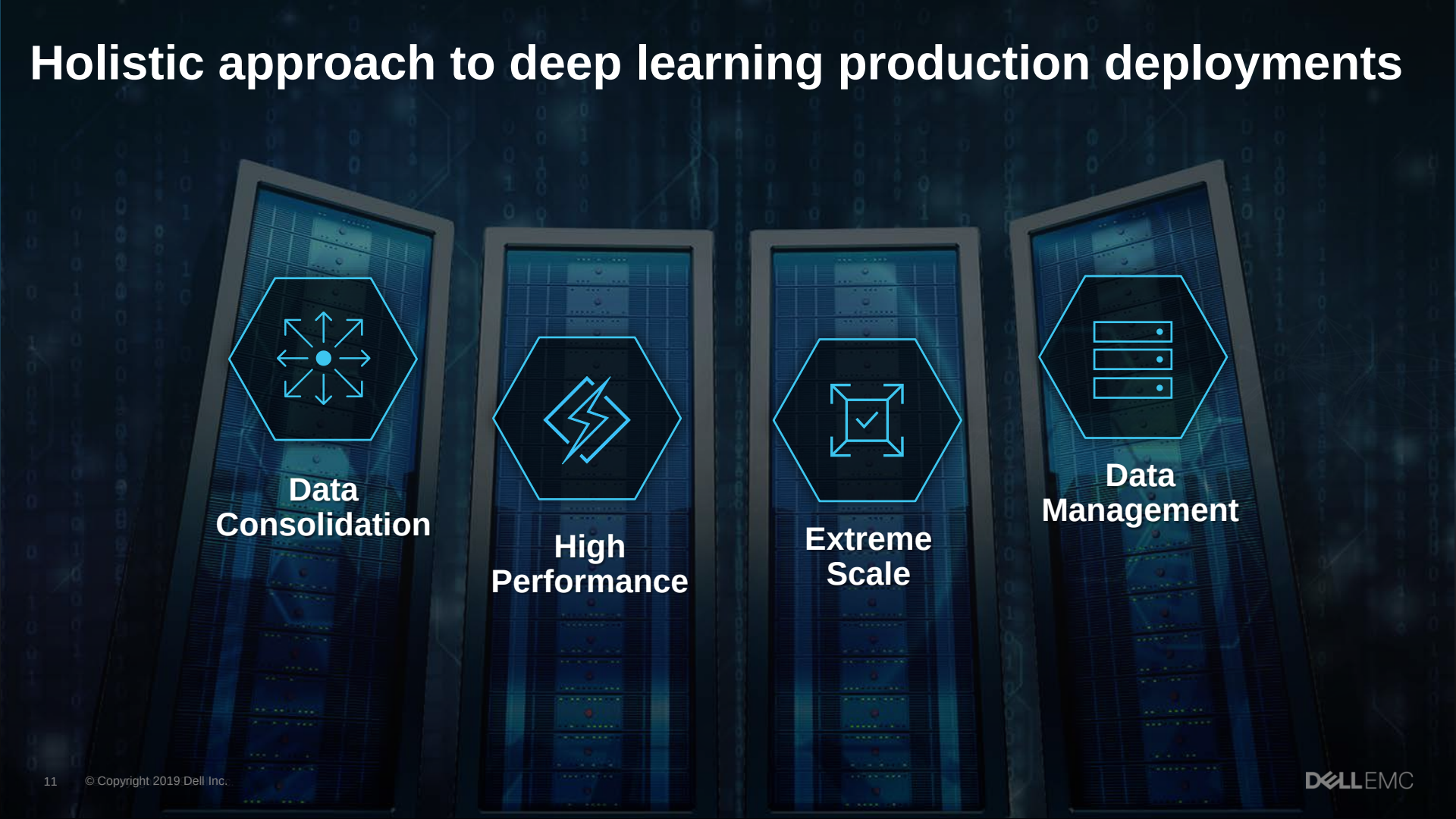
DATA MANAGEMENT

DATA PROTECTION

DATA SECURITY

DATA COMPLIANCE

# Holistic approach to deep learning production deployments



# DELL EMC Ready Solution for AI, Deep Learning with NVIDIA

## A distributed architecture built to scale-out AI


- Pre-integrated Solution
  - PowerEdge Servers with V100 Tesla GPUs
  - 4-way high speed GPU NVLink Interconnect
  - All-Flash Isilon
  - Data Scientist Portal and Bright Cluster Manager
  - Open Source DL Packages: Tensorflow, Caffe2, MxNet etc.
- Software Implementation Services from Dell EMC

**Launched Globally 2018**

# DELL EMC Ready Solution for AI with NVIDIA

## A distributed architecture built to scale-out AI


### Dell EMC PowerEdge Servers

Head nodes for cluster management

Worker nodes optimized to scale out


PowerEdge R740xd head node


PowerEdge C4140 worker nodes

### Dell EMC Networking

- Ethernet used to manage the cluster
- Infiniband for maximum throughout connectivity

Dell EMC S3048-ON Ethernet Switch

Mellanox SB7800 Infiniband Switch

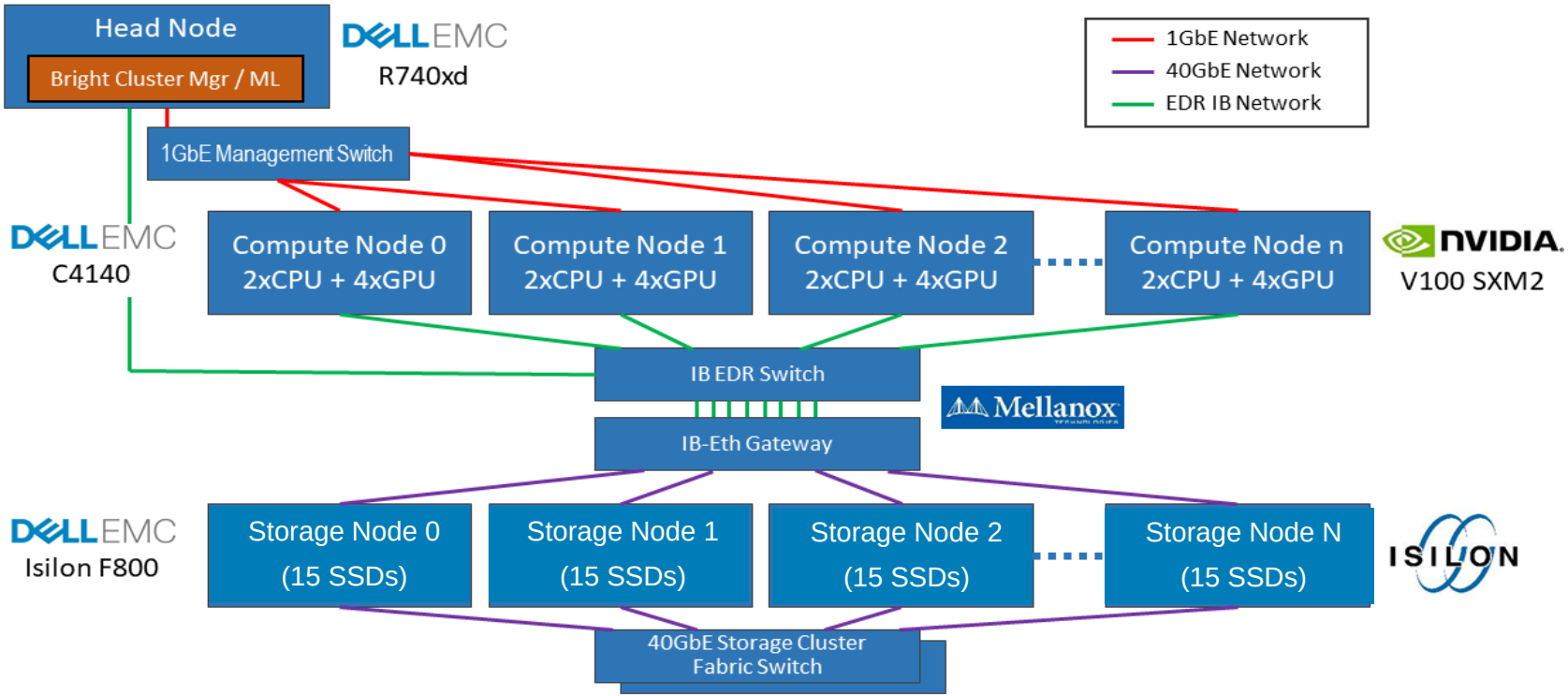
### Dell EMC Isilon Storage

Up to 250K IOPS & 15 GB/s bandwidth

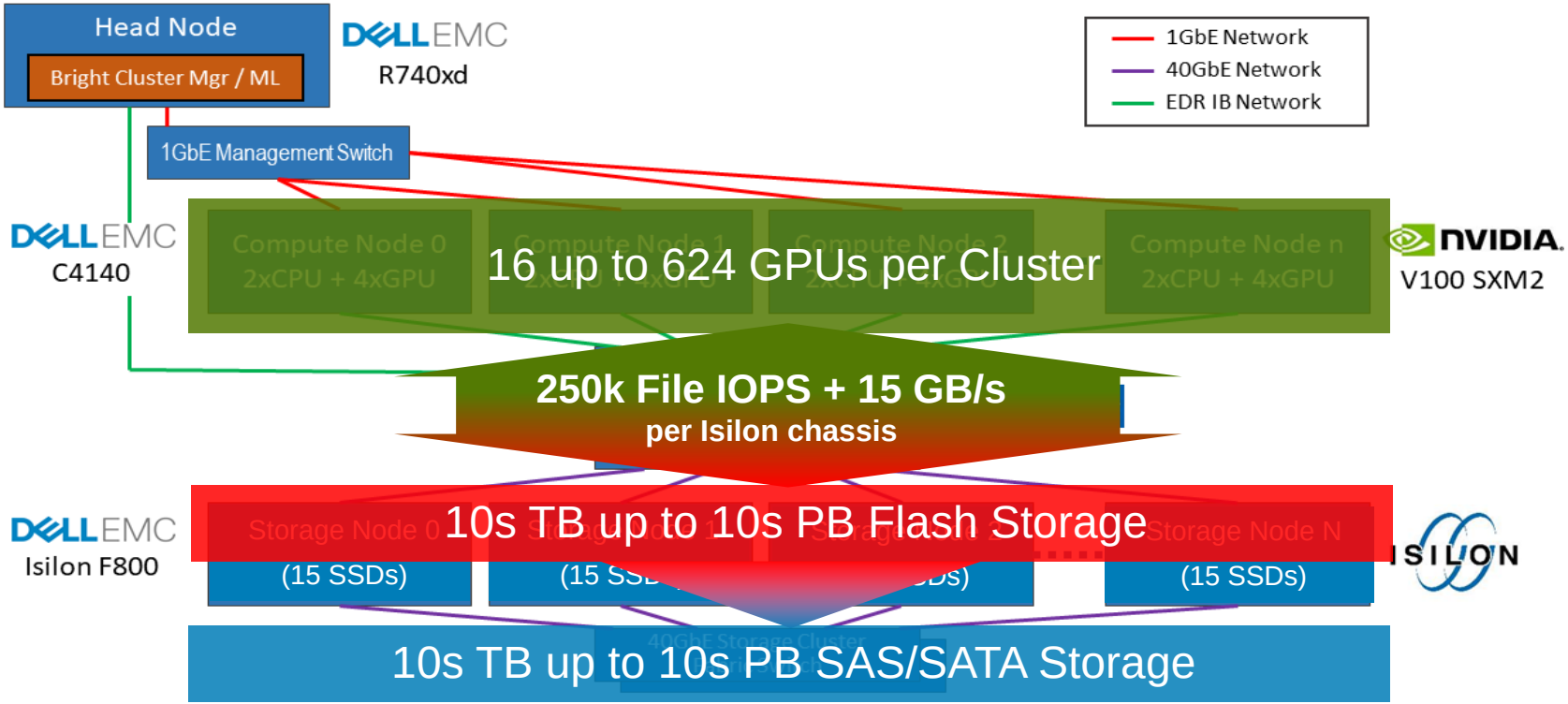
Stores 96-924 TB capacity per chassis


Isilon F800 all-flash scale-out NAS

# DELL EMC Ready Solution for AI with NVIDIA

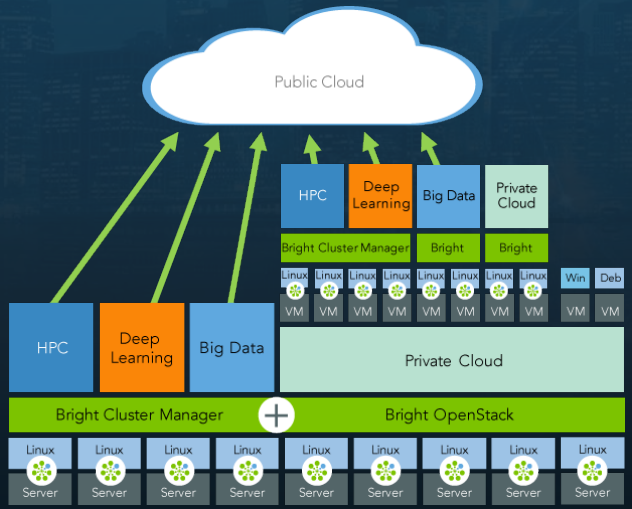


# DELL EMC Ready Solution for AI with NVIDIA



# Bright Cluster Manager

**Bright Cluster Manager** provides a choice of machine learning frameworks and libraries to simplify deep learning projects.



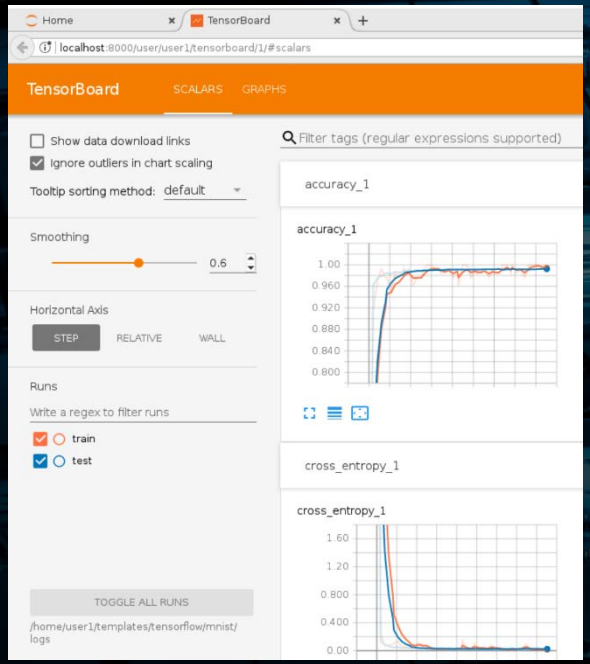

**Open source frameworks**
TensorFlow, MX Net, CNTK, Theano, Torch, Caffe/Caffe 2

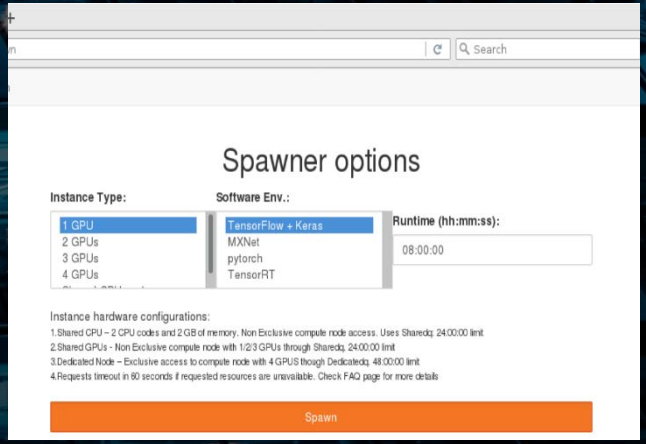
**Neural network libraries**
MLPython, CaffeOnSpark, cuDNN, cuBLAS, NCCL, Keras, GIE…

# Data Science Portal

- Ease of use
  - Spawner for Jupyter Hub
  - Integrated into
    - Slurm Scheduler
    - LDAP for user management
    - Module environment
    - Python2, Python 3 and R support
    - Tensorboard
    - Terminal CLI environment
  - Templates for different ecosystems
  - Support for NGC containers
  - Singularity support

[Walkthrough Video]




# Ready Solution for AI with NVIDIA: Benchmark

**Image Classification with TensorFlow and ImageNet Data Set**


- Data set super-sampled to 1.4 Terabytes to avoid caching
- 4 PowerEdge C4140 nodes with 4 V100 Tesla GPUs each
- 4 node Isilon F800 Chassis
- Horovod used to distribute across multiple compute nodes
- Both FP16 and FP32 Floating Point Precision

- Average GPU Utilization = 95%
- Max Disk I/O Throughput achieved = 15 GBps

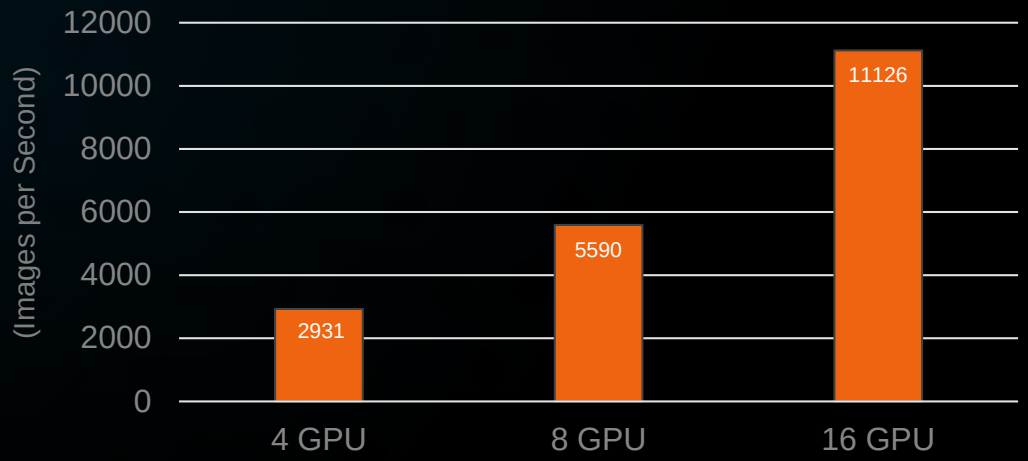
*projected from empirical results

| Storage Performance Demanded | Benchmark Required | GPUs per Isilon F800 Node* |
|---|---|---|
| Low | ResNet50, FP32 | 60 |
| Med | ResNet50, FP16 | 30 |
| High | AlexNet | 13 |

# Ready Solution for AI with NVIDIA: Benchmark results

**Image Classification with TensorFlow and ImageNet Data Set**





# Announcing today : Dell EMC Isilon with NVIDIA DGX-1 Solution

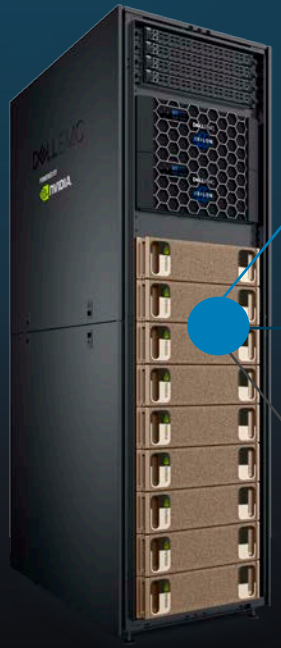
## For customers currently looking for 8-way GPU Interconnect

- Pre-integrated Solution
  - DGX-1 Servers and Software
  - 8-way high-speed GPU NVLink interconnect
  - All-Flash Isilon
- Implementation Services provided by VAR partners

**AMER Launch** through WWT, FusionStorm, Presidio, Insight, Sirius

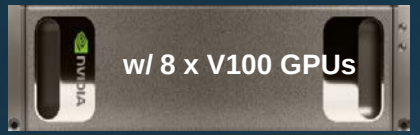
# Dell EMC Isilon with NVIDIA DGX-1

## For customers currently looking for 8-way GPU Interconnect



### NVIDIA GPU Acceleration

- 8-way GPU with High Speed NVLink Interconnect
- Cloud-based container registry for Deep Learning software




NVIDIA DGX-1

### Networking

- Ethernet used to connect the cluster
- 100G – 40G conversion as needed

Dell EMC S5232 Switches
(OR equivalent Switch)

### Dell EMC Isilon Storage

- Up to 250K IOPS & 15 GB/s bandwidth per Chassis
- Stores 96-924 TB capacity per chassis


Isilon F800 all-flash scale-out NAS

# Dell EMC Deep Learning Solution Portfolio with NVIDIA

## Ready Solution for AI with NVIDIA

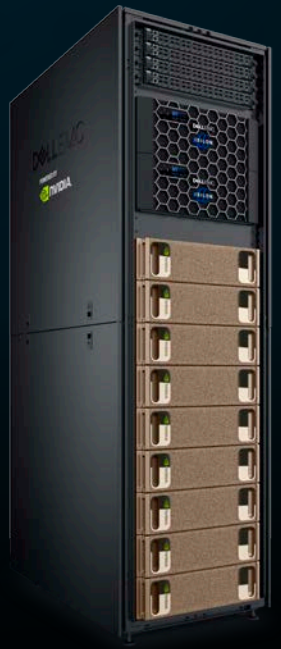

## NVIDIA DGX-1 with Isilon Solution

NVIDIA

| Good for | Many Common Deep Learning Workloads | Workloads needing 8-way GPU Interconnect |
|---|---|---|

# Dell EMC Isilon with NVIDIA DGX-1: Benchmark

**Image Classification with TensorFlow and ImageNet Data Set**



- Data set super-sampled to 22TB to avoid caching
- 9 DGX-1 nodes (72 GPUs total)
- 8 Isilon F800 nodes in 2 chassis
- 40 GigE to Isilon

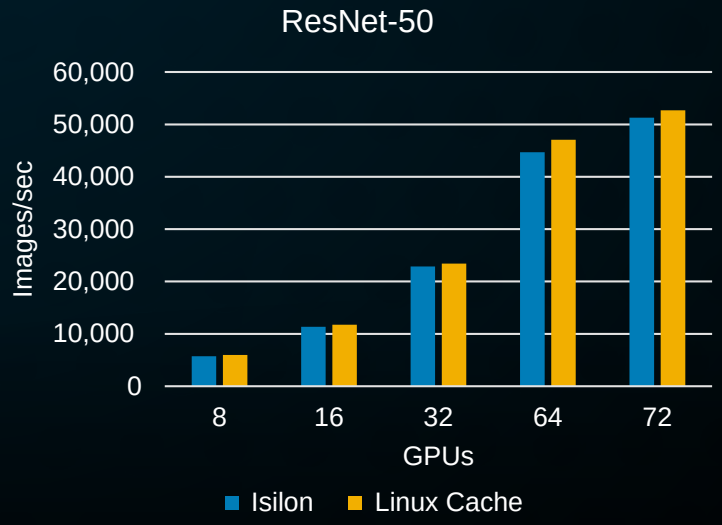
- All GPUs > 97% Utilization
- 96% of local memory throughput with Isilon
- Linear Scaling from 8 to 32 to 72 GPUs

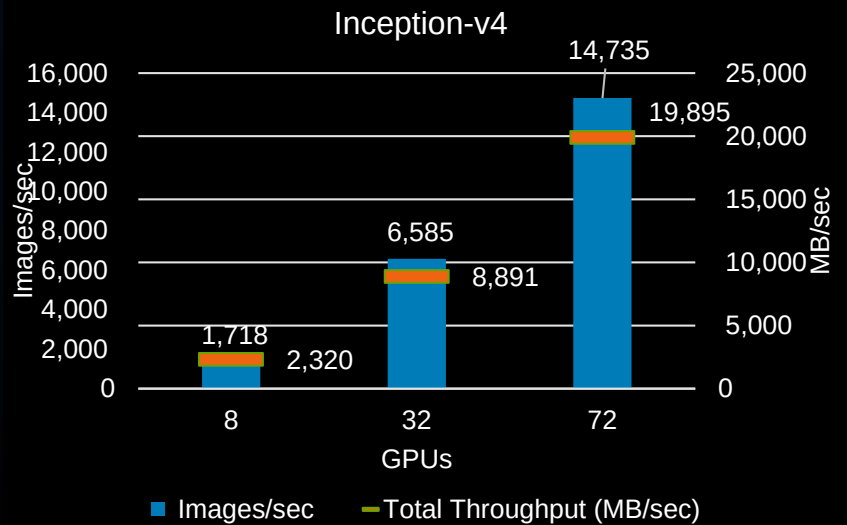
| Storage Performance Demanded | Benchmark | V100 GPUs per Isilon F800 Node |
|---|---|---|
| Low | Training, Small Images, ResNet50 | 30 |
| Medium | Inference, Small Images, ResNet50 | 20 |
| High | Training, Large Images, Inception v4 | 9 |

# Isilon with DGX-1: Benchmark Results

**Training: Image Classification with TensorFlow and 22 TB ImageNet**

Training → Inferencing → Run-time Scoring



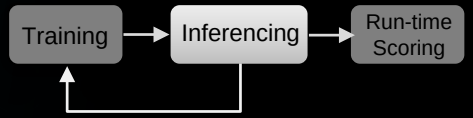


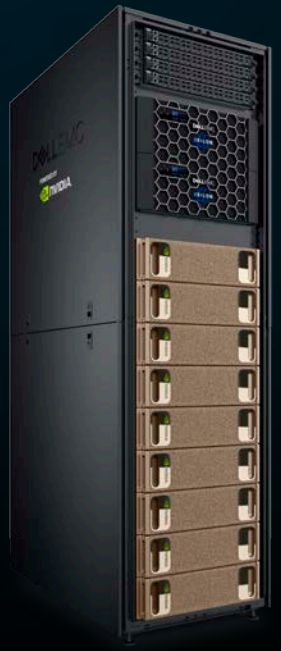
+

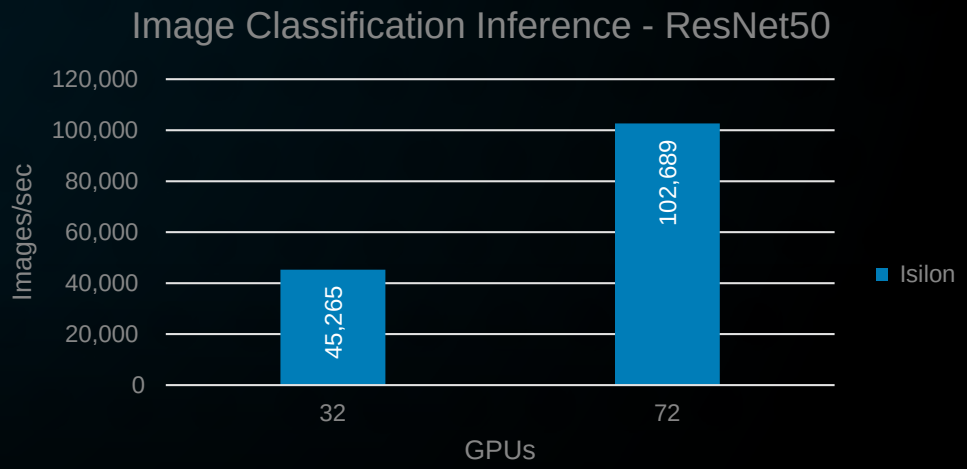

- 97% GPU utilization or higher
- Linear Scaling from 8 to 32 to 72 GPUs
- Delivers up to 19.9 GB/s

# Isilon with DGX-1: Benchmark results

**Inferencing: Image Classification with TensorFlow and ImageNet Data Set**






Inferencing

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

# The bottomline:

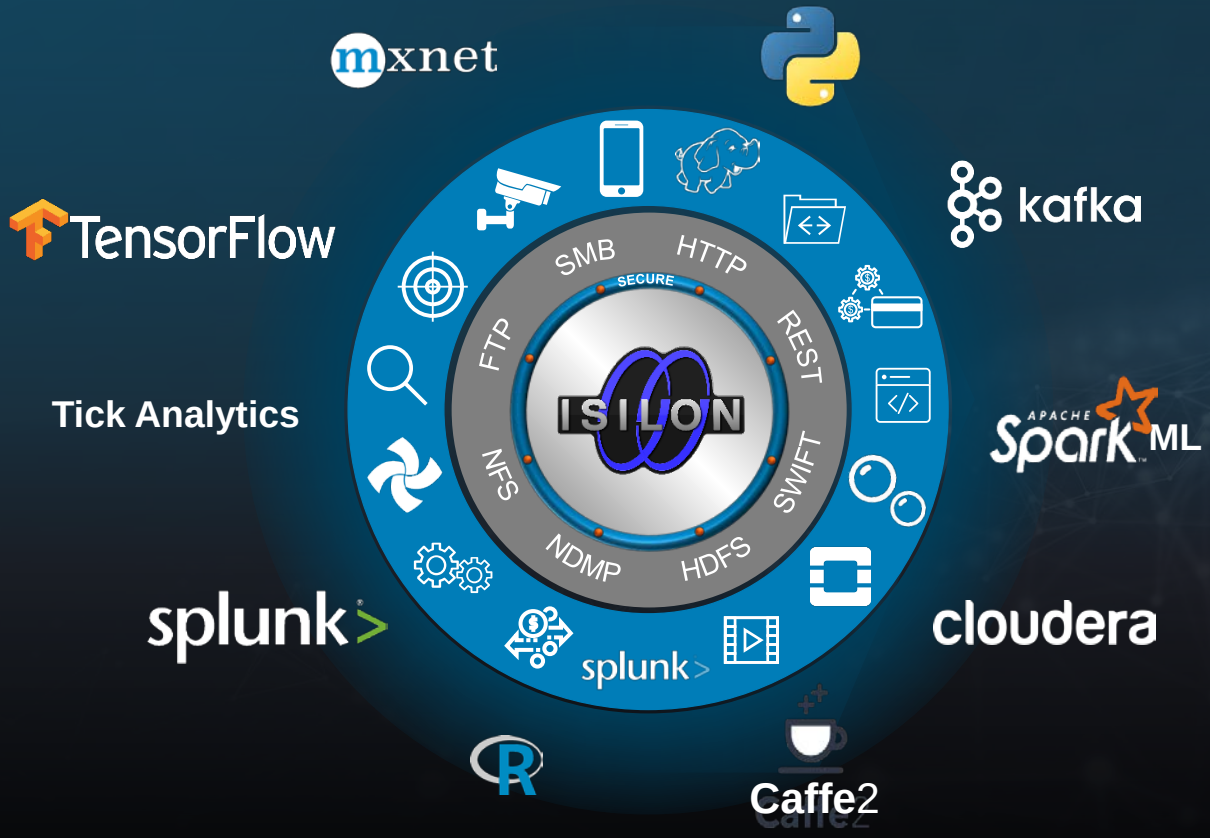
## Deep Learning I/O bottlenecks eliminated at any scale


***"We must rely on AI to help make sense of it all. Dell EMC Isilon is a critical component of how we push the science forward by giving us a simple scale-out solution to manage and consume Petabytes of data and to expedite genome processing from weeks to hours."***

**- James Lowey, CIO TGEN**

- Faster training and validation of AI models
- Higher model accuracy
- Improve data science productivity
- Maximize ROI of compute investments

# Consolidate data in the data lake = Bring Deep Learning closer to IT




- **Minimize cost and time to market** with in-place AI
- **Improve IT re-use and agility** with ability to work with any compute or application

# Enterprise data management

## Bringing Production BI robustness to Deep Learning

DATA MANAGEMENT


DATA PROTECTION

DATA SECURITY

DATA COMPLIANCE

# Bringing deep learning to existing Isilon deployments

**Ready Solution for AI**
Deep Learning with NVIDIA


All Flash

Automatic Tiering

An Industry First

**Dell EMC Isilon with NVIDIA DGX-1**

All Flash

Automatic Tiering


**Existing Isilon Clusters**

# Data pipelines simplified

# Accelerate ADAS/AD development with AI

## Isilon powers the journey to fully Autonomous Driving


A single FLiR operating at 2800 MBit/s travelling at 60km/h over the course of 200,000km will produce:

**3K+**
Hours of data

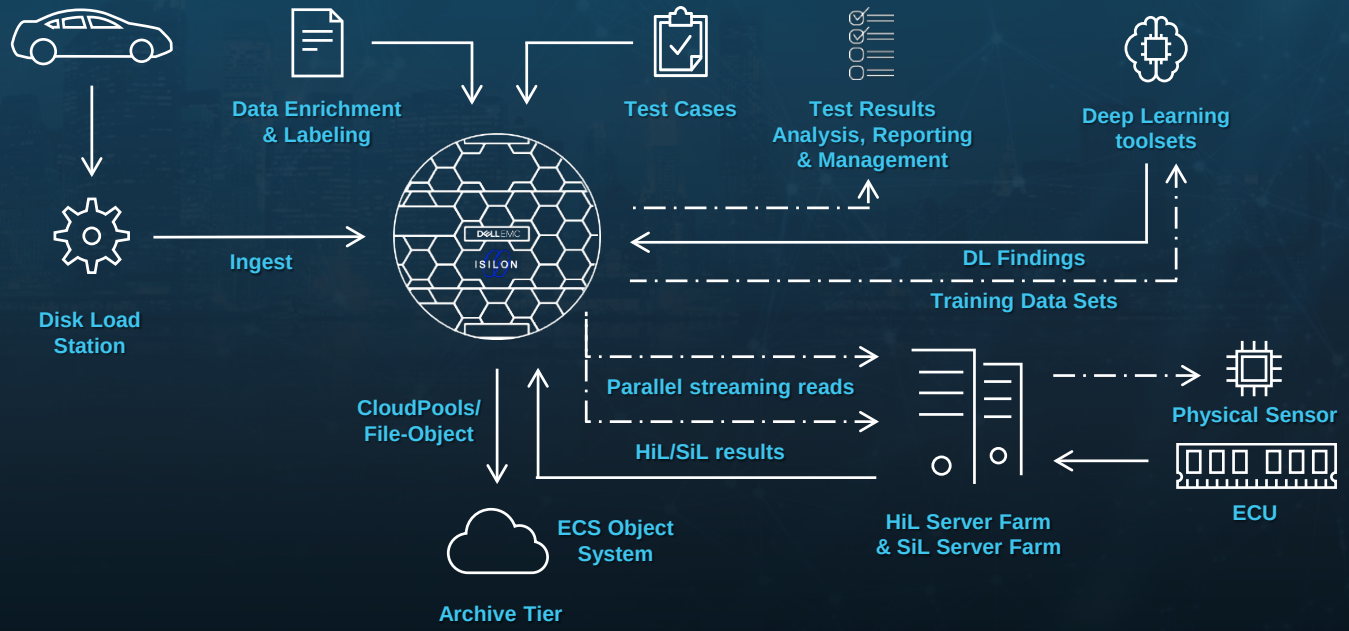
**1260**
GB per hour

**4.2PB**
Over the course of the project for ONE sensor

# Isilon-based end to end ADAS solution

## CAPTURED DATA

- Video
- Radar
- Lidar
- GPS
- Ultrasonic
- Vehicle Data
- And More



# Revolutionize patient care with AI

## Isilon powers life-saving precision medicine

Patient genomic data

Cohort data

EMR

IoT health devices

Reference genomic data

Mapping a single human genome requires analysis of a large dataset. Imagine doing this for thousands of patients.

**4TB**

Size of genome being mapped per person

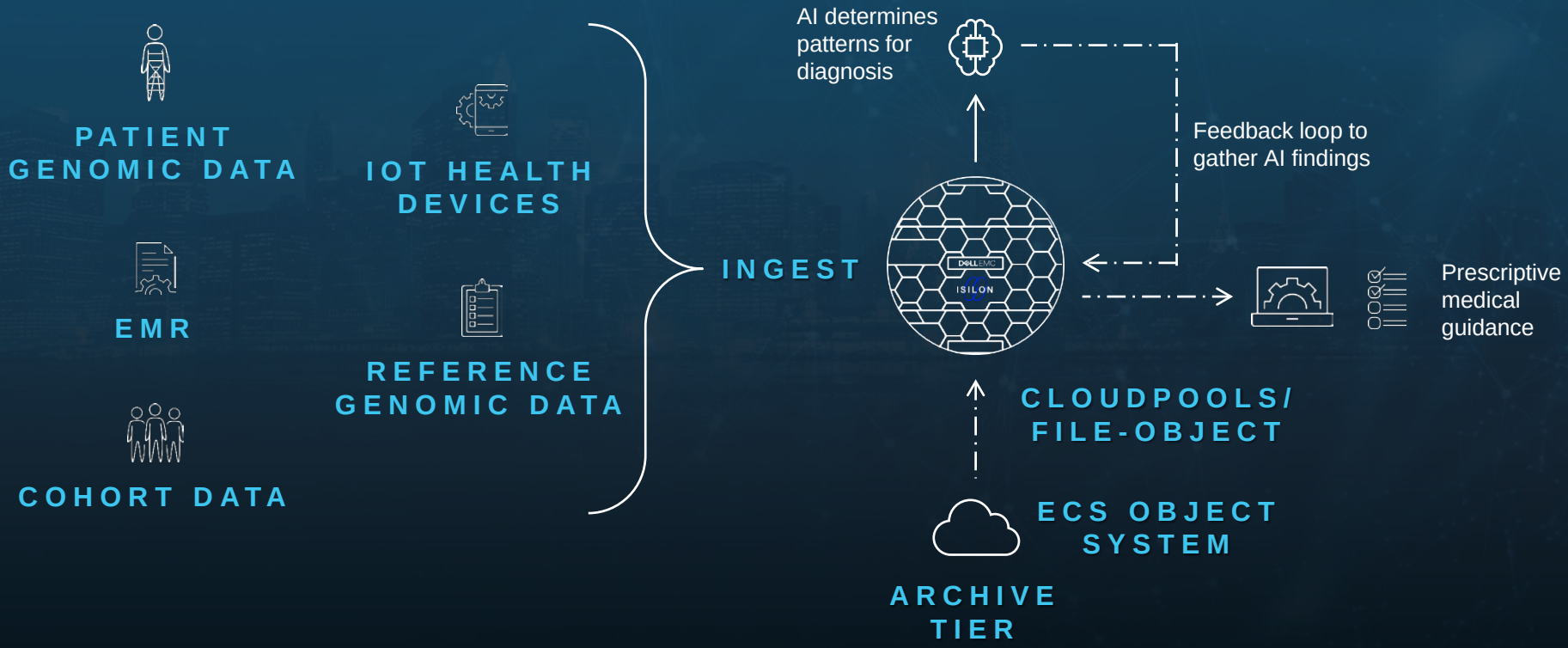
**6Gbps**

of genomic data analysis throughput

**<24HRs**

required to fully map

# Empowering precision medicine with AI



# Improve manufacturing predictability with AI

## Isilon powers AI-driven pre-emptive quality and maintenance activities

Vibration sensors

Metal purity analysis

Moisture detector

Thermometers

Microphones

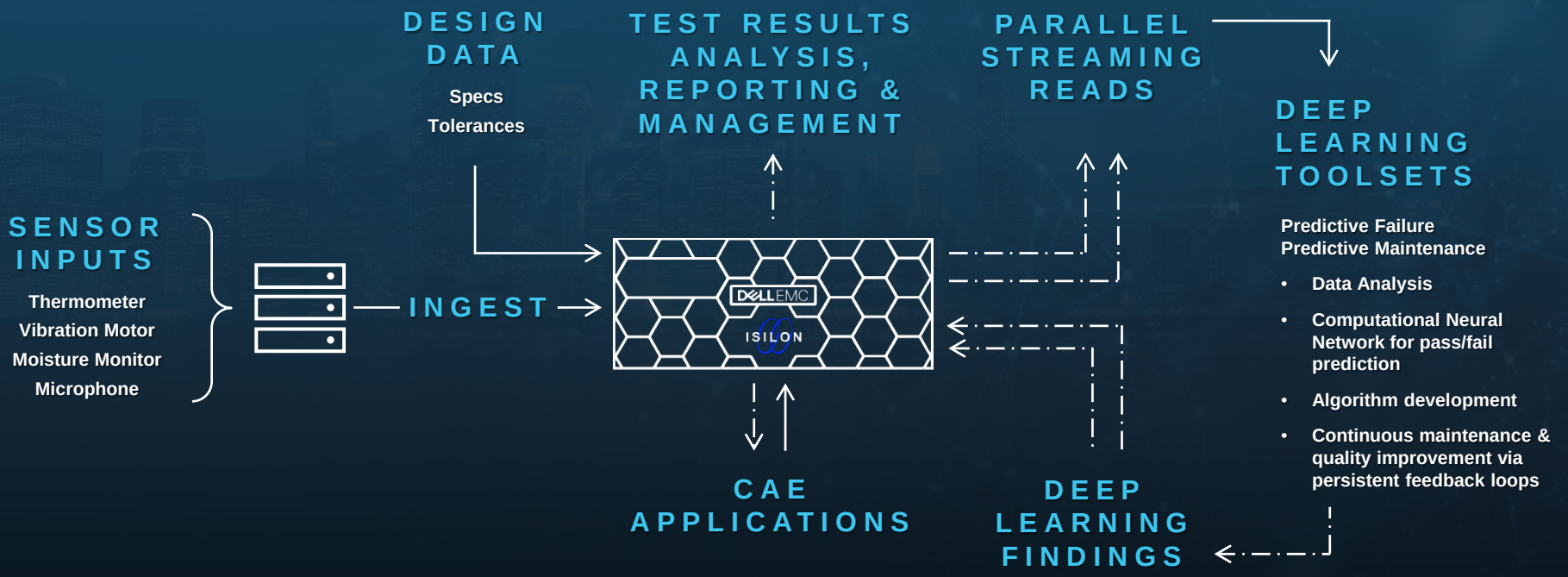
Utilize AI to make sense of a massive amount of sensor data to increase yield, improve product quality and reduce downtime

**1000s**
of IOT Sensors

**24/7**
data collection

**TBs to PBs Scale**
required as data grows exponentially

# End-end AI solutions for manufacturing



# Holistic approach to deep learning at scale

Data Consolidation

High Performance

Extreme Scale

Enterprise Features

# Dell EMC + NVIDIA partnership in AI

- Market Alignment
  - Common channel partners
  - Thought Leadership Sessions and Conferences
- Solution Alignment
  - Ready Solution for AI with NVIDIA
  - Dell EMC Isilon with NVIDIA DGX-1 Solution
  - Industry Vertical Solutions

Dell EMC: One of NVIDIA's largest OEM partner



LIFE SCIENCES


FINANCIAL


FEDERAL


ADAS/AUTO


MEDIA AND ENTERTAINMENT


VIDEO SURVEILLANCE


O & G


HEALTHCARE

DELL EMC + NVIDIA

DELL EMC @ NVIDIA GTC GPU TECHNOLOGY CONFERENCE

- Booth # 1311
- 1:1 with Solution Experts @ Hilton

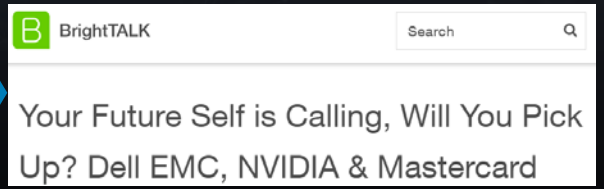
## 6 Demo Stations

| | |
|---|---|
| AI READY SOLUTIONS | HPC SOLUTIONS |
| VIRTUAL DESKTOP | DELL WORKSTATION |
| DATA SCIENCE WORKSTATION | INTELLIGENT VIDEO ANALYTICS |

## Websites

- Deep Learning Solutions and case studies with Isilon
- Dell EMC Solutions for Machine Learning and Deep Learning

Upcoming BrightTalk (Apr 1st)






DELL EMC


DELL EMC