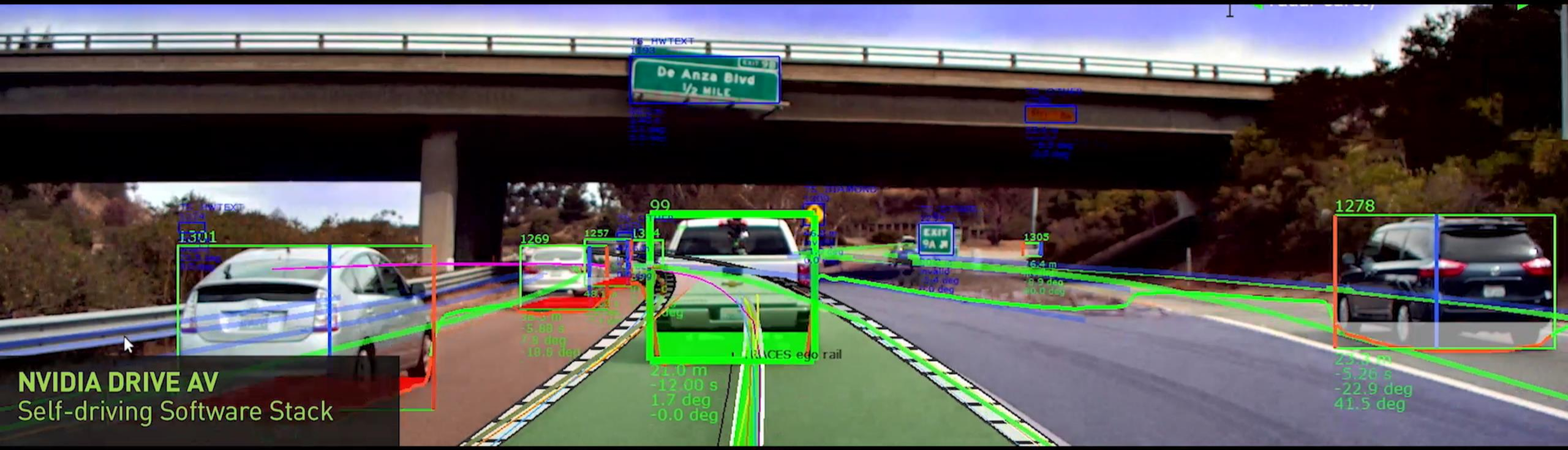
IBM SPECTRUM STORAGE FOR AI WITH NVIDIA DGX

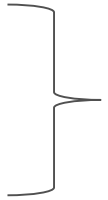
DGX REFERENCE ARCHITECTURE SOLUTION

Dr. Adolf Hohl, SA AUTO Datacenter EMEA


De Anza Blvd
1/2 MILE
1301
1269
1257
99
1278
1305
21.0 m
-12.00 s
1.7 deg
-0.0 deg
-5.26 s
-22.9 deg
41.5 deg
NVIDIA DRIVE AV
Self-driving Software Stack

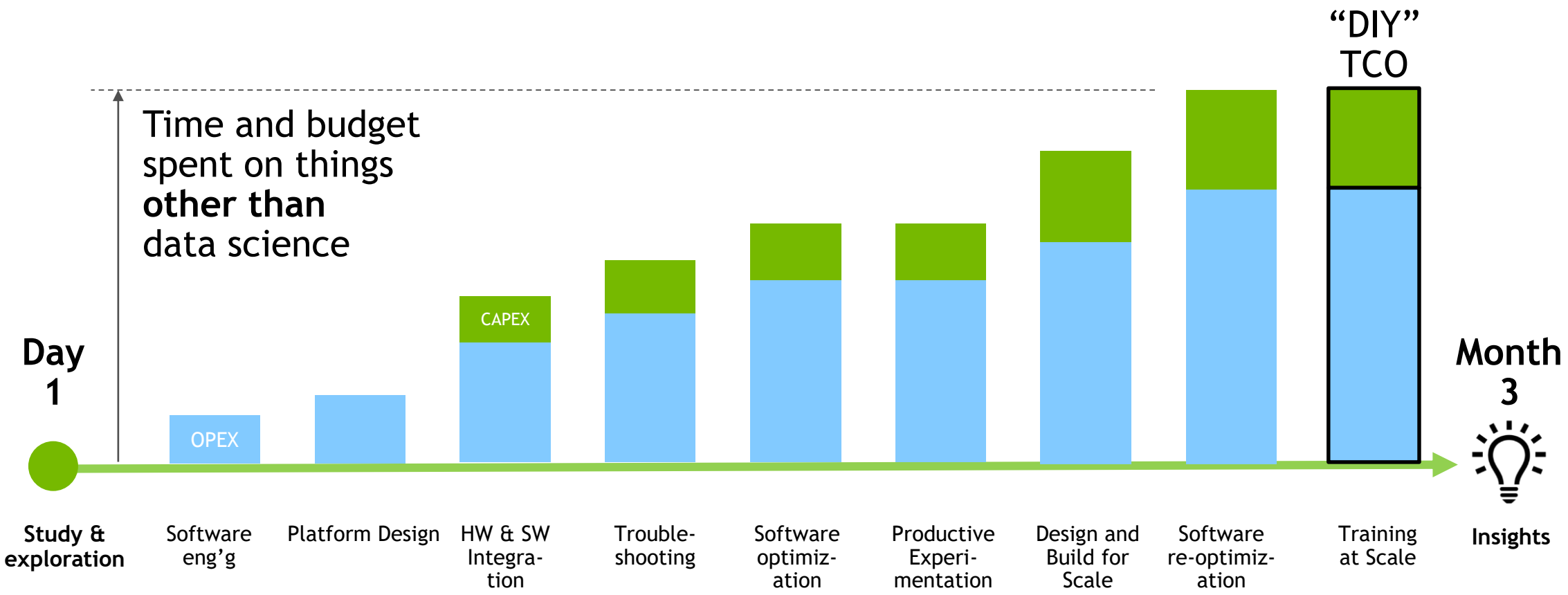
HOW DO WE TRAIN THESE NETWORKS?

- SINGLE GPU CODE is a dying specie
- All our AV DL code is made for MULTIGPU and scalable :
  - Runs on Single GPU
  - Runs on Multi GPU
  - Runs on Multi Nodes with Multiple GPUs
- We use a Cluster for DL Training
- Just ONE codebase
- Just ONE way to orchestrate



I talked about these in a previous IBM Meetup (https://www.youtube.com/watch?v=8xj4CK4ZUMQ)

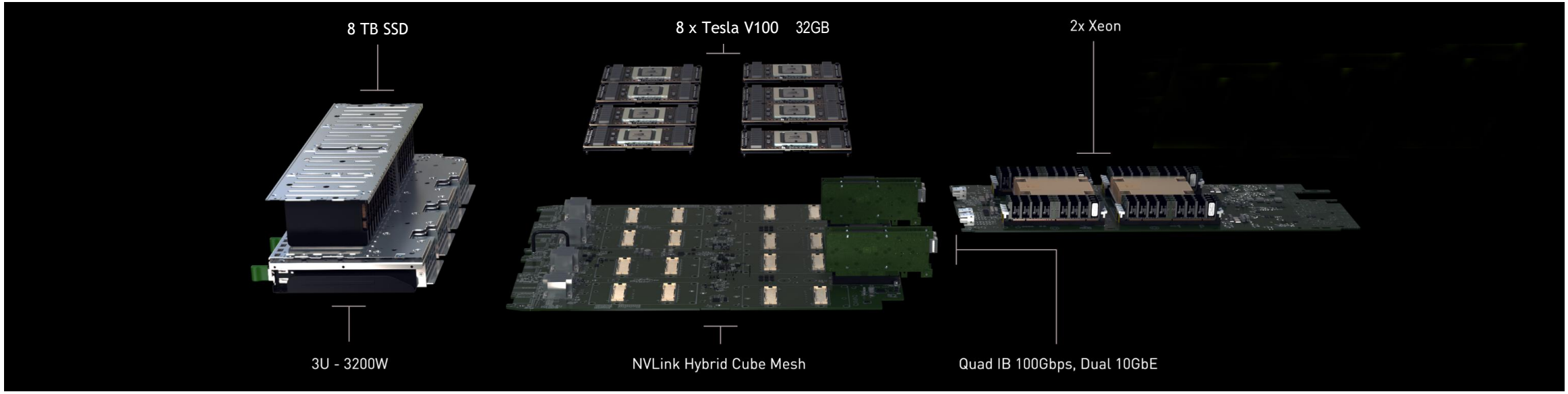
THE TRUE TCO OF AN AI PLATFORM



1. Designing and Building an AI Compute Platform – from Scratch

NVIDIA DGX-1: THE ESSENTIAL TOOL OF AI

Fastest Start, Effortless Productivity, Revolutionary Performance

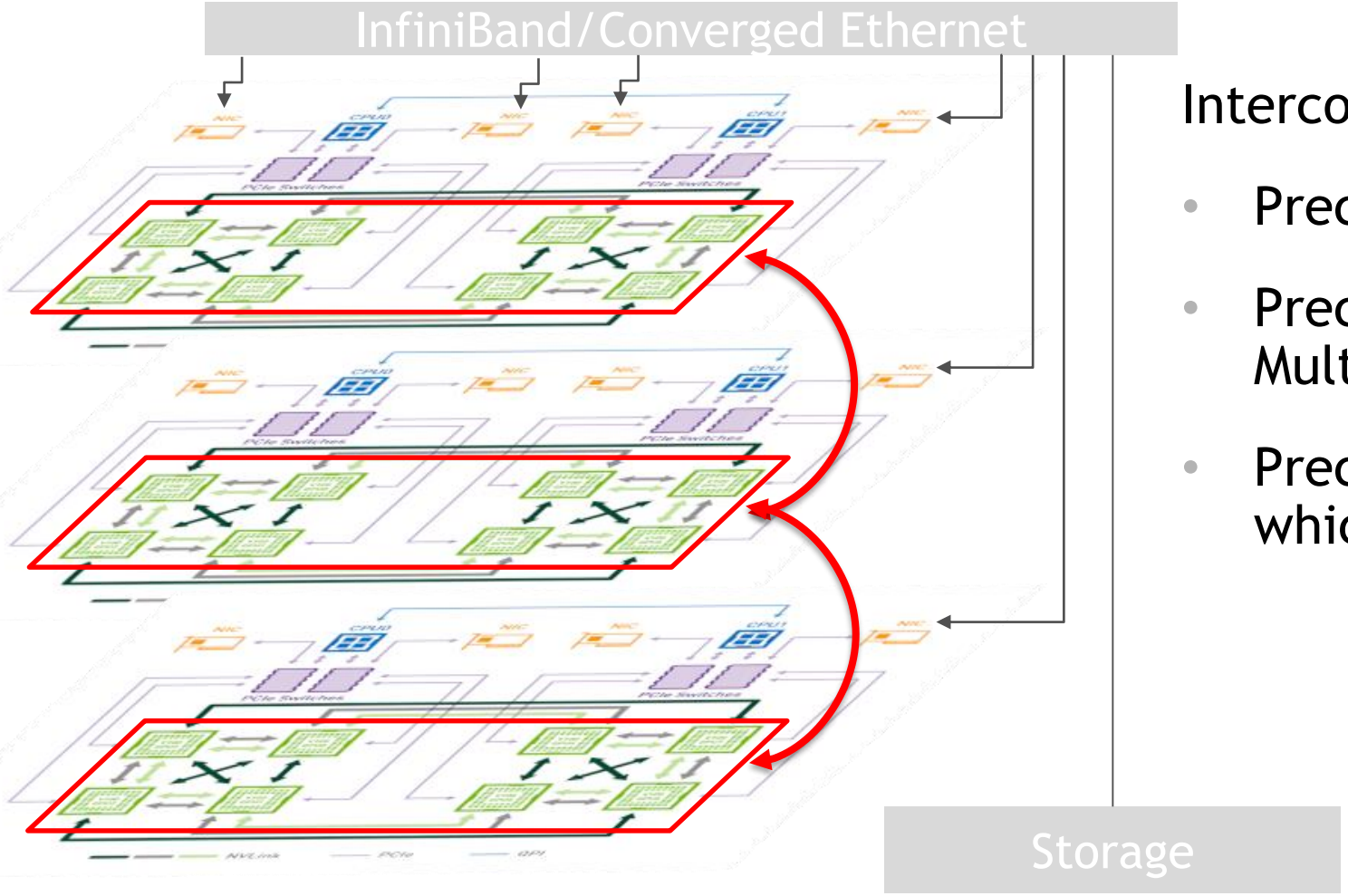


1 PFLOPS | 8x Tesla V100 32GB | 300 GB/s NVLink Hybrid Cube Mesh

2x Xeon | 8 TB RAID 0 | Quad 100Gbps, Dual 10GbE | 3U — 3500W

STACKING DGX

Aggregating Ressources – Scaling Out

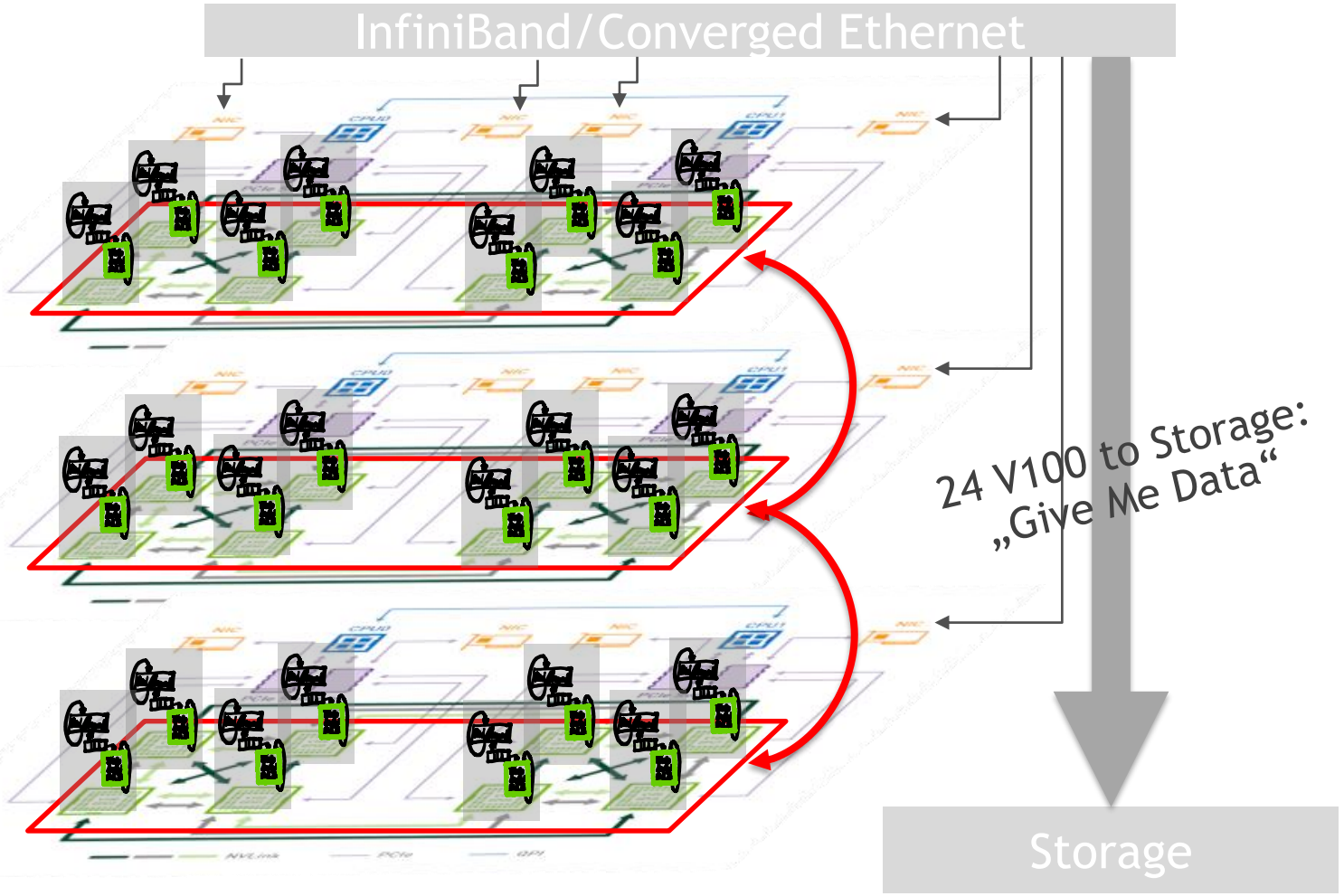


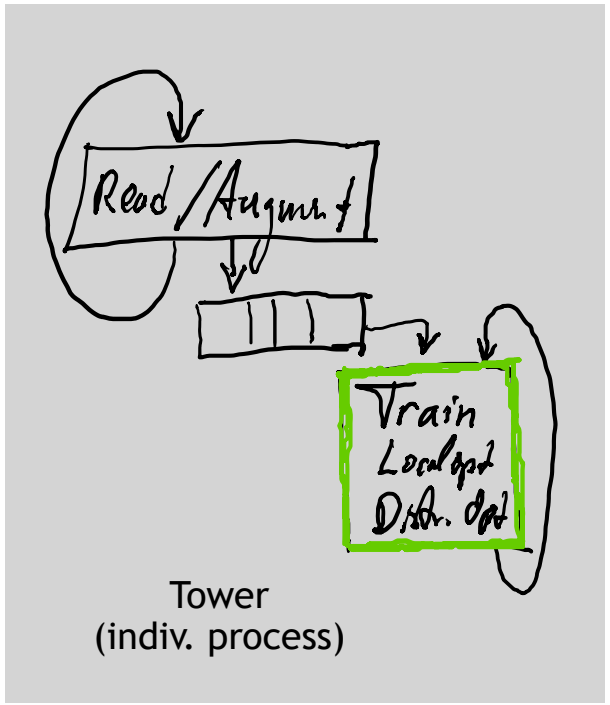
Interconnected Nodes

- Precondition to Scale
- Precondition for effective MultiNode-MultiGPU scaling
- Precondition to aggregate ressources which were left over

SCALING WITH HOROVOD

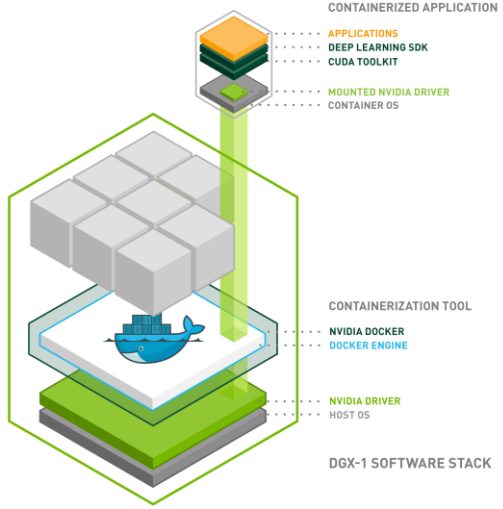
One Process per GPU – One Datapipeline per GPU





SOFTWARE STACK TO SCALE OUT

- NVIDIA GPU CLOUD (NGC)
  - Ready to scale
  - Optimized
  - MPI, Horovod
  - NCCL
  - ngc.nvidia.com



- IBM PowerAI
- ibmcom/powerai
  - Ready to scale
  - Optimized
  - hub.docker.com/r/ibmcom/powerai/

IBM SPECTRUM STORAGE FOR AI WITH NVIDIA DGX

The Engine to Power Your AI Data Pipeline

HARDWARE

- **NVIDIA DGX-1 | up to 9**x DGX-1 Systems
- **IBM Spectrum Scale NVMe Appliance|** 40GB/s per node, 120GB/s in 6RU| 300TB per node
- **NETWORK: Mellanox SB7700 Switch |** 2x EDR IB with RDMA

SOFTWARE

- **NVIDIA DGX SOFTWARE STACK** | NVIDIA Optimized Frameworks
- **IBM:** High performance, low latency, parallel file system
- **IBM:** Extensible and composable

IBM SPECTRUM STORAGE FOR AI WITH NVIDIA DGX: SCALABLE REFERENCE ARCHITECTURES

Scaling with NVIDIA DGX-1

- Start with a single IBM Spectrum Scale NVMe and a single DGX-1
- Grow capacity in a cost-effective, modular approach
- Each config delivers balanced performance, capacity and scale
- IBM Spectrum Scale NVME all-flash appliance is power efficient to allow maximum flexibility when designing rack space and addressing power requirements

3:1 Configuration

6:2 Configuration

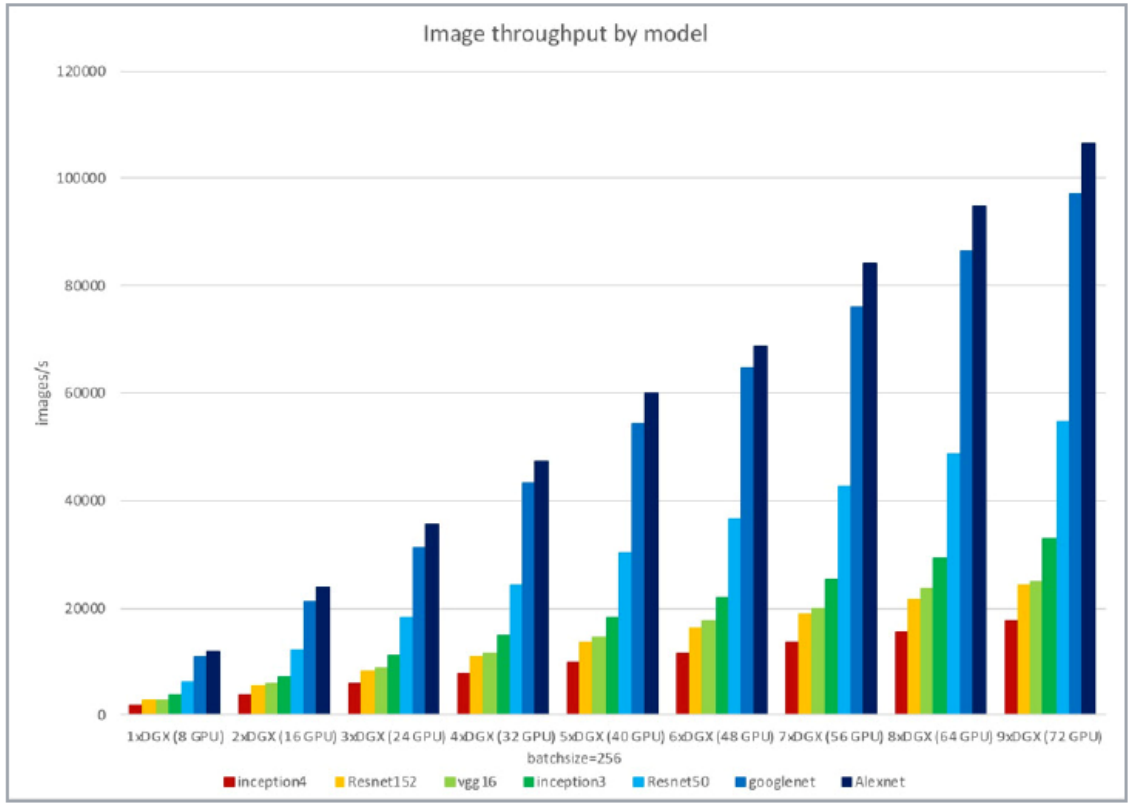
9:3 Configuration

IBM STORAGE WITH NVIDIA DGX: FULLY-OPTIMIZED AND QUALIFIED

Performance at Scale

For multiple DGX-1 servers, IBM Spectrum Scale on NVMe architecture demonstrates linear scale up to full saturation of all DGX-1 server GPUs

The multi-DGX server image processing rates shown demonstrate scalability for Inception-v4, ResNet-152, VGG-16, Inception-v3, ResNet-50, GoogLeNet and AlexNet models

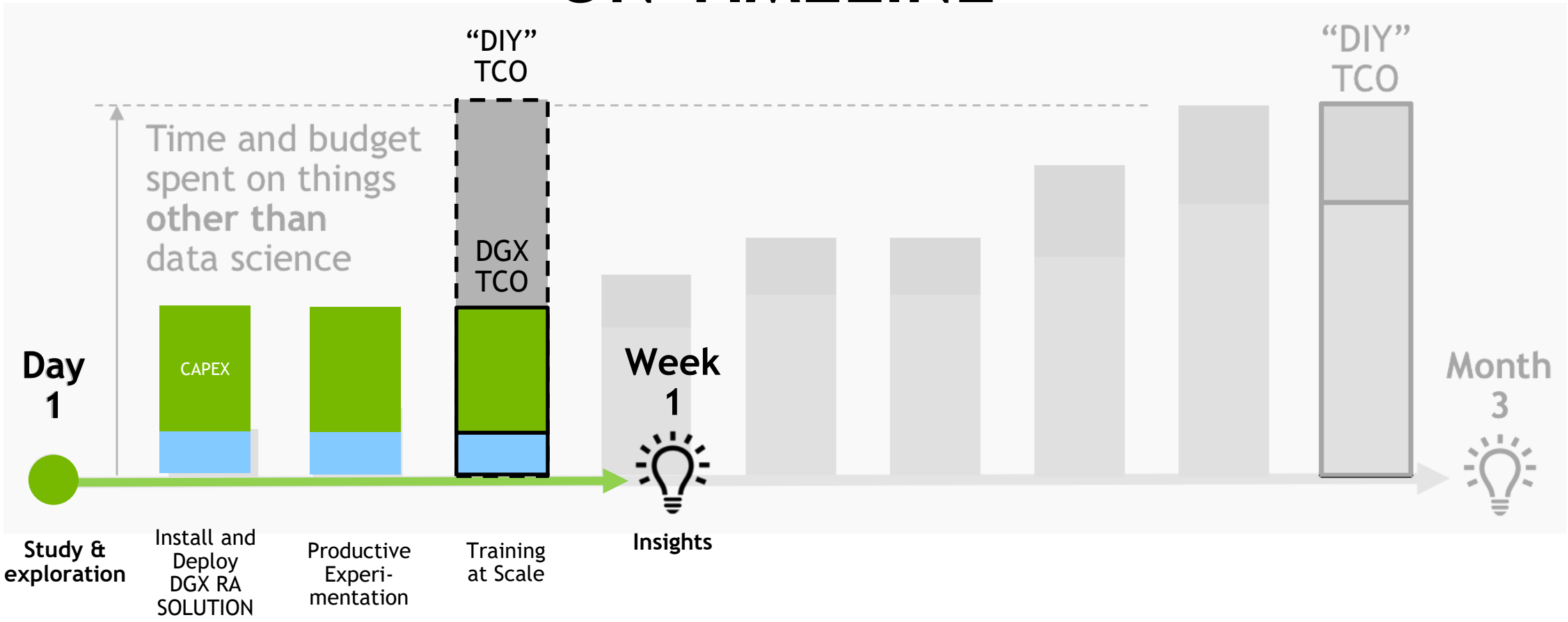


BUSINESS IMPACT OF IBM SPECTRUM STORAGE FOR AI WITH NVIDIA DGX

THE IMPACT OF IBM STORAGE + NVIDIA DGX ON TIMELINE

2. Deploying an Integrated, Full-Stack AI Solution using DGX Systems

THE IMPACT OF IBM STORAGE + NVIDIA DGX ON TIMELINE



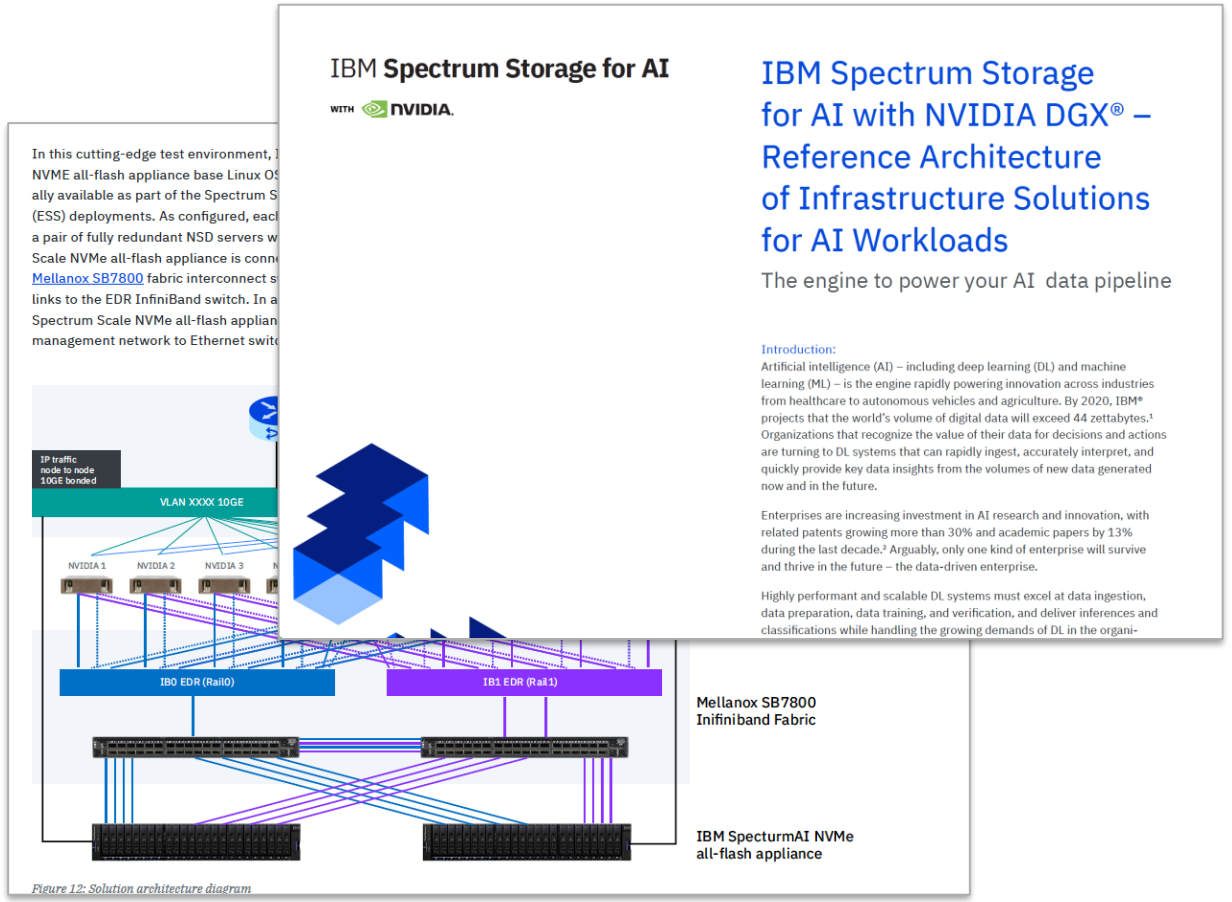
2. Deploying an Integrated, Full-Stack AI Solution using DGX Systems

IBM & NVIDIA REFERENCE ARCHITECTURE

Validated design for deploying DGX at-scale with IBM Storage

Download at
https://bit.ly/2GcYbgO

Learn more about
DGX RA Solutions at:
https://bit.ly/2OpXYeC




NVIDIA
IBM
Spectrum
Storage