AI from Edge to Cloud – It's real

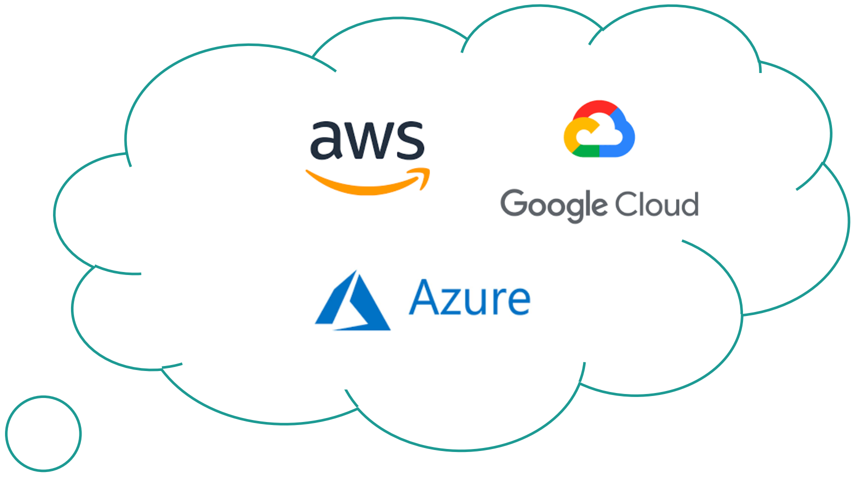
S9948 - AI from Edge to Cloud: How HPE and Seagate deliver Quality and Efficiency across the Manufacturing Supply Chain

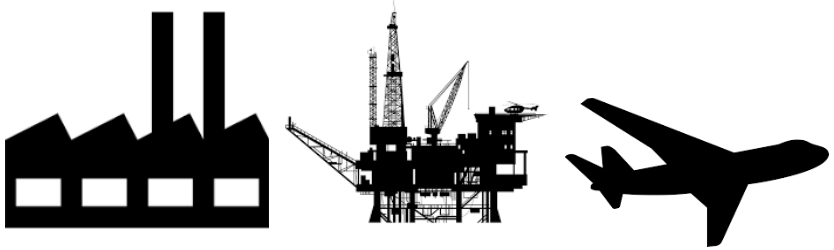
Bharath Ramesh, Head – WW IoT Product Management & Marketing, HPE
Bruce King, Data Science Technologist, Seagate Technology

18th March 2019

Edge is the NEW “Off Prem”

Cloud – The current “Off Prem”





Edge – The NEW “Off Prem”

More sensors
= More data
= More analytics
= Smarter Operations

Datacenter – The “On Prem”

© Copyright 2018 Hewlett-Packard Enterprise Corporation. The information contained herein is subject to change without notice.

The Edge is a spectrum of use cases that benefit from Deep Learning

* OT = Operational Technology

Manufacturing

Legacy Interworking/PLC bridging, Asset Management, ***Condition Monitoring/Predictive****, Augmented Reality,* ***Edge Video Analytics***

Energy & Utilities

Asset Management, ***Condition Monitoring/Predictive****, Augmented Reality,* ***Custom AI/Deep-Learning***

Transportation & Automotive

Connected Vehicles, Asset Management, ***Condition Monitoring/Predictive****,* ***Edge Video Analytics***

Campus, Branch, Retail

Edge Workspace Appliance for Workstation/Desktop/App delivery, ***Edge Video Analytics (Campus Surveillance/CCTV)***

OT* Edges **IoT** **Enterprise IT Edges**

Oil & Gas

Asset Management, ***Condition Monitoring/Predictive****, Augmented Reality,* ***Edge Video Analytics****,* ***Custom AI/Deep-Learning***

Defense & Intelligence

Connected Battlefield, Signals Intelligence, Edge Video Analytics, Custom AI/Deep-Learning

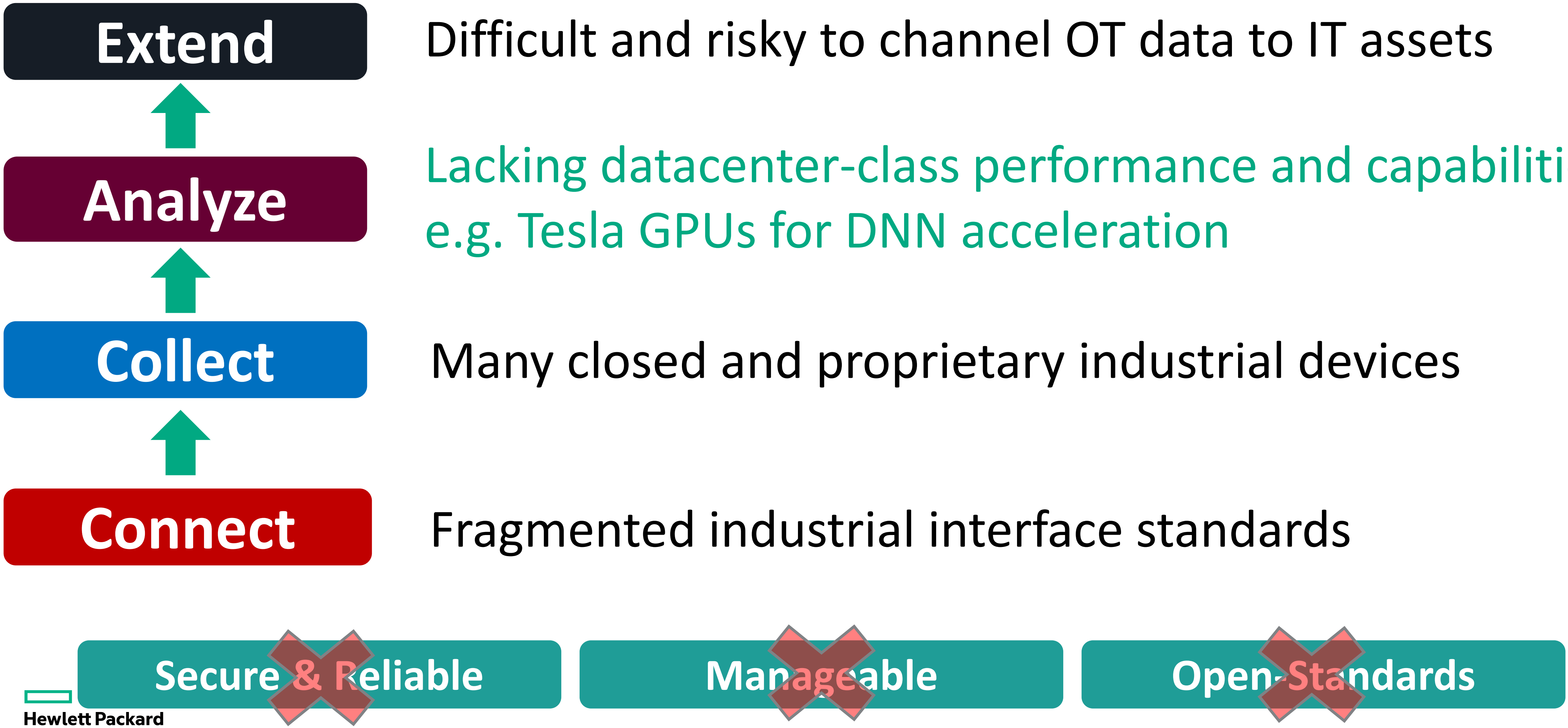
Distributed Telco, Media, Comms

Multi-Access Edge Computing (MEC), *5G, vRAN, vCPE, Edge CDN/Caching,* ***Video Transcoding***

Hewlett Packard Enterprise

© Copyright 2018 Hewlett-Packard Enterprise Corporation. The information contained herein is subject to change without notice.

Challenges at the Operations Edge



Extend — Difficult and risky to channel OT data to IT assets

Analyze — Lacking datacenter-class performance and capabilities e.g. Tesla GPUs for DNN acceleration

Collect — Many closed and proprietary industrial devices

Connect — Fragmented industrial interface standards

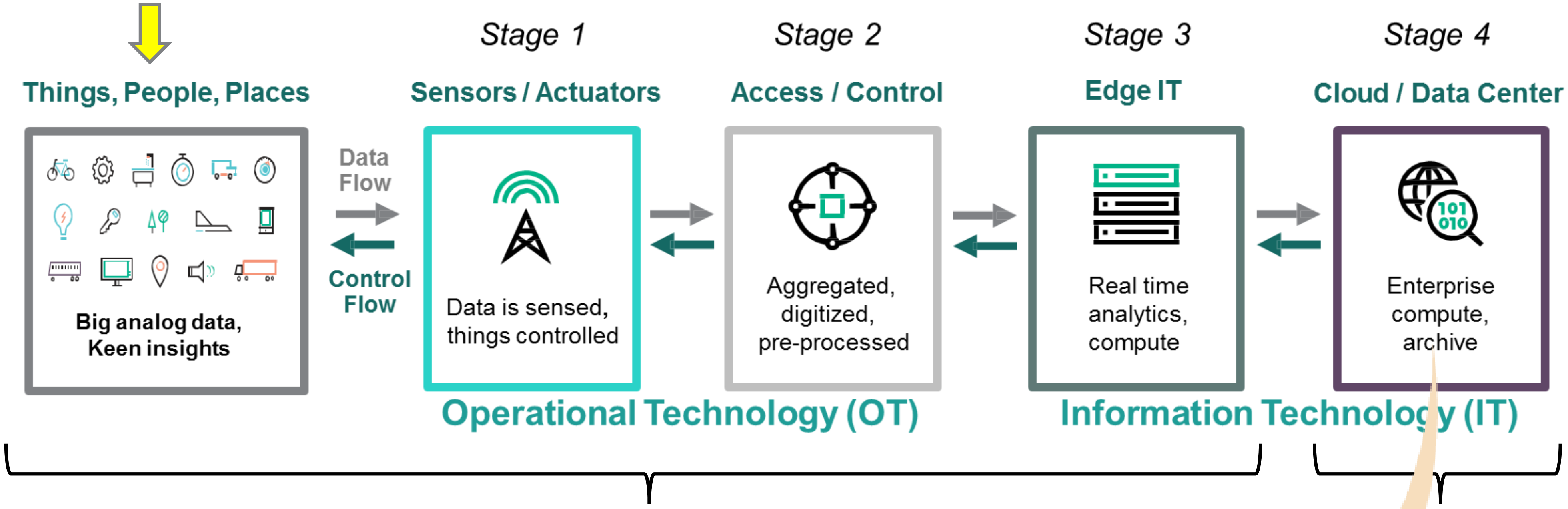
~~Secure & Reliable~~ | ~~Manageable~~ | ~~Open-Standards~~

Hewlett Packard Enterprise

© Copyright 2018 Hewlett-Packard Enterprise Corporation. The information contained herein is subject to change without notice.

Introducing the “Converged Edge System”

Emerging source of “Big Data”



Stage 1 · *Stage 2* · *Stage 3* · *Stage 4*

Things, People, Places

Big analog data,
Keen insights

Sensors / Actuators

Data is sensed,
things controlled

Access / Control

Aggregated,
digitized,
pre-processed

Edge IT

Real time
analytics,
compute

Cloud / Data Center

Enterprise
compute,
archive

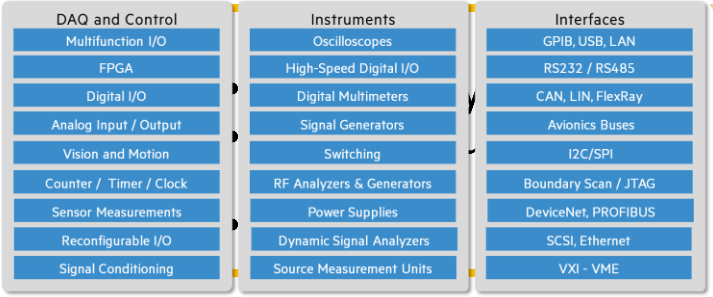
Operational Technology (OT)

Information Technology (IT)

The Edge

Data Center / Cloud

“Converge OT”



DAQ and Control	Instruments	Interfaces
Multifunction I/O	Oscilloscopes	GPIB, USB, LAN
FPGA	High-Speed Digital I/O	RS232 / RS485
Digital I/O	Digital Multimeters	CAN, LIN, FlexRay
Analog Input / Output	Signal Generators	Avionics Buses
Vision and Motion	Switching	I2C/SPI
Counter / Timer / Clock	RF Analyzers & Generators	Boundary Scan / JTAG
Sensor Measurements	Power Supplies	DeviceNet, PROFIBUS
Reconfigurable I/O	Dynamic Signal Analyzers	SCSI, Ethernet
Signal Conditioning	Source Measurement Units	VXI - VME

Converged Edge Systems

Pioneered by HPE!

“Enterprise class IT”

IT :

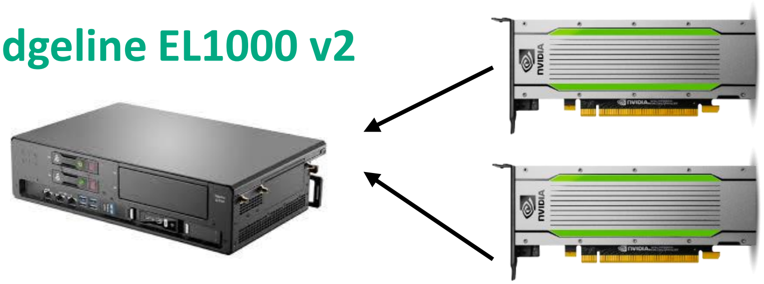
High performance compute
- High capacity storage
- Systems management

Hewlett Packard
Enterprise

© Copyright 2018 Hewlett-Packard Enterprise Corporation. The information contained herein is subject to change without notice.

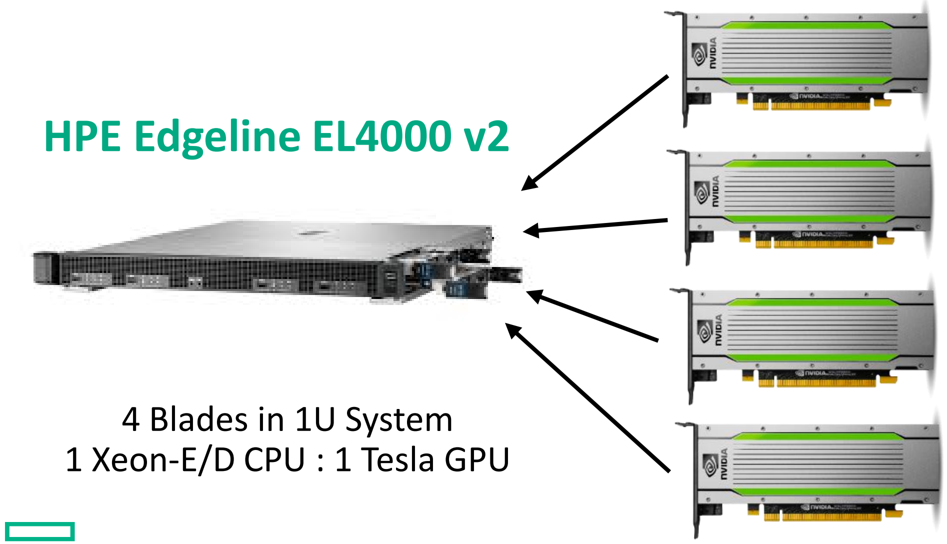
HPE Edgeline – Unleashing AI on “Big Data” problems at the Edge

HPE Edgeline EL1000 v2



1 Blade per System
1 Xeon-E/D CPU : 2 Tesla GPUs

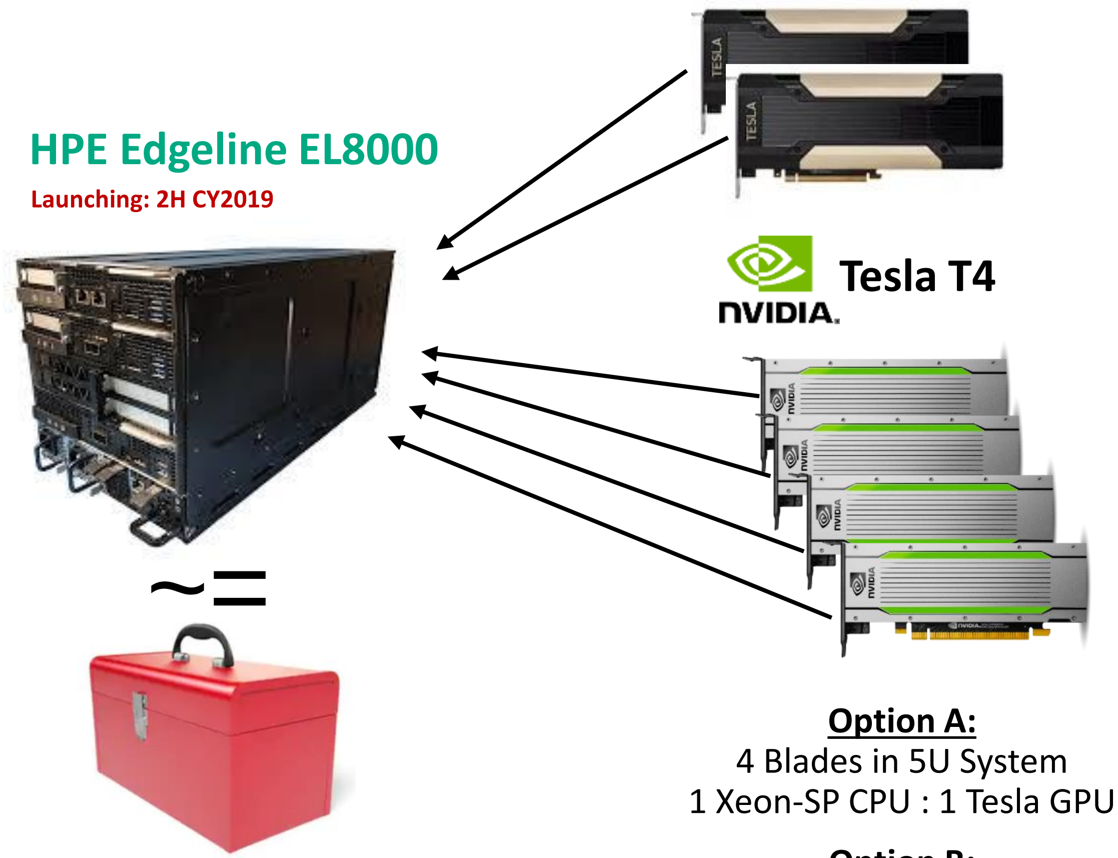
HPE Edgeline EL4000 v2



4 Blades in 1U System
1 Xeon-E/D CPU : 1 Tesla GPU

HPE Edgeline EL8000

Launching: 2H CY2019



222mm D x 220mm W x 431mm H
(17.5” D x 8.7” W x 16.9” H)

Rugged & Compact Enterprise-class Edge System
Perfect for the Operations Edge!

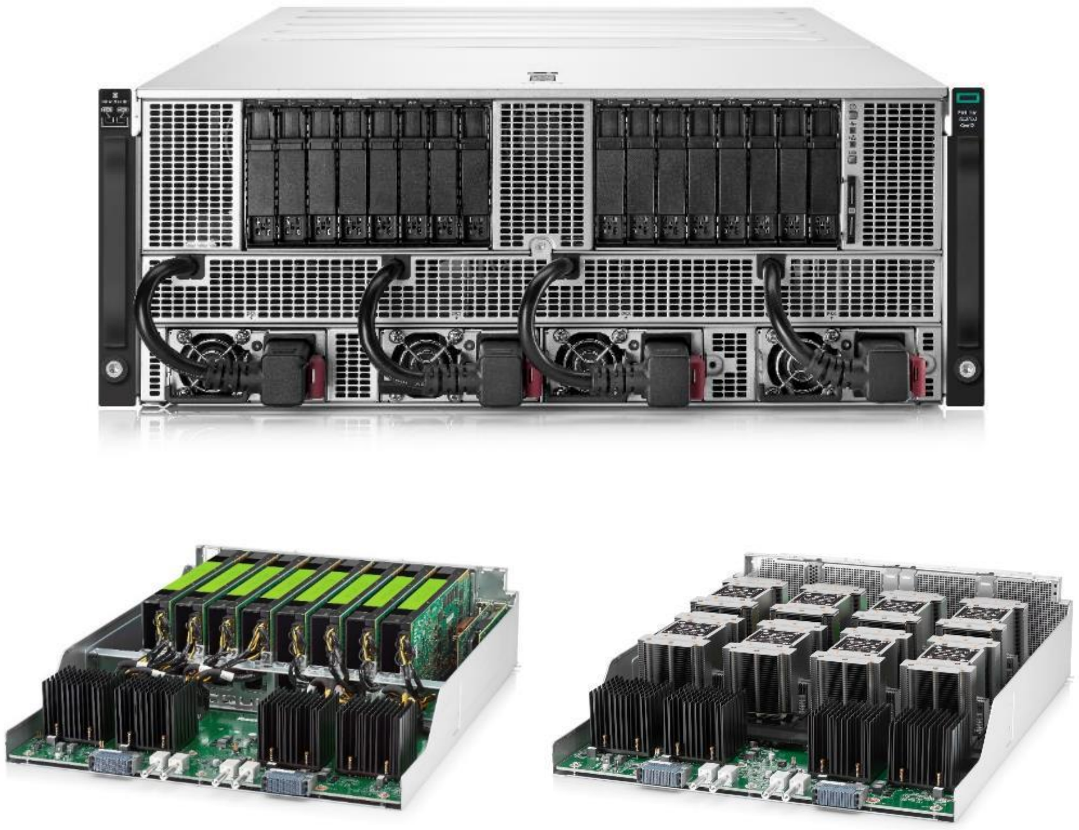
Option A:
4 Blades in 5U System
1 Xeon-SP CPU : 1 Tesla GPU

Option B:
2 Blades in 5U System
1 Xeon-SP CPU : 1-4 Tesla GPUs

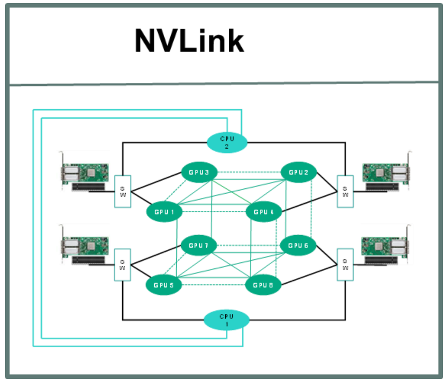
Option C:
Mix and Match

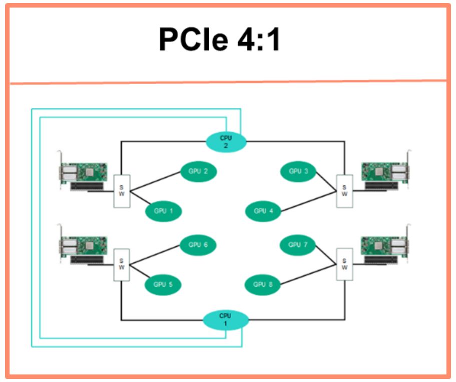
© Copyright 2018 Hewlett-Packard Enterprise Corporation. The information contained herein is subject to change without notice.

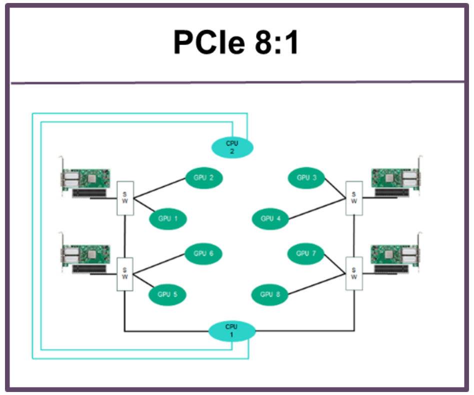
HPE Apollo 6500 Gen10 – Enterprise AI acceleration at the Core



Tesla V100 (PCIe & SMX2)
Tesla T4
Quadro RTX 6000 & RTX 8000

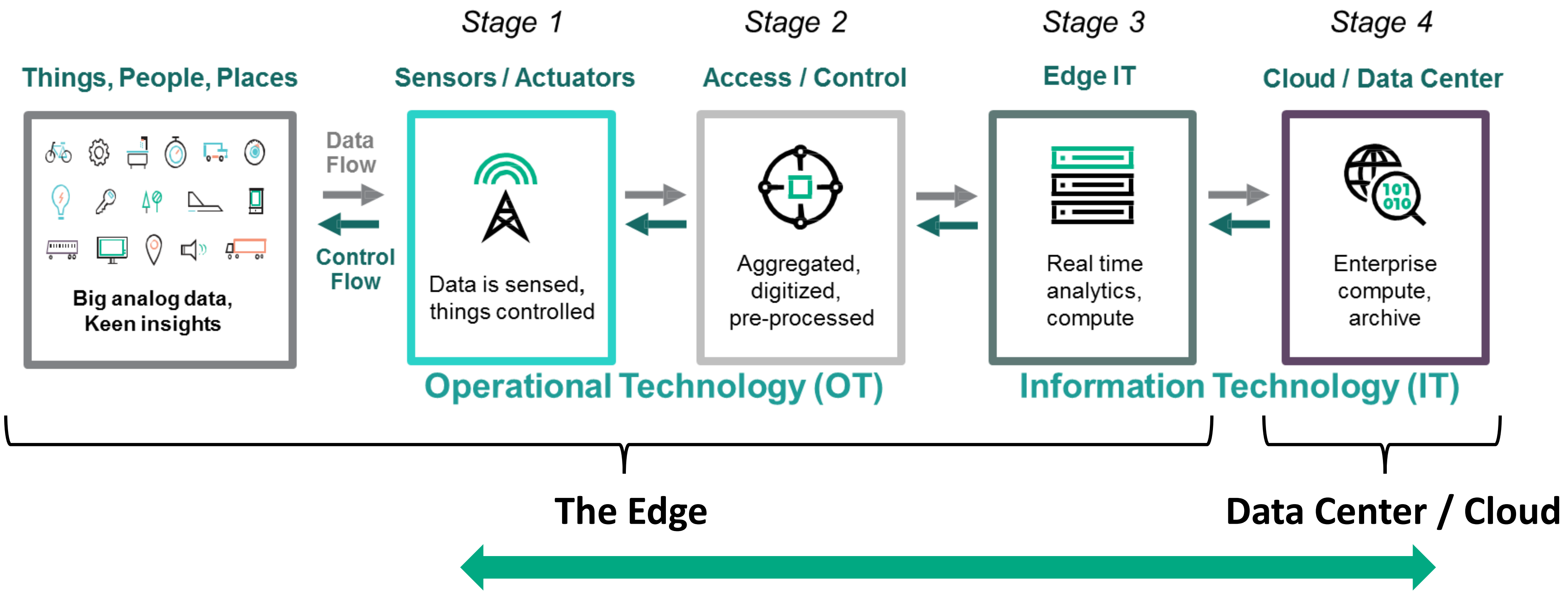

NVLink
CPU 2
GPU 3
GPU 2
SW
GPU 1
GPU 4
GPU 7
GPU 6
GPU 5
GPU 8
CPU 1


PCIe 4:1
CPU 2
GPU 2
GPU 3
SW
GPU 1
GPU 4
GPU 6
GPU 7
GPU 5
GPU 8
CPU 1


PCIe 8:1
CPU 2
GPU 2
GPU 3
SW
GPU 1
GPU 4
GPU 6
GPU 7
GPU 5
GPU 8
CPU 1

© Copyright 2018 Hewlett-Packard Enterprise Corporation. The information contained herein is subject to change without notice.

Technologies that can scale from Edge to Cloud



Things, People, Places

Big analog data, Keen insights

Data Flow

Control Flow

Stage 1

Sensors / Actuators

Data is sensed, things controlled

Stage 2

Access / Control

Aggregated, digitized, pre-processed

Stage 3

Edge IT

Real time analytics, compute

Stage 4

Cloud / Data Center

Enterprise compute, archive

Operational Technology (OT)

Information Technology (IT)

The Edge

Data Center / Cloud

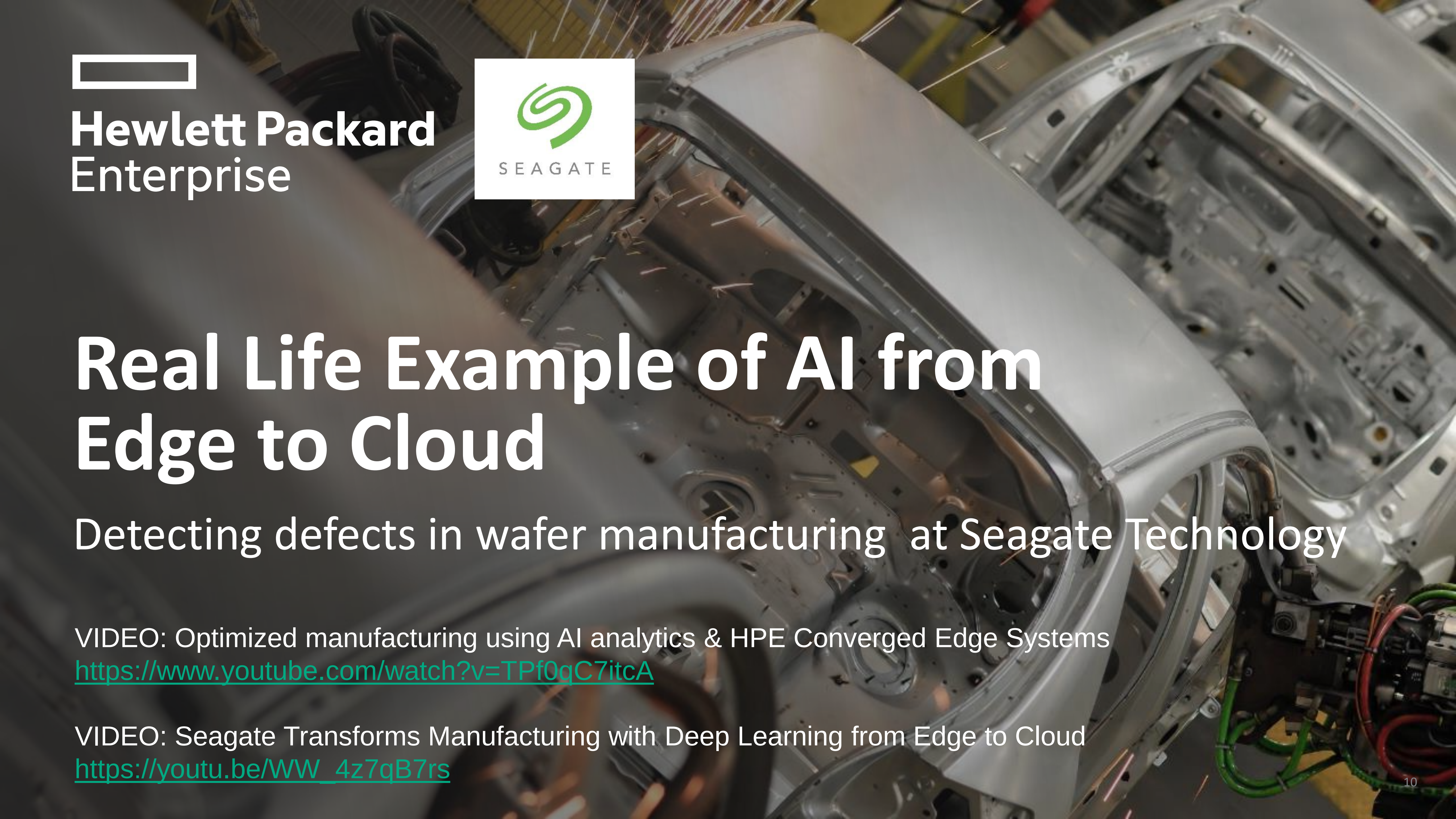
Vision of Edge-to-Cloud

- Identical Capabilities including Accelerators
- Identical Performance Levels
- Identical Software Stacks
- Management with a single-pane-of-glass
- Secure everywhere

Hewlett Packard Enterprise

Hewlett Packard Enterprise

© Copyright 2018 Hewlett-Packard Enterprise Corporation. The information contained herein is subject to change without notice.



Hewlett Packard
Enterprise

SEAGATE

Real Life Example of AI from Edge to Cloud

Detecting defects in wafer manufacturing at Seagate Technology

VIDEO: Optimized manufacturing using AI analytics & HPE Converged Edge Systems
https://www.youtube.com/watch?v=TPf0qC7itcA

VIDEO: Seagate Transforms Manufacturing with Deep Learning from Edge to Cloud
https://youtu.be/WW_4z7qB7rs

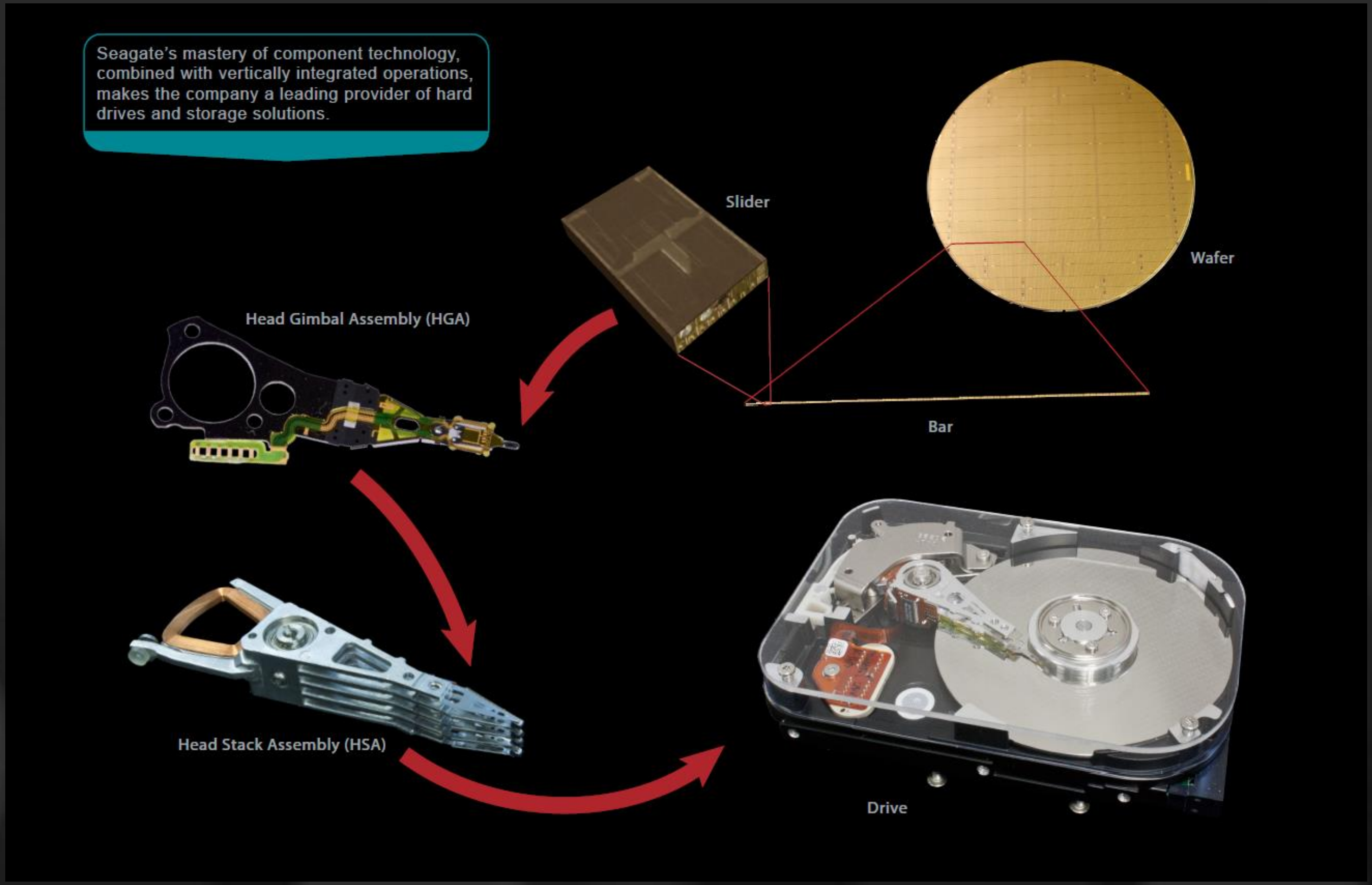
From Wafers to Heads

Each 200 mm wafer
contains 10,000’s of sliders.

A slider is also called a recording transducer

We make 1 Billion transducers per year

Operations in Normandale, Springtown, and in Korat, Wuxi, Singapore



Seagate’s mastery of component technology, combined with vertically integrated operations, makes the company a leading provider of hard drives and storage solutions.

Slider
Wafer
Head Gimbal Assembly (HGA)
Bar
Head Stack Assembly (HSA)
Drive

Data Generation and Analytics Constraints

Millions of Electron Microscope & Optical images per day

Real time sensor data from 100’s of manufacturing tools

Real Time Insights & Automated Controls demand low latency

High bandwidth costs (10TB/day) precludes public cloud

ML Training demands large datasets near GPU

Seagate on Edge
Project Athena

Edge Analytics in the Factory: Enable real time AI & ML

Challenge

Seagate factory's produce Billion+ recording transducers per year. ML/AI lab projects consistently show entitlement. How to scale low latency ML Inference?

- High data volume millions of images & thousands of factory tool sensors
- Long latencies and bandwidth cost to use Cloud analytics
- Need for real time analytics

Solution

Apply Machine Learning to transducer images and sensor data to find anomaly's and defects early and reliably.

- Seagate All-Flash for Training Data Storage
- NVidia GPU and TensorFlow

- Future: Real Time Inference at the Edge

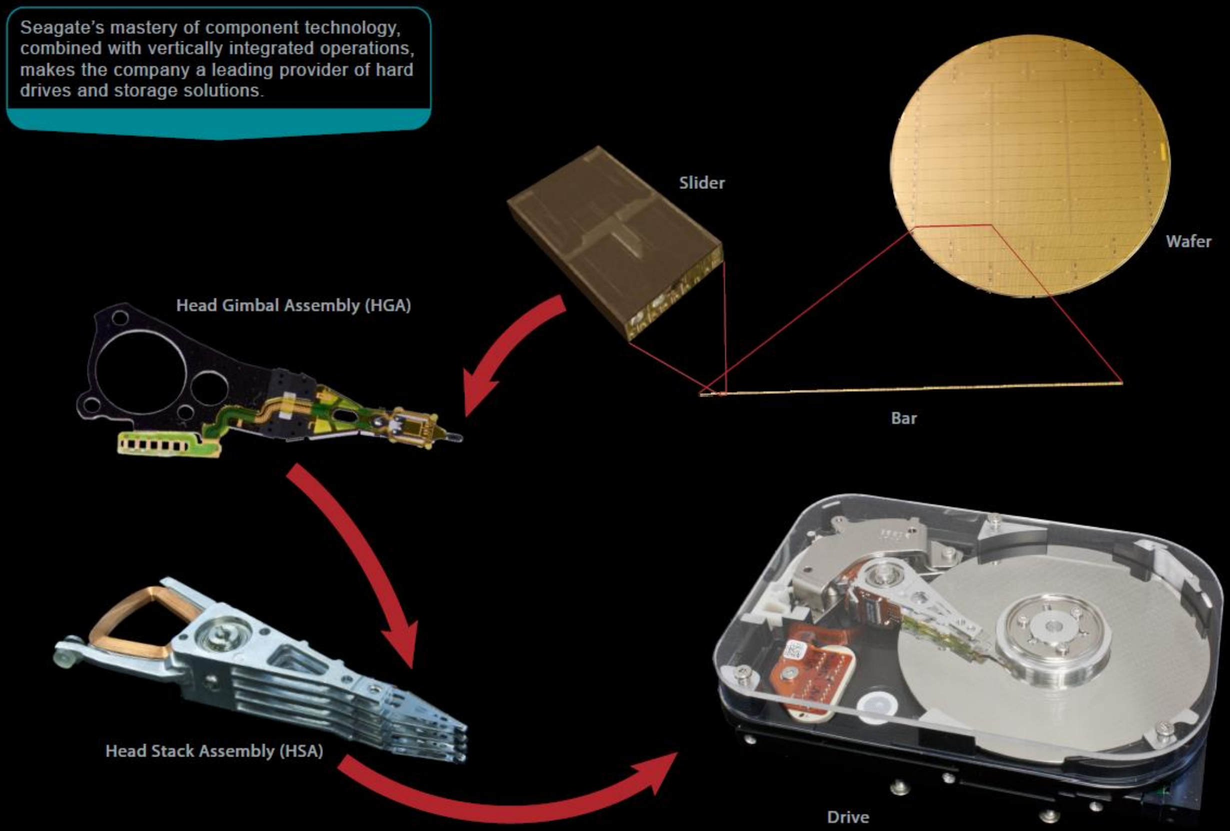
Benefits

Reduced test time, increased efficiency and improve throughput.

- Reduced Bandwidth Costs
- Faster test cycles
- Increased Utilization

Athena Pilot Inference Dataflow

Wafer Process: ~ 100 SEM or Optical steps

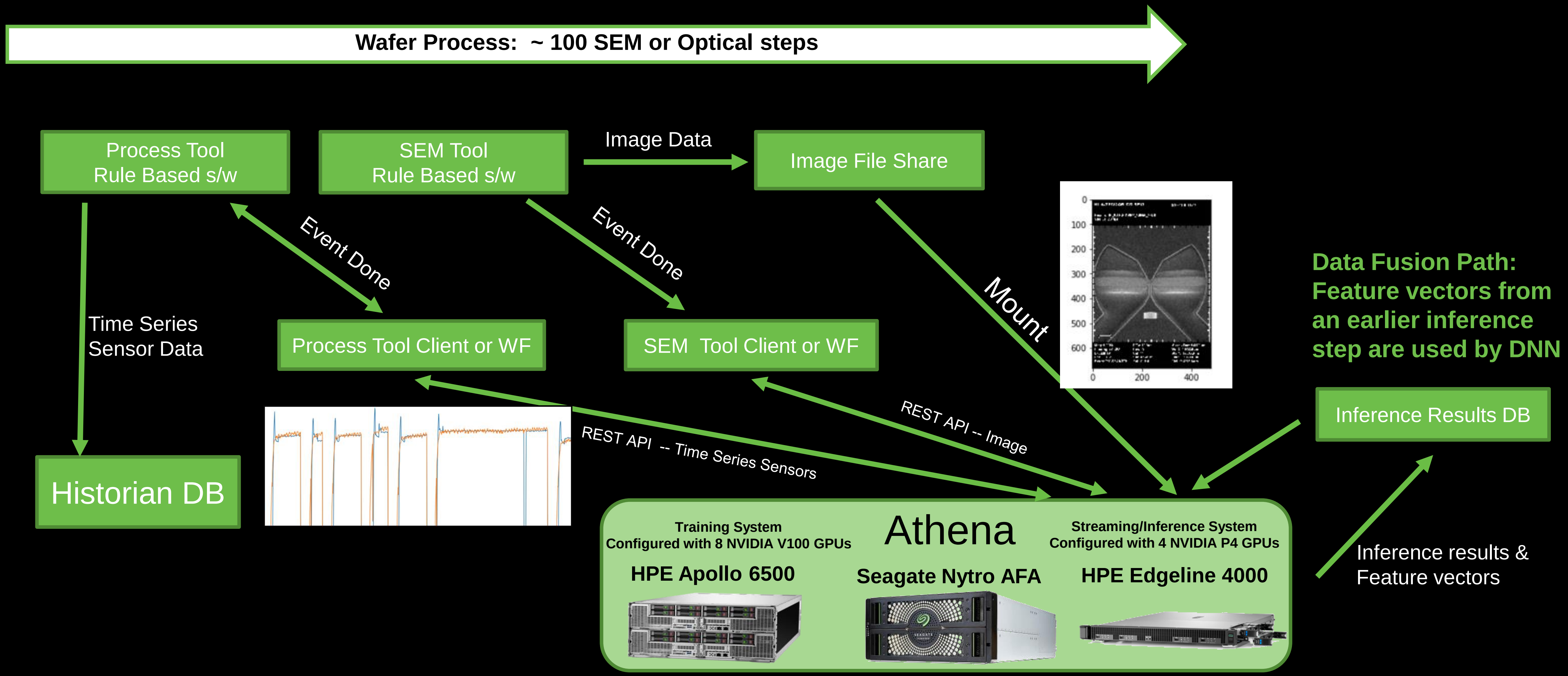


Seagate's mastery of component technology, combined with vertically integrated operations, makes the company a leading provider of hard drives and storage solutions.

Slider
Wafer
Head Gimbal Assembly (HGA)
Bar
Head Stack Assembly (HSA)
Drive

Seagate Confidential

Athena Pilot Inference Dataflow



Wafer Process: ~ 100 SEM or Optical steps

Process Tool
Rule Based s/w

SEM Tool
Rule Based s/w

Image Data

Image File Share

Event Done

Event Done

Mount

Time Series
Sensor Data

Process Tool Client or WF

SEM Tool Client or WF

Data Fusion Path: Feature vectors from an earlier inference step are used by DNN

Inference Results DB

REST API -- Image

REST API -- Time Series Sensors

Historian DB

Athena

Training System
Configured with 8 NVIDIA V100 GPUs

HPE Apollo 6500

Seagate Nytro AFA

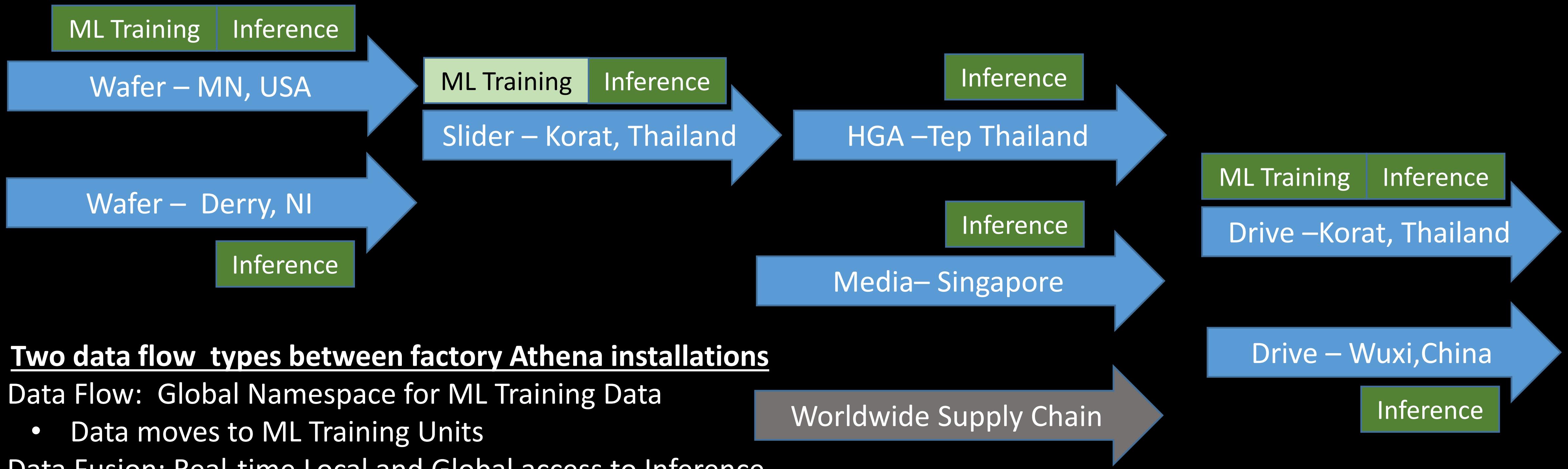
Streaming/Inference System
Configured with 4 NVIDIA P4 GPUs

HPE Edgeline 4000

Inference results &
Feature vectors

Seagate on Edge
Project Athena

Machine Learning Data Flow through the Vertical Supply Chain



ML Training | Inference

Wafer – MN, USA

Wafer – Derry, NI

Inference

ML Training | Inference

Slider – Korat, Thailand

Inference

HGA –Tep Thailand

Inference

Media– Singapore

Worldwide Supply Chain

ML Training | Inference

Drive –Korat, Thailand

Drive – Wuxi,China

Inference

Two data flow types between factory Athena installations

- Data Flow: Global Namespace for ML Training Data
  - Data moves to ML Training Units
- Data Fusion: Real-time Local and Global access to Inference results at successive Inference points
- Trained Model Files flow from ML Training to Inference

Two Athena version: with and without ML Training Hardware (HPE Apollo 6500)

Seagate on Edge
Project Athena

Edge Analytics enables new capabilities – The outlook for 2020

Expert Assistant

SME trains AI/ML Agent to help do work that augments human abilities

Interactive ML tools will allow Machines to learn from SME’s

SME Assistant will find “interesting” relationships and anomaly’s in huge datasets.

Machine assisted AI/ML model building

Data Fusion/Digital Twins

Image, Time-Series Sensor, etc. data from across the vertical supply chain is integrated for ML/AI

Pervasive AI/ML Athena Real-time Edge Inference

“Fuse” Results from Local Edge Inference

Global storage of inference results

System Optimization

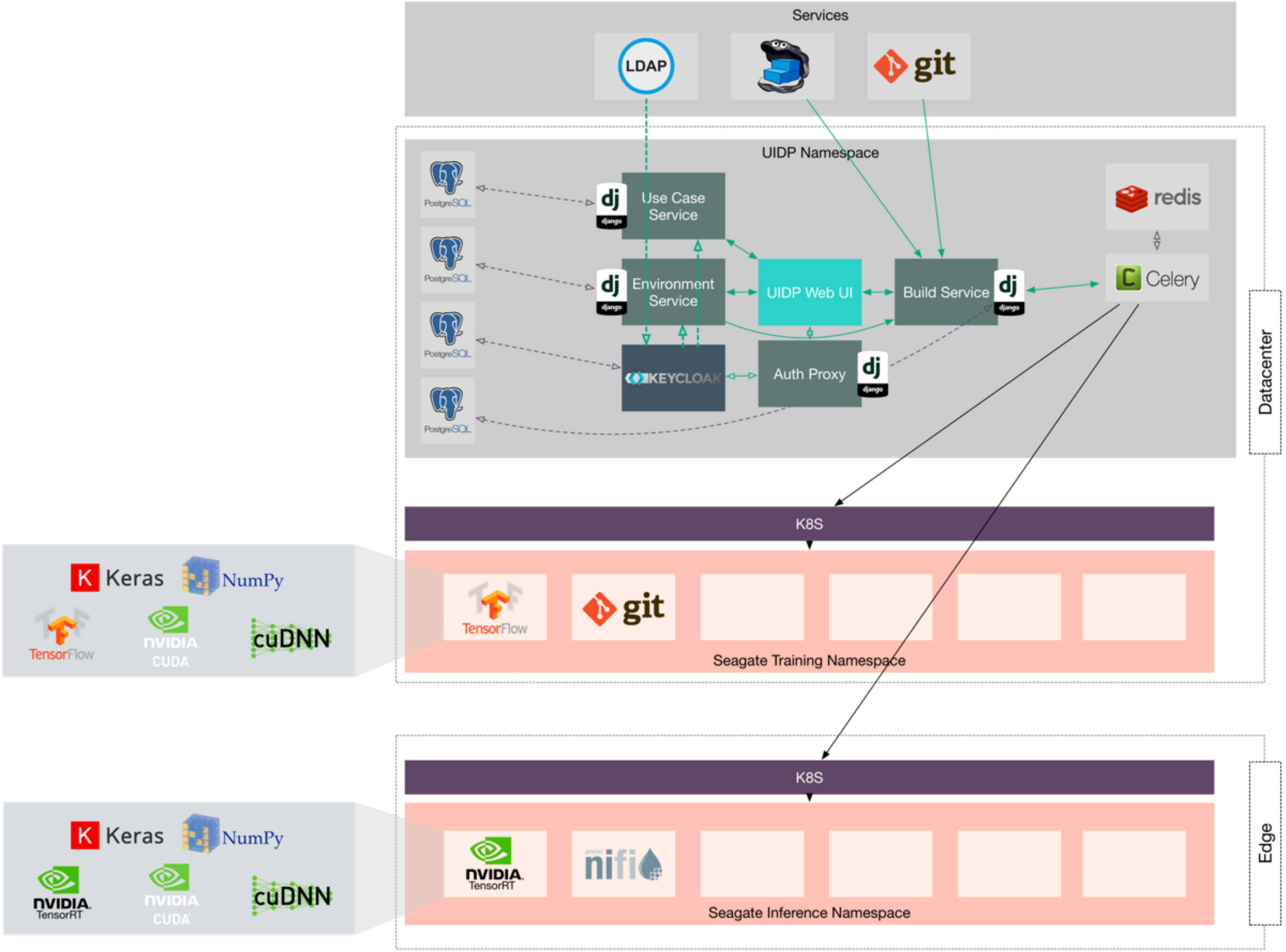
ML/AI Applied to Supply-Chain “Fused Data”

Reduce cost by replacing “defense in depth” with systemic understanding

Improve performance and quality

Recognize autonomous car or robotics analogy

Example Use Case Architecture


Services
LDAP
git
UIDP Namespace
PostgreSQL
PostgreSQL
PostgreSQL
PostgreSQL
dj django
Use Case Service
dj django
Environment Service
UIDP Web UI
Build Service
dj django
redis
Celery
KEYCLOAK
Auth Proxy
dj django
Datacenter
K8S
TensorFlow
git
Seagate Training Namespace
Keras
NumPy
TensorFlow
NVIDIA CUDA
cuDNN
K8S
NVIDIA TensorRT
nifi
Seagate Inference Namespace
Edge
Keras
NumPy
NVIDIA TensorRT
NVIDIA CUDA
cuDNN

Hewlett Packard Enterprise

SEAGATE

Thank You