NetApp

NetApp ONTAP AI & HCI

全方位人工智慧及超融合解決方案

Major Chuang
Sr. Solution Engineer

2018.10.12

Agenda

1) NetApp Data Fabric
2) NetApp ONTAP AI
3) NetApp HCI

NetApp & Data Fabric

Our Purpose

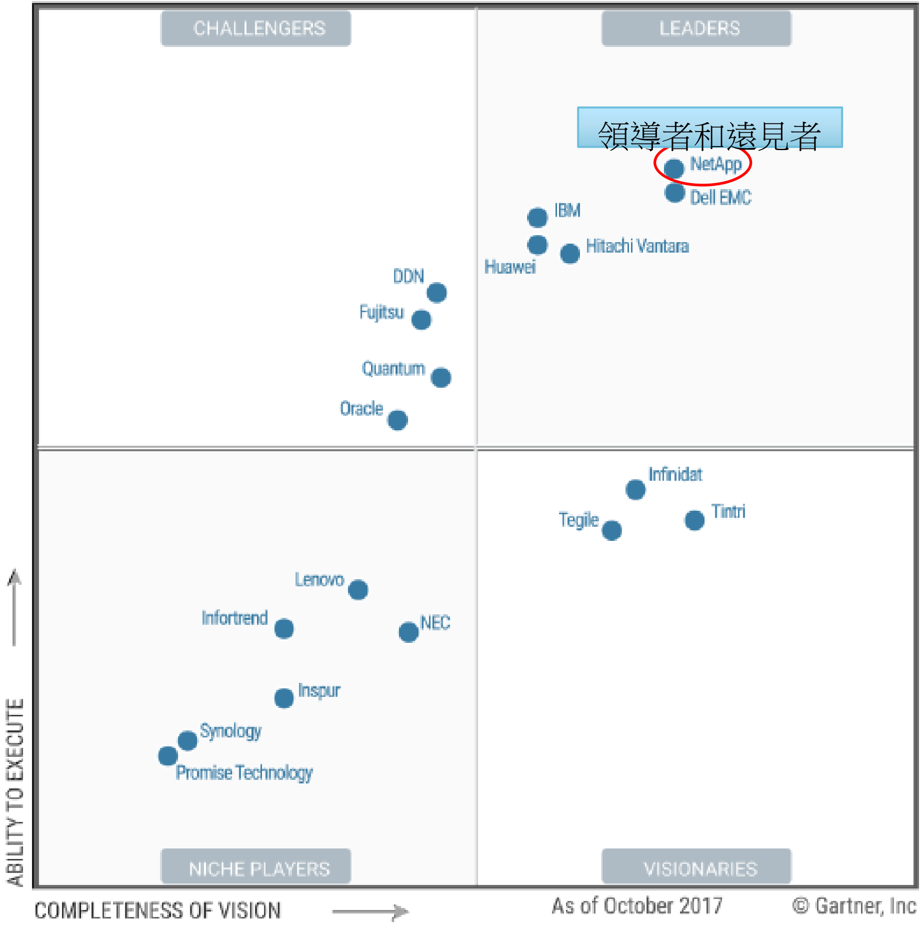
Empowering our customers to change the world with data

2017 Gartner 通用儲存魔力象限

NetApp 再次被 Gartner 2017 Magic Quadrant 的通用儲存磁碟陣列評為領導者

Magic Quadrant

Figure 1. Magic Quadrant for General-Purpose Disk Arrays

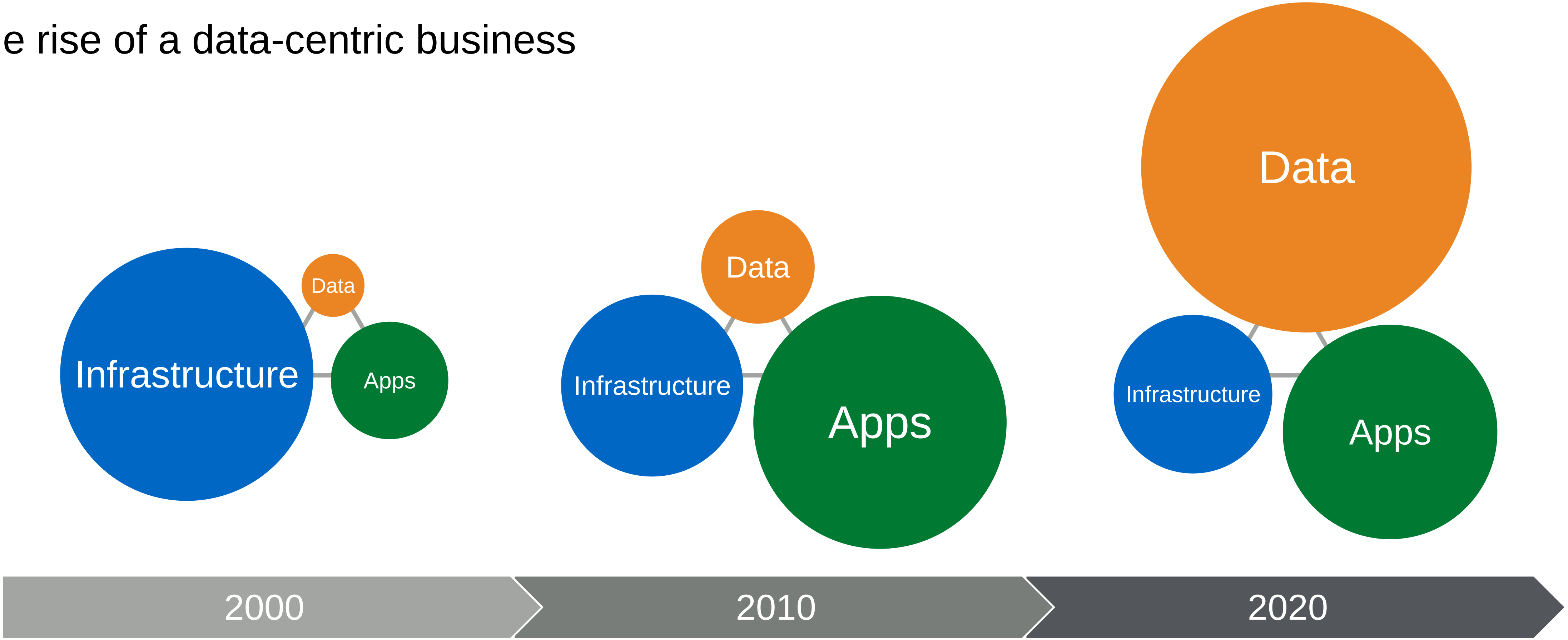


Source: Gartner (October 2017)

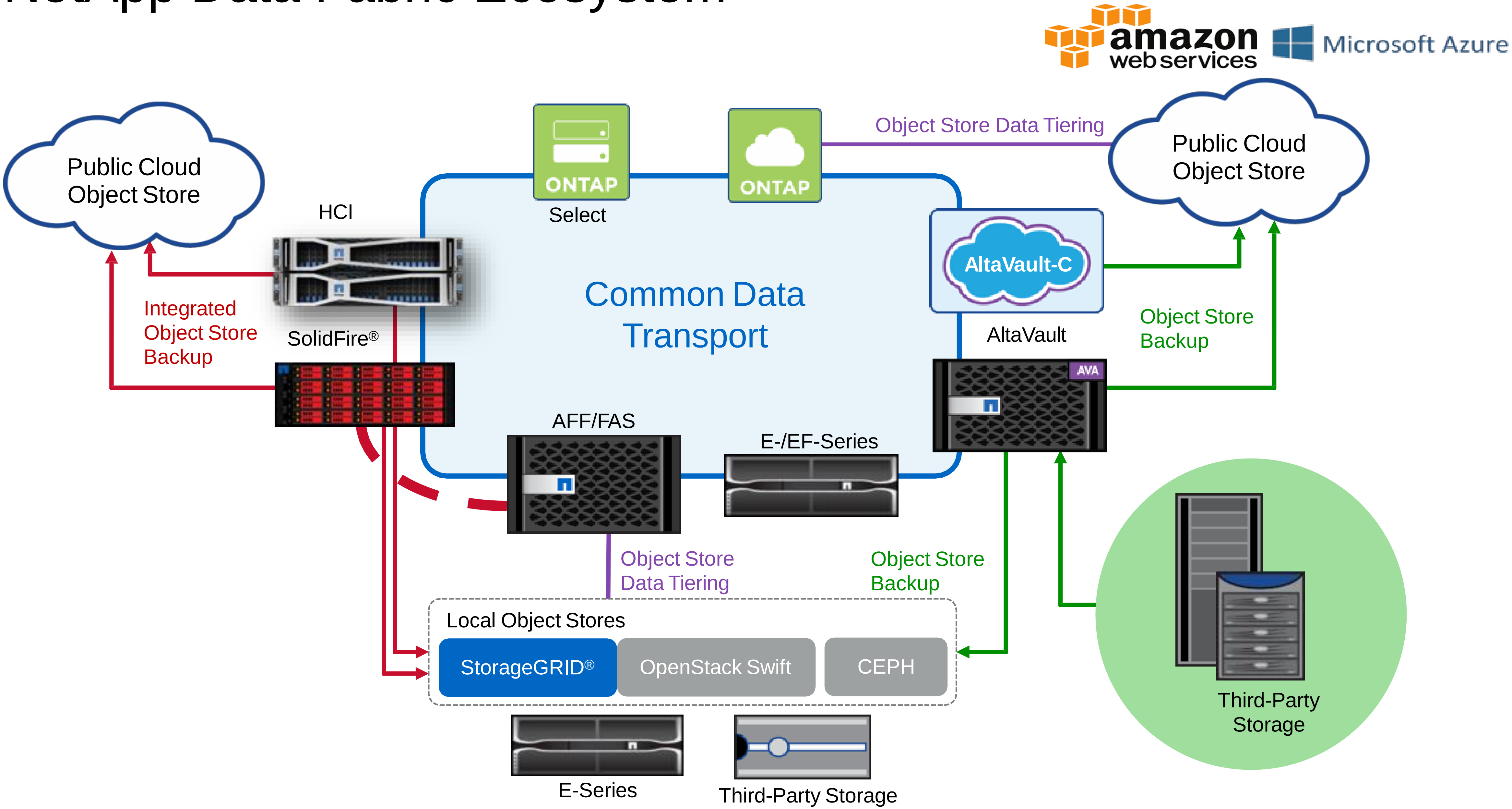
- 2017 年10月31日，Gartner 在全球發佈《通用磁碟陣列存儲魔力象限分析報告》
- NetApp 再一次位居領導者與遠見者位置
- Gartner :「NetApp 利用 Data Fabric 願景成功提升了敏捷靈活性，從存儲與資料管理角度説明客戶探索混合雲潛力……。」

The unstoppable drive toward data management

- The rise of a data-centric business



The NetApp Data Fabric Ecosystem



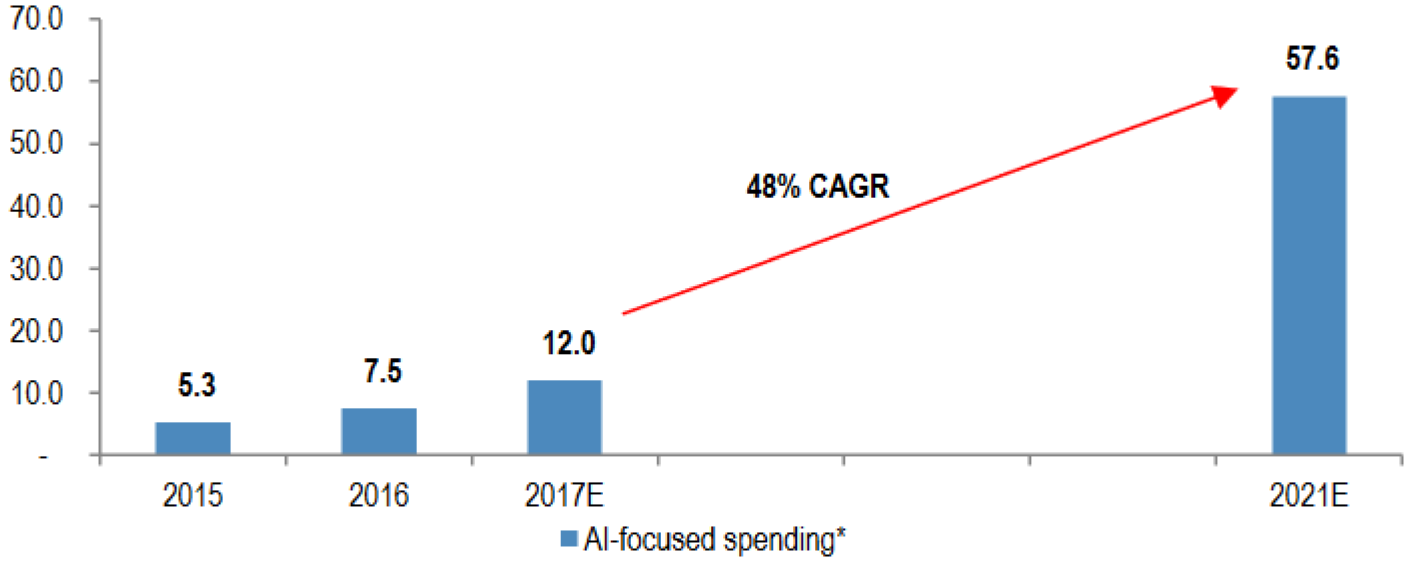
NetApp

NetApp ONTAP AI

Global AI Focused Spending

Big spend put there but most are “invisible” strategic initiatives

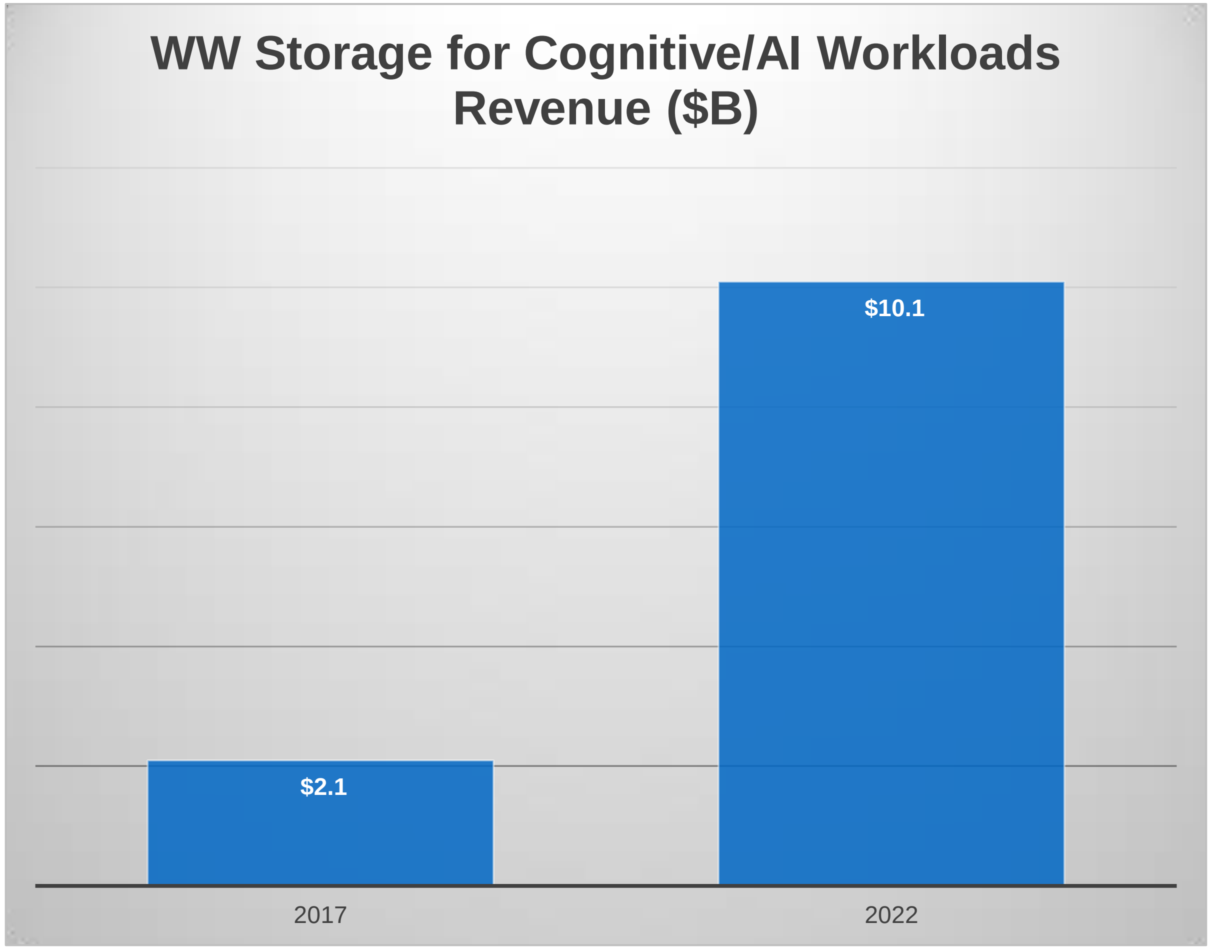
Figure 2: Global AI-focused spending* ($, bn)

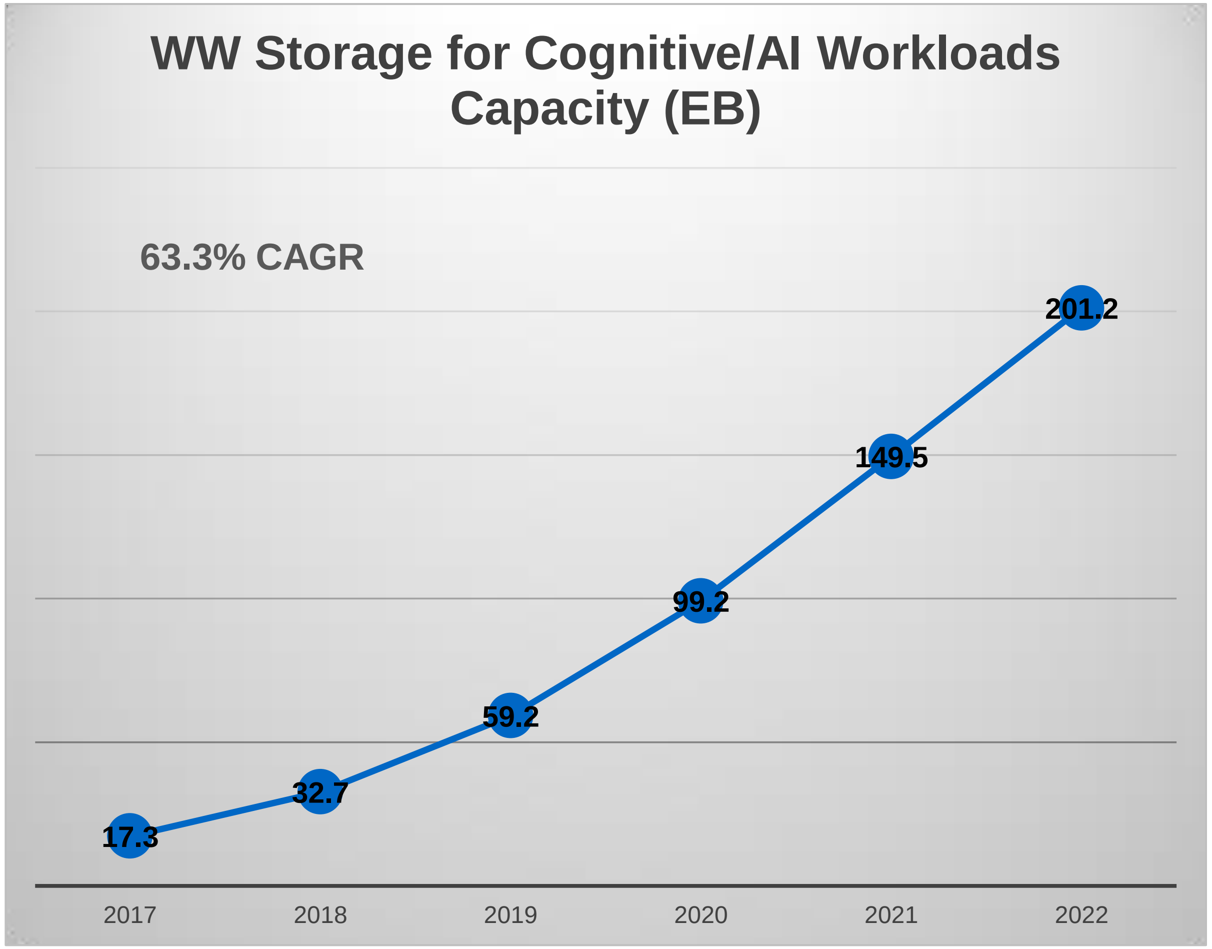


Source: AI-spending estimates from IDC. *Includes AI-focused spending on hardware, software (applications + software platforms), and services (IT consulting & system implementation).

Storage Revenue and Capacity

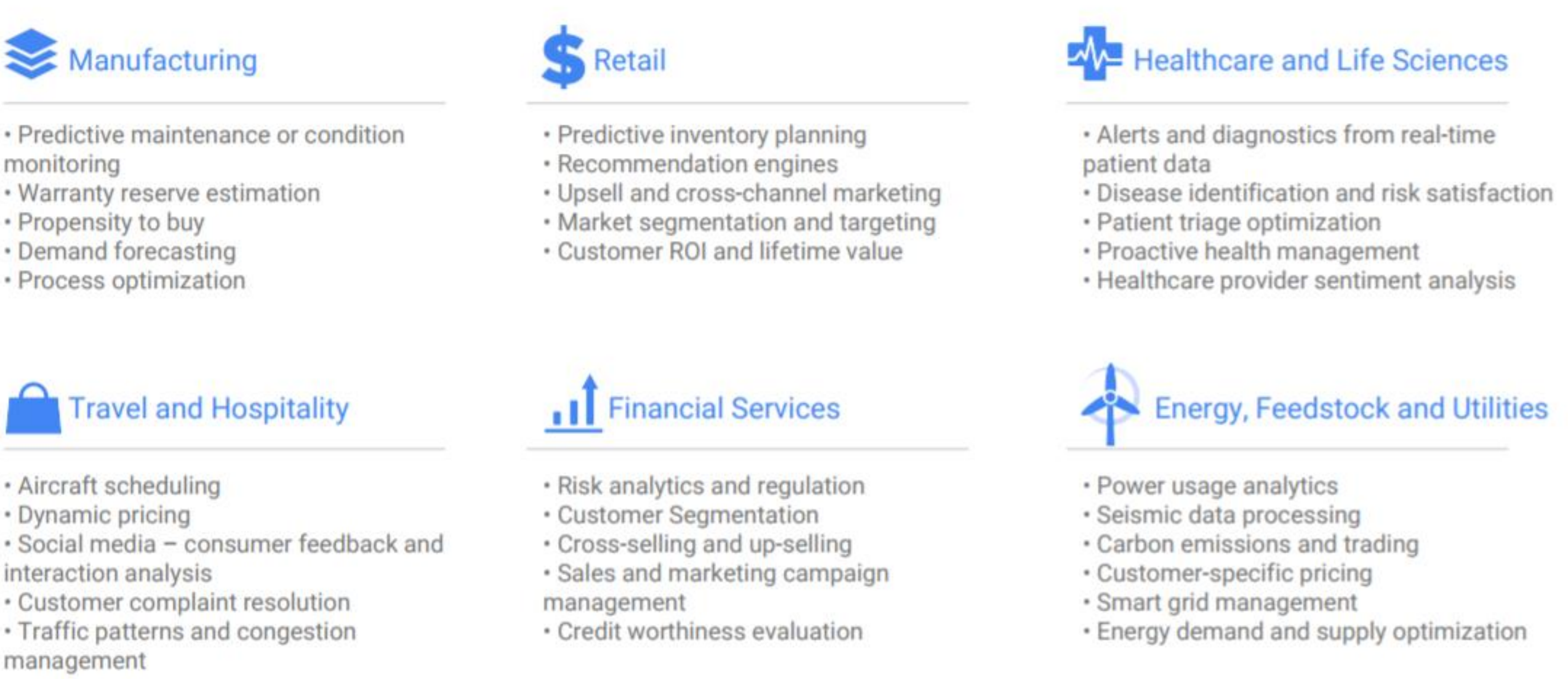
Cognitive/AI workloads





Source: IDC WW Storage for Cognitive/AI Workloads Forecast, 2017-2022

AI/ML Use Cases



Manufacturing

- Predictive maintenance or condition monitoring
- Warranty reserve estimation
- Propensity to buy
- Demand forecasting
- Process optimization

Retail

- Predictive inventory planning
- Recommendation engines
- Upsell and cross-channel marketing
- Market segmentation and targeting
- Customer ROI and lifetime value

Healthcare and Life Sciences

- Alerts and diagnostics from real-time patient data
- Disease identification and risk satisfaction
- Patient triage optimization
- Proactive health management
- Healthcare provider sentiment analysis

Travel and Hospitality

- Aircraft scheduling
- Dynamic pricing
- Social media – consumer feedback and interaction analysis
- Customer complaint resolution
- Traffic patterns and congestion management

Financial Services

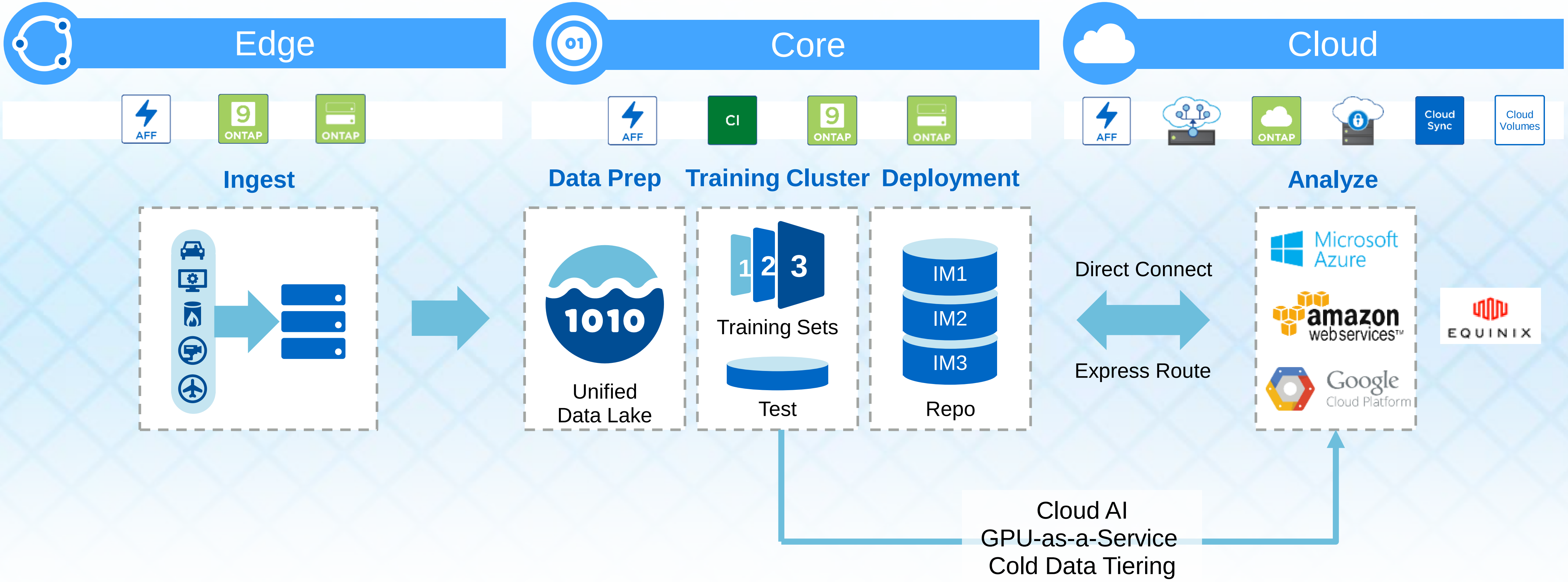
- Risk analytics and regulation
- Customer Segmentation
- Cross-selling and up-selling
- Sales and marketing campaign management
- Credit worthiness evaluation

Energy, Feedstock and Utilities

- Power usage analytics
- Seismic data processing
- Carbon emissions and trading
- Customer-specific pricing
- Smart grid management
- Energy demand and supply optimization

NetApp Edge to Core to Cloud Data Pipeline

Future-proof and ultra-high-performance



Edge

Core

Cloud

Ingest

Data Prep

Training Cluster

Deployment

Analyze

Unified Data Lake

Training Sets

Test

Repo

Direct Connect

Express Route

Cloud AI
GPU-as-a-Service
Cold Data Tiering

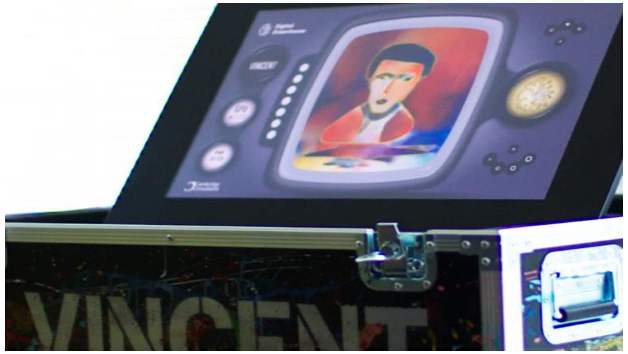
NVIDIA DGX-1 with NetApp

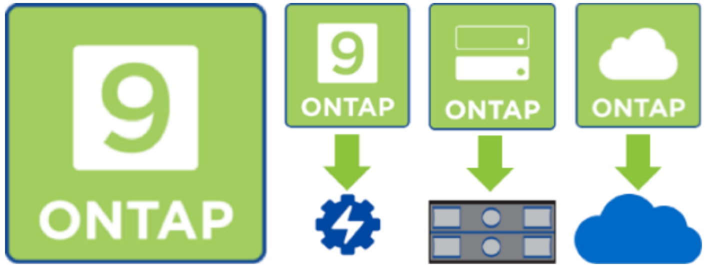
Digital Greenhouse

“Vincent,” is a breakthrough in machine learning that completes a drawing started with a human sketch. Completed ‘works of art’ combine a user’s sketch with the digested sum of art since the renaissance, as if Van Gogh, Cézanne, and Picasso were inside the machine, producing art to order.

Digital Greenhouse by Cambridge Consultants.

Built using NetApp ONTAP with highly scalable NFS single namespace and NVIDIA DGX-1.





NetApp ONTAP AI

Accelerate your AI data pipeline for deep learning

SIMPLE

Eliminates design complexity and guesswork

Partners deliver complete solution

INTEGRATED

Intelligently manage your data across Edge, Core & Cloud

Deploy AI Frameworks with confidence

TensorFlow Chainer Caffe2 PYTORCH Caffe mxnet PaddlePaddle

POWERFUL

Scale without limits

Start small and grow non-disruptively

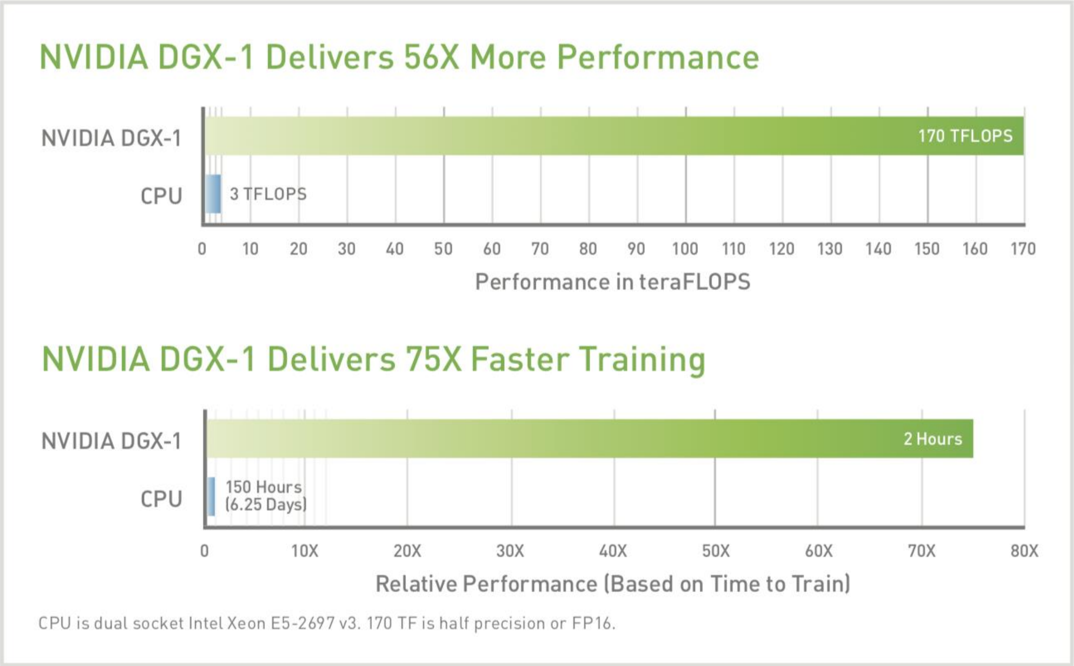
NVIDIA DGX-1 GPU AI Compute Platform

Supercomputer in a box

The NVIDIA® DGX-1TM is the world’s first purpose-built system optimized for deep learning, with fully integrated hardware and software that can be deployed quickly and easily. The revolutionary performance of the DGX-1TM significantly accelerates training time, making it the world’s first deep learning supercomputer in a box.

Replace 400 traditional servers with 1 DGX-1

Replace 800 traditional servers with 1 DGX-2

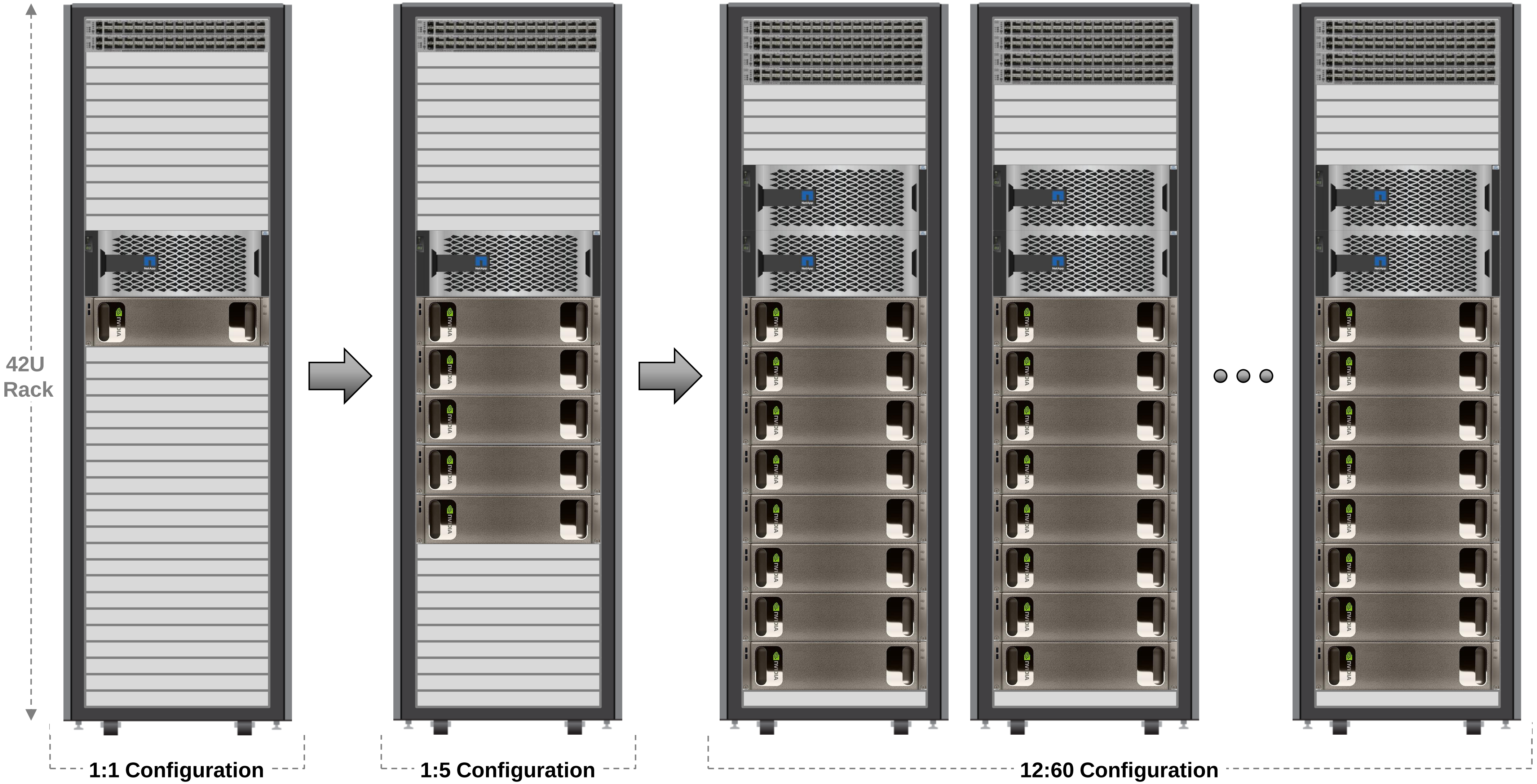

NVIDIA DGX-1 Delivers 56X More Performance

NVIDIA DGX-1 Delivers 75X Faster Training

CPU is dual socket Intel Xeon E5-2697 v3. 170 TF is half precision or FP16.

GPU:

NetApp Rack Scale AI
Scale from 1:1 to 12:60 Storage:AI Config



42U Rack

1:1 Configuration

1:5 Configuration

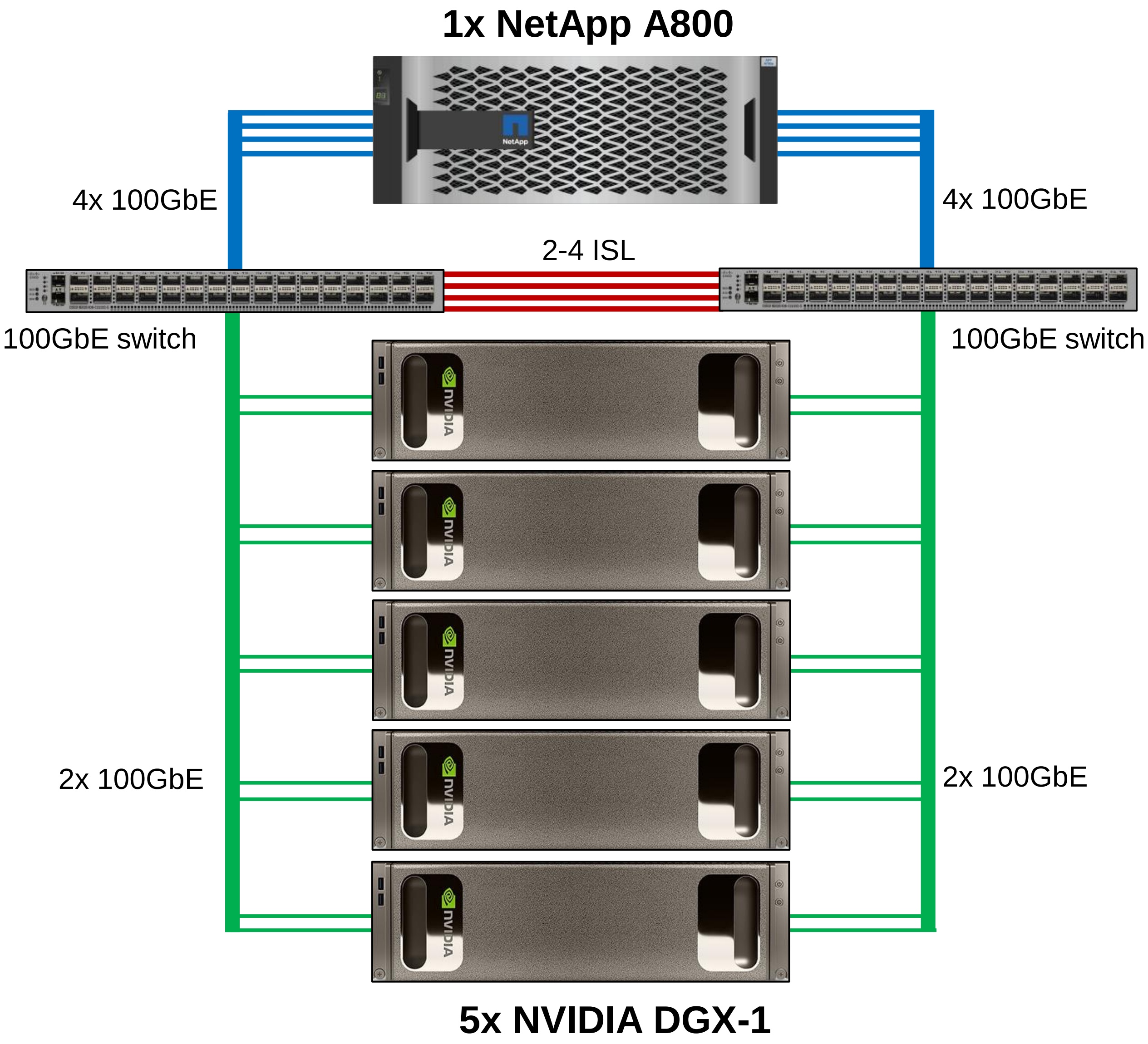
12:60 Configuration

* Based on 35kW racks

* Based on performance requirements, DL model used, size of datasets the compute:storage ratio can change

NetApp

1:5 configuration* - Network connectivity



** Based on performance requirements, DL model used, size of datasets the compute:storage ratio can change*

NetApp

AFF A800: The World’s Fastest Data Platform for AI

500µs
latency

1M
IOPS

25GB/s
throughput

24-node Cluster

11.4M IOPS

300GB/s throughput; **4x** higher than competitor

AI/DL Training : Start Small, Scale Big

Configurations with A800/A700s and NVIDIA DGX-1

# of A800 Storage Systems	# of DGX-1 Servers	Throughput	Images/Sec	Typical Raw Capacity	Raw Capacity w/ Expansion
1 HA pair	5	25GB/s	250K	364.8TB	6.2PB
4 HA pairs	20	100GB/s	1000K	1.5PB	24.8PB
12 HA pairs	60	300GB/s	3000K	4.4PB	74.8PB

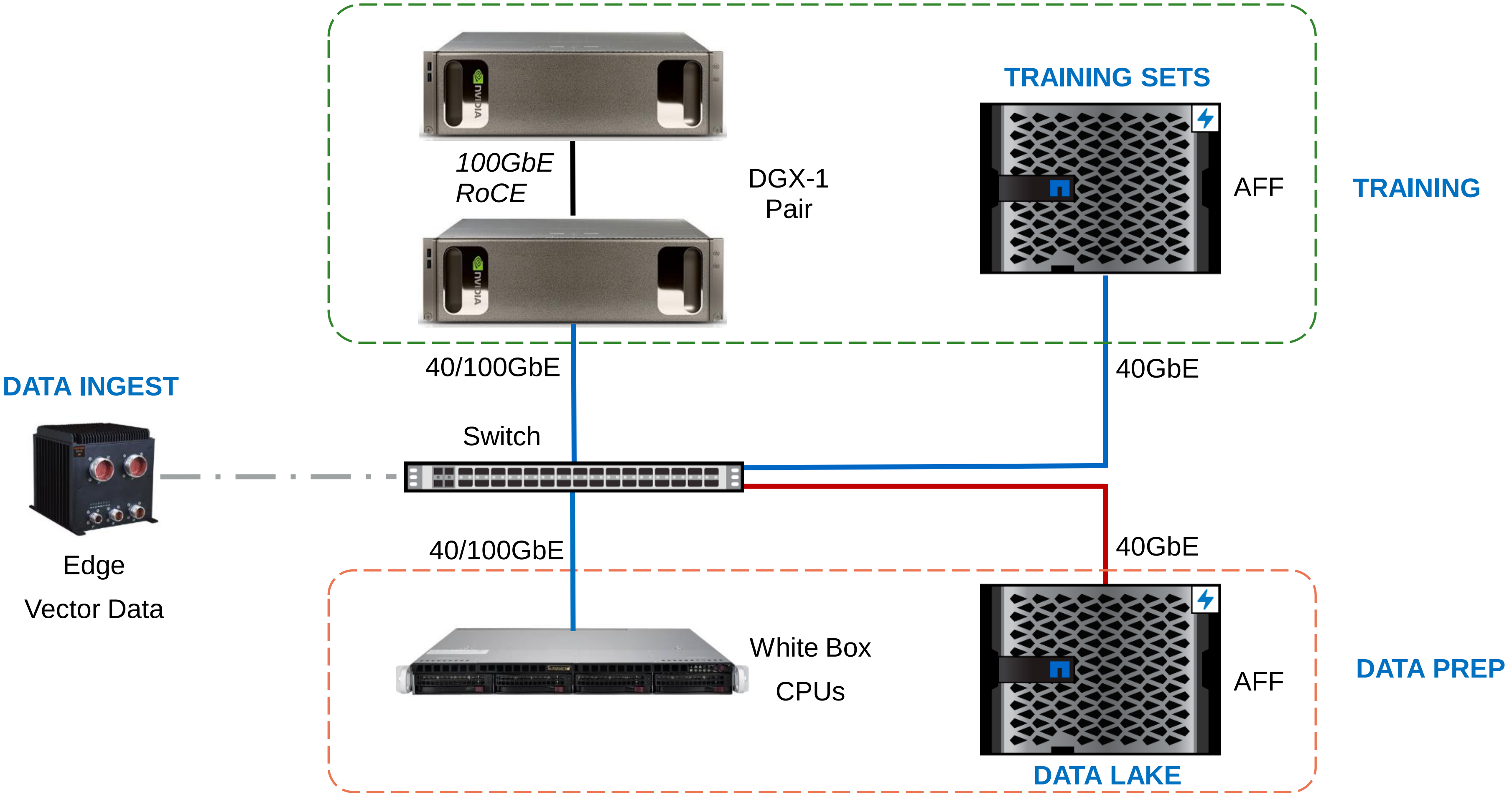
# of A700s Storage Systems	# of DGX-1 Servers	Throughput	Images/Sec	Typical Raw Capacity	Raw Capacity w/ Expansion
1 HA pair	4	18GB/s	180K	367.2TB	3.3PB
4 HA pairs	16	72GB/s	720K	1.5PB	13.2PB
12 HA pairs	48	216GB/s	2160K	4.4PB	39.7PB

NOTES:

- Based on ONTAP 9.4 performance metrics
- AlexNet model with average image size of 100KB
- Each DGX-1 capable of processing 50K images per second *(tensorflow.org)*
- Can scale down to A300 if lower end

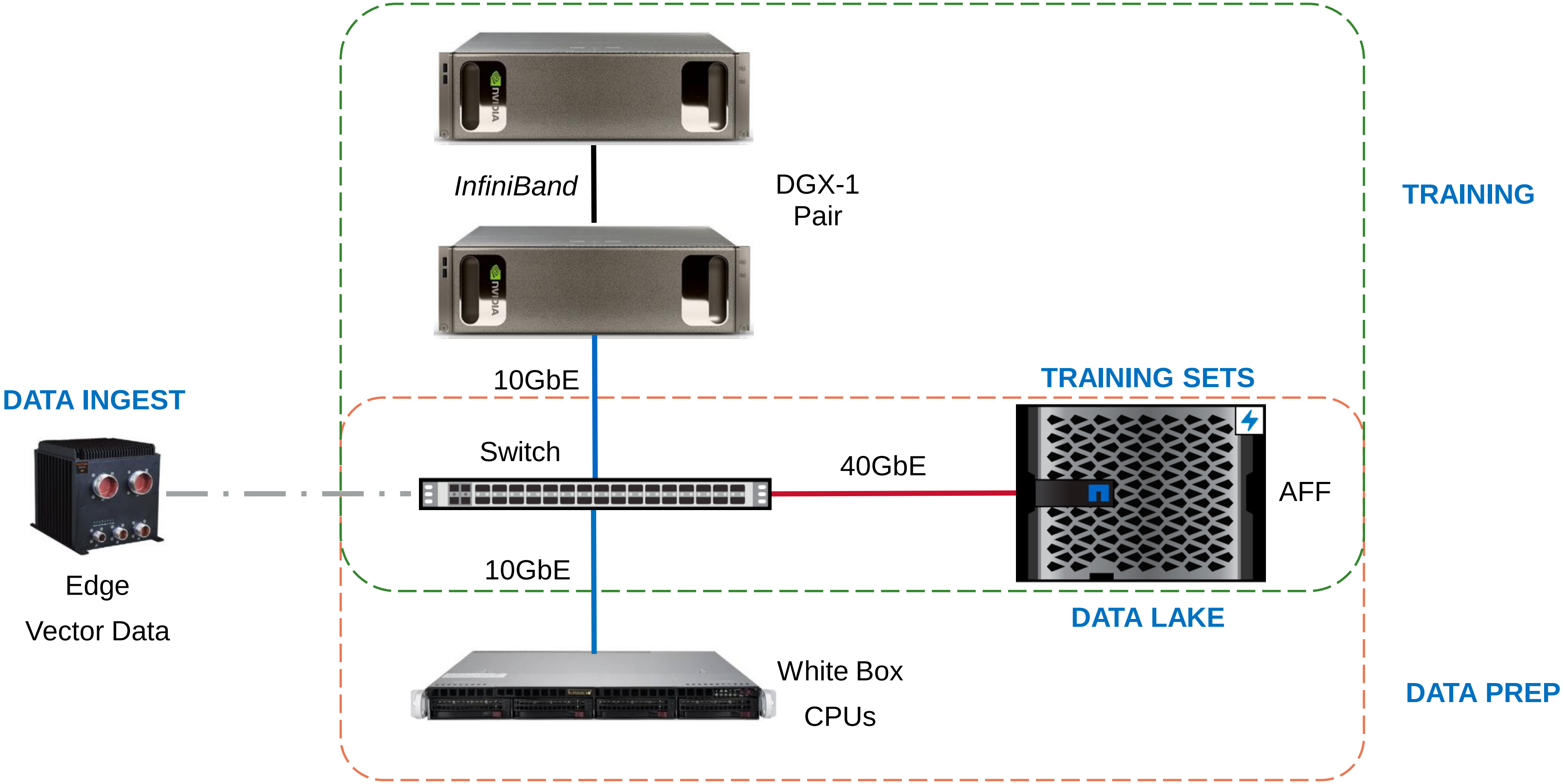
Sample AI and Deep Learning Training Cluster

NVIDIA DGX-1 100GbE RoCE v2 cluster, 40/100GbE AI Data Platform



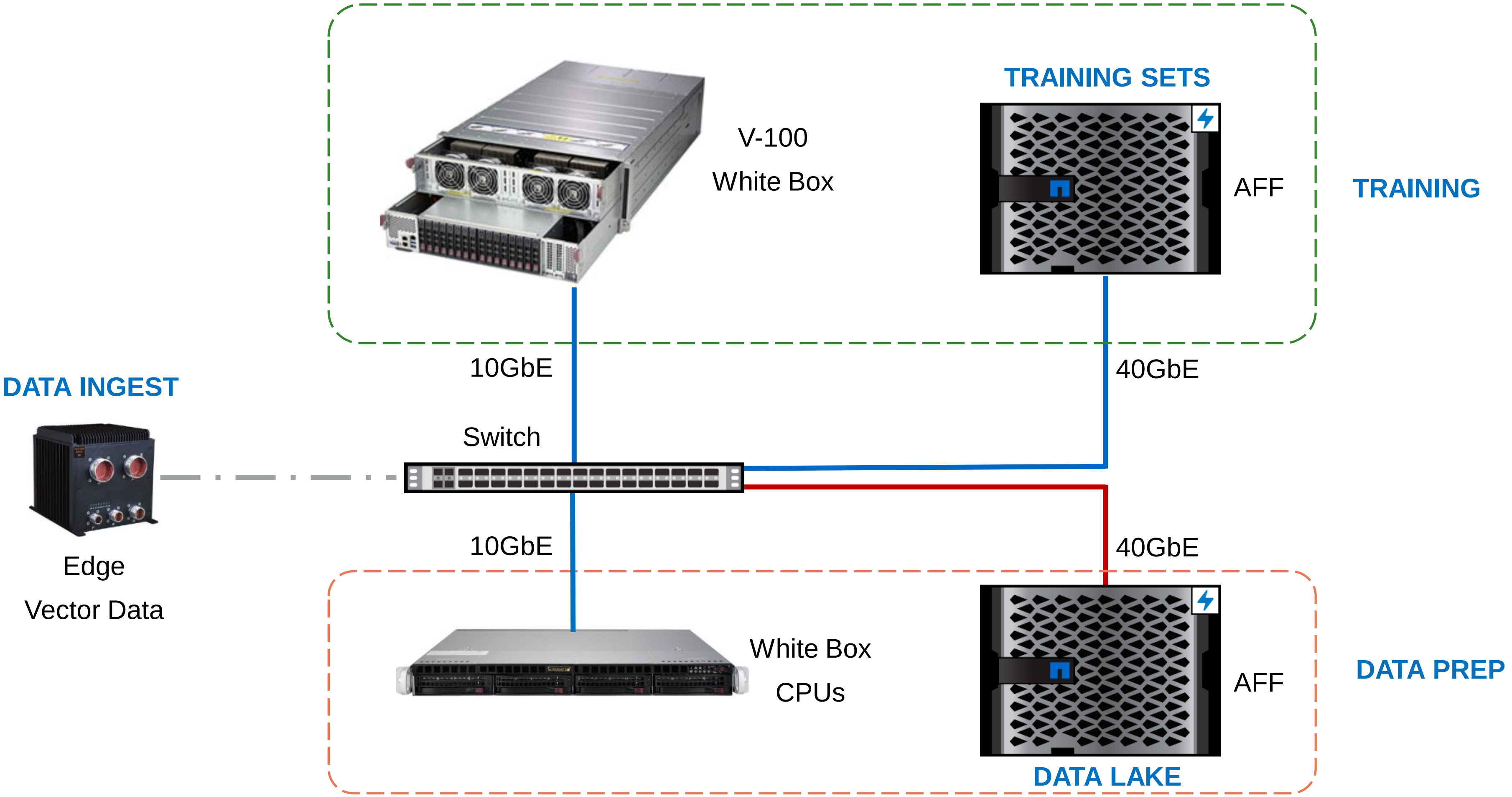
Sample AI and Deep Learning Training Cluster

In-place AI and Deep Learning with Data Lake



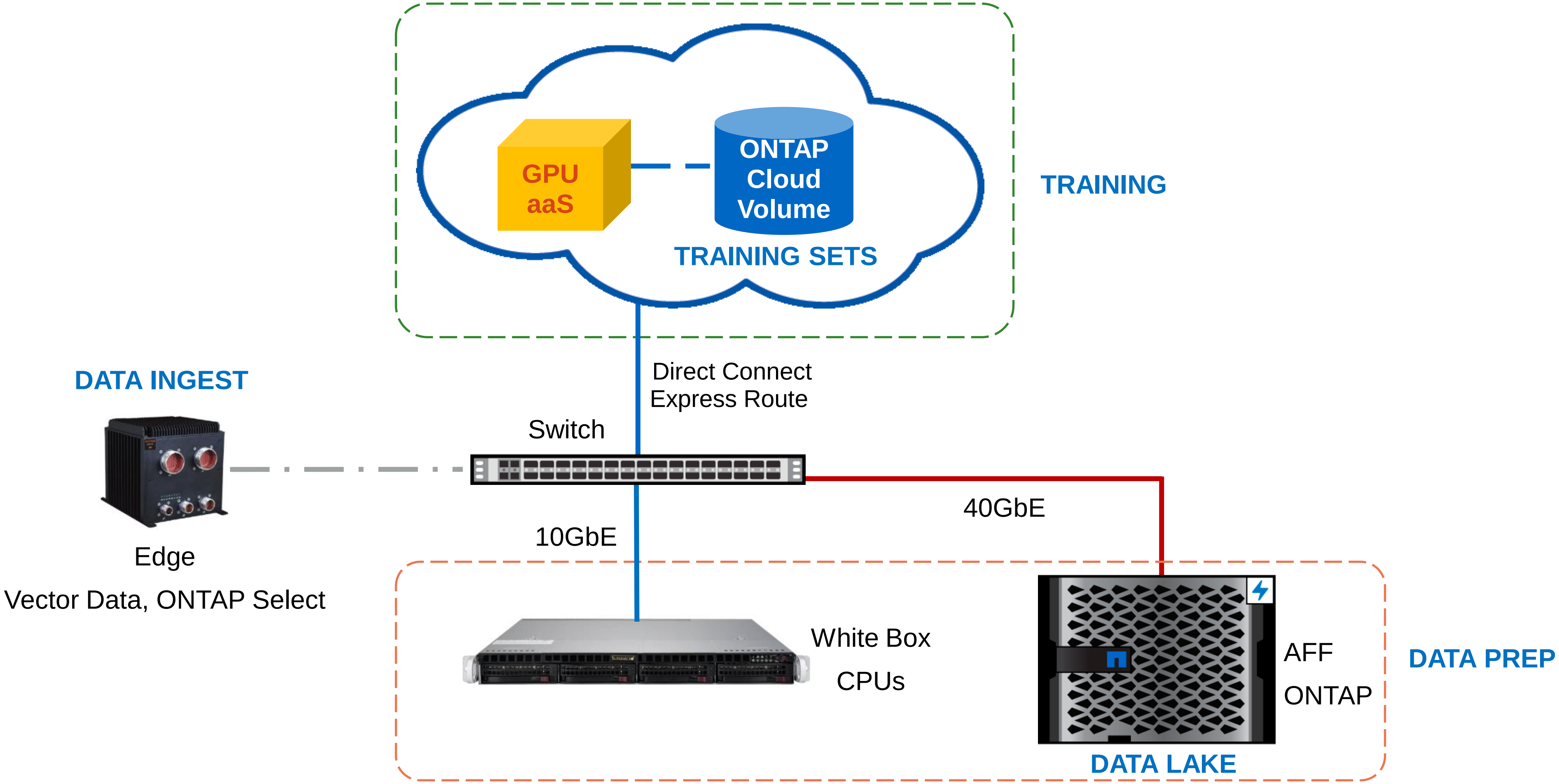
Sample AI and Deep Learning Training Cluster

White box with GPU based AI Compute Solution



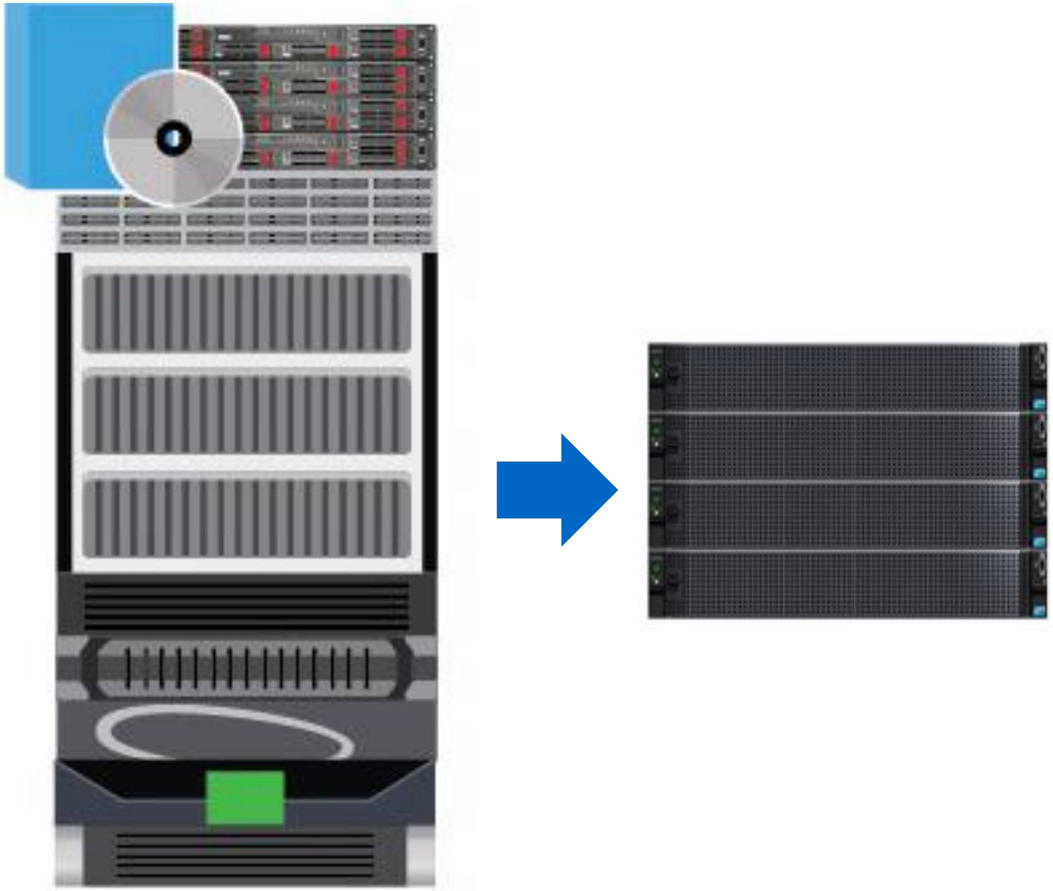
Sample Cloud AI and Deep Learning Training Cluster

Cloud based AI with Data Lake on-prem



NetApp HCI

Hyper Converged Infrastructure 1.0



What it is today…

- Consisting of software-defined compute, networking and storage
- Easier to manage; intuitive hypervisor-aware storage
- Shared-core approach for low entry point
- Rapid response to the business via fewer integration points
- Simple and rapid deployment and management
- Pay-as-you-go expansion and economics
- One size fits all architecture

NetApp HCI. Enterprise-Scale.

Guaranteed Performance

Deliver All Your Applications with Confidence

Flexibility & Scale

Scale On Your Terms

Automated Infrastructure

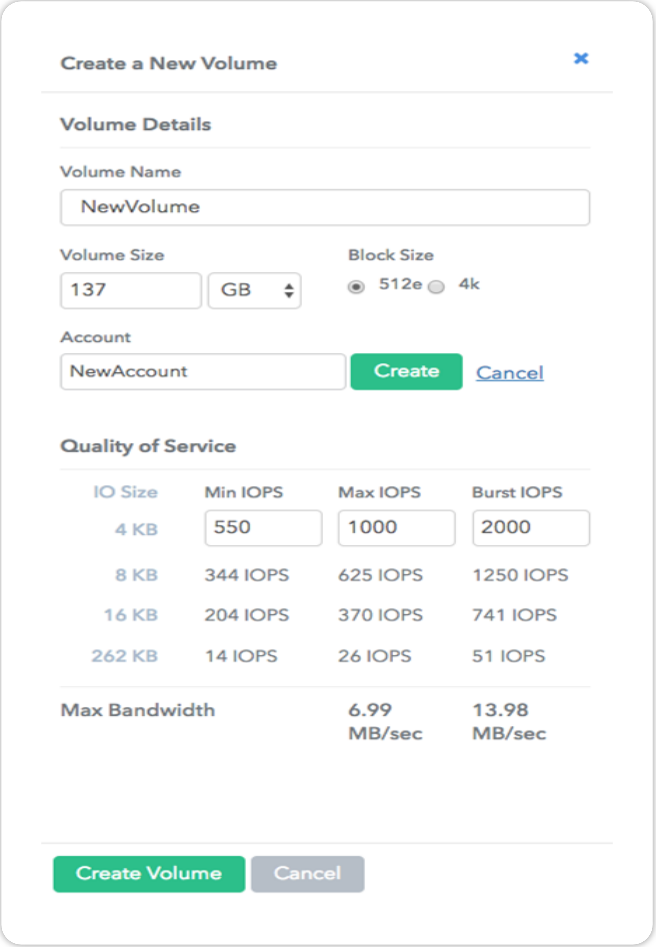
Transform & Empower Your IT Operations

NetApp Data Fabric

Unleash the Power of Data to Achieve a New Competitive Advantage

Consolidate Mixed Workloads

Unique Quality of Service Capabilities



Volume Details

Volume Name: NewVolume

Volume Size: 137 GB

Block Size: ● 512e ○ 4k

Account: NewAccount

Quality of Service

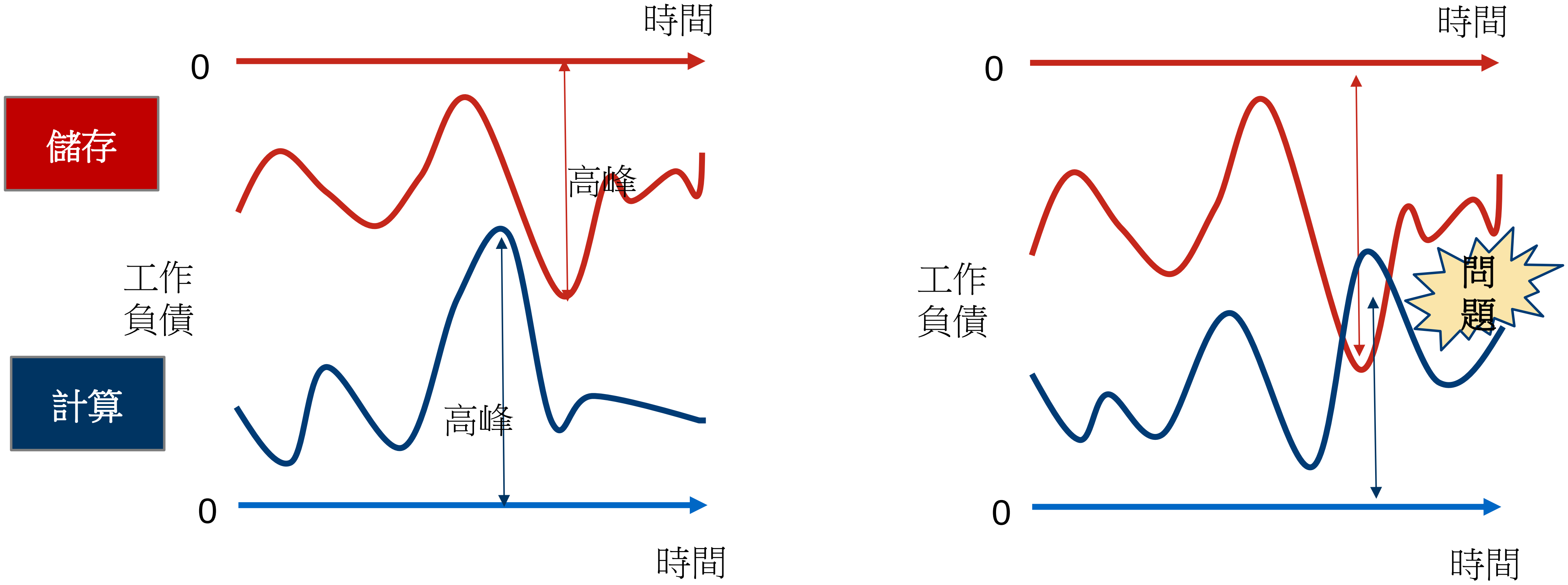
IO Size	Min IOPS	Max IOPS	Burst IOPS
4 KB	550	1000	2000
8 KB	344 IOPS	625 IOPS	1250 IOPS
16 KB	204 IOPS	370 IOPS	741 IOPS
262 KB	14 IOPS	26 IOPS	51 IOPS
Max Bandwidth		6.99 MB/sec	13.98 MB/sec

Dynamically **Allocate, Manage** and **Guarantee** performance independent of capacity

Define/enforce **Min, Max** and **Burst** settings for each application/volume

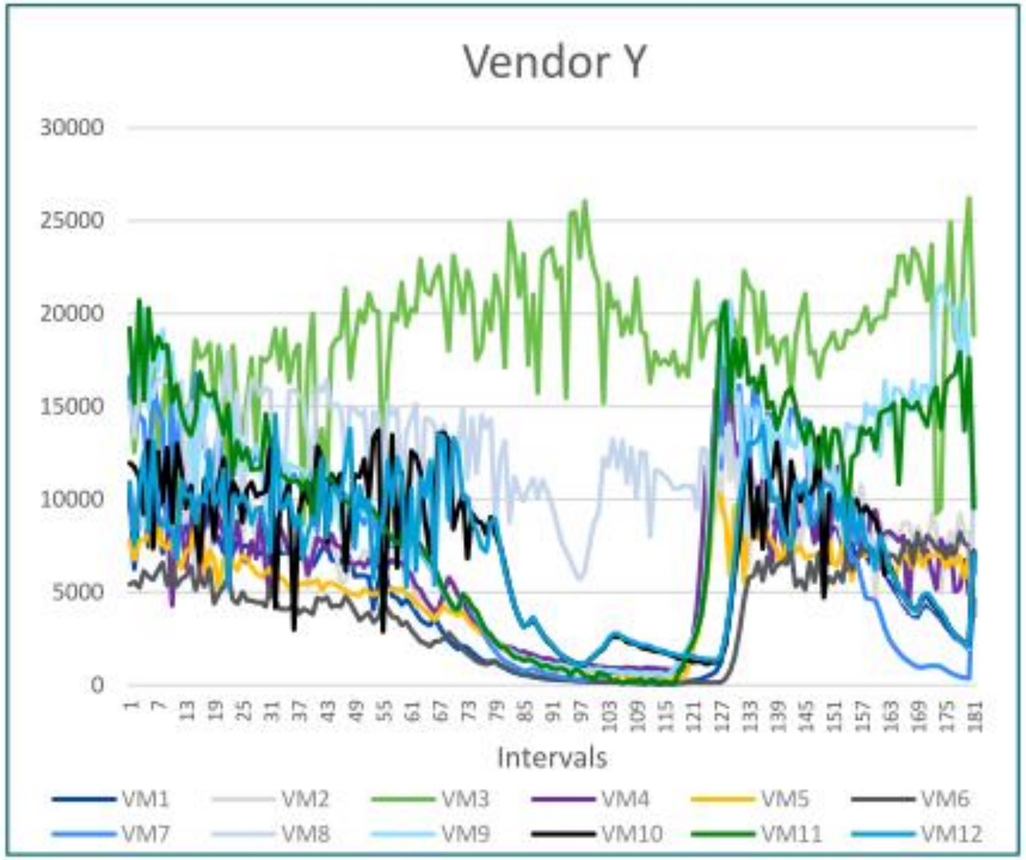
*Quantifying the Economic Value of a SolidFire Deployment – ESG Whitepaper, February 2015

HCI 1.0 版的資源分享 (上)

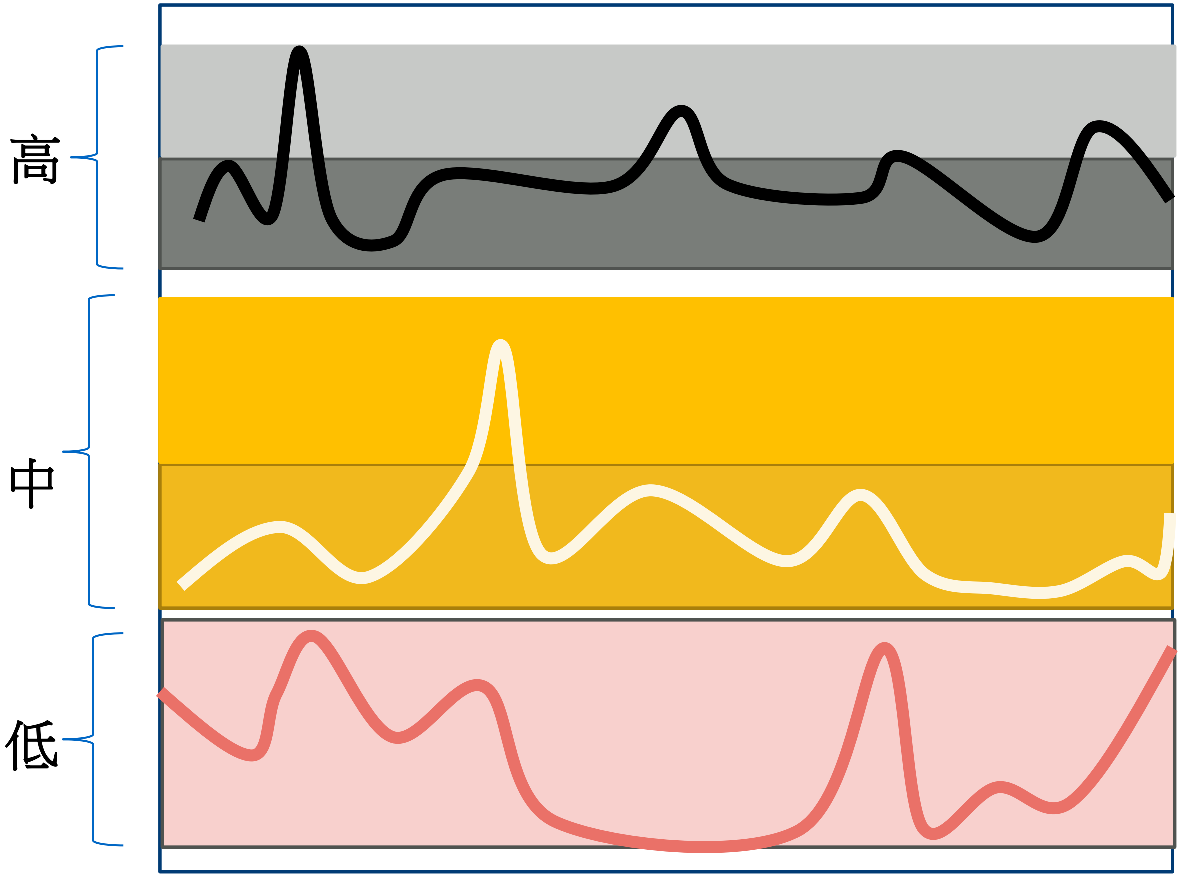


NetApp

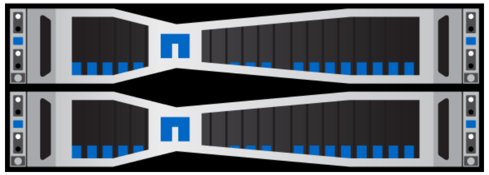
HCI 1.0 版的資源分享 (下)



Source: Cisco, 2018



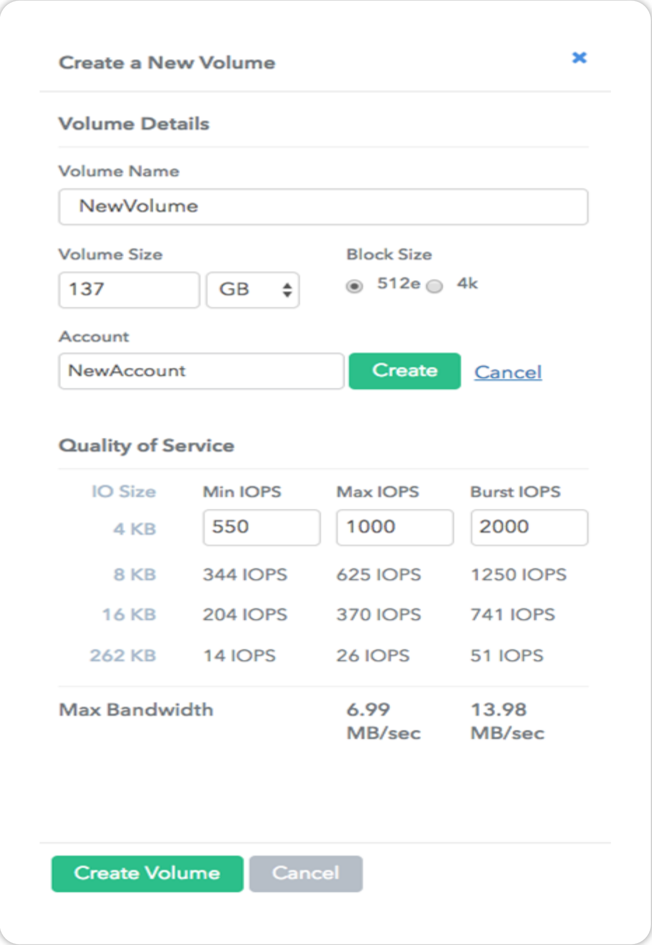
HCI 2.0



NetApp

Consolidate Mixed Workloads

Unique Quality of Service Capabilities



Create a New Volume ×

Volume Details

Volume Name: NewVolume

Volume Size: 137 GB

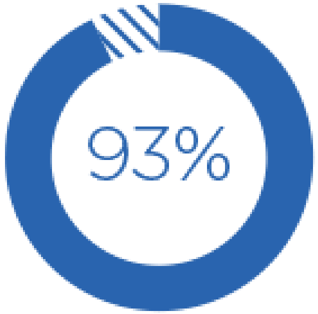
Block Size: ● 512e ○ 4k

Account: NewAccount

Quality of Service

IO Size	Min IOPS	Max IOPS	Burst IOPS
4 KB	550	1000	2000
8 KB	344 IOPS	625 IOPS	1250 IOPS
16 KB	204 IOPS	370 IOPS	741 IOPS
262 KB	14 IOPS	26 IOPS	51 IOPS
Max Bandwidth		6.99 MB/sec	13.98 MB/sec

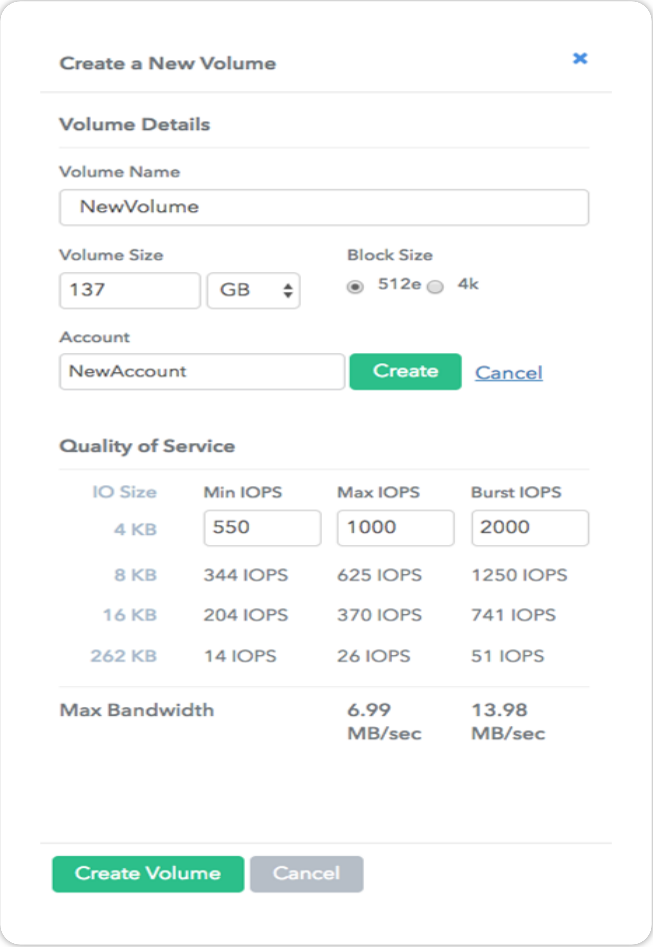
ELIMINATES



of traditional performance related storage problems*

Consolidate Mixed Workloads

Unique Quality of Service Capabilities



Create a New Volume

Volume Details

Volume Name: NewVolume

Volume Size: 137 GB

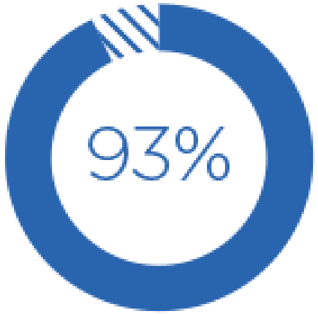
Block Size: ● 512e ○ 4k

Account: NewAccount

Quality of Service

IO Size	Min IOPS	Max IOPS	Burst IOPS
4 KB	550	1000	2000
8 KB	344 IOPS	625 IOPS	1250 IOPS
16 KB	204 IOPS	370 IOPS	741 IOPS
262 KB	14 IOPS	26 IOPS	51 IOPS
Max Bandwidth		6.99 MB/sec	13.98 MB/sec

ELIMINATES

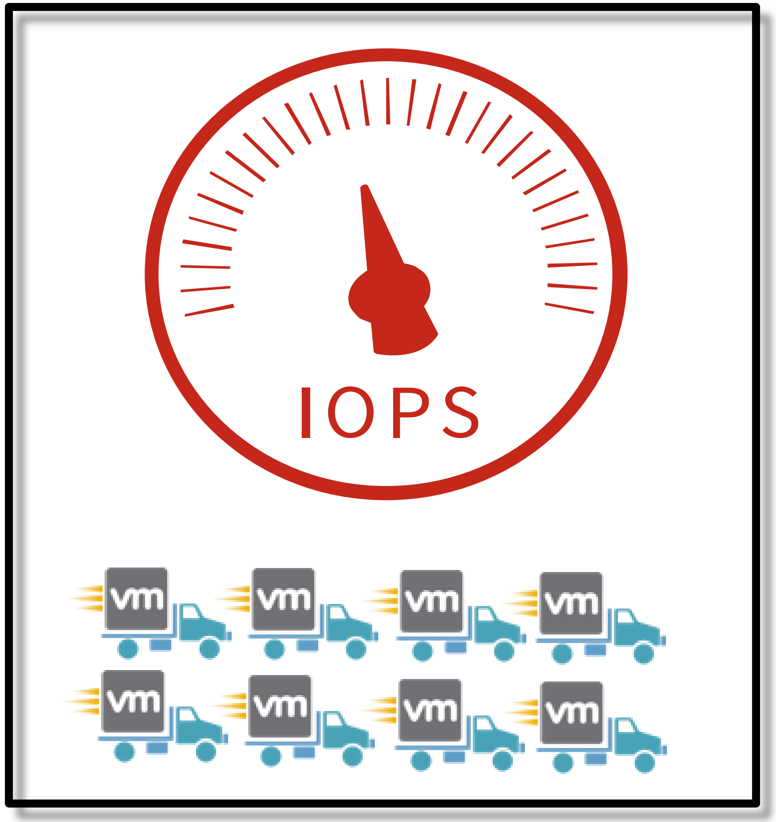


of traditional performance related storage problems*

Provide granular control at VM level

- Prevent any VM from impacting the performance of another

Without Control



With Control (VVols)

Flexibility & Scale

Scale on Your Terms

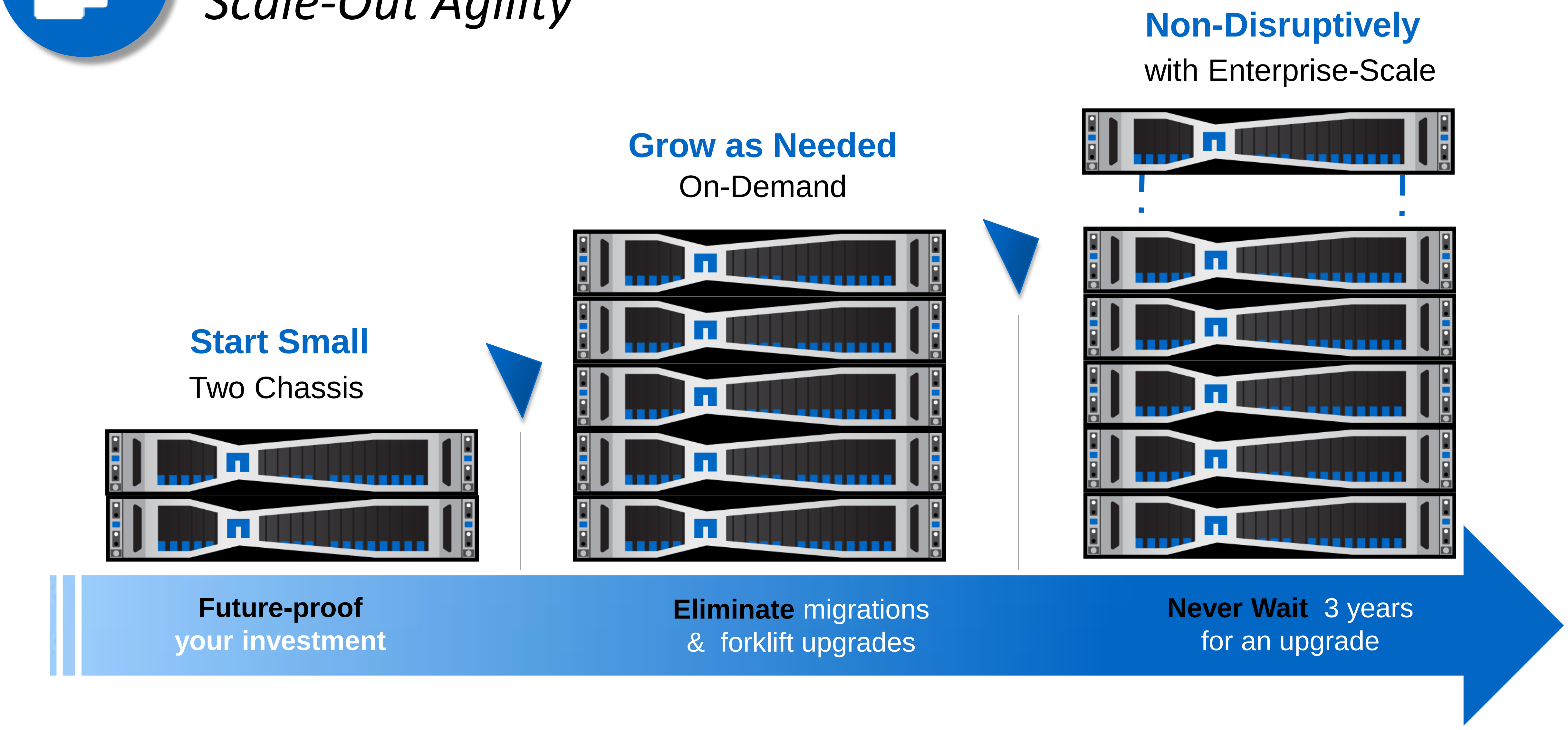
Optimize & Protect Existing Investments

Scale Compute & Storage Independently

Eliminate “HCI Tax”

Optimize & Protect Existing Investments

Scale-Out Agility



Start Small
Two Chassis

Grow as Needed
On-Demand

Non-Disruptively
with Enterprise-Scale

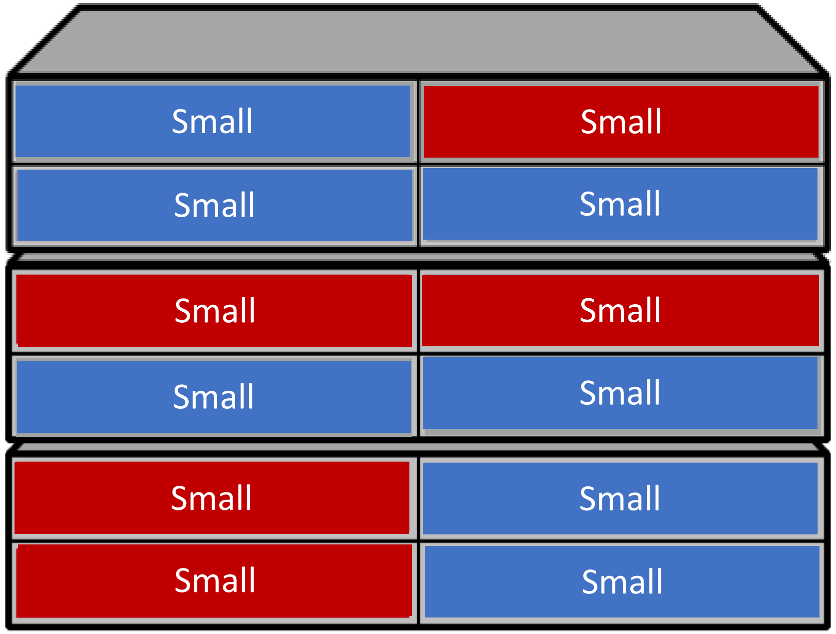
Future-proof
your investment

Eliminate migrations
& forklift upgrades

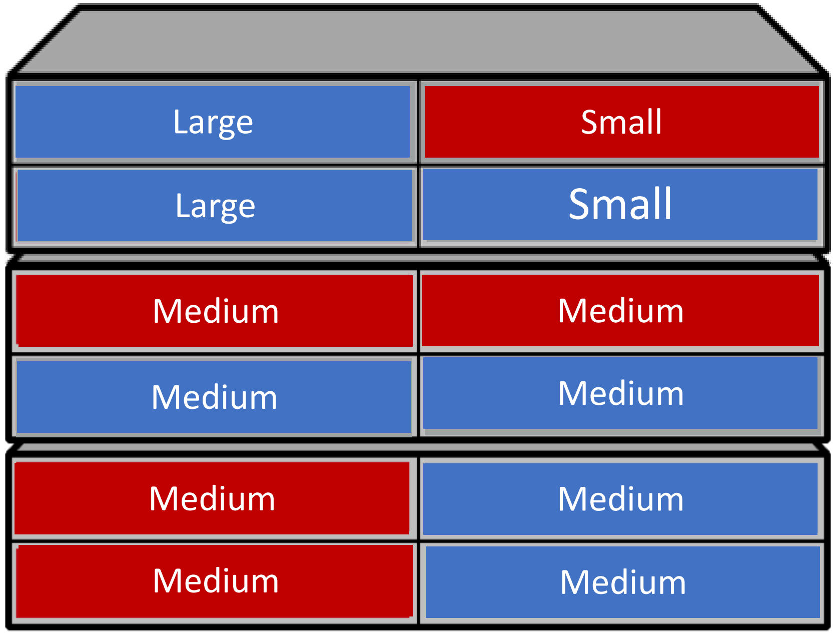
Never Wait 3 years
for an upgrade

Independently Scale Compute & Storage

Mix and Match to Fit Your Needs



Small	Small
Small	Small
Small	Small
Small	Small
Small	Small
Small	Small



Large	Small
Large	Small
Medium	Medium
Medium	Medium
Medium	Medium
Medium	Medium

Compute Node Components

- Specifications per node

	Small	Medium	Large
RU	1RU, half-width	1RU, half-width	1RU, half-width
Cores for VM's	16	24	36
CPU	Intel 2620 - 2.1G	Intel 2650 - 2.2G	Intel 2695 - 2.1G
Memory	384 GB	512 GB	768 GB
Boot Device	2 x 240GB MLC	2 x 240GB MLC	2 x 240GB MLC
Base Networking	4x 10/25 GbE SFP 28 + 2x 1GbE RJ45	4 x 10/25 GbE SFP 28 + 2x 1GbE RJ45	4x 10/25 GbE SFP 28 + 2x 1GbE RJ45

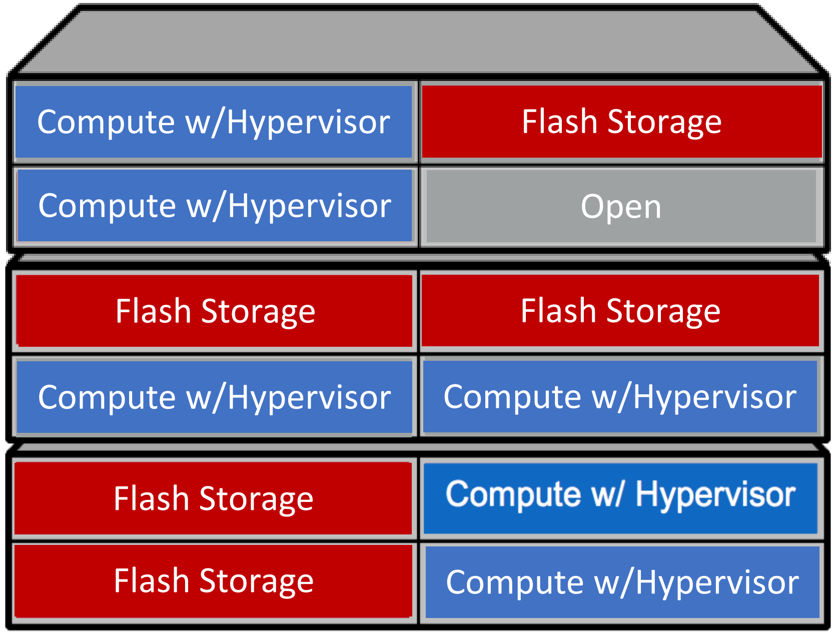
Storage Node Components

- Specifications per node

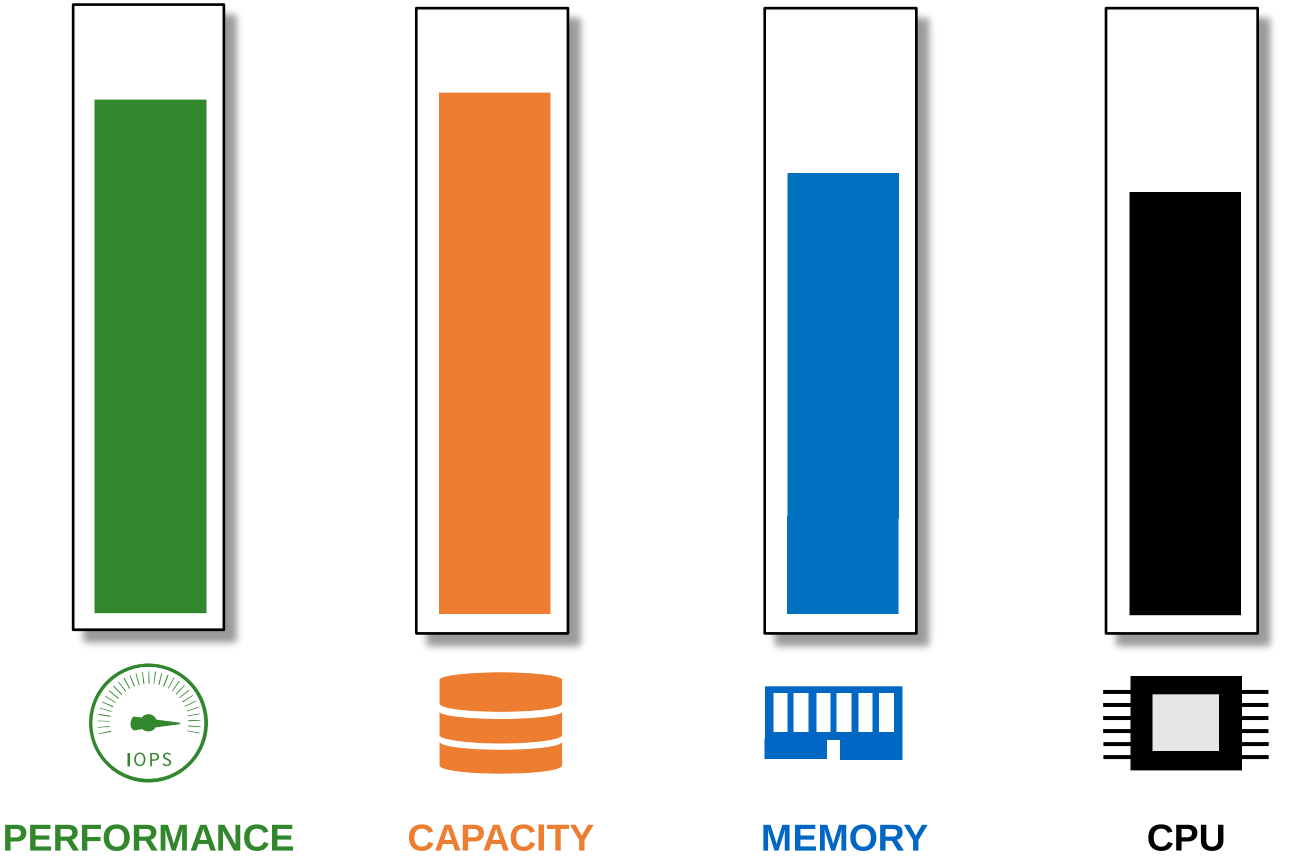
	Small	Medium	Large
RU	1RU, half-width	1RU, half - width	1RU, half - width
IOPS	50,000	50,000	100,000
Boot Device	1 x 240GB MLC	1 x 240GB MLC	1 x 240GB MLC
Base Networking	4 x 10/25 GbE SFP 28 + 2x 1GbE RJ45	4 x 10/25 GbE SFP 28 + 2x 1GbE RJ45	4 x 10/25 GbE SFP 28 + 2x 1GbE RJ45
SSD	6 x 480 GB	6 x 960 GB	6 x 1.92 TB
Effective Block Capacity*	5.5TB – 11TB	11TB – 22TB	22TB – 44TB

Compute and storage scaled independently

By node

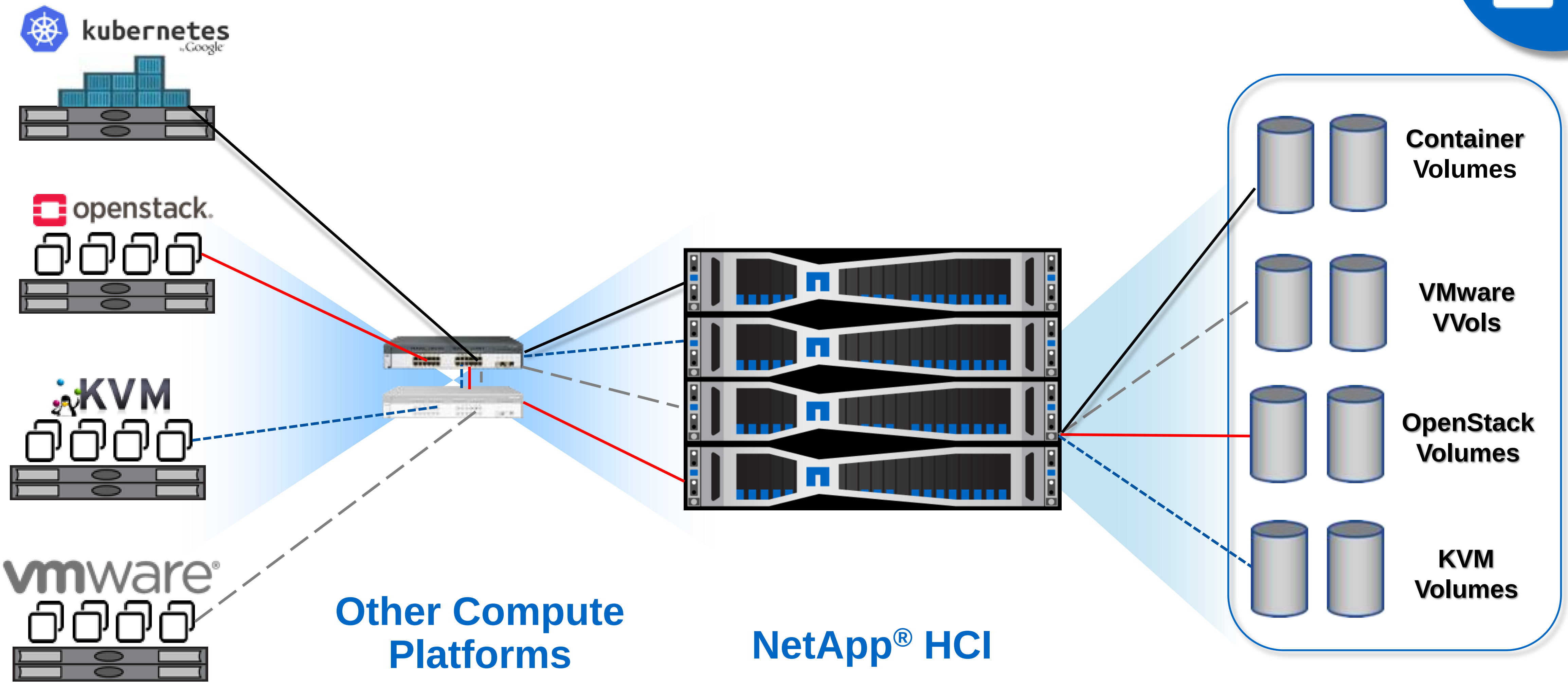


Compute w/Hypervisor	Flash Storage
Compute w/Hypervisor	Open
Flash Storage	Flash Storage
Compute w/Hypervisor	Compute w/Hypervisor
Flash Storage	Compute w/ Hypervisor
Flash Storage	Compute w/Hypervisor



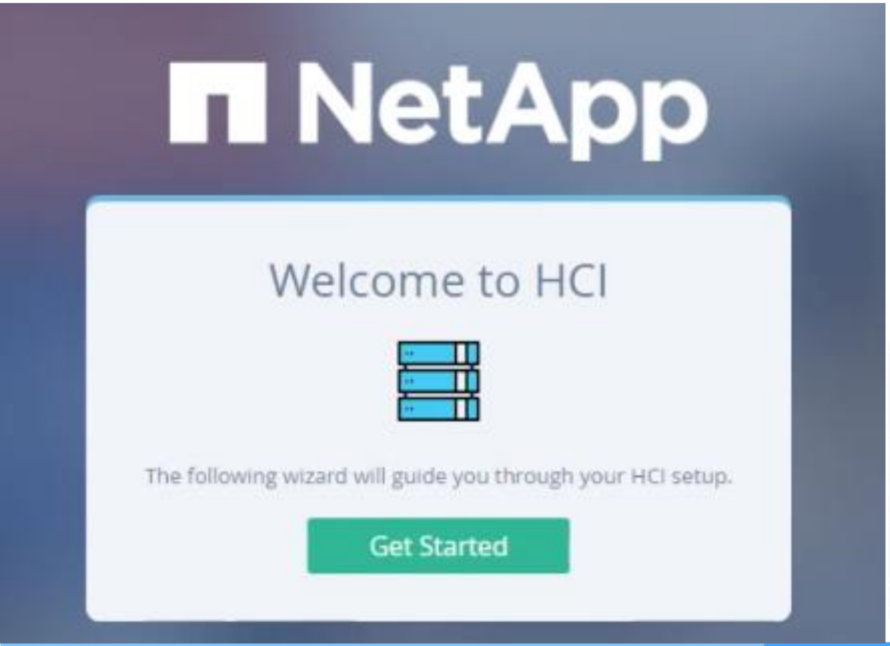
Open storage model

Flexibility to integrate external compute systems with NetApp HCI storage targets



Day 0: get up and running in 30 minutes

Intuitive deployment engine reduces 400+ inputs < 30


NetApp

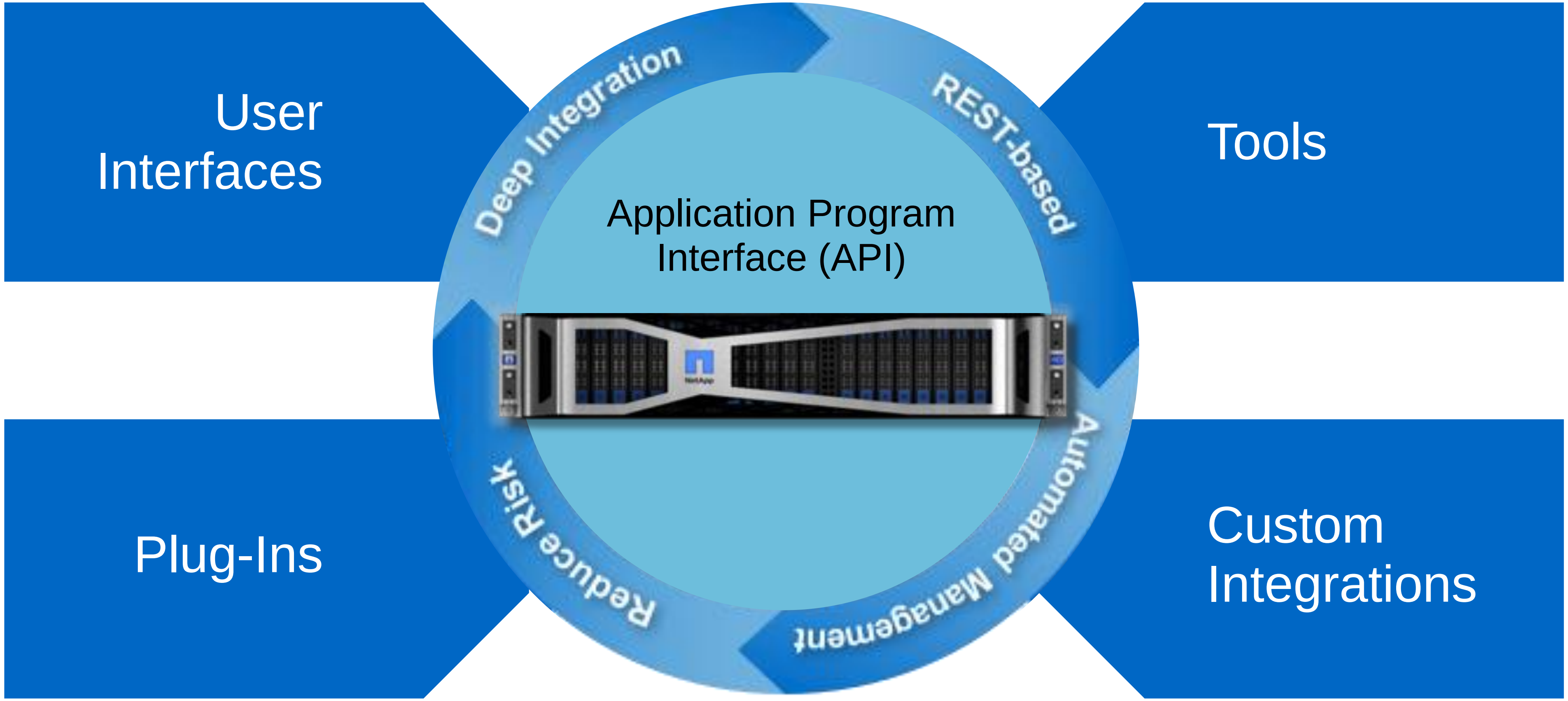
Welcome to HCI

The following wizard will guide you through your HCI setup.

Initializing → Configuring → Building → Finishing

Day 1+: simplified management

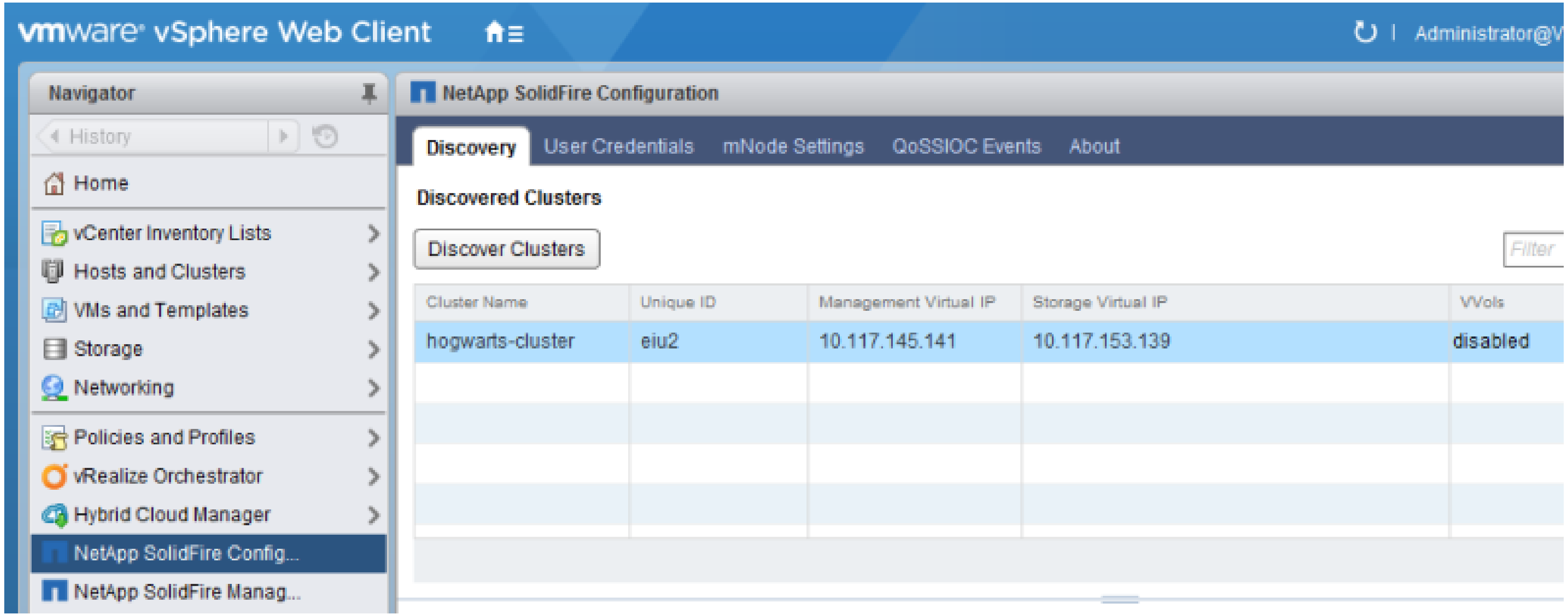
Comprehensive set of robust APIs



Simplified Operations and Management

Leverage VMware vCenter for day-to-day operational tasks

- 95% of operations performed from vCenter, including acknowledgment hardware alerts



Discovered Clusters

Cluster Name	Unique ID	Management Virtual IP	Storage Virtual IP	VVols
hogwarts-cluster	eiu2	10.117.145.141	10.117.153.139	disabled

Key Takeaways

- NetApp Data Fabric
- NetApp ONTAP AI
- NetApp HCI

Thank You