



# Lenovo NetApp携手NVIDIA

## 加速人工智能之旅

联想凌拓产品管理与市场营销高级总监 林佑声  
联想凌拓产品经理 尹伟  
英伟达高级客户经理 吴强



联想凌拓

**数据** 至关重要  
DATA IS KING

互联网时代+智能时代

数据量几何级数增长



# 2020年



中国1000强中  
»» 50% 的企业



全球2000强中  
»» 50% 的企业

## 需要整体数字化转型

需要创造数字化增强产品、服务和体验



# 联想凌拓

## 智能数据管理 解决方案及服务

致力于加速企业实现数字化转型



# 我们的愿景

建设一个  
由数据智能驱动的未来



# 我们的使命



## 满足

企业业务需求



## 加速

智能化数字化转型



## 推动

商业创新



## 适合

本地市场

领先的高性能智能数据管理解决方案 + 强大的产品组合服务网络和研发能力

# 我们的承诺



高效便捷



随机应变



安心可靠



# 我们的优势



最大化的一站式  
全方位解决方案

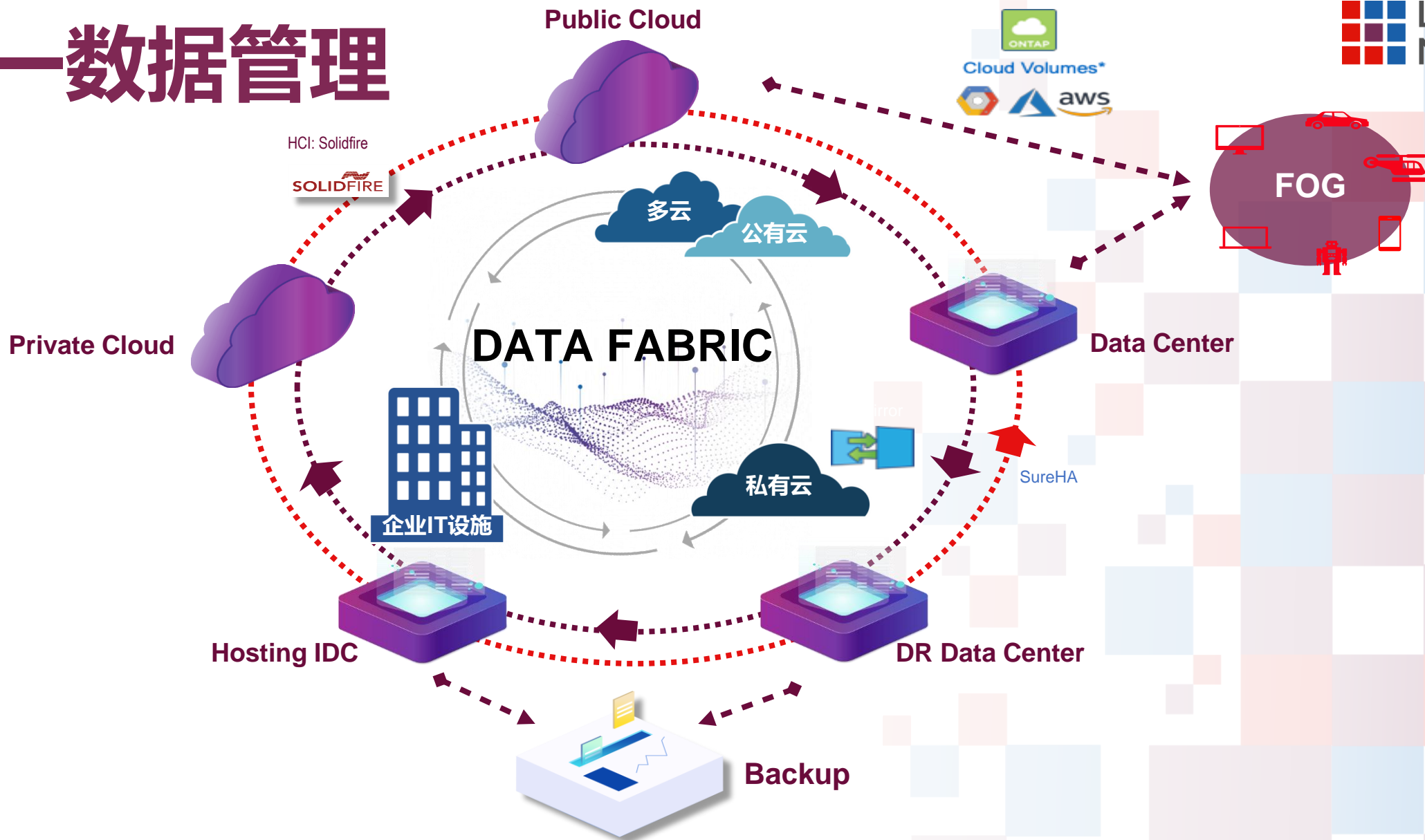


双品牌策略



本地优化

# 统一数据管理



# 全新存储产品家族 – 双品牌策略

实现完整的端到端产品组合 与众不同的科技和数据管理

## Lenovo品牌 ThinkSystem DE / DM



### Lenovo 统一存储/高性能块存储

- ONTAP 技术平台/高性能块存储平台
- 全闪/混合存储
- 云分层技术
- Lenovo XClarity 支持
- 联想品牌及服务

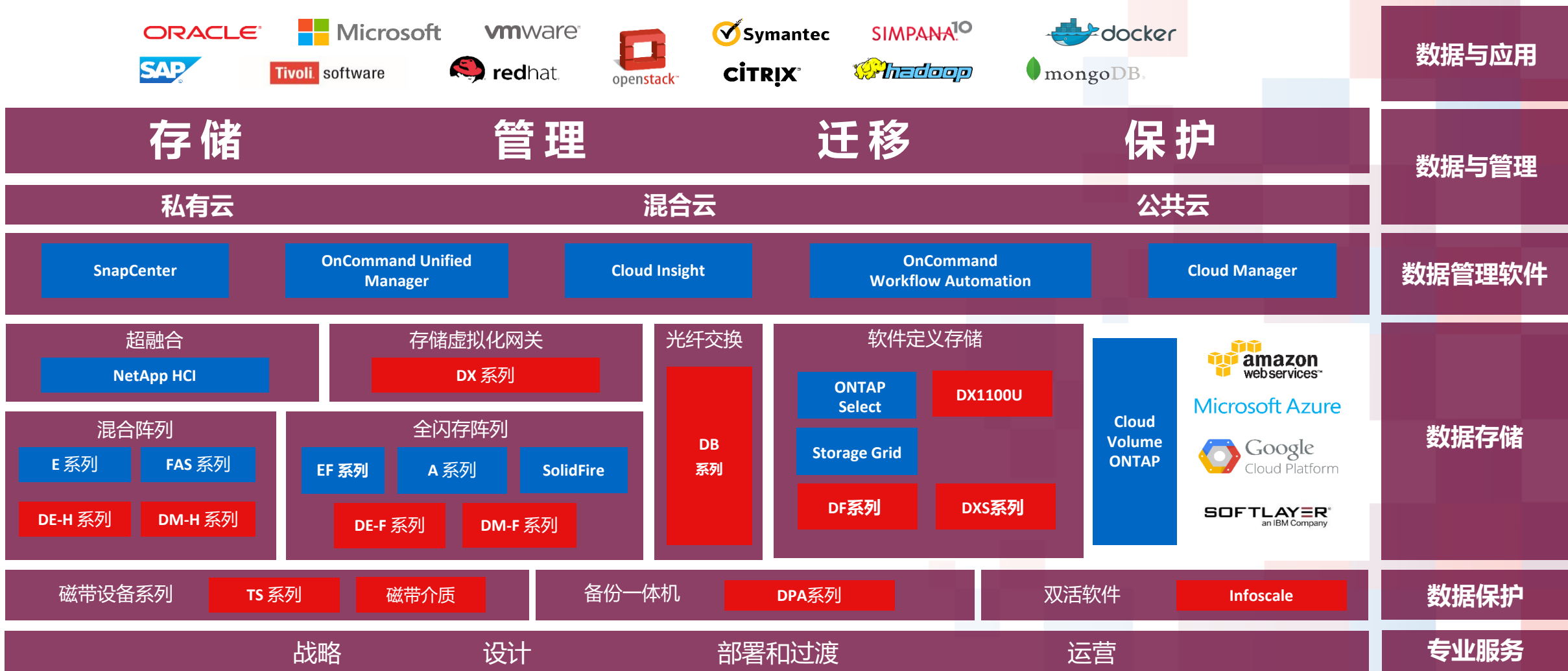
## NetApp 品牌



### NetApp全产品线产品

- ONTAP
- Data Management
- Software Define
- NetApp品牌

# 我们的产品家族





# 统一存储FAS/DM产品组合



NetApp™

C190

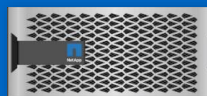


AFF A220



All SAN Array

AFF A300



AFF A400



All SAN Array

AFF A700



All SAN Array

AFF A700s



AFF A800



Lenovo

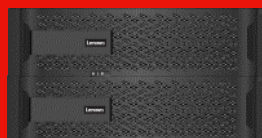
DM5000F



DM7000F



DM7100F



NetApp™

FAS2720



FAS2750



FAS8200



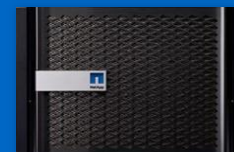
FAS8300



FAS8700



FAS9000



Lenovo

DM3000H



DM5000H



DM7000H



DM7100H



# 入门级光纤存储DE/E系列产品组合



E2800



E2800



EF280



E5700



EF570



EF600



Lenovo

DE2000H



DE4000H



DE4000F



DE6000H

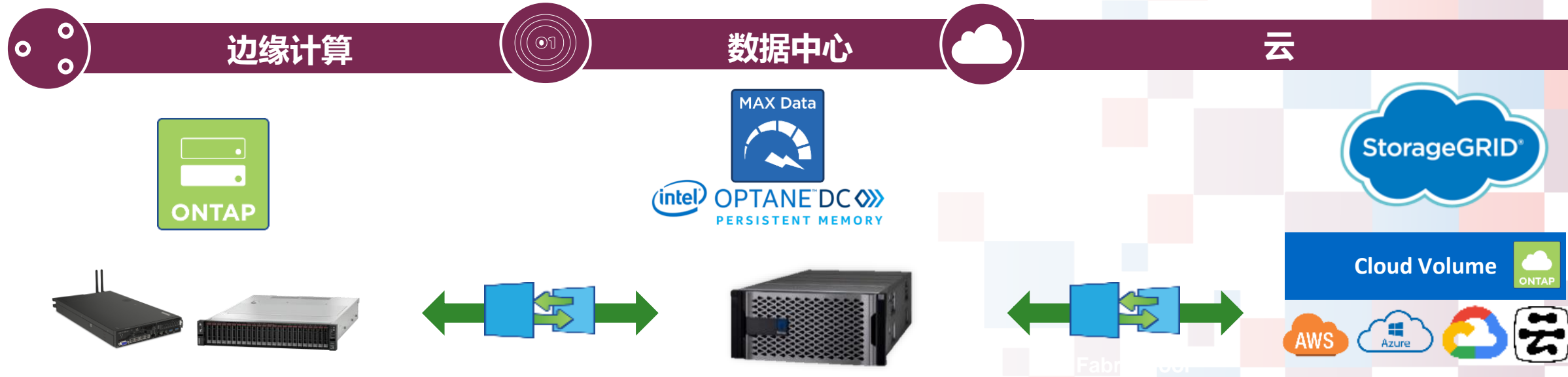


DE6000F



# 软件定义解决方案

## 端到端的软件定义技术实现数据赋能



**ONTAP Select 一体机**  
边缘及分支机构数据管理  
灵活

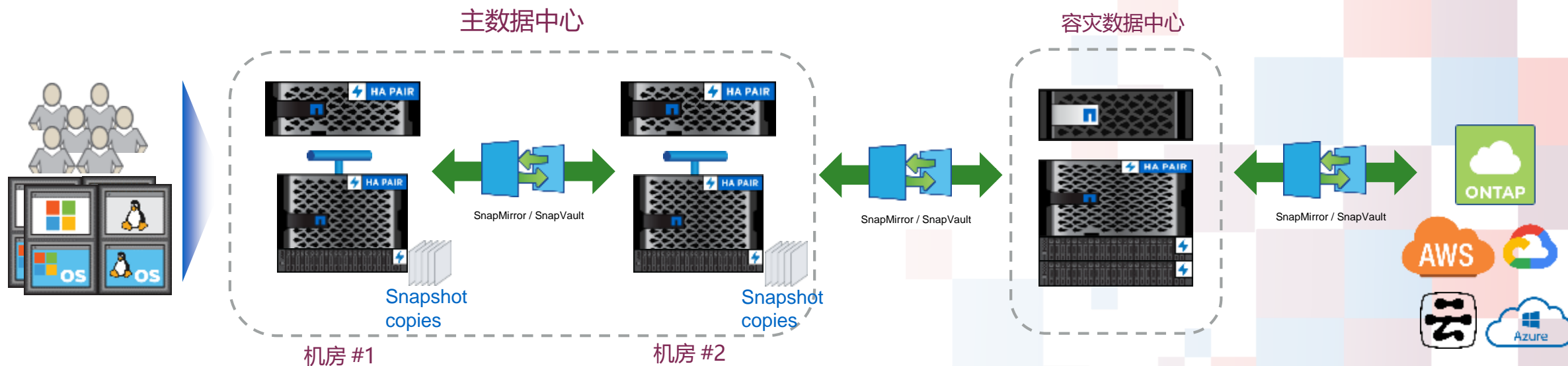
**MAX Data**  
内存数据管理  
业界领先

**新型全闪存储**  
数据管理平台  
极致

**StorageGRID**  
私有云存储平台  
海量

**软件定义云存储**  
由ONTAP管理的  
公有云存储

# 存储整合灾备解决方案



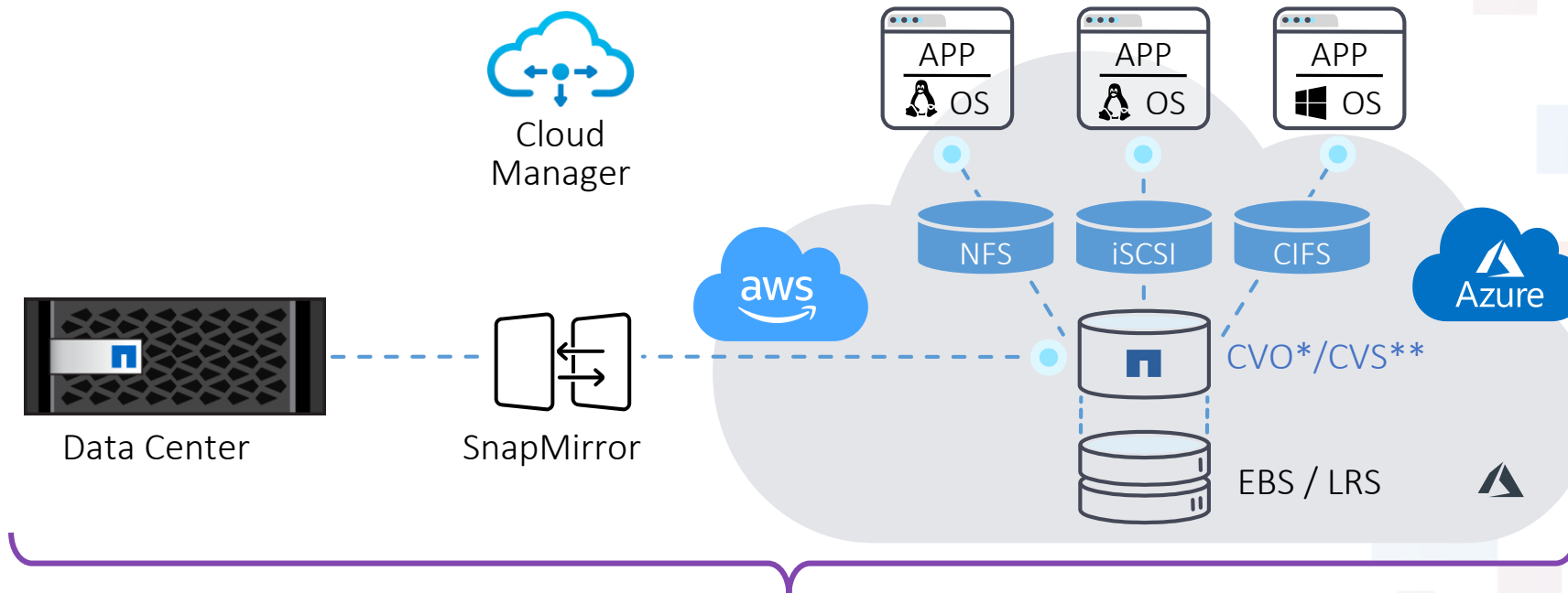
## SnapCenter插件

## SnapCenter插件



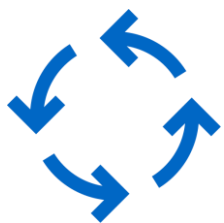


# 混合云解决方案



- 自由平滑数据迁移
- 跨平台统一数据管理
- 高效数据保护
- 备份/灾备简单上云
- 开发/运维直入云端
- 省钱的企业级云存储
- 高能低耗存储分层

容灾



提升算力



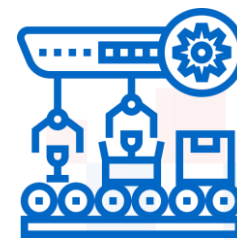
开发运维



备份归档

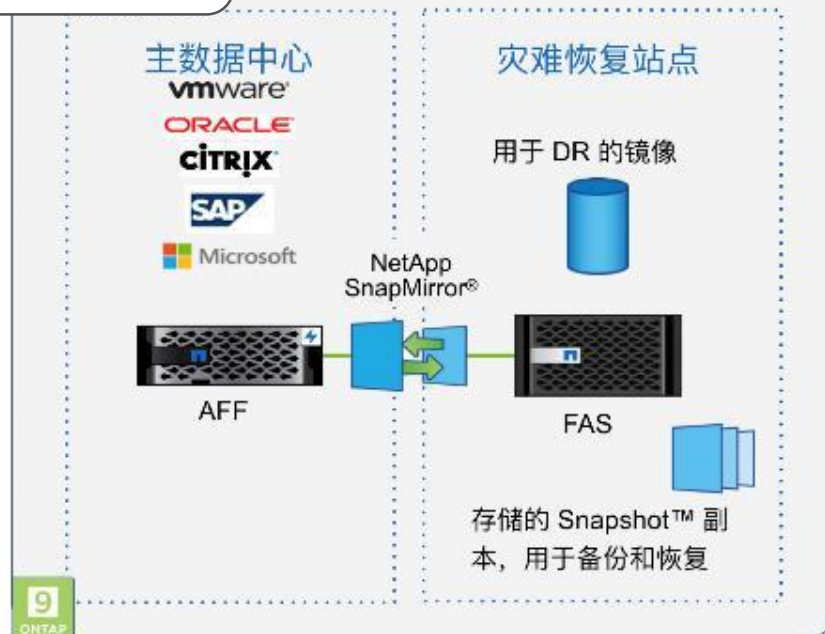


生产创造



# 数据中心灾备解决方案

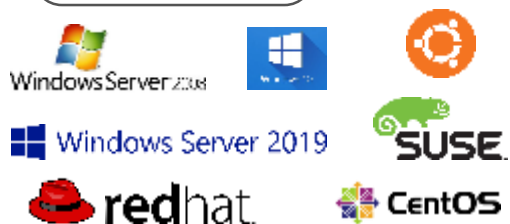
基于存储灾备方案



更多应用支持



单独服务器支持



第三方硬件支持



备份到专用磁  
盘备份设备



备份到磁带长  
期保留数据



备份到云

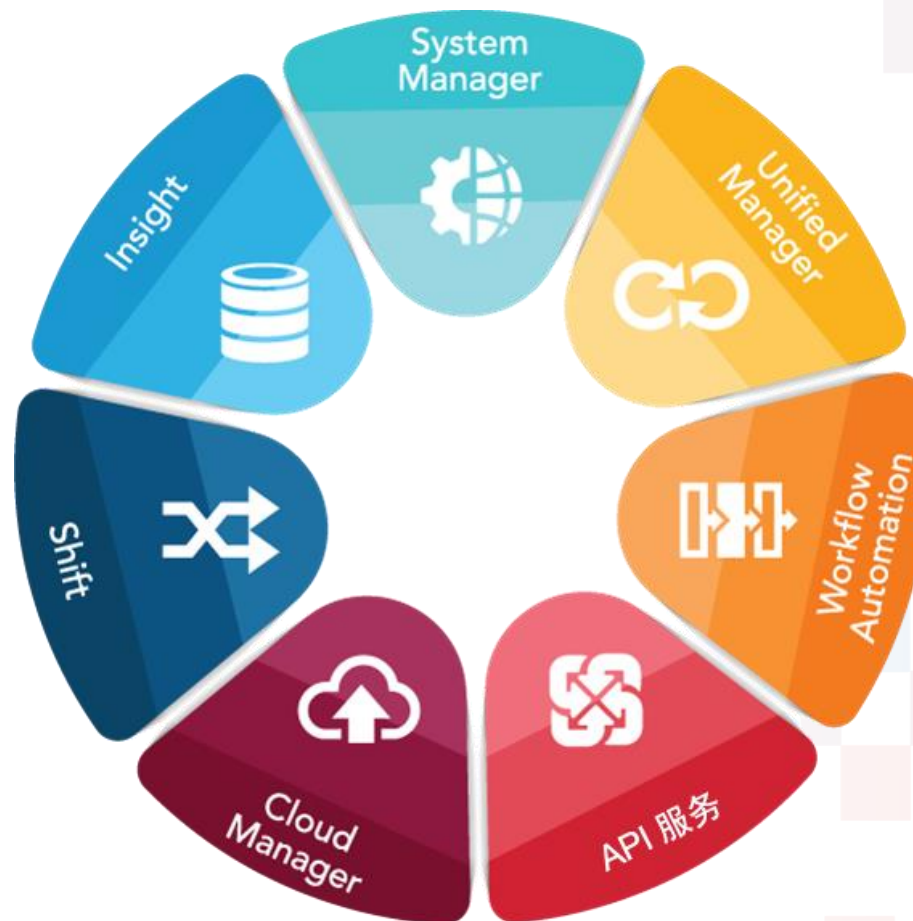


**DPA备份一体机**  
**9000/12000/24000**

# 整合管理平台

## Data Fabric

- 实施软件定义的数据中心
- 过渡到 SLO 驱动型管理
- 支持融合基础架构/混合型企业



## 管理简化

- 简化常见 workflows
- 全面掌握存储基础架构运行状况
- 管理多供应商基础架构

Lenovo  
xClarity

## Cloud Insight

# 奖项及荣誉



最佳数据管理解决方案奖  
IT运维网



全国石油石化企业信息化产品技术创新奖  
中国石油企业协会



闪存存储系统企业奖  
DOIT



数字中国数字化转型  
创新先锋奖  
E企研究院



软件定义存储创新平台奖  
至顶网

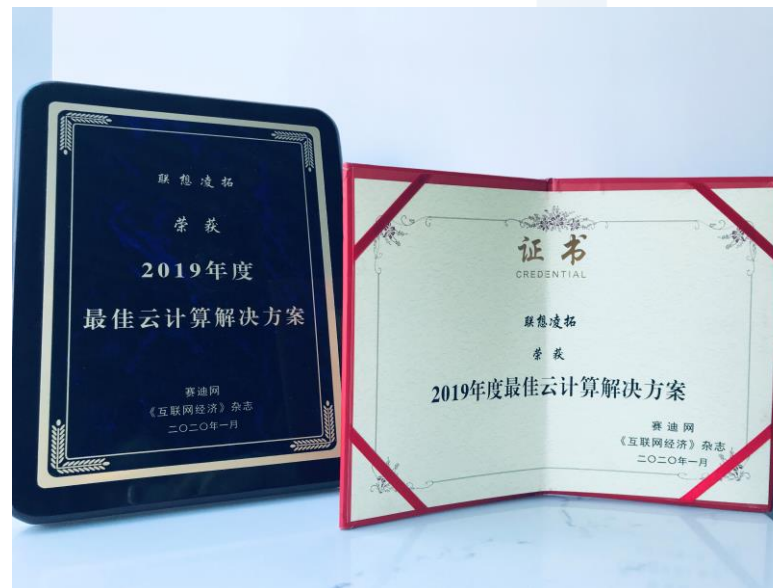
# 奖项及荣誉



商业领袖人物奖  
数字商业时代



中国IT行业  
领军企业奖  
51CTO



最佳云计算解决方案奖  
赛迪网



年度新锐数据管理  
解决方案提供商奖  
企业网



# 智慧数据构建智能世界

## Building a World of Digital Intelligence



# 谢谢!

智慧数据构建智能世界

# NVIDIA企业级产品 DGX的应用及案例



Stephen Wu





# NVIDIA

- > Founded in 1993
- > Jensen Huang, Founder & CEO
- > 13,227 employees
- > \$144B market cap; \$11.7B revenue in FY19

“World’s Most Admired Companies”

— Fortune

“50 Smartest Companies: #1”

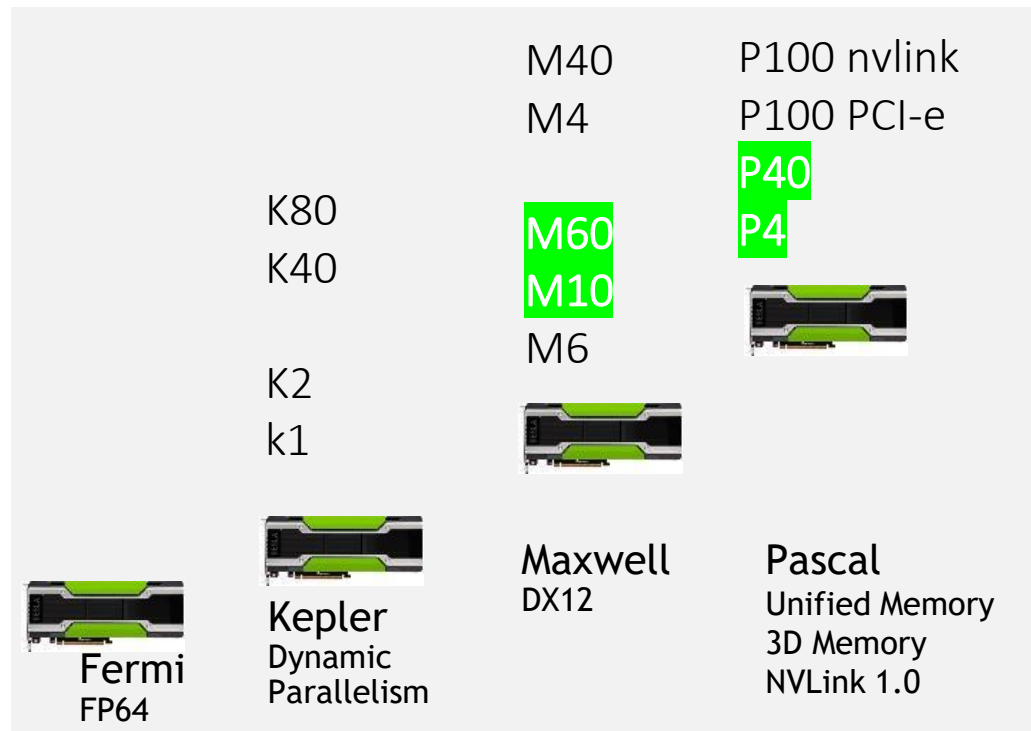
— MIT Tech Review

“#1 Top CEO in the World”

— Harvard Business Review

“Most Innovative Companies”

# TESLA / DGX 路线图



**NVIDIA GTC20**

全球首播  
NVIDIA CEO 黄仁勋  
主题演讲

解读人工智能、深度学习、自动驾驶、机器人及专业图形领域最新突破

北京时间 5月14日 21:00  
扫码观看主题演讲

## CLOUD-SCALE AI

### NGC



Optimized Platform  
with the Highest  
AI Efficiency

## AI WORKSTATION

### DGX Station



with



AI Workstation for  
Data Science Teams

## AI DATA CENTER

### DGX-1



with



The Essential  
Instrument for AI  
Research

### DGX-2

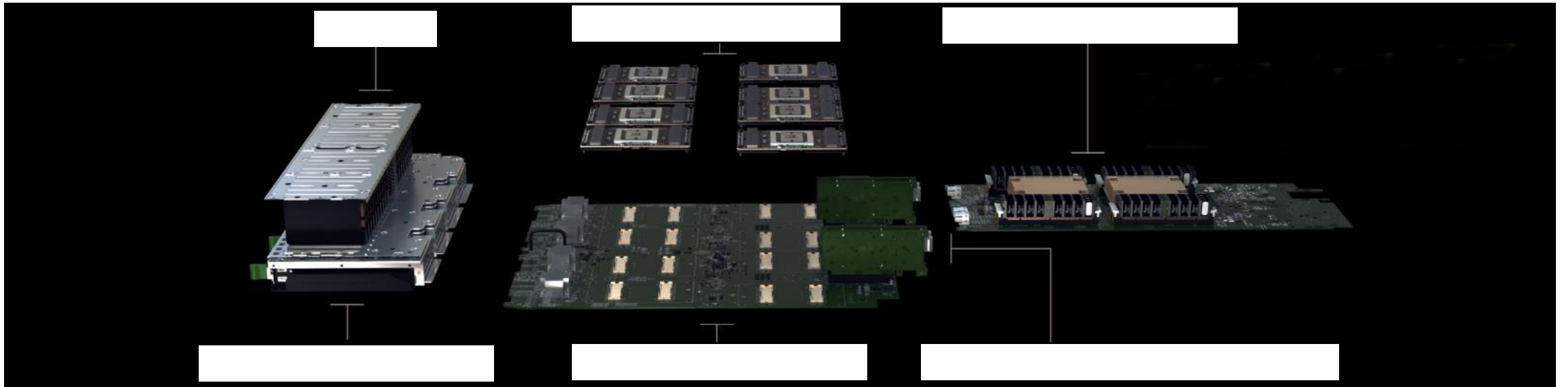


with



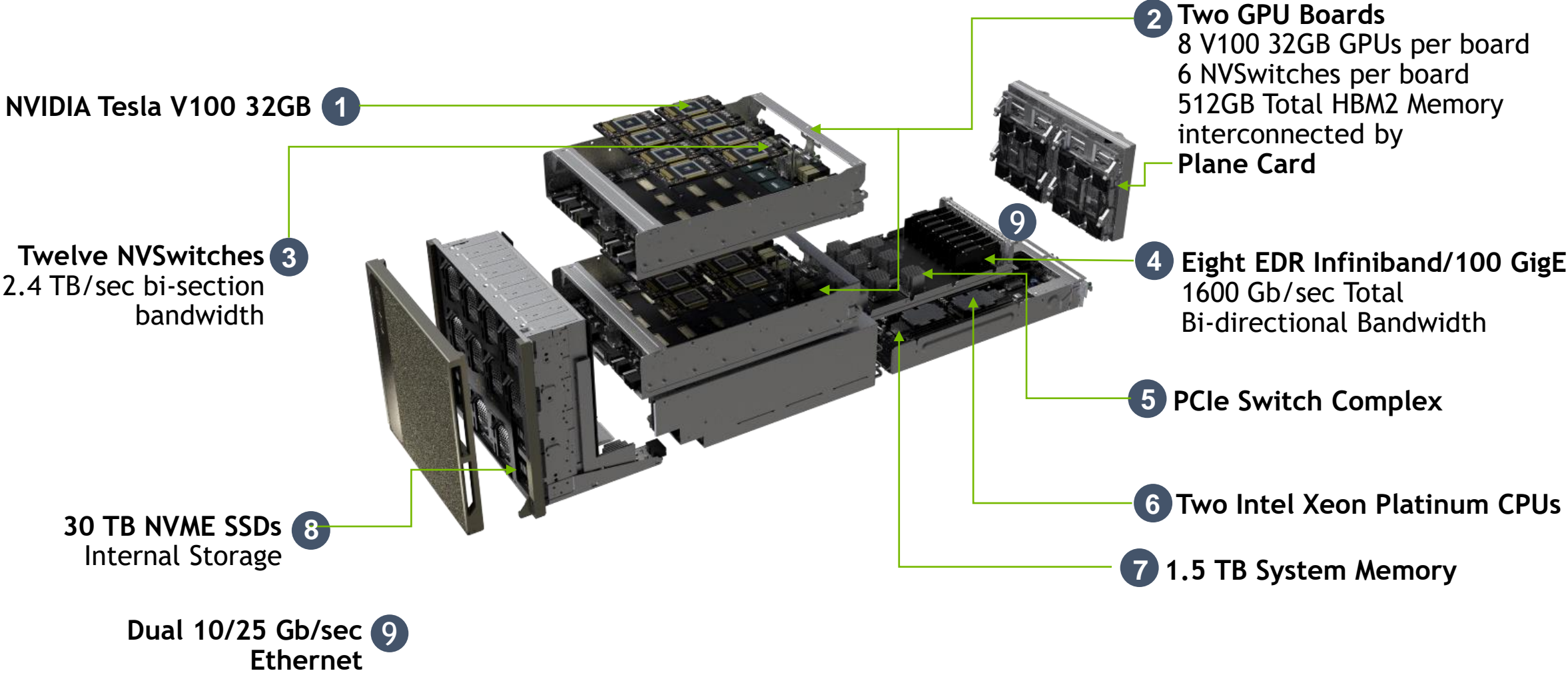
The World's Most  
Powerful AI System for  
the Most Complex AI  
Challenges

## Highest Performance, Fully Integrated HW System

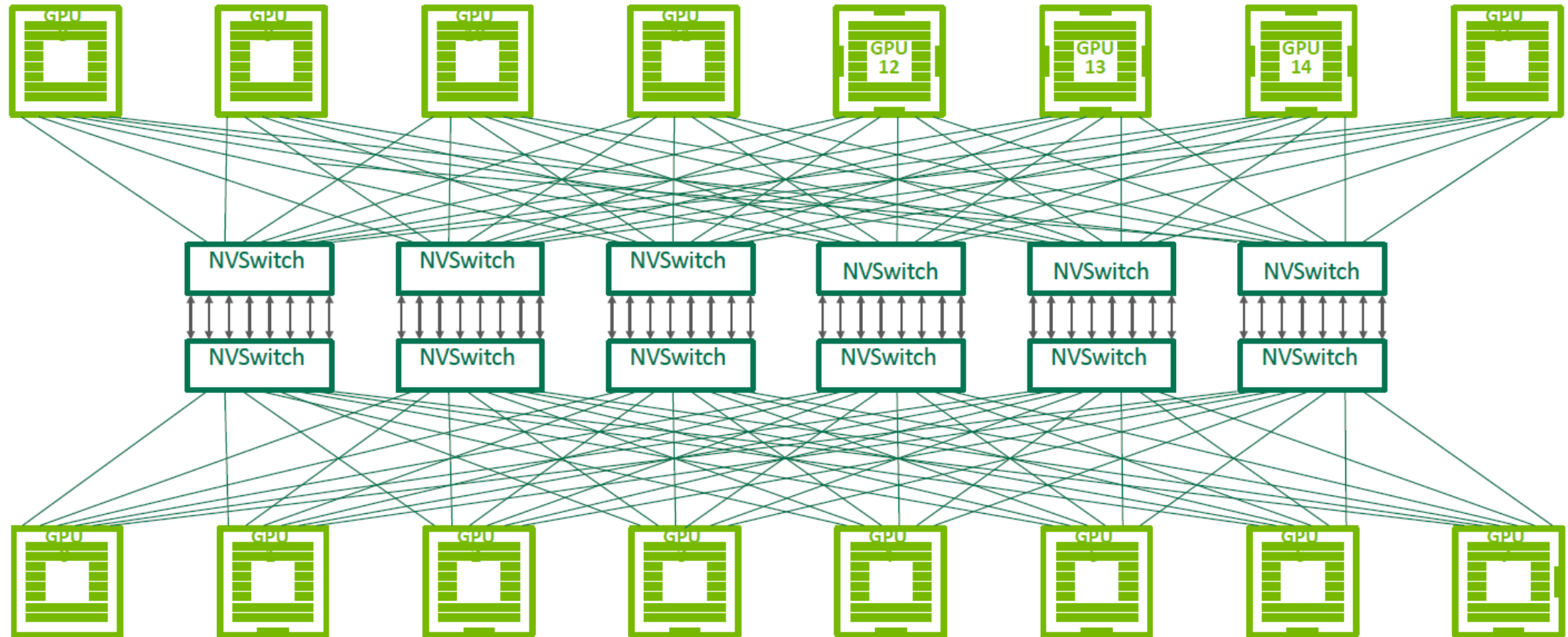


**1 PetaFLOPS** | **8x Tesla V100 32GB** | **300 Gb/s** NVLink Hybrid Cube Mesh  
2x Xeon | 7 TB RAID 0 | Quad IB/Ethernet 100Gbps, Dual 10GbE | 3U – 3500W

# designed to train the previously impossible



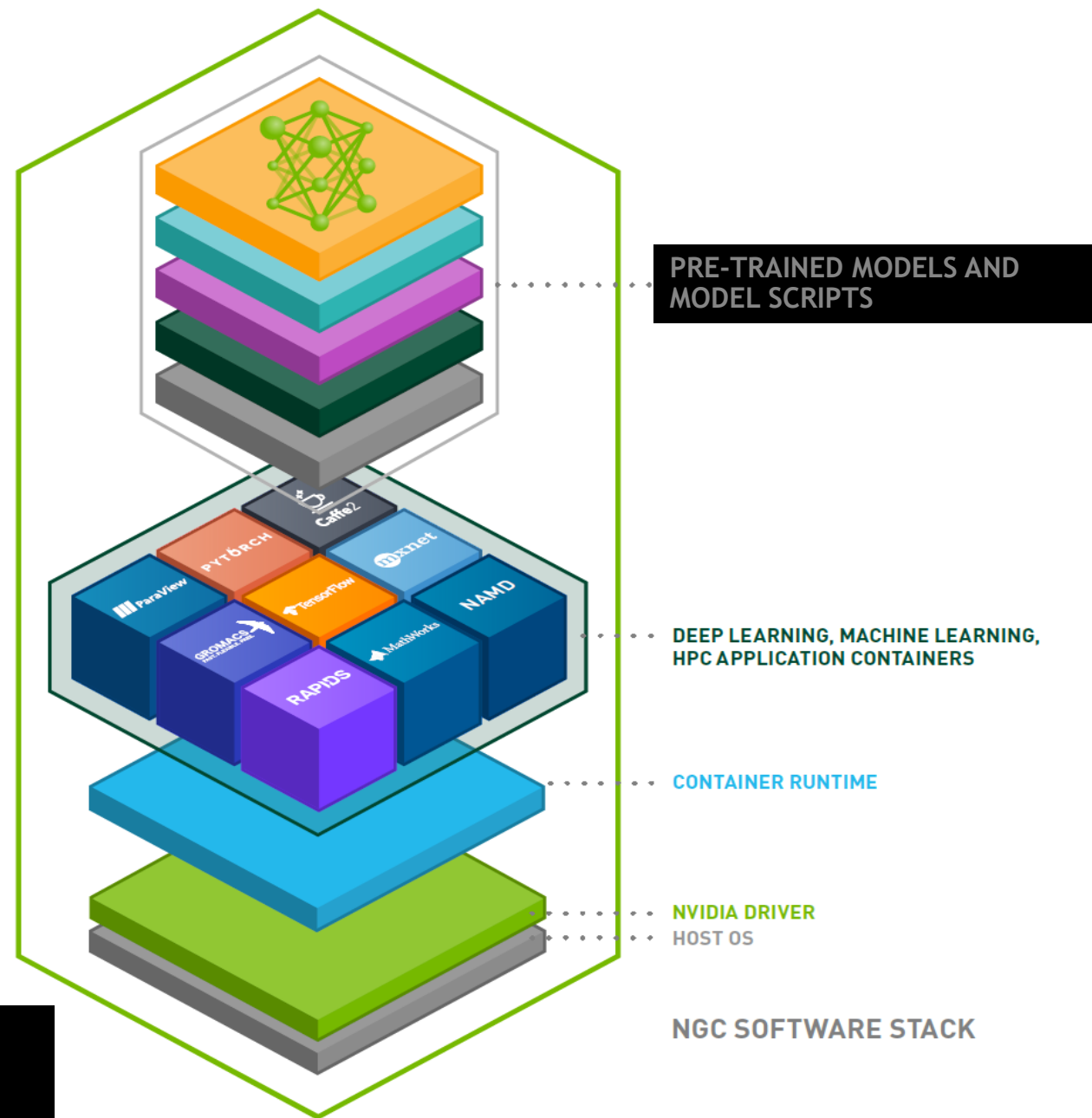
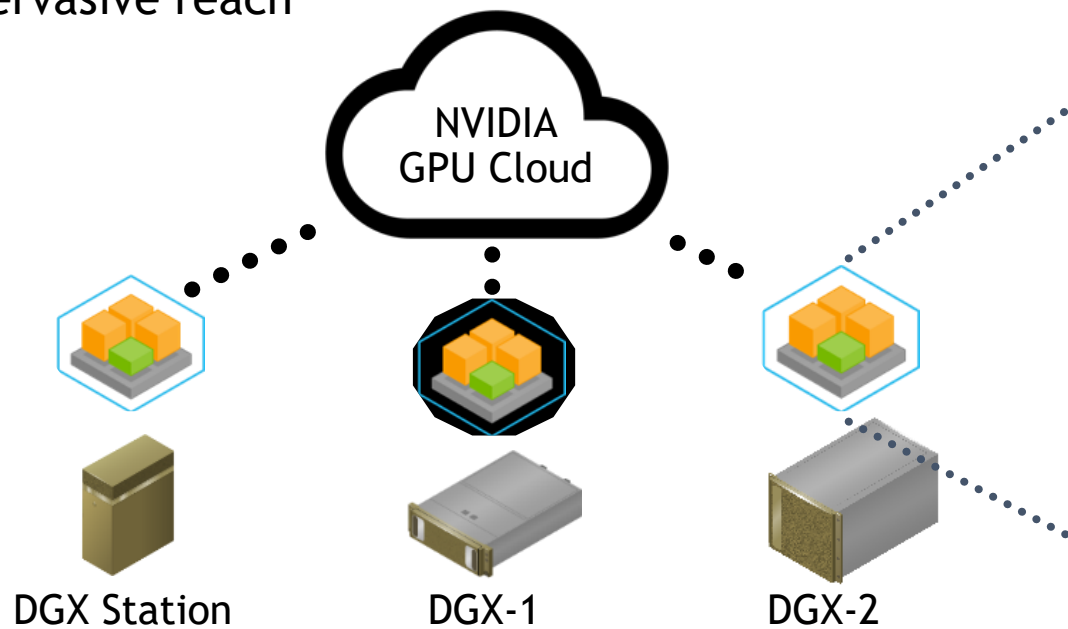
# FULL NON-BLOCKING BANDWIDTH



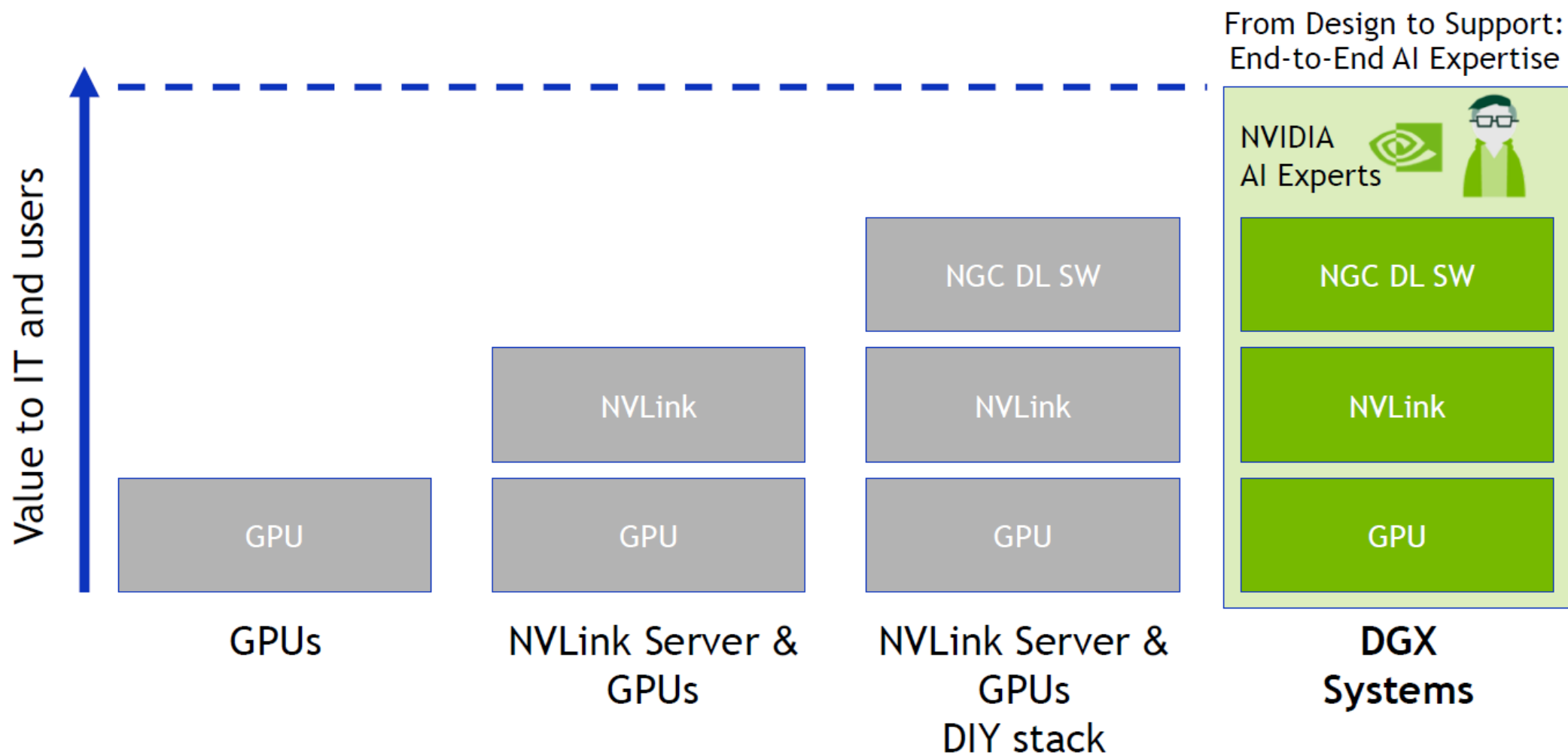
Single, unified stack for deep learning frameworks

Predictable execution across platforms

Pervasive reach



# DGX SYSTEMS – THE VALUE OF INTEGRATED SOLUTIONS AND AI EXPERTISE





Simplify, Accelerate, and Scale the Data Pipeline for Deep Learning



# ANNOUNCING NVIDIA DGX SUPERPOD

AI LEADERSHIP REQUIRES  
AI INFRASTRUCTURE LEADERSHIP

Test Bed for Highest Performance Scale-Up Systems

- 9.4 PF on HPL | ~200 AI PF | #22 on Top500 list
- <2 mins To Train RN-50

Modular & Scalable GPU SuperPOD Architecture

- Built in 3 Weeks
- Optimized For Compute, Networking, Storage & Software

Integrates Fully Optimized Software Stacks

- Freely Available Through NGC

Autonomous Vehicles | Speech AI | Healthcare | Graphics | HPC



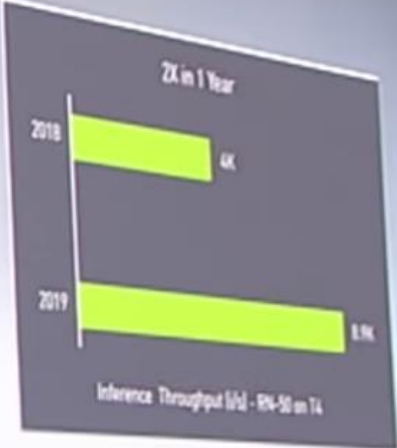
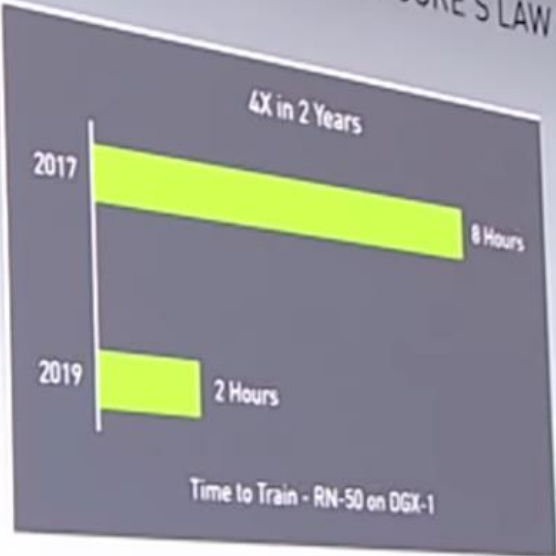
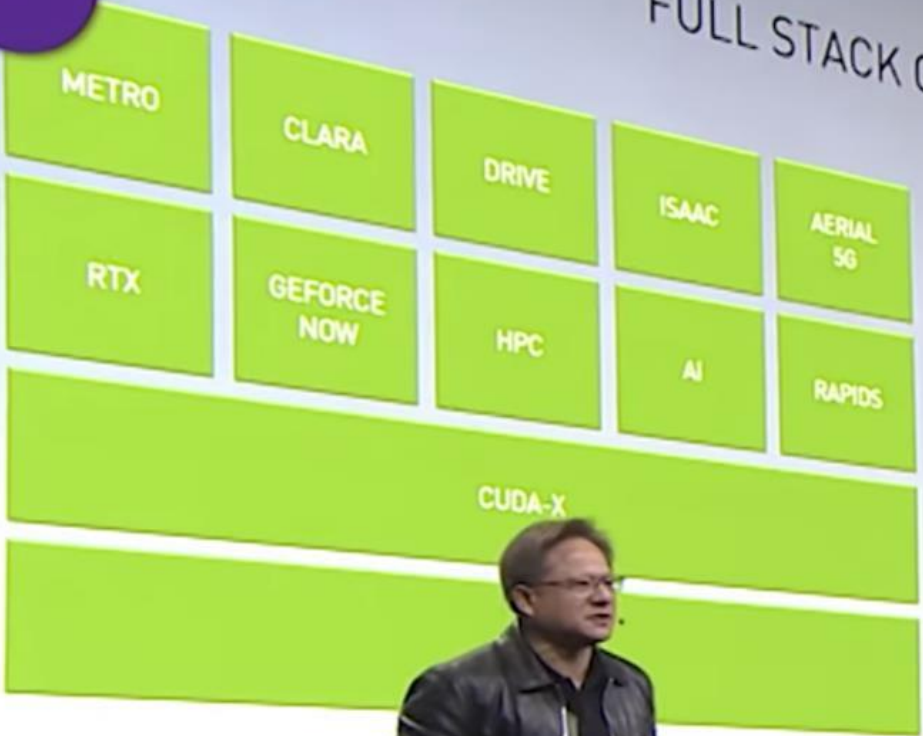
- 96 DGX-2H
- 10 Mellanox EDR IB per node
- 1,536 V100 Tensor Core GPUs
- 1 megawatt of power

# DGX 相关软件应用

500+

# FULL STACK OPTIMIZATION FOR SUPER MOORE'S LAW SPEEDUP

TensorRT 6.0  
cuDNN 7.6  
cuBLAS 10.2  
DALI 0.15  
NCCL 2.5  
  
CUDA 10.2



Inference Throughput on T4 (RTX) 2018 Model @ Inference with int8  
2019 Model @ Inference with int8

# ACCELERATION SOFTWARE ON NGC

## Ready-to-run GPU Optimized Software, Anywhere

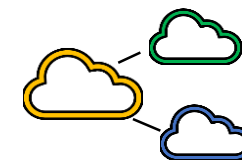
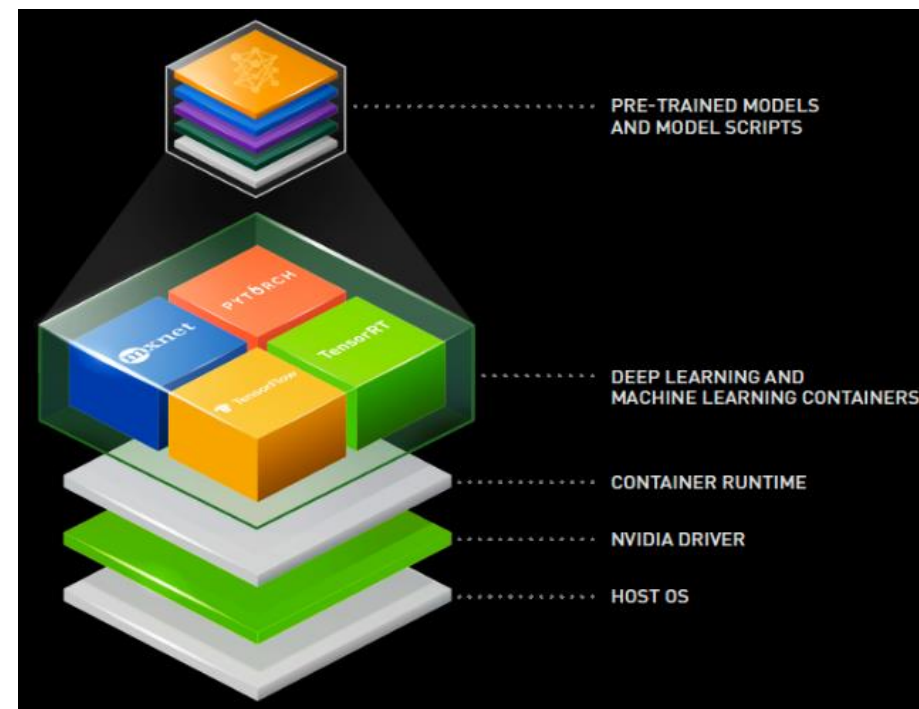
79+ Containers

27+ Model Training Scripts



60+ Pre-trained Models

Industry Workflows



# NGC.NVIDIA.COM

ACCELERATED SOFTWARE

SETUP

ALL CONTENT TYPES




CONTAINERS

MODELS

MODEL SCRIPTS

HELM CHARTS




Publisher: All

### CUDA

CUDA is a parallel computing platform and programming model that enables dramatic increases in computing performance by harnessing the power of the NVIDIA GPUs.




**latest**  
built by NVIDIA 12/31/19

### Wide & Deep for Tensor...

GPU accelerated inference for the Wide & Deep Recommender model




built by unknown 12/24/19

### Wide & Deep TensorFlow

Pretrained weights for the Wide & Deep model.




**2**  
built by unknown 12/24/19

### TensorFlow

TensorFlow is an open-source software library for high-performance numerical computation. Its flexible architecture allows easy deployment of computation ...




**19.12-tf1-py3**  
built by NVIDIA 12/20/19

### TensorRT Inference Serv...

TensorRT Inference Server provides a data center inference solution optimized for NVIDIA GPUs. It maximizes inference utilization and performance on GPUs via ...




**19.12-py3-clientsdk**  
built by NVIDIA 12/20/19

### DIGITS

The NVIDIA Deep Learning GPU Training System (DIGITS) puts the power of deep learning into the hands of engineers and data scientists.




**19.12-tensorflow-py3**  
built by NVIDIA 12/20/19

### PyTorch

PyTorch is a GPU accelerated tensor computational framework with a Python front end. Functionality can be easily extended with common Python libraries ...




**19.12-py3**  
built by NVIDIA 12/20/19

### TensorRT

NVIDIA TensorRT is a C++ library that facilitates high-performance inference on NVIDIA graphics processing units (GPUs). TensorRT takes a trained network, which ...




**19.12-py3**  
built by NVIDIA 12/20/19

### MXNet

MXNet is a deep learning framework that allows you to mix the flavors of symbolic programming and imperative programming to maximize efficiency and ...

**19.12-py3**  
built by NVIDIA 12/20/19

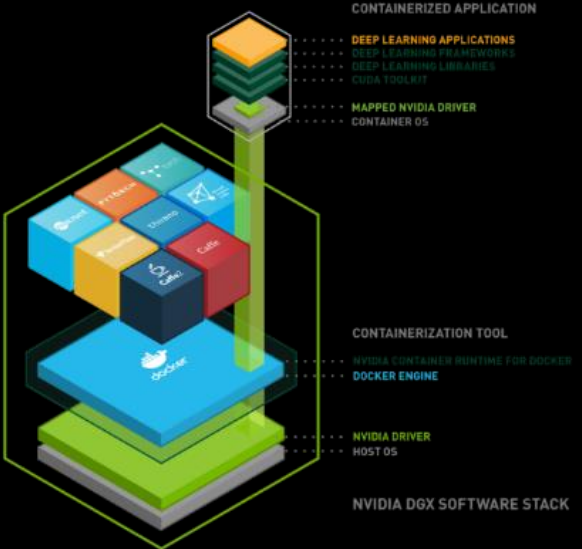
### Kaldi

Kaldi is an open-source software framework for speech processing.

**19.12-py3**  
built by NVIDIA 12/20/19

# Kubernetes on NVIDIA GPUs

- Caffe2
- CNTK
- GAMESS
- GROMACS
- DeepStream
- DIGITS
- NAMD
- NVCaffe
- PaddlePaddle
- PyTorch
- RELION
- TensorFlow
- Kinetica
- MATLAB
- OmniSci (MapD)
- RAPIDS
- TensorRT
- TensorRT Inference
- Server



10 containers

October 2017

## SOFTWARE ON THE NGC CONTAINER REGISTRY

79 containers

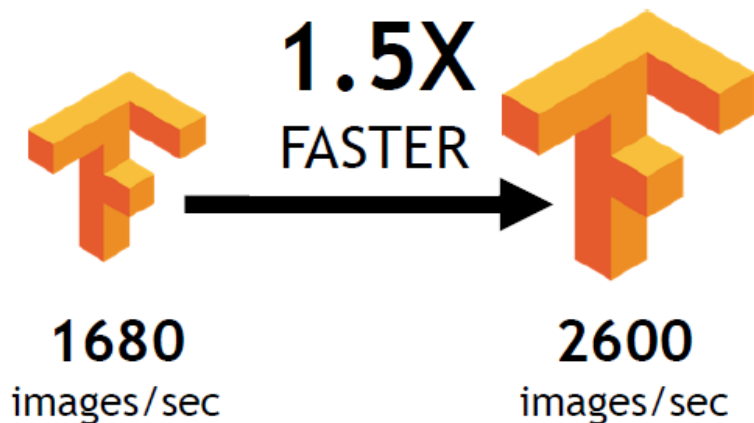
Jan. 2020

# 实际的性能提升

GLOBAL TECHNOLOGY FIRM  
SPECIALIZING IN DIGITAL MEDIA

Home-grown  
“optimized”  
TensorFlow stack

DGX+NGC  
TensorFlow  
stack

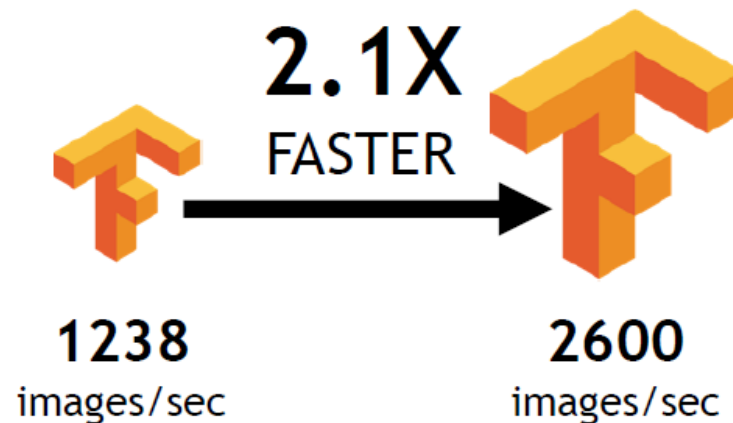


ResNet50 Training

WORLD-LEADING MEDICAL  
RESEARCH CENTER

Home-grown  
TensorFlow stack

DGX+NGC  
TensorFlow  
stack



ResNet50 Training



# DGX 医疗教育相关行业案例



## AI PLATFORM TO ACCELERATE CANCER RESEARCH

To speed advances in the fight against cancer, the Cancer Moonshot initiative unites the Department of Energy, the National Cancer Institute and other agencies with researchers at Oak Ridge, Lawrence Livermore, Argonne, and Los Alamos National Laboratories. NVIDIA is collaborating with the labs to help accelerate their AI framework, CANDLE, as a common discovery platform, with the goal of achieving 10X annual increases in productivity for cancer researchers.



Image: Graphical recording created at the Cancer Moonshot Summit, June 29, 2016.  
Image credit: [Adi Leigh Brown](#)/White House Cancer Moonshot Task Force.

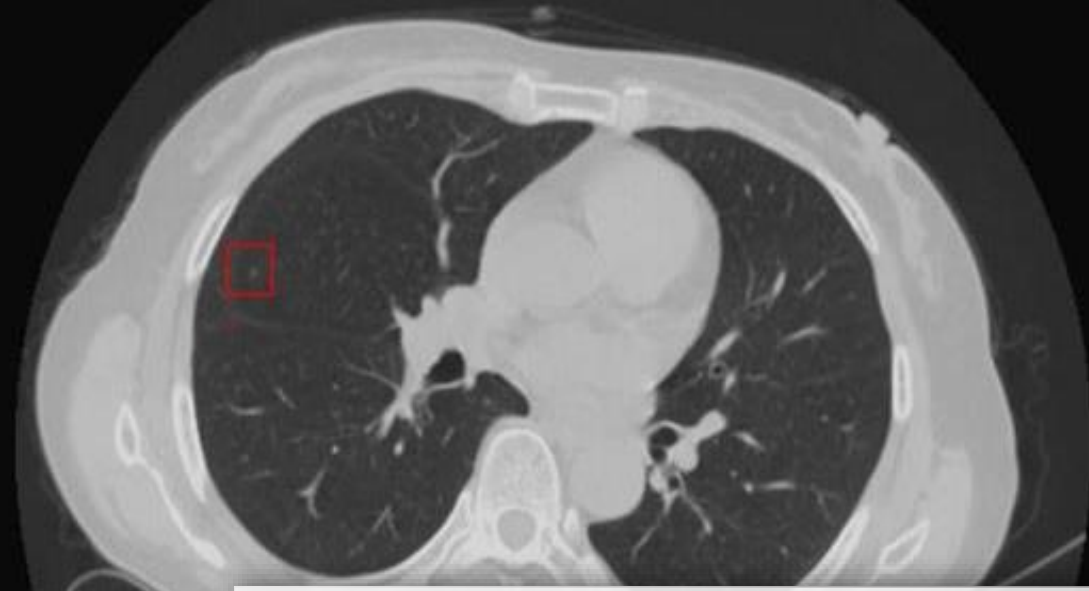
# Lung Nodule Detection



- Completed Tasks
- Step 1 - Upload Medical Images
  - Step 2 - Deep Learning Analysis
  - Step 3 - Nodule Confirmation
- Upcoming Tasks
- Step 4 - Create Report

**1515046947**  
2015-07-06/09:24:10

Pos:0, 0, 130, 265  
Thickness:1.250000 mm  
Patient Name:



WindowCenter = -500  
WindowWidth = 1500

# Lung Nodules Screening Report

## Patient Information

Exam Date:20150706  
Exam Time:092410.375564  
Patient Name:

## Observation

Found a nodule at slice 125-127 Nodule short diameter is 2.23mm. The nodule long diameter is 2.23 mm  
Found a nodule at slice 128-131 Nodule short diameter is 2.23mm. The nodule long diameter is 2.23 mm

## List of Detected Nodules

Index	Measurement	Location
1	2.23,2.23	87-91
2	4.34,6.93	97-98
3	2.23,2.23	125-127
4	2.23,2.23	128-131
5	2.23,2.23	140-142
6	10.29,11.46	148-149
7	6.84,11.52	160
8	10.79,11.95	179
9	7.42,11.13	198-199

# AI TOOL SPEEDS UP CT SCAN SCREENING

CT scans help Radiologists diagnose lung cancer, but it can take up to 15-20 minutes to scrutinize one image series.


InferVISION's GPU-powered InferRead CT Lung system automatically identifies and labels lung nodules in under 30 seconds. When using NVIDIA Tesla T4 GPUs for inference, InferVISION achieved speedups of ~4X over enterprise grade CPU. This automation is helping radiologists reduce workloads and file diagnostic reports with confidence.



- 联合武汉同济、中南医院和深圳三院合作开发，业界首家新冠肺炎AI系统
- 快速筛查、智能预警 加速医生诊断、增强诊断信心、减少院内感染分享
- 精准分析、快速评估 AI对肺部CT影像自动提取密度、形态、体积等信息，辅助医生评估患者肺部受累情况、帮助医生量化判断病人病，为诊断及治疗提供依据
- 全自动智能病情及疗效评估 提供全自动前后片对比功能，详细对比患者前后病灶变化，提供量化分析，为病情变化和疗效评估提供依据
- 区域疫情实时监控 区域内医疗机构统一部署后，AI系统自动汇总各级医院数据、形成肺炎报告及分布、形成上下联动的区域疫情监测与防控信息
- 使用的NVIDIA技术：GPU、RAPIDS、CLARA SDK等
- 单体医院：2U2卡服务器（500套CT薄层日处理量）

# GPU-accelerated

## 武汉同济医院肺炎 AI 系统 上线试用报告



夏黎明 (武汉同济医院放射科主任医师)  锁定

 本词条由华中科技大学同济医学院附属同济医院 认证。

夏黎明, 男, 华中科技大学同济医学院附属同济医院放射科主任, 教授、主任医师, 博士生导师, 同济医院名医《放射学实践》杂志编辑部主任《放射学实践》杂志常务副主编, 《中华放射学杂志》编委、《中华解剖与临床杂志》编委、《临床放射学杂志》常务编委、《中国CT和MRI杂志》编委、《磁共振成像》编委教育背景。参与十一五国家科技支撑计划课题, 主持国家自然科学基金、省级课题4项, 参与国家级、省级课题8项, 课题鉴定2项, 获省、市级科技进步二、三等奖4项, 结题项目3项, 在研课题1项。

中文名称	夏黎明	临床职称	主任医师
毕业院校	武汉医学院	教学职称	教授
		执业地点	武汉同济医院



项目名称	武汉同济医院肺炎特别版 AI 系统试用项目		
系统部署时间	2020年1月29日	试运行起始时间	2020年1月30日
系统对接方案	<input type="checkbox"/> CT 主控台直接发送 DICOM 影像数据到推想服务器 <input type="checkbox"/> PACS 转发 DICOM 影像数据到推想服务器 <input checked="" type="checkbox"/> AI 服务器主动从 PACS 数据库读取 DICOM 影像数据 <input type="checkbox"/> 双方协定数据信息 API 接口		
产品版本	InferRead 肺炎特别版 V	影像脱敏	<input checked="" type="checkbox"/> 是 <input type="checkbox"/> 否
接入 CT 设备范围	医院全部	接入 CT 影像数量	约日均 500-700 例
AI 阅片端安装范围	胸组全部、科室局部	使用医生范围	医师、规培医师
供货厂商	北京推想科技有限公司		
科室负责人	夏黎明	联系方式	136 0717 6908
系统情况说明	肺炎特别版 AI 产品, 针对肺炎以及病毒性肺炎影像进行处理。可以将疑似感染区域快速的进行体积和密度的测量及分析, 协助医生精准评估病情; 随后推想 AI 还将增加患者的病程评估及转归分析功能, 通过自动比对 CT 影像在形态、密度、大小等的变化, 从而快速完成随访和病程发展的评估。 推想人工智能处理的时间在 10 秒左右, 在阅片端提示病例的异常, 帮助医生快速的诊断肺炎患者、提升工作效率, 减少患者交叉感染的风险。		
科室使用医生	试用签字:  2020 年 1 月 31 日 		

公司名称: 北京推想科技有限公司

目标用户: 医院、医疗机构、各级卫健委

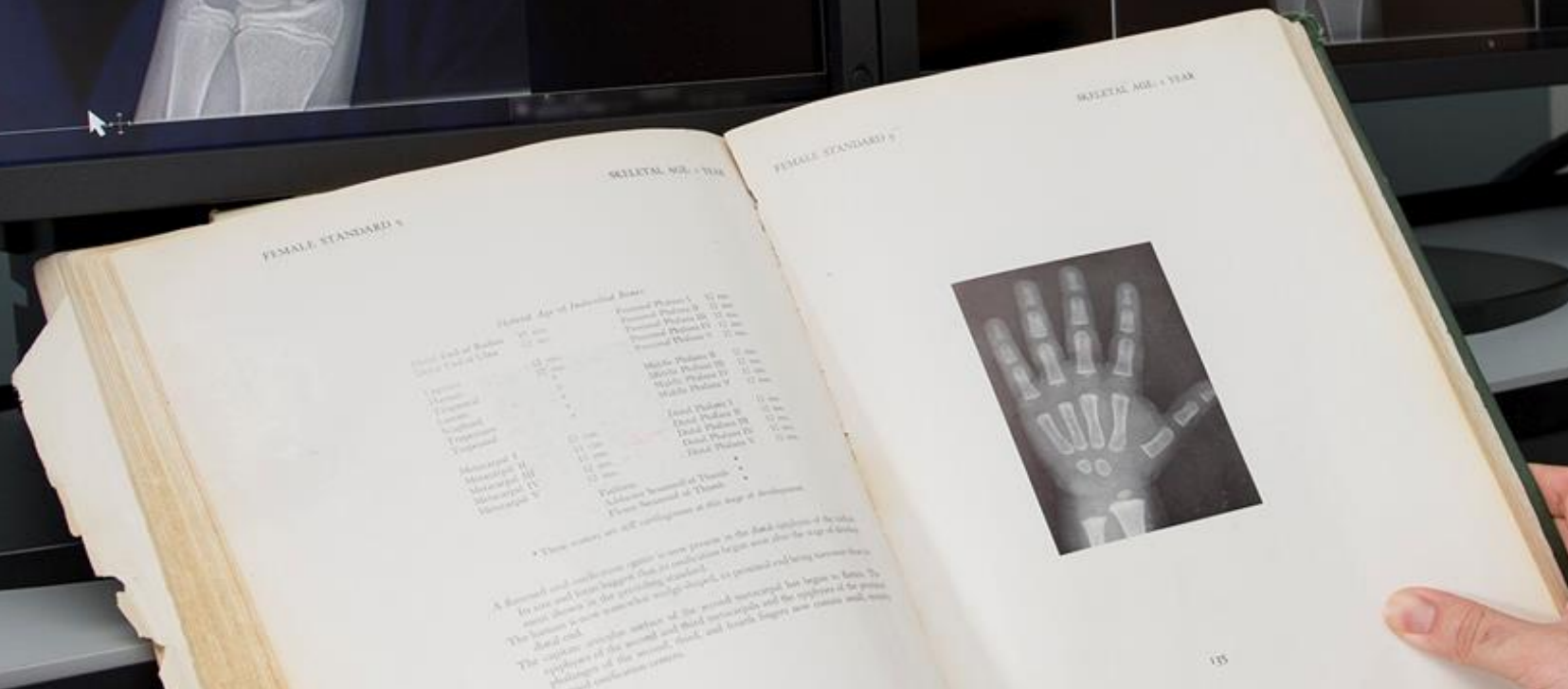
成功案例: 武汉同济医院、武汉中南医学、武汉金银潭医院、深圳三院

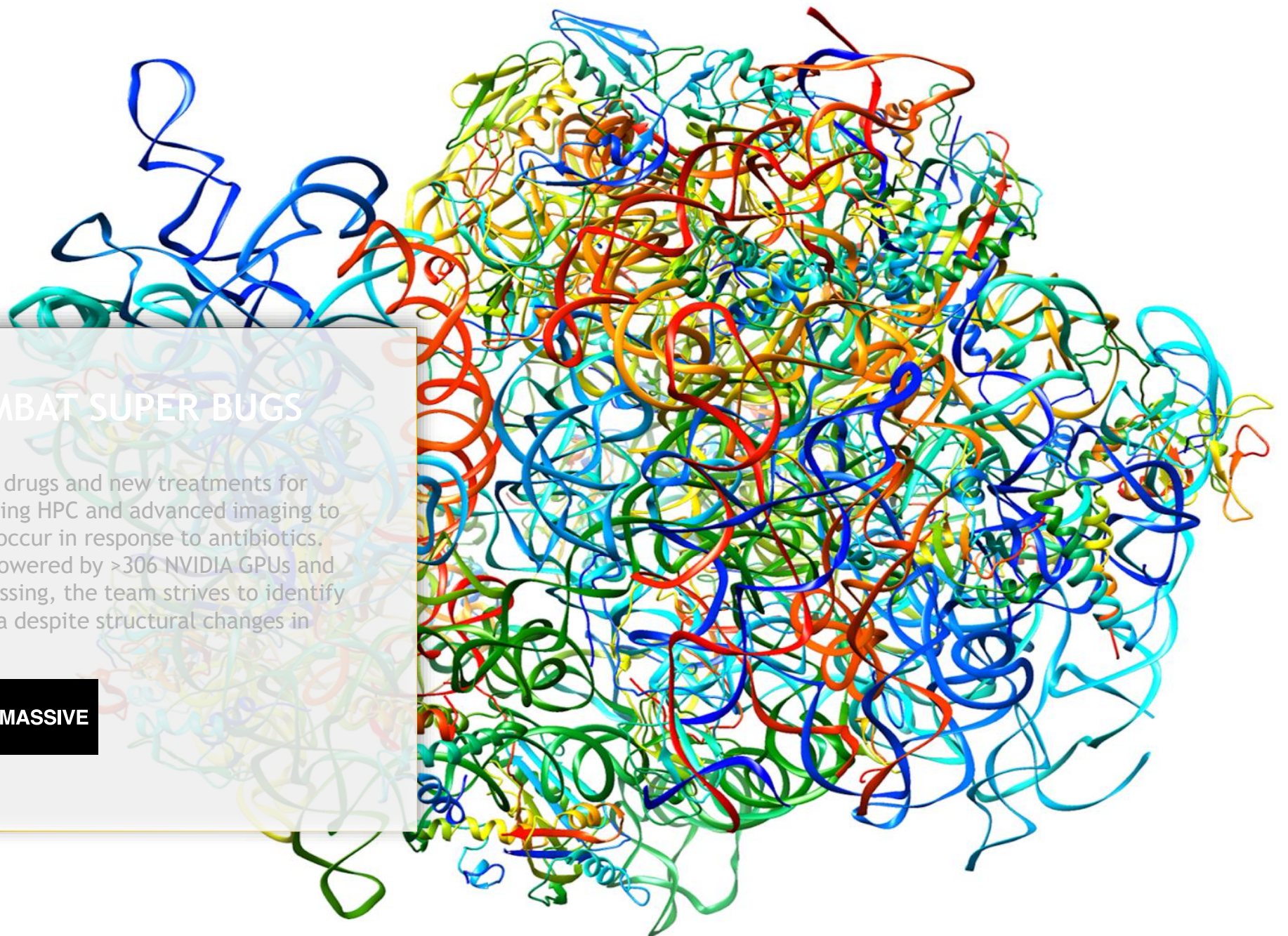


# DETECTING GROWTH PROBLEMS IN CHILDREN

Detecting growth-related problems in children requires calculating their bone age. But it's an antiquated process that requires radiologists to match X-rays with images in a 1950s textbook.

Massachusetts General Hospital, which conducts the largest hospital-based research program in the US, developed an automated bone-age analyzer built on the NVIDIA DGX-1. The system is 99% accurate and delivers test results in seconds versus days.



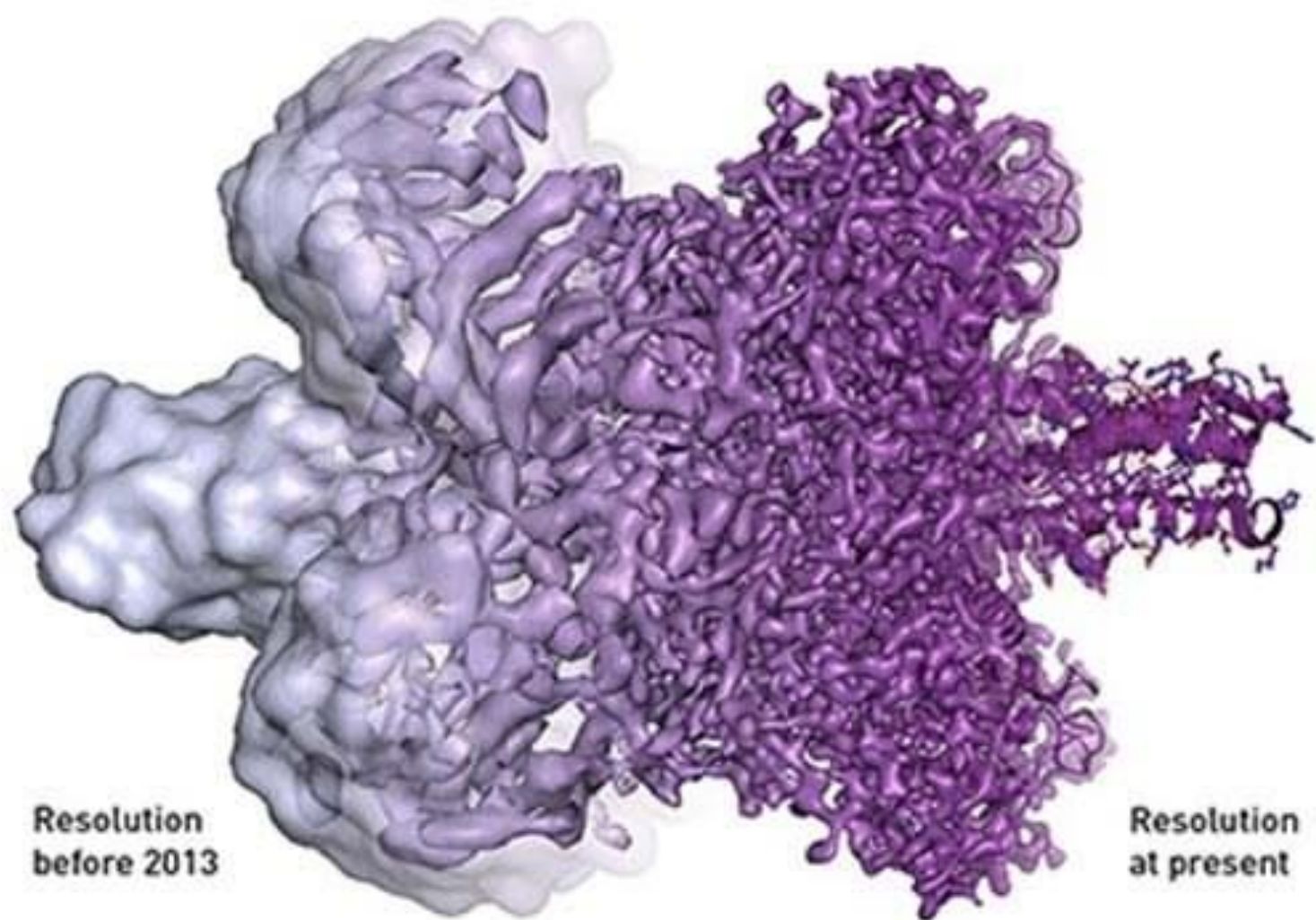


## SUPER DRUGS TO COMBAT SUPER BUGS

In the race to design more effective drugs and new treatments for diseases, Australian scientists are using HPC and advanced imaging to visualize changes in ribosomes that occur in response to antibiotics. With MASSIVE's M3 supercomputer powered by >306 NVIDIA GPUs and 6 DGX-1V's to accelerate data processing, the team strives to identify new drugs that are lethal to bacteria despite structural changes in ribosomes.



Image: One of Dr Matt Belousoff and Professor Trevor Lithgow ribosome structures illustrating complicated details that can be determined using MASSIVE and the Titan Krios at the Ramaciotti Centre for Cryo Electron Microscopy.



Resolution  
before 2013

Resolution  
at present



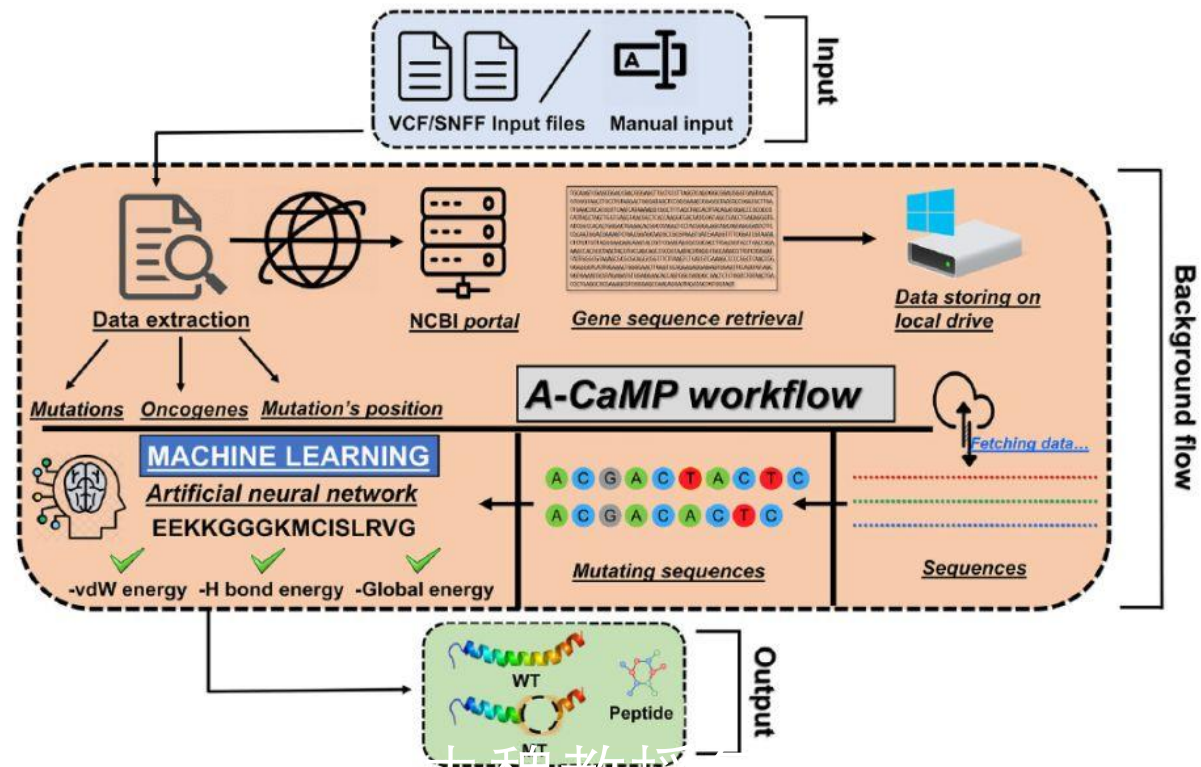
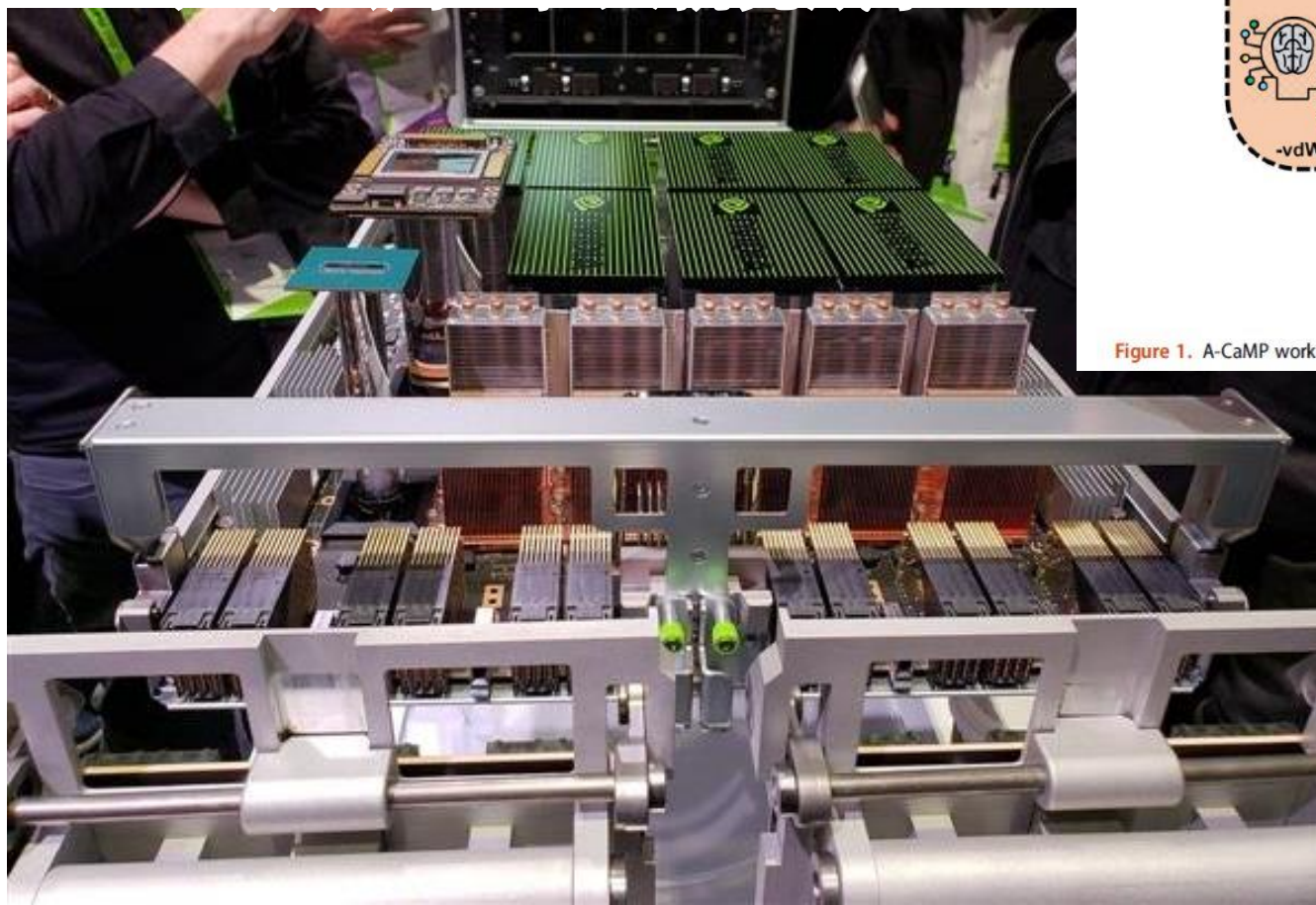
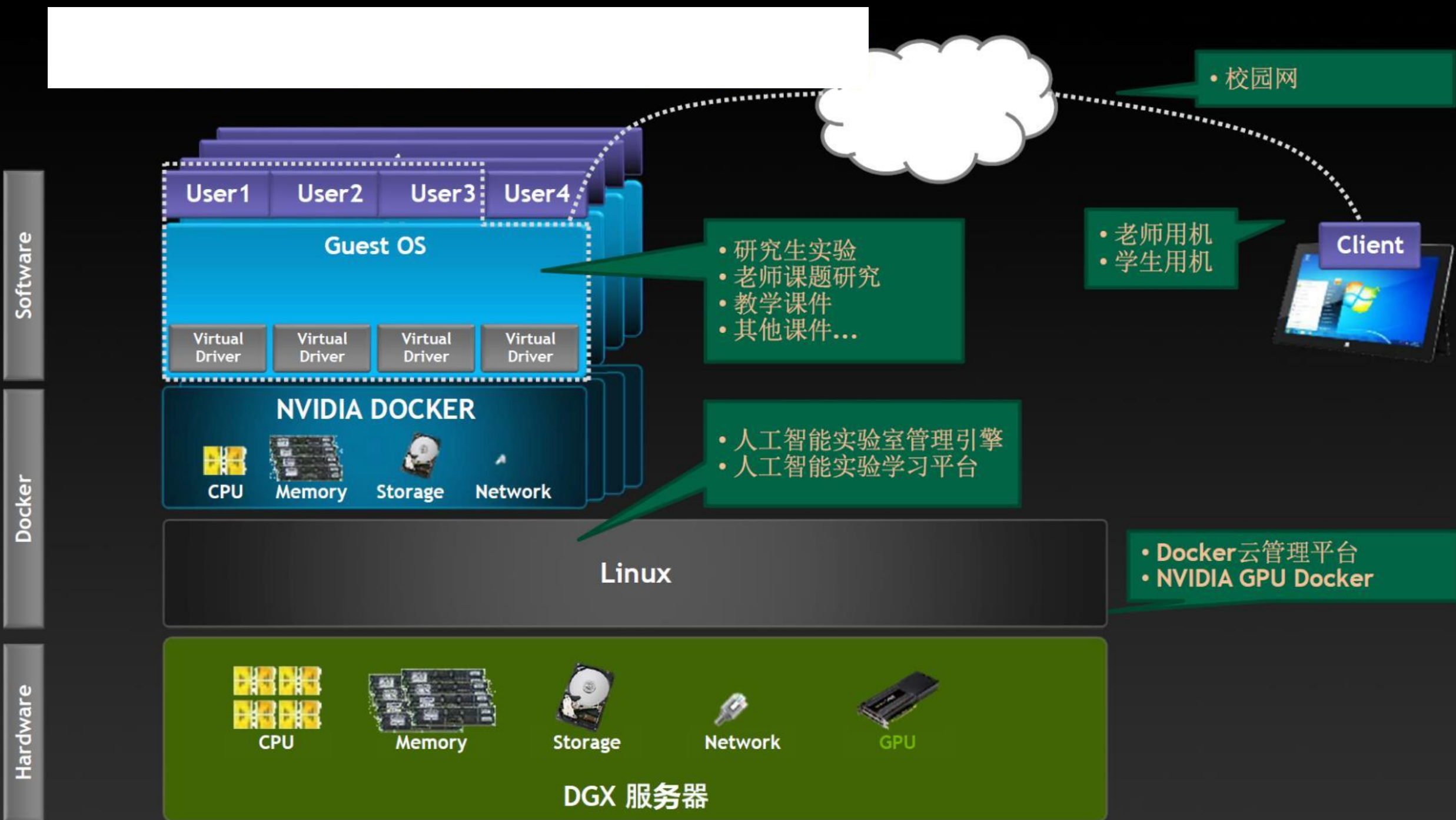
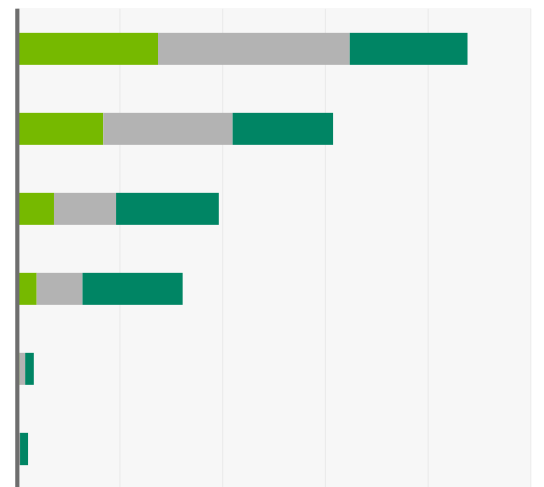
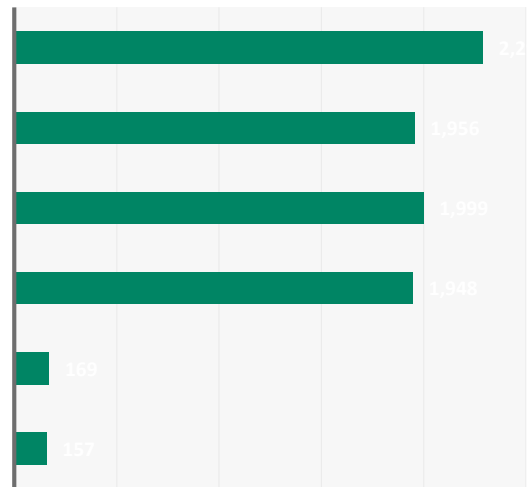
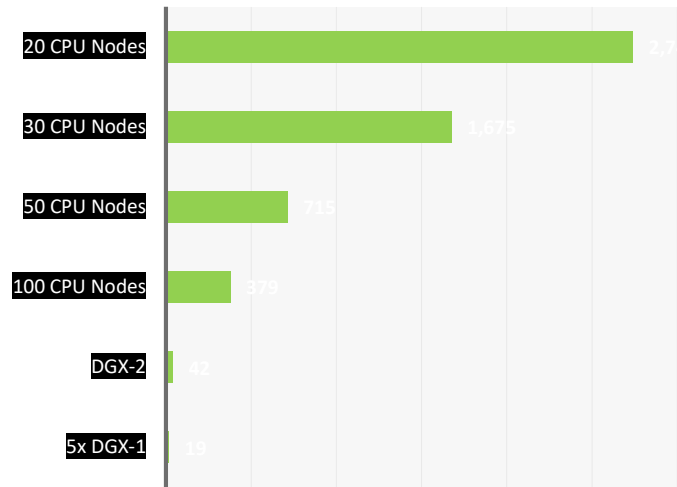


Figure 1. A-CaMP workflow. A flowchart depicting A-CaMP accessibility, peptide prediction through GUI interface, and the automated background operations.





CPU nodes (



**NVIDIA**®



联想凌拓



# Lenovo NetApp携手NVIDIA

## 加速人工智能之旅

联想凌拓产品经理 尹伟

英伟达高级客户经理 吴强

# AI市场分析

# 新基建：前所未有的经济刺激举措



特高压电网



高铁与城市轨道交通



新能源汽车充电桩



新型智慧城市与园区



5G基站与网络



大数据中心



人工智能



工业互联网

图中的数字是预计2020年投资额，8大新基建2020年投资总额约2.11万亿元人民币



# 人工智能(AI), 机器学习(ML)与深度学习(DL)

## ARTIFICIAL INTELLIGENCE

Techniques that enable computers to mimic human intelligence.

包括所有让计算机模拟人类行为的技术



1950's

## 人工智能：目标

## MACHINE LEARNING

Ability to learn without being explicitly programmed

无需明确编程的学习能力



1980's

## 机器学习及深度学习：方法

## DEEP LEARNING

Learning based on Deep Neural Network

使用多层神经网络学习数据中的底层特性



2010's

# 人工智能AI拥有巨大机会

**40%**

of Digital  
Transformation  
initiatives will use  
AI services by 2019

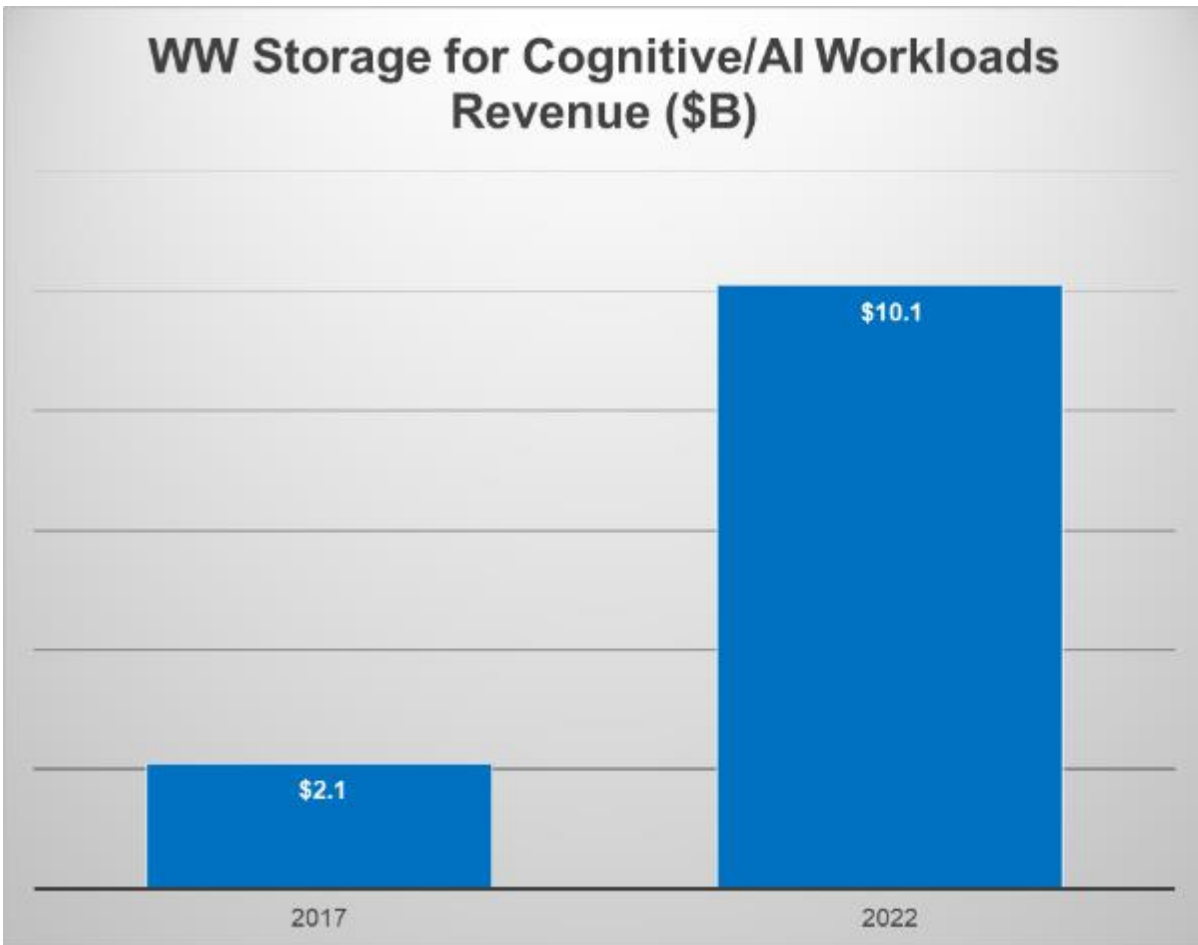
**50%**

of enterprise  
infrastructure will  
employ AI by 2021

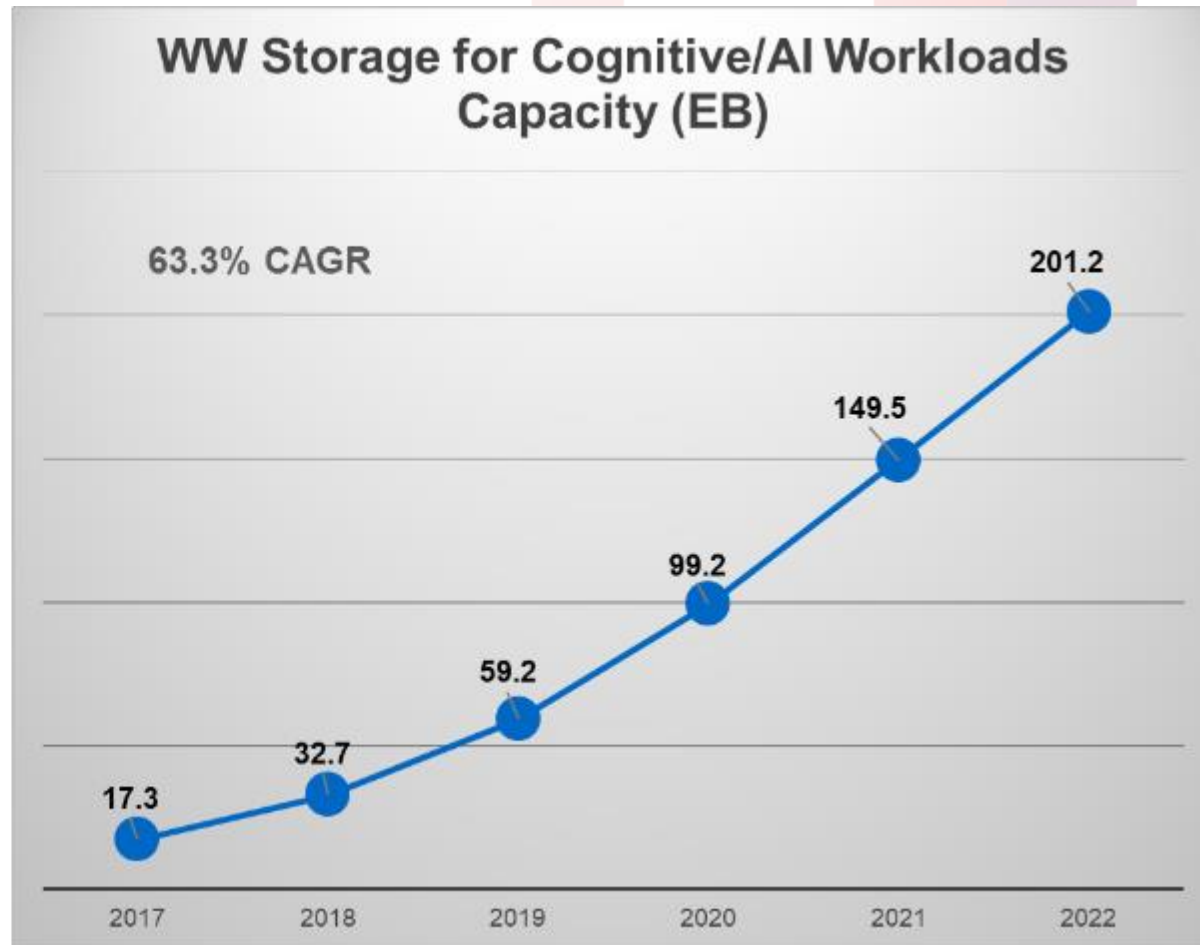
Source IDC – Storage Workloads : Understanding The 2nd Chapter, July 2018

# 存储在人工智能AI领域的机会

WW Storage for Cognitive/AI Workloads Revenue (\$B)



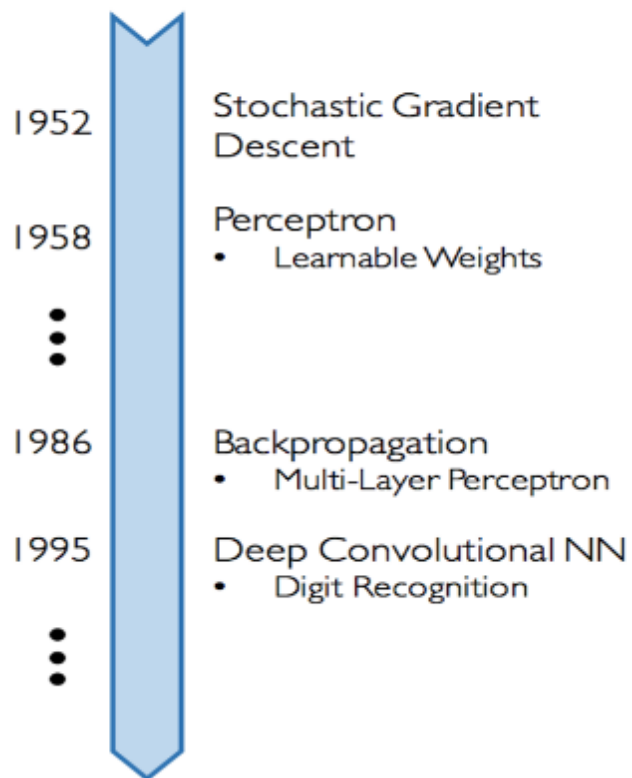
WW Storage for Cognitive/AI Workloads Capacity (EB)



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

# 现在是AI最好的机会

## Why Now?



## DIGITAL TRANSFORMATION

Neural Networks date back decades, so why the resurgence?

### 1. Big Data

- Larger Datasets
- Easier Collection & Storage

IMAGENET



WIKIPEDIA  
The Free Encyclopedia



### 2. Hardware

- Graphics Processing Units (GPUs)
- Massively Parallelizable



### 3. Software

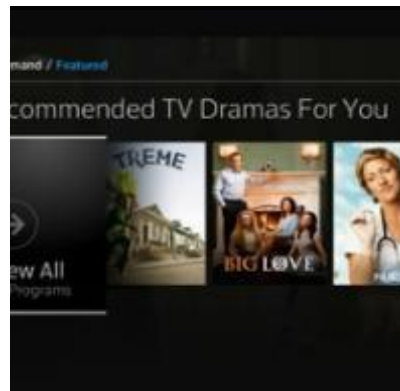
- Improved Techniques
- New Models
- Toolboxes



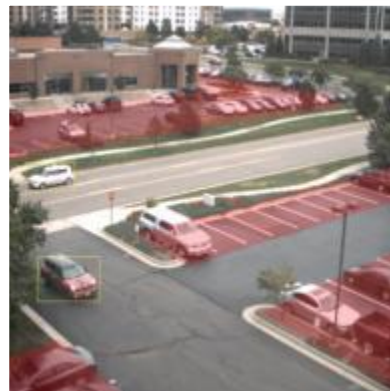
# 人工智能助力各行各业的业务转型



生产制造



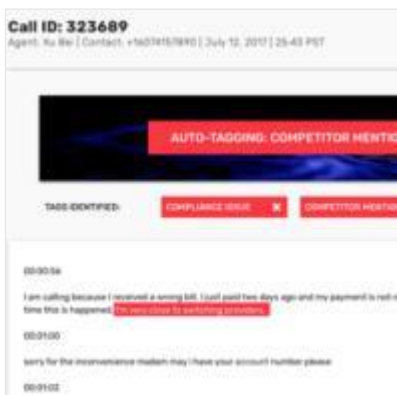
媒体



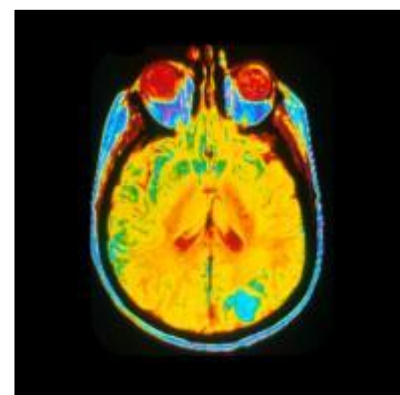
智慧城市



娱乐



零售



医疗健康



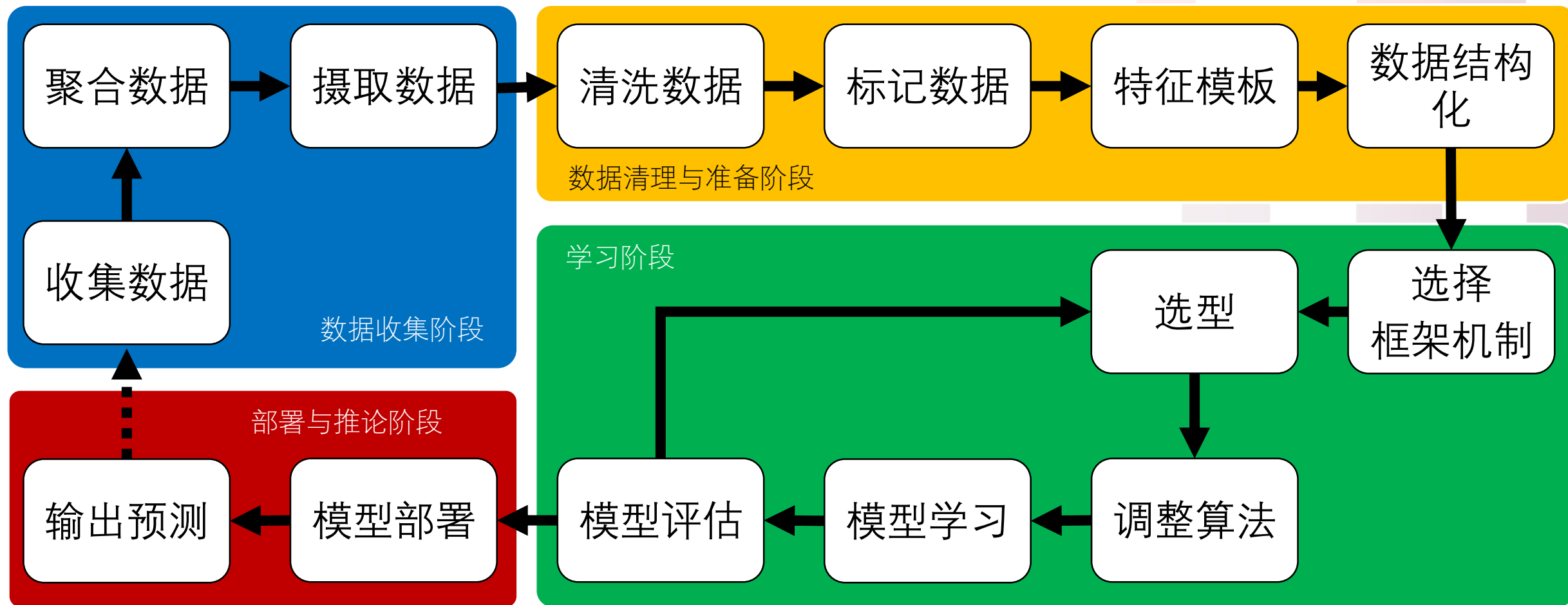
商业



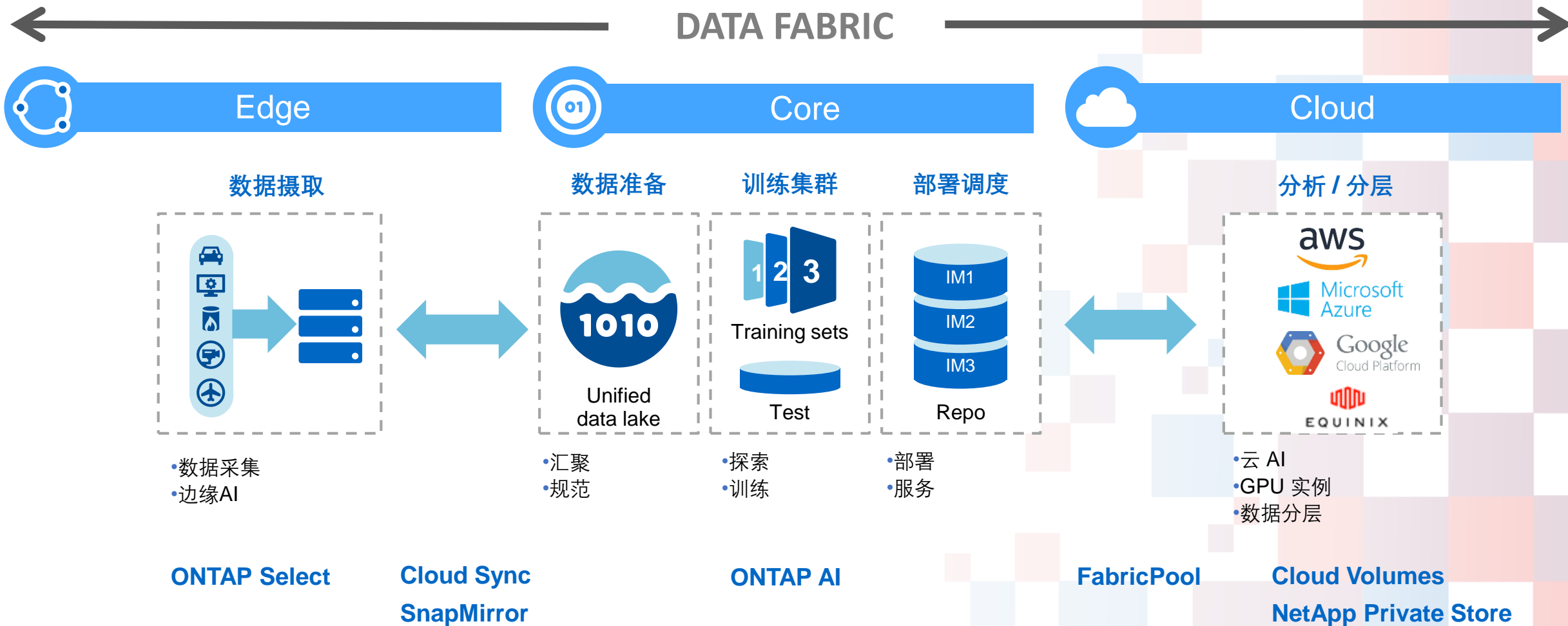
运输交通

# 解决方案介绍

# 人工智能与机器学习 workflows 的四个阶段



# Data Fabric助力AI与机器学习的应用





# ONTAP AI – AI加速器

经过验证



1个DGX 到 100个DGX

企业级性能

**300GB/s**

4-6x Performance  
vs. Competition

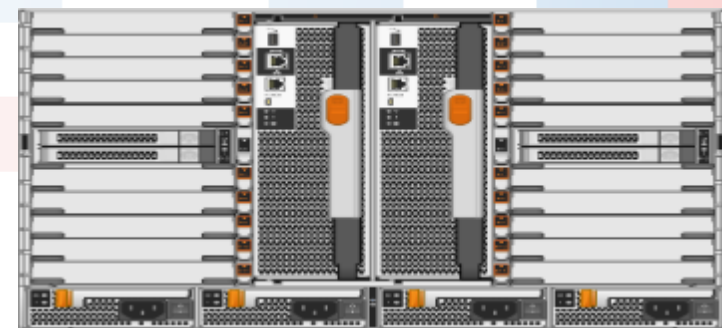
**#1 NFS**

DGX-1 Uses NFS by Default

**79PB**

> 25x Raw Storage Capacity  
vs. Competition

全闪存存储



# 经过市场验证的AI解决方案架构

## Hardware

- NVIDIA DGX-1 supercomputers
- NetApp AFF A800 cloud-connected all flash
- Cisco Nexus 3232C 100GbE switches

## Software

- NVIDIA GPU Cloud Deep Learning Stack
- NetApp ONTAP 9
- Trident, dynamic storage provisioner

## Support

- Tech. partner collaboration through TSANet's integrated support framework

### 简洁SIMPLE

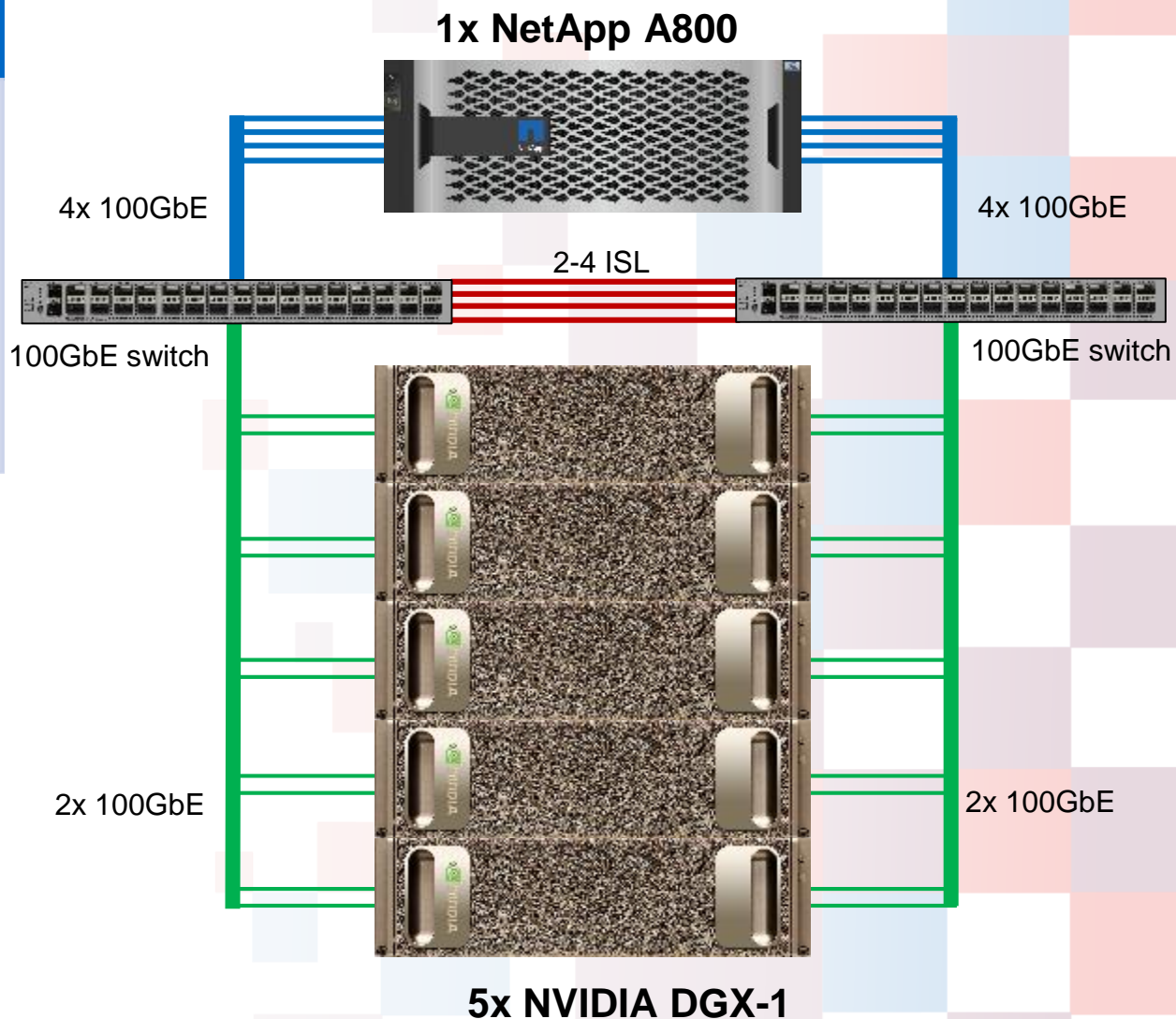
无需为猜测你的深度学习架构  
交付完整的深度学习平台

### 整合INTEGRATED

整合管理你的数据，无论它在终端，核心，云端  
自信的交付AI框架

### 高效POWERFUL

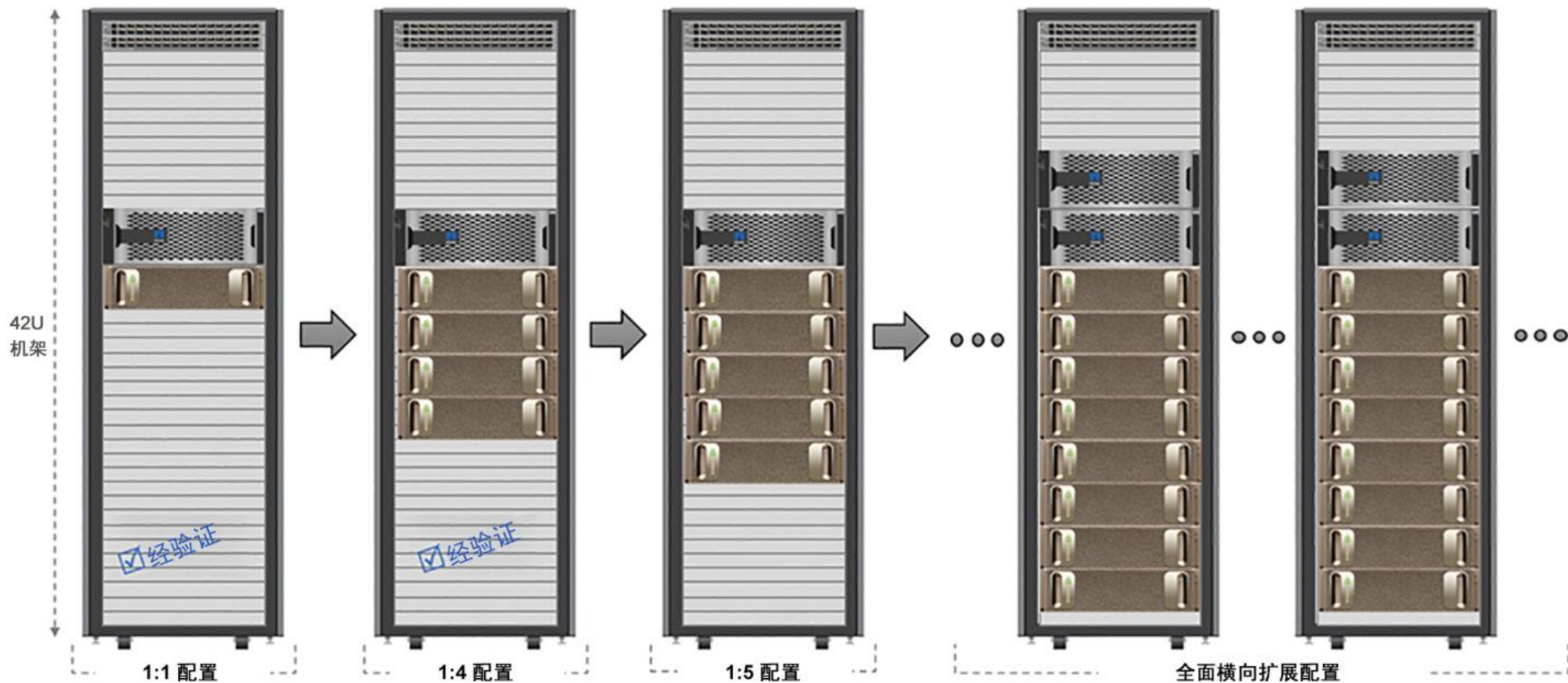
大规模扩展的深度学习架构  
从小起步，无中断扩展



# 适用于深度学习的机架级架构

从 1:1 横向扩展到 1:5 甚至更高配置

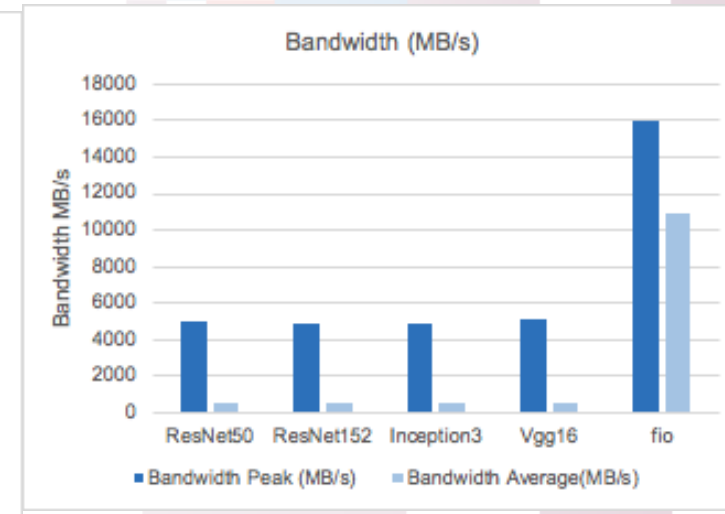
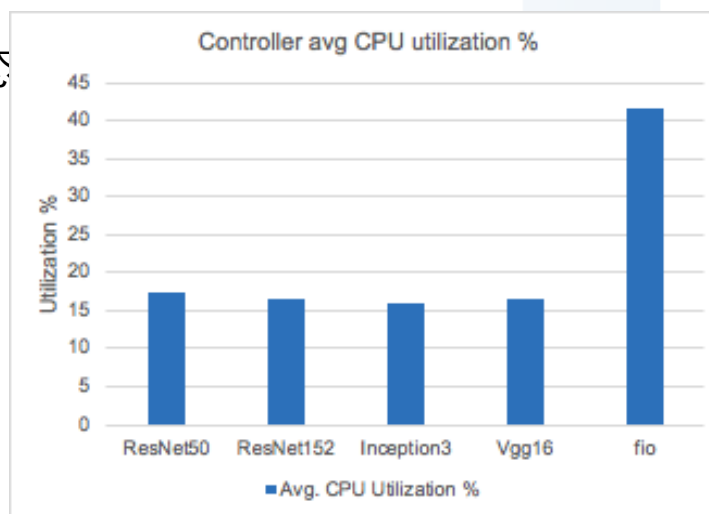
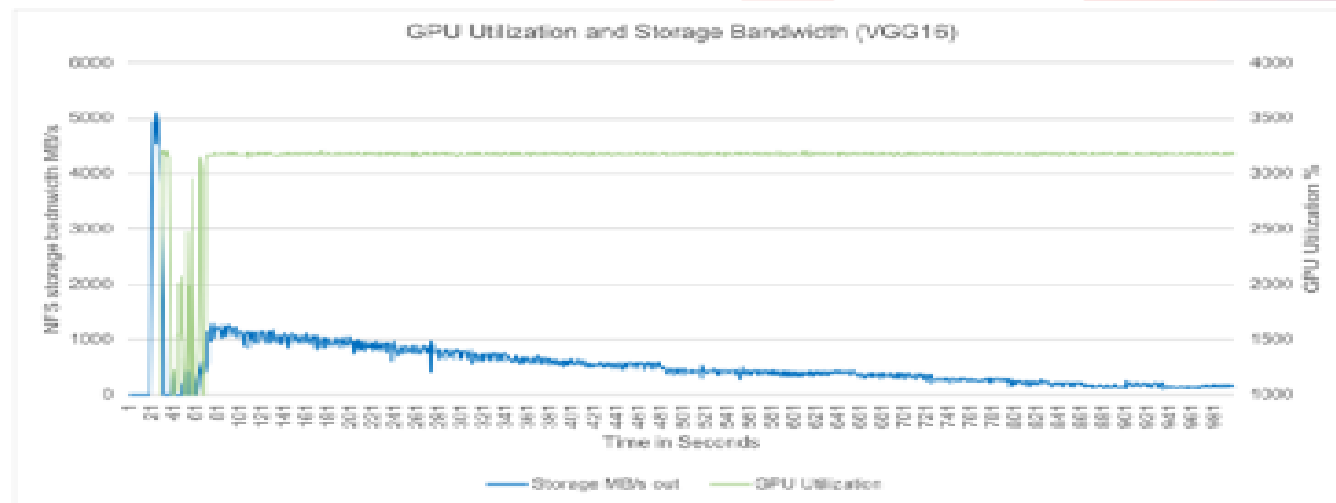
\*机架密度根据每个机架 35kW 电源计算得出



# 无与伦比的性能指标

在整个训练阶段，GPU 利用率保持在接近 100% 的水平，这表明 A800系统在保持高训练速度的同时，能够足够快速地为 GPU 馈送数据。

- 关键指标
  - 4 台DGX-1 GPU服务器+ 1台A800
    - 峰值带宽需求: 5GB/s
    - 持续带宽需求: 4GB/s
    - 平均存储时延: ~600us
    - 所有32个GPU保持持续忙碌状态 >95%
    - 存储CPU利用率: ~18%
    - A800可提供 25GB/s 读带宽



# AI/DL:从小到大, 无缝扩展

配置NetApp AFF和NVIDIA DGX-1



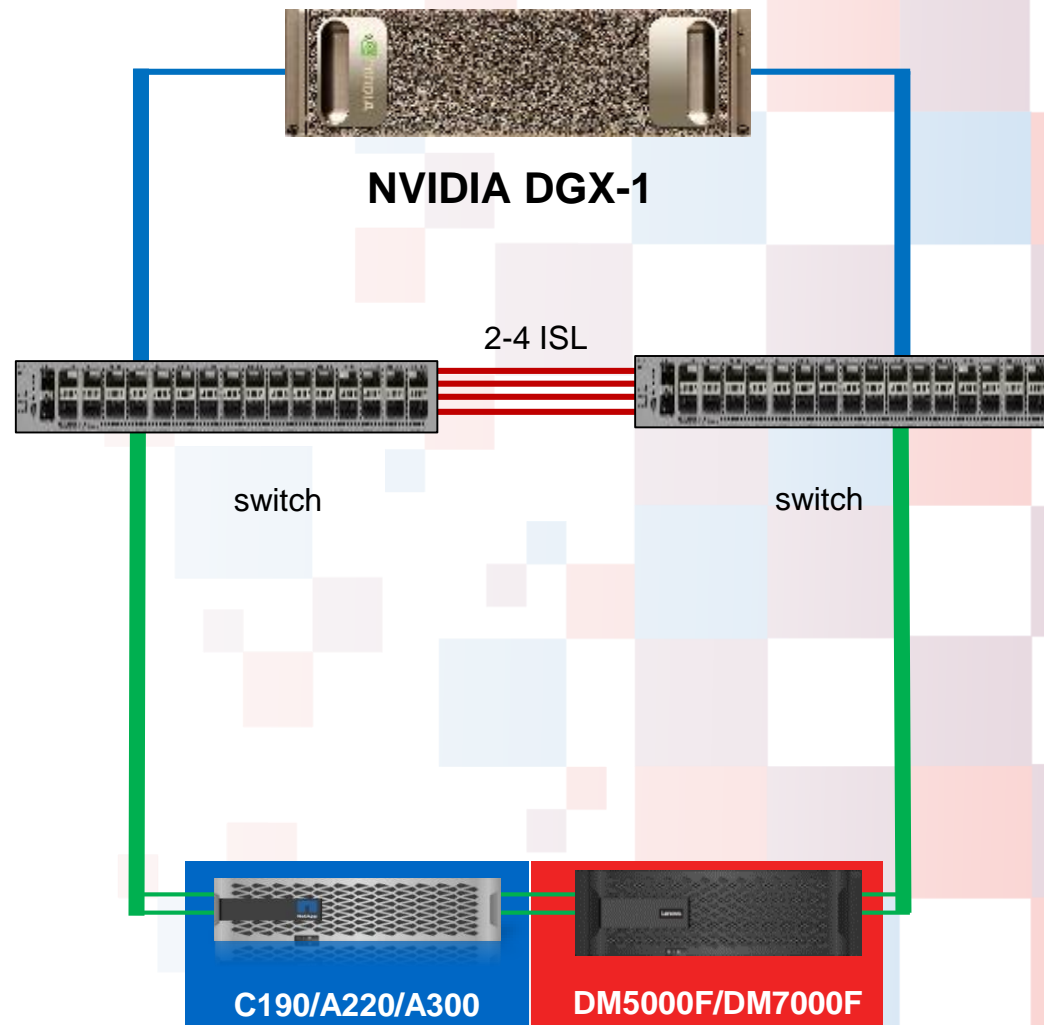
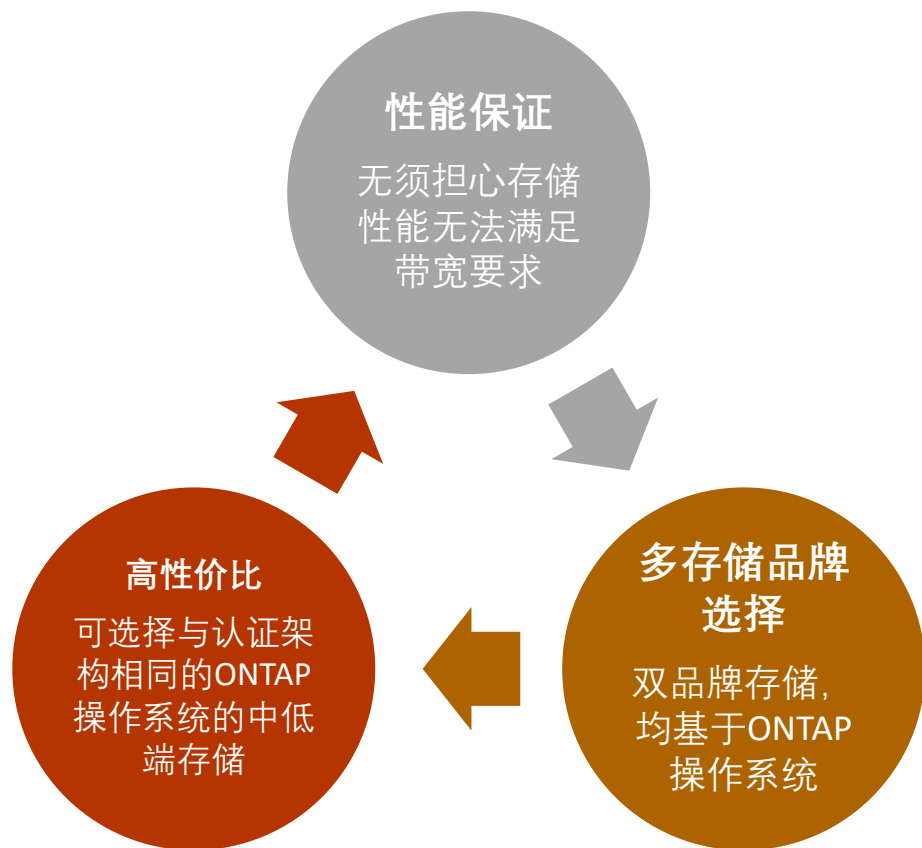
		Throughput **	Raw Capacity (Typical)	Raw Capacity (w/ Expansion)	Connectivity
AFF A800	1 HA pair *	25GB/s	364.8TB	6.6PB	100GbE
	4 HA pairs	100GB/s	1.5PB	26.4PB	
	12 HA pairs	300GB/s	4.4PB	<b>79.2PB</b>	
AFF A700s	1 HA pair	18GB/s	367.2TB	6.6PB	40GbE
	4 HA pairs	72GB/s	1.5PB	26.4PB	
	12 HA pairs	216GB/s	4.4PB	<b>79.2PB</b>	
AFF A300	1 HA pair	9.7GB/s	182.4TB	11.7PB	40GbE
	4 HA pairs	38.8GB/s	729.6TB	46.8PB	
	12 HA pairs	116GB/s	2.2PB	<b>140.4PB</b>	
AFF A220	1 HA pair	5.3GB/s	91.2TB	4.4PB	10GbE
	4 HA pairs	21.2GB/s	364.8TB	17.6PB	
	12 HA pairs	63.6GB/s	1.1PB	<b>52.8PB</b>	



\* 1 AFF = 1 HA pair = 2 Nodes

\*\* Based on ONTAP 9.4, NAS workloads

# 易捷AI解决方案架构



# 数据协议的选择

- 性能
- 延时
- 可扩展性
- 单一命名空间
- 数据移动性
- 易于管理

Question: What is your choice of file system for the AI/DL Data Tier	Answers:
NFS	32.38%
HDFS	20.00%
S3	15.24%
GPFS	10.48%
Ext4	7.62%
Lustre	6.67%
ZFS	4.76%
XFS	2.86%

# DGX-1 外部存储

存储系统网络可选择万兆/IB网络

以下的表格是我们基于深度学习框架的通用IO访问模式针对存储系统的参考推荐，**该推荐只作为参考使用**

应用场景	充足的读缓存	DGX缓存能力	推荐网络类型	网络文件系统选择
数据分析	NA		10Gbe	对象存储，NFS，或其他并发读及小文件性能优的存储
HPC	NA		10/40/100Gbe IB	NFS，或其他支持大量客户端，单点存储性能优的HPC并行存储系统
DL 256x256 图片	Yes	63 million images	10Gbe	NFS或小文件读写效率高的存储
DL 1080p 图片	Yes	13 million images	10/40Gbe IB	高性能NFS，HPC存储系统，高并发
DL 4K 图片	Yes	5 million images	40Gbe IB	高性能NFS，HPC存储系统，高并发，单节点3GB/s+
DL 无压缩图片	Yes	1 million images	IB 40/100Gbe	高性能NFS，HPC存储系统，高并发，单节点3GB/s+
DL 不缓存数据集	no	NA	IB 10/40/100Gbe	性能同上, 总的性能需要满足所有应用并发使用的需求



# DGX-1 网络配置

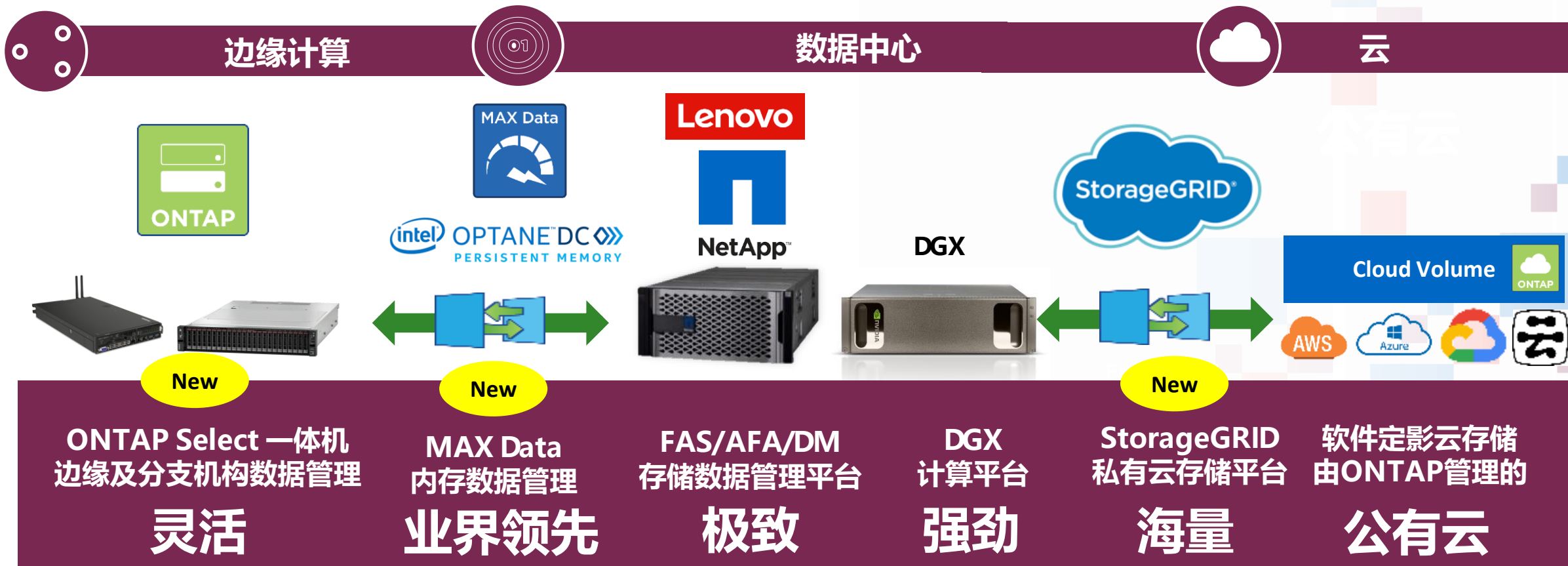
网络类型	网络带宽	网口数量	网络类型	备注
BMC	100/1000 Mb/s	1	远程管理网	RJ45
以太网	100/1000/10000 Mb/s	2	业务/数据网	标配RJ45电口 万兆光线需光转电模块
IB网	100 Gb/s	4	业务/数据网	建议接入2个或以上



OCP Mezz 10GbE 2-Port SFP+ card  
NVIDIA P/N 965-0028-000



# 构建从边缘到核心到云的数据流

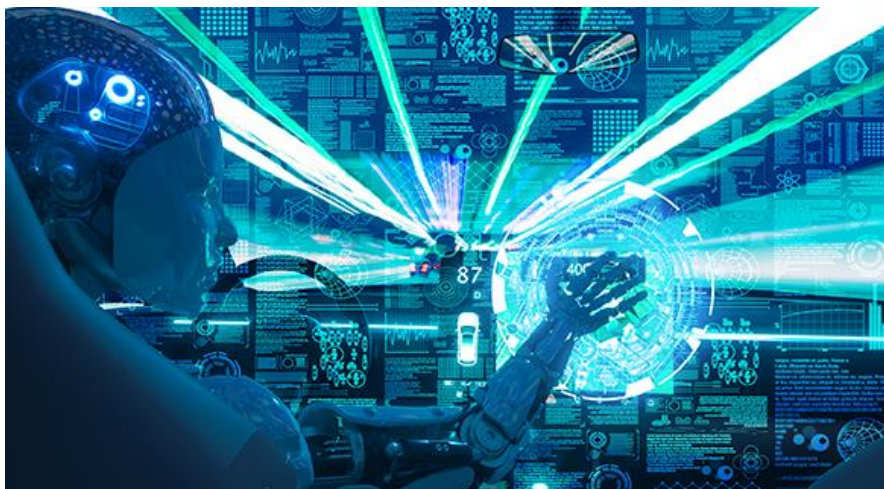


# DGX-1和DGX Station工作站 产品预定火热进行中

人工智能：由 NVIDIA 提供支持

当混合云数据服务领导企业与人工智能计算领域的领导者强强联合，您才是最大的赢家！

立刻订购 NVIDIA® DGX-1™和DGX Station™工作站



人工智能与深度学习解决方案详情，欢迎登陆联想凌拓官网：[LINK](#)

# 联系我们

联想凌拓企业官微



联想凌拓Bilibili企业公众号



官方网站: [www.lenovonetapp.com](http://www.lenovonetapp.com)

官方微信: 联想凌拓

服务热线: 400 828 3001 (呼叫中心)

400 116 0099 (销售热线)

# 谢谢!

智慧数据构建智能世界