AMD | NVIDIA | Lenovo NetApp

联想助力互联网智能转型

- 吴强 英伟达高级合作伙伴经理

NVIDIA

> Founded in 1993

> Jensen Huang, Founder & CEO

> 13,227 employees

> $321B market cap;   **Q3 Rev:** $4.7B YoY: 57%

"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"
— Fast Company

TESLA / DGX 路线图

DGX-1

DGX-Station

DGX-2

DGX-A100
DGX STATION A100

K80
K40

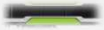
K2
k1

Kepler
Dynamic
Parallelism

2012

M40
M4

M60
M10
M6



Maxwell
DX12

2014

P100 nvlink
P100 PCI-e
P40
P4

Pascal
Unified Memory
3D Memory
NVLink 1.0

2016

V100 PCI-e

V100 NV-link

Volta
Unified Memory
Tensor Core
NVLink 2.0

2017

Tesla T4

Turing
RT Core +
Tensor Core

2018

V100s

RTX6000p
RTX8000p

RT Core +
Tensor Core

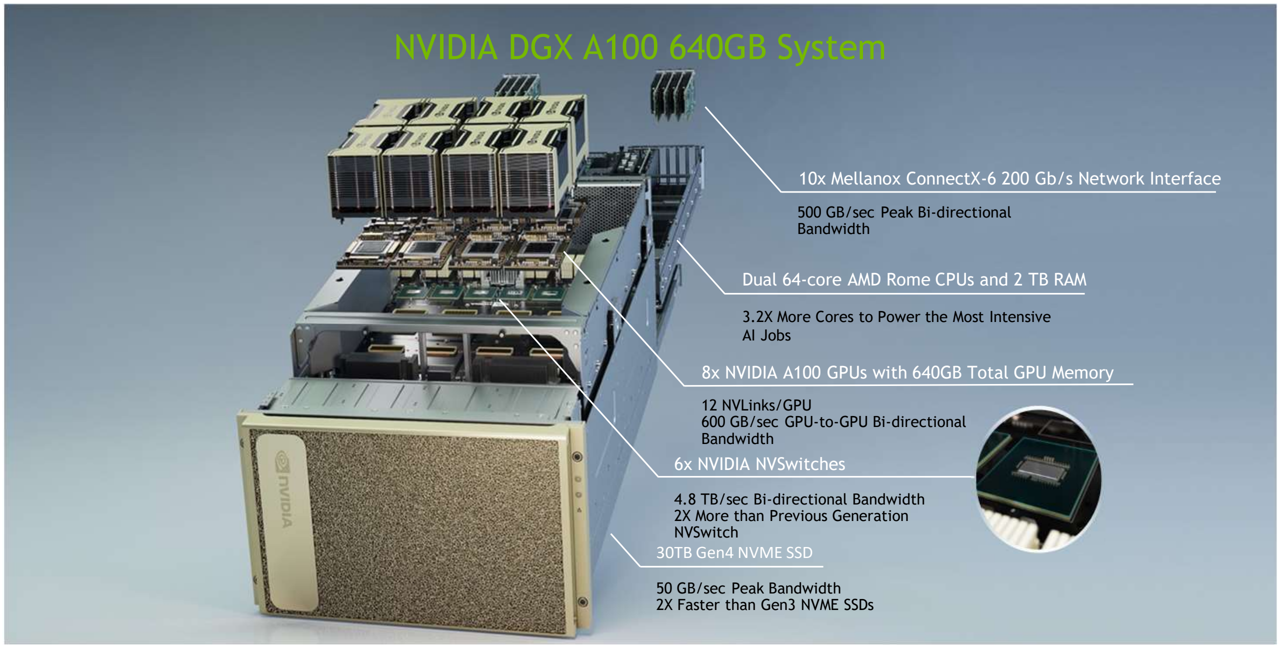
2019

A100

A40 PCI-e
RTX A6000

2020

GAME-CHANGING PERFORMANCE FOR INNOVATORS



NVIDIA DGX A100 640GB System

10x Mellanox ConnectX-6 200 Gb/s Network Interface

500 GB/sec Peak Bi-directional Bandwidth

Dual 64-core AMD Rome CPUs and 2 TB RAM

3.2X More Cores to Power the Most Intensive AI Jobs

8x NVIDIA A100 GPUs with 640GB Total GPU Memory

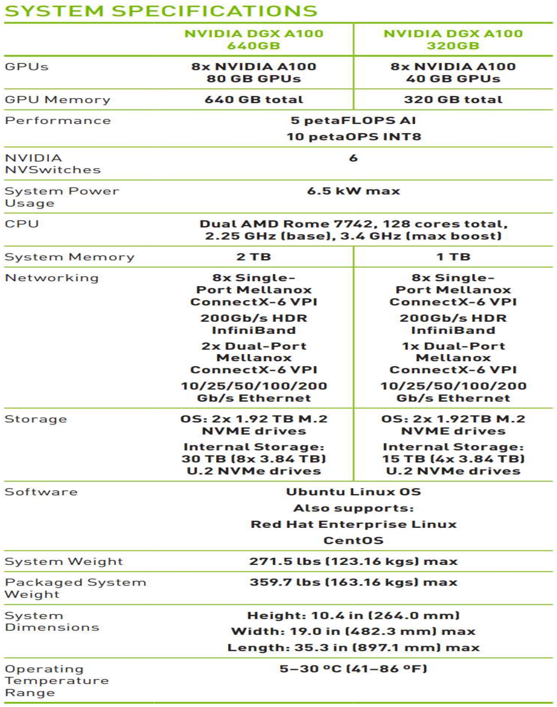
12 NVLinks/GPU
600 GB/sec GPU-to-GPU Bi-directional Bandwidth

6x NVIDIA NVSwitches

4.8 TB/sec Bi-directional Bandwidth
2X More than Previous Generation NVSwitch

30TB Gen4 NVME SSD

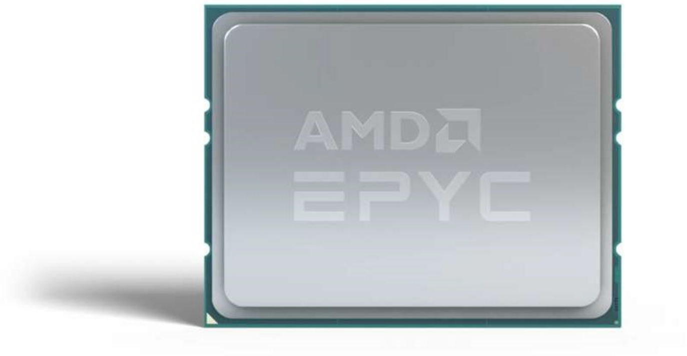
50 GB/sec Peak Bandwidth
2X Faster than Gen3 NVME SSDs



SYSTEM SPECIFICATIONS

	NVIDIA DGX A100 640GB	NVIDIA DGX A100 320GB
GPUs	8x NVIDIA A100 80 GB GPUs	8x NVIDIA A100 40 GB GPUs
GPU Memory	640 GB total	320 GB total
Performance	5 petaFLOPS AI 10 petaOPS INT8	
NVIDIA NVSwitches	6	
System Power Usage	6.5 kW max	
CPU	Dual AMD Rome 7742, 128 cores total, 2.25 GHz (base), 3.4 GHz (max boost)	
System Memory	2 TB	1 TB
Networking	8x Single-Port Mellanox ConnectX-6 VPI 200Gb/s HDR InfiniBand 2x Dual-Port Mellanox ConnectX-6 VPI 10/25/50/100/200 Gb/s Ethernet	8x Single-Port Mellanox ConnectX-6 VPI 200Gb/s HDR InfiniBand 1x Dual-Port Mellanox ConnectX-6 VPI 10/25/50/100/200 Gb/s Ethernet
Storage	OS: 2x 1.92 TB M.2 NVME drives Internal Storage: 30 TB (8x 3.84 TB) U.2 NVMe drives	OS: 2x 1.92TB M.2 NVME drives Internal Storage: 15 TB (4x 3.84 TB) U.2 NVMe drives
Software	Ubuntu Linux OS Also supports: Red Hat Enterprise Linux CentOS	
System Weight	271.5 lbs (123.16 kgs) max	
Packaged System Weight	359.7 lbs (163.16 kgs) max	
System Dimensions	Height: 10.4 in (264.0 mm) Width: 19.0 in (482.3 mm) max Length: 35.3 in (897.1 mm) max	
Operating Temperature Range	5–30 ºC (41–86 ºF)	

双路CPU部分：AMD EPYC™ 7742

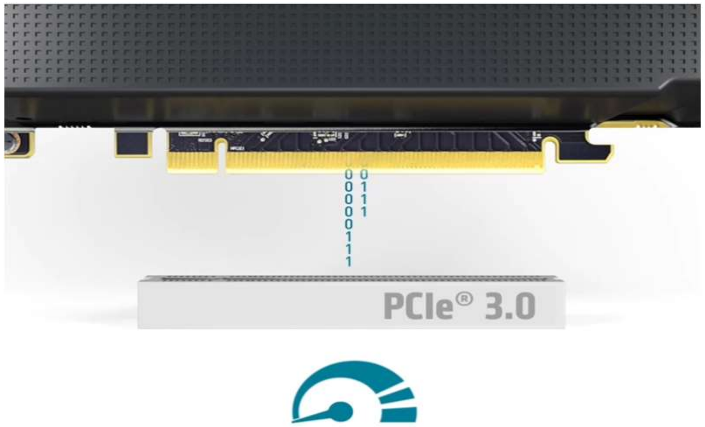


CPU 核心数量: 64
线程数量:128
基准时钟频率:2.25GHz

最大加速时钟频率:最高可达 3.4GHz
三级缓存:256MB
封装:SP3

支持的CPU插槽数:1P/2P
PCI Express 版本 :PCIe 4.0 x128
默认 TDP/TDP :225W

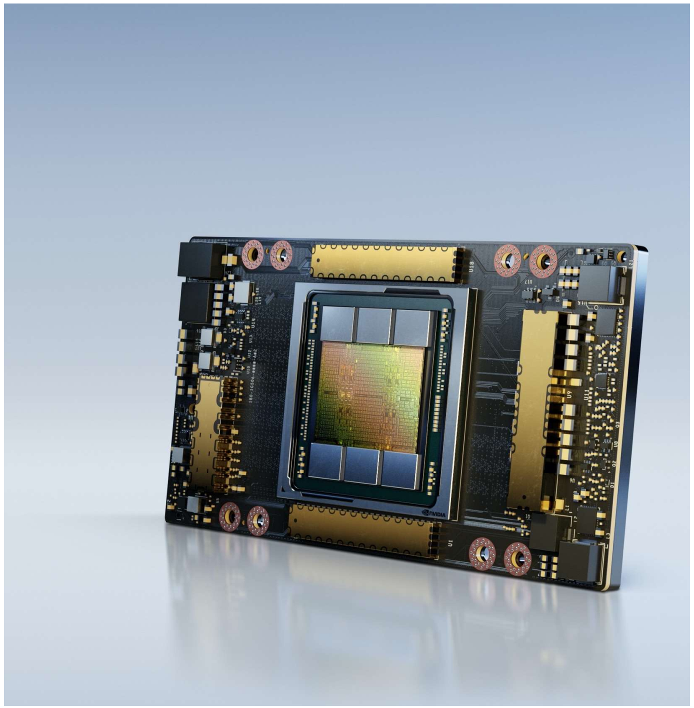
最高内存速度：Up to 3200MHz
内存类型：DDR4
内存通道：8
内存带宽（每路）：204.8 GB/s



NVIDIA A100 GPU

Greatest Generational Leap – 20X Volta

	Peak	Vs	Volta
FP32 TRAINING	312	TFLOPS TF32	20X
INT8 INFERENCE	1,248	TOPS	20X
FP64 HPC	19.5	TFLOPS	2.5X
MULTI INSTANCE GPU			7X GPUs



SUPERCHARGED AI SUPERCOMPUTING WITH A100 80GB

World’s Fastest GPU with World’s Fastest Memory

A100 80GB Throughput vs A100 40GB

2X	2X	3X
Simulation Quantum Espresso	**Big Data Analytics** 10 TB Retail Benchmark	**AI Training** DLRM Recommender

1.25X	1.25X
MIG Inference RNN-T Speech Recognition	**Energy Efficiency** Shatters 25 GF/W

适用于办公室的服务器级解决方案

数据中心外的数据中心技术

- AI 时代的工作组服务器
- 没有数据中心但拥有数据中心性能
- 可以放置在任何地方的 AI 设备
- 运行更大的模型，更快地获得答案

2.5 PFLOPS AI

320 GB GPU MEMORY

唯一具有 4 路 NVLink 和多实例 GPU（MIG）的工作组服务器

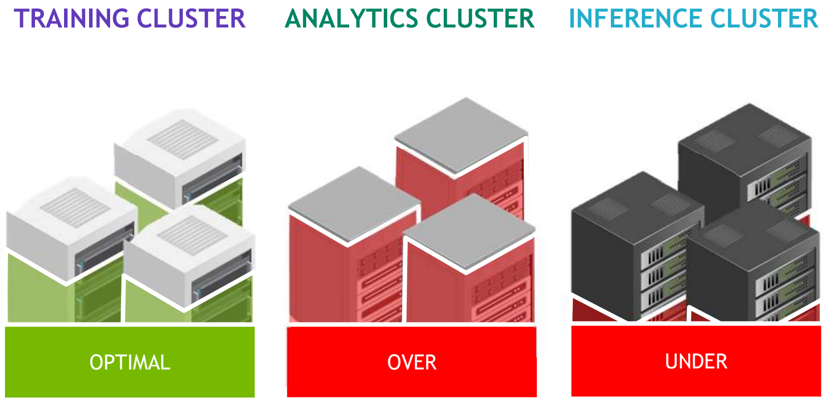
	DGX Station A100 320GB	DGX Station A100 160GB
GPUs	4x NVIDIA A100 Tensor Core GPUs	
GPU Memory (total)	320GB	160GB
Performance	2.5 petaFLOPS AI; 5 petaOPS INT8	
System Memory	512GB DDR4 RDIMM, 3200MT/s	
Storage	OS: 1 x 1.92TB M.2 NVME Data:1 x 7.68TB U.2 NVME	
CPU	AMD® Epyc® CPU 7742, 2.25GHz to 3.4GHz, 64 cores/128 threads, PCIe Gen4	
Networking	Dual 10GBASE-T (RJ45)	
Display GPU	4GB, 4x Mini DisplayPort	
Acoustics	<37dB	
Cooling	Custom refrigerant cooling system for GPUs and CPU	
System Power (max)	1,5kW	
Management	AST2500, IPMI, Redfish	
System Dimensions	518 D x 256 W x 639 H (mm)	
Operating Temp.	5°C to 35°C (41°F to 95°F)	

ELASTIC AI INFRASTRUCTURE WITH DGX A100

DGX A100 with MIG Delivers New Agility for Today’s Enterprise Data Center

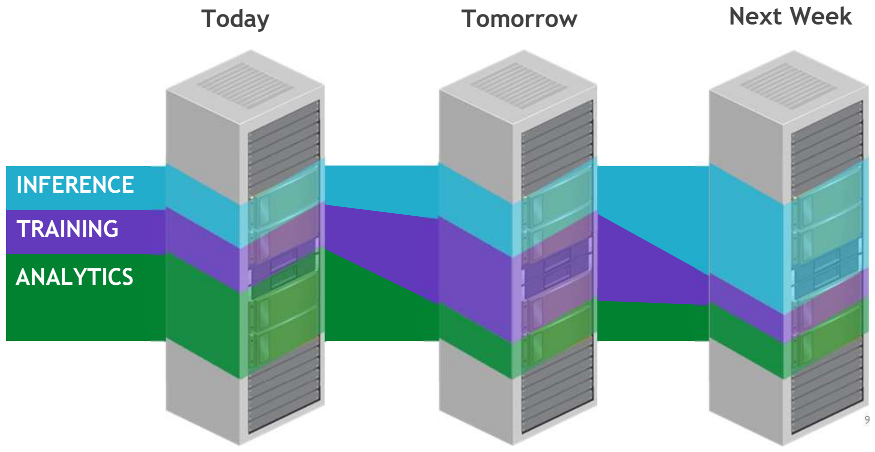
Traditional Infrastructure is Constrained

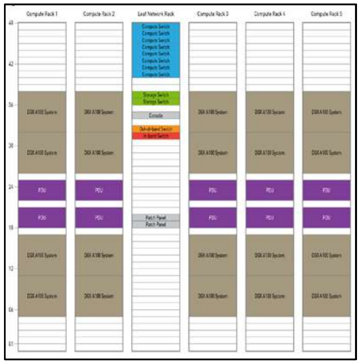
Infrastructure silos starve AI workloads or waste capacity

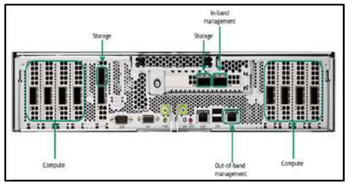


DGX A100 Infrastructure is Agile

DGX A100 infrastructure uses MIG to allocate GPU resources to workloads





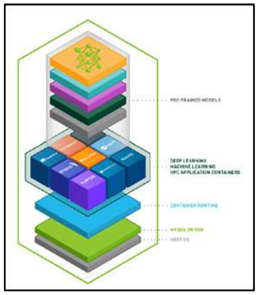


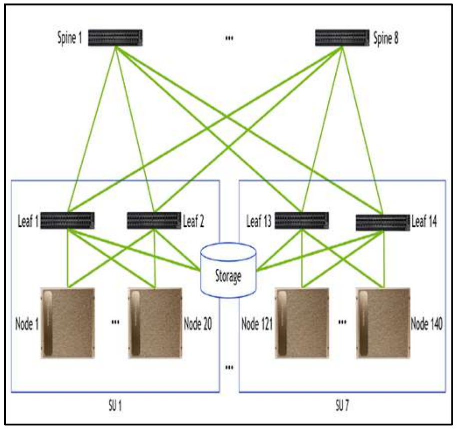
INTELLIGENTLY ADAPTED AND INTEGRATED

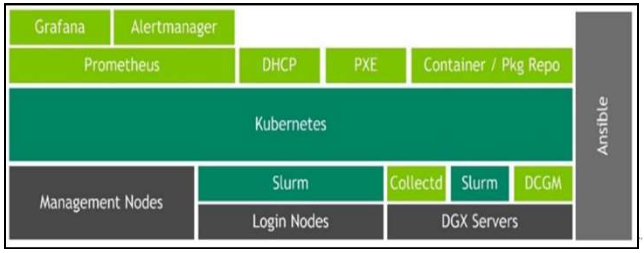
Flexible deployments tailored to the unique needs of your environment

DGX SuperPOD Solution for Enterprise:

- Your partner to help your IT team navigate:
  - AI workflow tools to speed time-to-insight
  - Customized data center design, optimized for you
  - Flexible deployment options tailored to your environment
- NVIDIA professional services in combination with select NVIDIA SuperPOD partners will speed your deployment experience







DGX SUPERPOD DEPLOYMENTS AT NVIDIA

#1 on MLPerf for commercially available systems

#5 on TOP500 (63 PetaFLOPS HPL)

#1 on Green500 (26.2 GigaFLOPS/watt)

Fastest Industrial System in U.S.

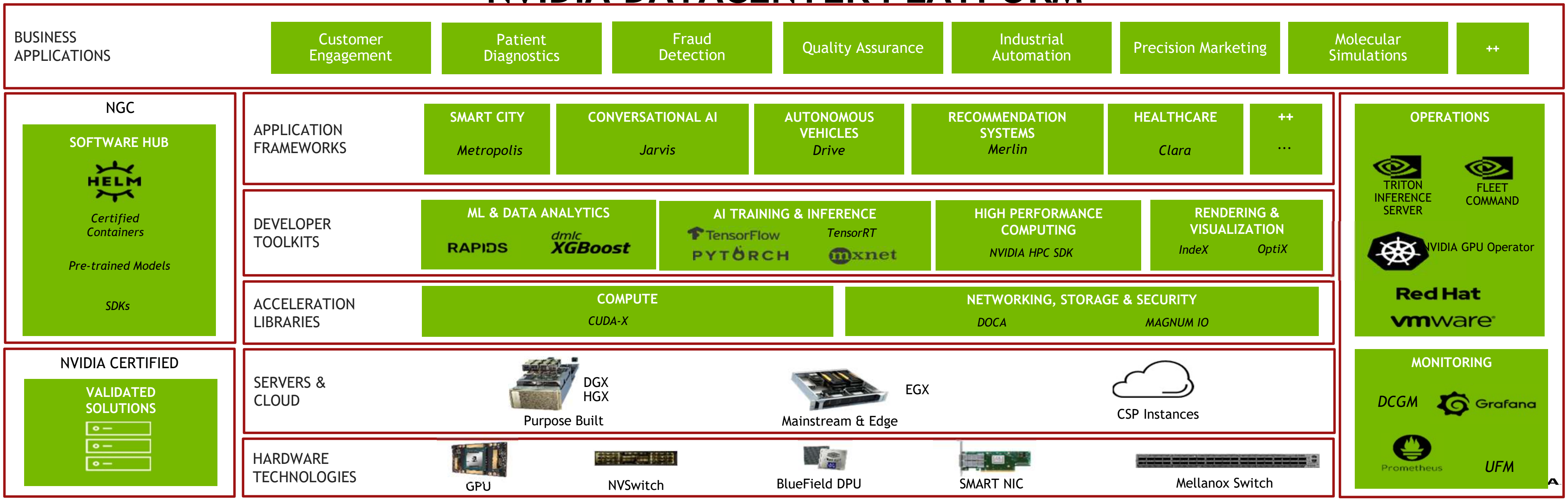
Both are built with the NVIDIA DGX SuperPOD arch:

- NVIDIA DGX A100 and NVIDIA Mellanox IB
- NVIDIA’s decade of AI experience

Selene Configuration:

- 4,480 NVIDIA A100 Tensor Core GPUs
- 560 NVIDIA DGX A100 640GB systems
- 850 Mellanox 200G HDR IB switches
- 14 PB of all-flash storage
- 2.8 ExaFLOPS of AI performance
- Built in 3 weeks

NVIDIA DATACENTER PLATFORM



BUSINESS APPLICATIONS

- Customer Engagement
- Patient Diagnostics
- Fraud Detection
- Quality Assurance
- Industrial Automation
- Precision Marketing
- Molecular Simulations
- ++

NGC

SOFTWARE HUB

HELM

Certified Containers

Pre-trained Models

SDKs

NVIDIA CERTIFIED

VALIDATED SOLUTIONS

APPLICATION FRAMEWORKS

- SMART CITY *Metropolis*
- CONVERSATIONAL AI *Jarvis*
- AUTONOMOUS VEHICLES *Drive*
- RECOMMENDATION SYSTEMS *Merlin*
- HEALTHCARE *Clara*
- ++ …

DEVELOPER TOOLKITS

- ML & DATA ANALYTICS: RAPIDS, dmlc XGBoost
- AI TRAINING & INFERENCE: TensorFlow, *TensorRT*, PYTORCH, mxnet
- HIGH PERFORMANCE COMPUTING: *NVIDIA HPC SDK*
- RENDERING & VISUALIZATION: *IndeX*, *OptiX*

ACCELERATION LIBRARIES

- COMPUTE: *CUDA-X*
- NETWORKING, STORAGE & SECURITY: *DOCA*, *MAGNUM IO*

SERVERS & CLOUD

- DGX HGX: Purpose Built
- EGX: Mainstream & Edge
- CSP Instances

HARDWARE TECHNOLOGIES

- GPU
- NVSwitch
- BlueField DPU
- SMART NIC
- Mellanox Switch

OPERATIONS

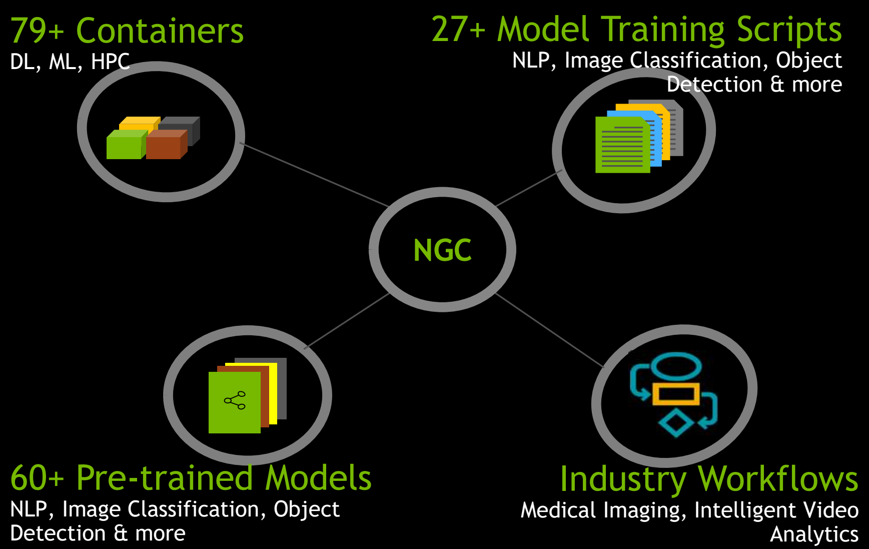
- TRITON INFERENCE SERVER
- FLEET COMMAND
- NVIDIA GPU Operator
- Red Hat
- vmware

MONITORING

- *DCGM*
- Grafana
- Prometheus
- *UFM*

ACCELERATION SOFTWARE ON NGC

http://ngc.nvidia.com



NGC

79+ Containers

DL, ML, HPC

27+ Model Training Scripts

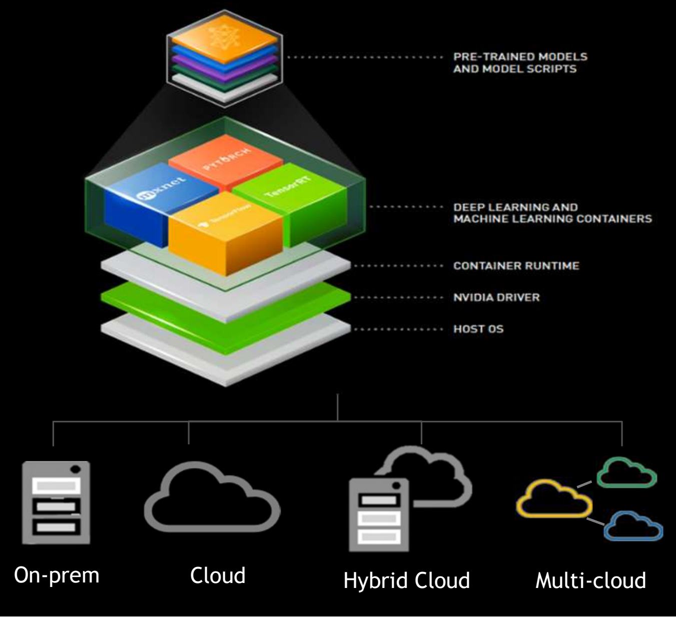
NLP, Image Classification, Object Detection & more

60+ Pre-trained Models

NLP, Image Classification, Object Detection & more

Industry Workflows

Medical Imaging, Intelligent Video Analytics


PRE-TRAINED MODELS AND MODEL SCRIPTS
DEEP LEARNING AND MACHINE LEARNING CONTAINERS
CONTAINER RUNTIME
NVIDIA DRIVER
HOST OS
On-prem
Cloud
Hybrid Cloud
Multi-cloud

```
  Hi Jennifer, how can I help you
Jennifer : How much
```

RECOMMENDERS — THE ENGINE OF THE INTERNET

Billions of Users - Trillions of Items

100’s of millions of etail items -> Amazon & Alibaba recommenders

1000’s of movies -> Netflix recommender

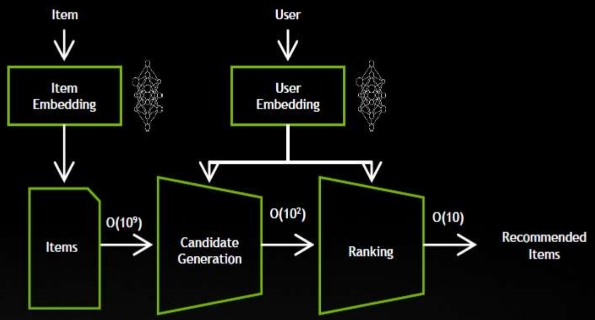
10’s of millions of songs -> Spotify & iTunes recommenders

10’s of millions of books -> Amazon recommender

Billions of Tik Tok & YT videos -> TT & YT recommender

Billions of websites -> Google search rank

So much news!!! -> Google & FB news recommenders



Item

User

Item Embedding

User Embedding

Items

$O(10^9)$

Candidate Generation

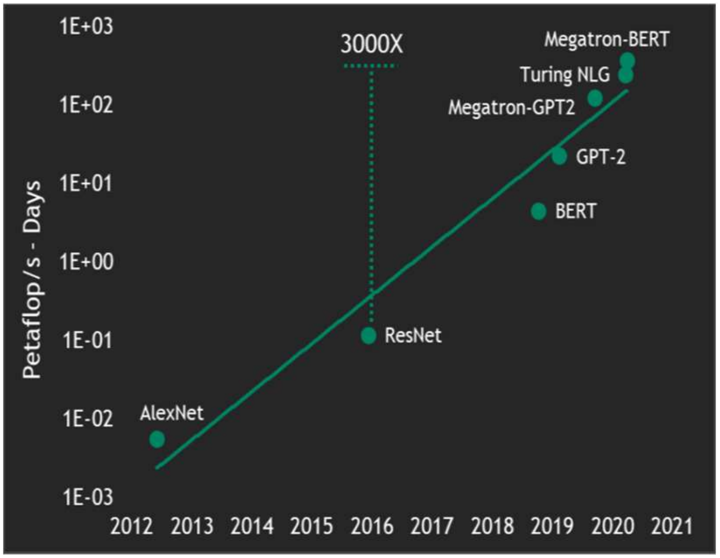
$O(10^2)$

Ranking

$O(10)$

Recommended Items

KEY CHALLENGES IN CONSUMER INTERNET



EXPLOSION OF AI MODELS & COMPLEXITY

- Number of models ranging from CNN, RNN, Transformer and new species like Wide & Deep
- Network complexity grew 10,000x in 7 years (peta flops)

INCREASING DATA ONSLAUGHT

- Datasets continuing to increasing dramatically
- Multiple sources, different formats, varying quality

CUMBERSOME DEPLOY & MANAGE WORKFLOWS

- Scaling infrastructure monitoring & management is a challenge
- Software deployment, versioning and updates are time consuming

GPU在互联网行业的主要应用领域

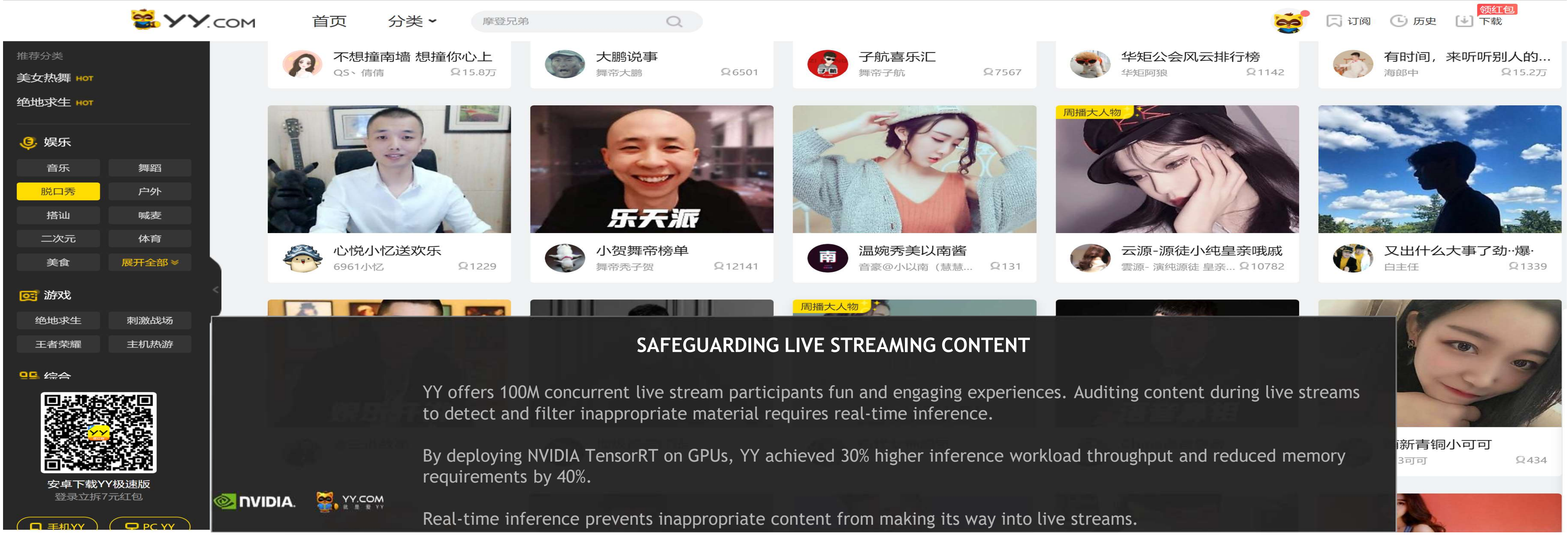
Content Understanding	Conversational AI	Recommender and Search	Video Processing	Machine/Drive	Data Analysis
• 内容审核 • 内容标签/分类检测 • 语音输入 • 语音识别 • 文本翻译 • 人脸/物体检测/识别 • OCR	• 智能终端 • 智能客服 • 在线会议/社交 • 语音助手 • 各类语音合成 • 智能问答 • 虚拟主播	• 推荐/搜索/广告 CTR • 召回、排序 • 内容标签挖掘和向量化	• 编解码 • 渲染特效 • 质量增强 • 视频编辑 • 生成特效 • AR/VR • Cloud Gaming	• 自动驾驶 • 智能配送 • 物流机器人 • 智能货架 • 仿真模拟	• Spark • 数据预处理 • 特征提取 • 传统机器学习 • 图计算 • 数据可视化
CV, ASR, NLP	**ASR，TTS，NLP**	**CTR、NLP, CV**	**Codec、CG、GAN、CV**	**Jetson、CV、speech、CG**	**Bigdata/Spark、xgBoost/ML**

主要类别互联网公司业务分布

	Content understanding	Conversational AI	Recommender and search	Video Processing	Machine/Drive	Data Analysis
云服务2B	****	***		**		**
电商	****	**	*****	*	***	**
社交/效率办公	***	***	**	**		*
娱乐/视频	*****	*	*****	****		*
教育/音频	***	**	*	***		*
生活服务	****	**	**	*	***	**
资讯/搜索	****	**	*****	**		**
金融	**	*	*			***
手机	****	**	**	**		*

REAL-TIME SPEECH SERVICES AT SCALE

WeChat, a leading Chinese social media platform with ~1B users, wanted to improve its speech to context services. But as the company deployed its new acoustic model, its CPU-only servers were unable to effectively run the new version. WeChat deployed servers equipped with Tesla P4 GPU inference accelerators and increased speech inference throughput by 2.5X and in-model accuracy by 20% — all while staying within its low latency budget.

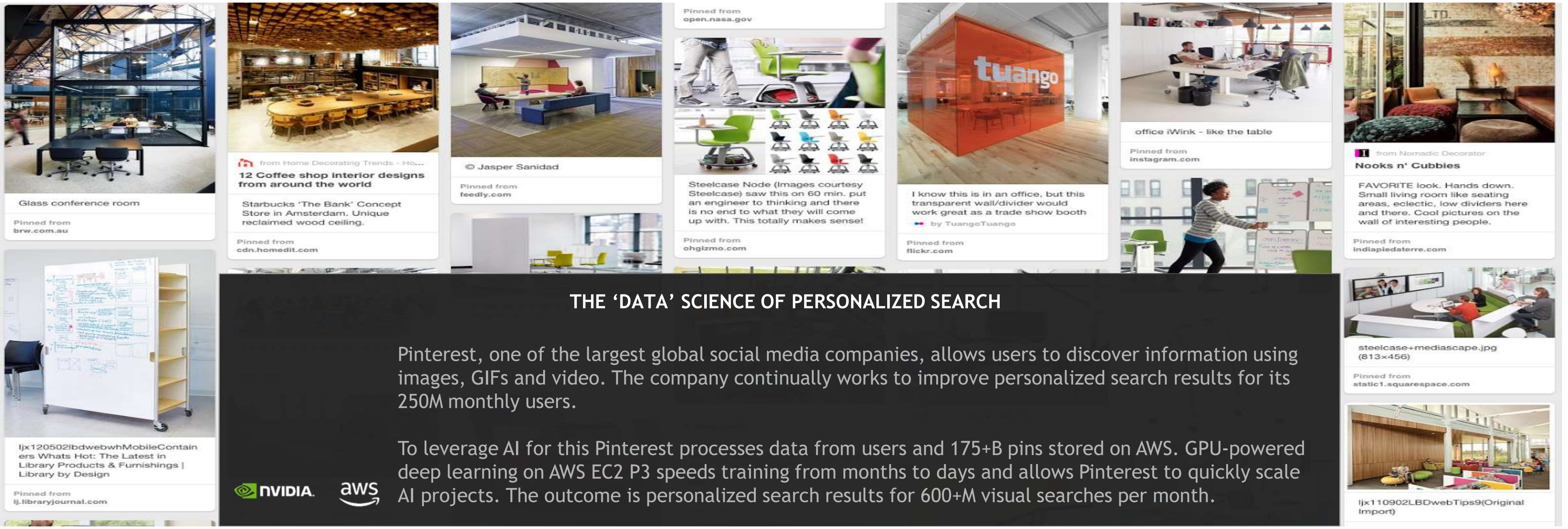


SAFEGUARDING LIVE STREAMING CONTENT

YY offers 100M concurrent live stream participants fun and engaging experiences. Auditing content during live streams to detect and filter inappropriate material requires real-time inference.

By deploying NVIDIA TensorRT on GPUs, YY achieved 30% higher inference workload throughput and reduced memory requirements by 40%.

Real-time inference prevents inappropriate content from making its way into live streams.



Glass conference room

Pinned from
brw.com.au

ljx120502lbdwebwhMobileContainers Whats Hot: The Latest in Library Products & Furnishings | Library by Design

Pinned from
lj.libraryjournal.com

from Home Decorating Trends - Ho...

12 Coffee shop interior designs from around the world

Starbucks 'The Bank' Concept Store in Amsterdam. Unique reclaimed wood ceiling.

Pinned from
cdn.homedit.com

© Jasper Sanidad

Pinned from
feedly.com

Pinned from
open.nasa.gov

Steelcase Node (Images courtesy Steelcase) saw this on 60 min. put an engineer to thinking and there is no end to what they will come up with. This totally makes sense!

Pinned from
ohgizmo.com

I know this is in an office, but this transparent wall/divider would work great as a trade show booth

by TuangoTuango

Pinned from
flickr.com

office iWink - like the table

Pinned from
instagram.com

from Nomadic Decorator

Nooks n' Cubbies

FAVORITE look. Hands down. Small living room like seating areas, eclectic, low dividers here and there. Cool pictures on the table wall of interesting people.

Pinned from
indiapiedaterre.com

steelcase+mediascape.jpg (813×456)

Pinned from
static1.squarespace.com

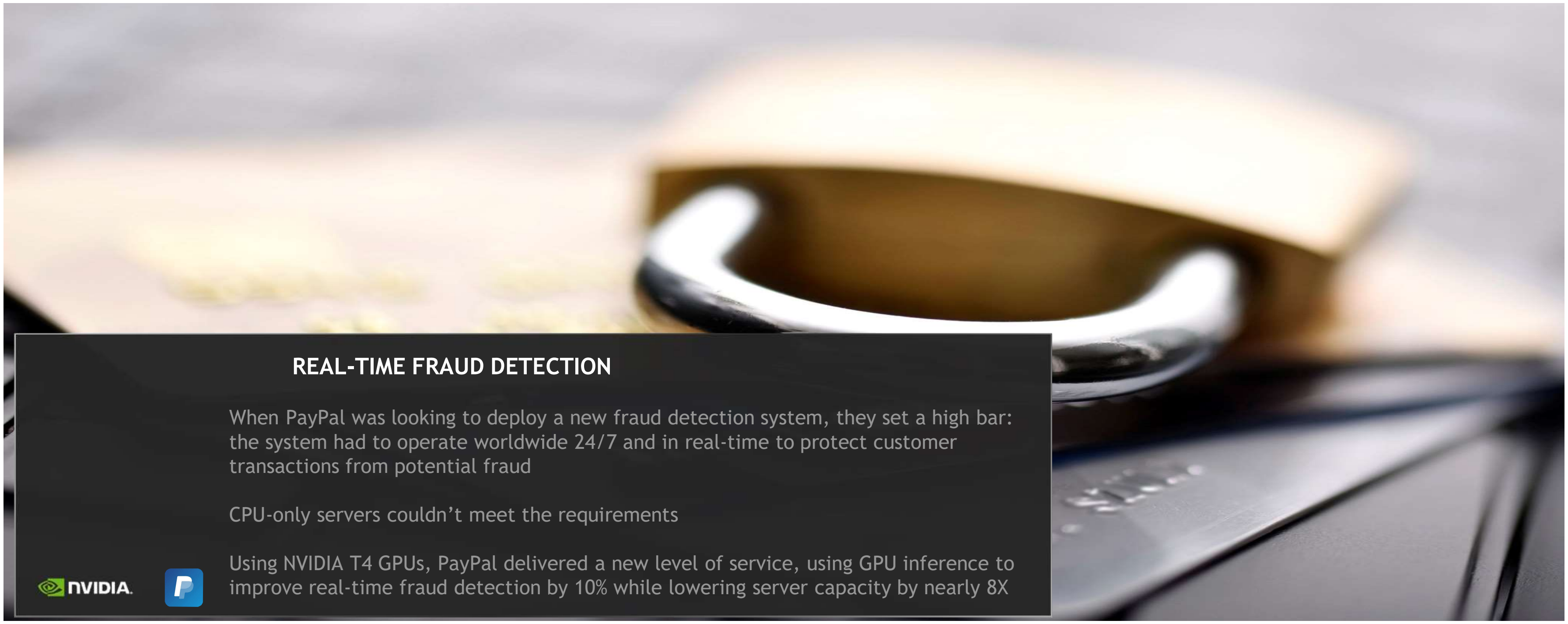
ljx110902LBDwebTips9(Original Import)

THE ‘DATA’ SCIENCE OF PERSONALIZED SEARCH

Pinterest, one of the largest global social media companies, allows users to discover information using images, GIFs and video. The company continually works to improve personalized search results for its 250M monthly users.

To leverage AI for this Pinterest processes data from users and 175+B pins stored on AWS. GPU-powered deep learning on AWS EC2 P3 speeds training from months to days and allows Pinterest to quickly scale AI projects. The outcome is personalized search results for 600+M visual searches per month.

NVIDIA aws



REAL-TIME FRAUD DETECTION

When PayPal was looking to deploy a new fraud detection system, they set a high bar: the system had to operate worldwide 24/7 and in real-time to protect customer transactions from potential fraud

CPU-only servers couldn’t meet the requirements

Using NVIDIA T4 GPUs, PayPal delivered a new level of service, using GPU inference to improve real-time fraud detection by 10% while lowering server capacity by nearly 8X

云游戏行业应用场景

剑灵
流放之路
NBA 2K ONLINE 2
FIFA ONLINE 4
QATAR AIRWAYS
FORTNITE
堡垒之夜
腾讯游戏
Tencent Games
START

AMD | NVIDIA | Lenovo NetApp

联想助力互联网智能转型

thanks.

Stephen(吴强)@NVIDIA

北京 朝阳

扫一扫上面的二维码图案，加我微信