



# MODERN ENTERPRISE COMPUTING

Liwei Zhao | Solutions Architecture & Engineering | May 2019

# EVOLUTION OF COMPUTING



**PC Internet**  
WinTel, Yahoo!  
1 billion PC users

1995



**Mobile-Cloud**  
iPhone, Amazon AWS  
2.5 billion mobile users

2005

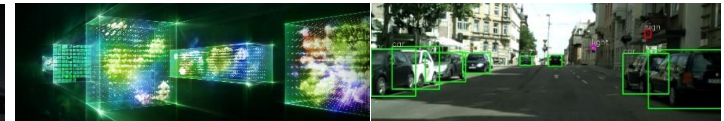
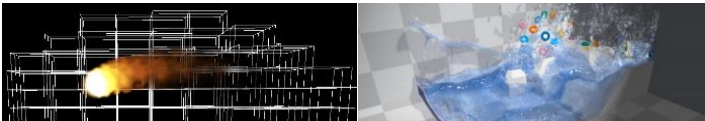
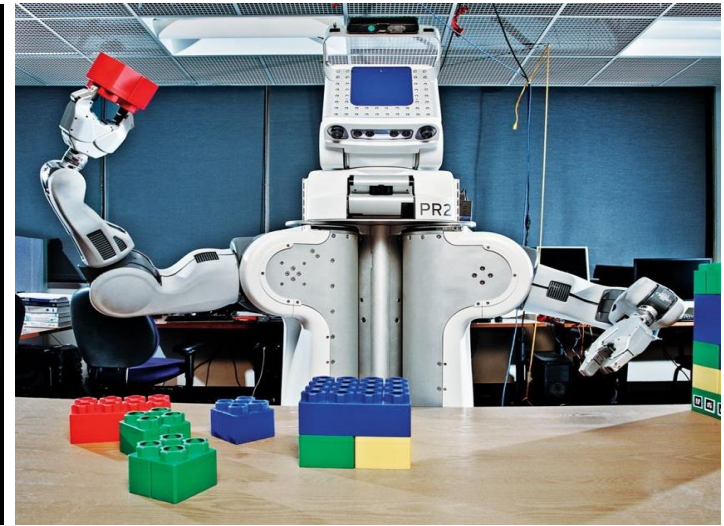


**AI & IOT**  
Deep Learning, GPU  
100s of billions of devices

2015

# NVIDIA

## “THE AI COMPUTING COMPANY”



Computer Graphics

GPU Computing

Artificial Intelligence

# BEYOND MOORE'S LAW

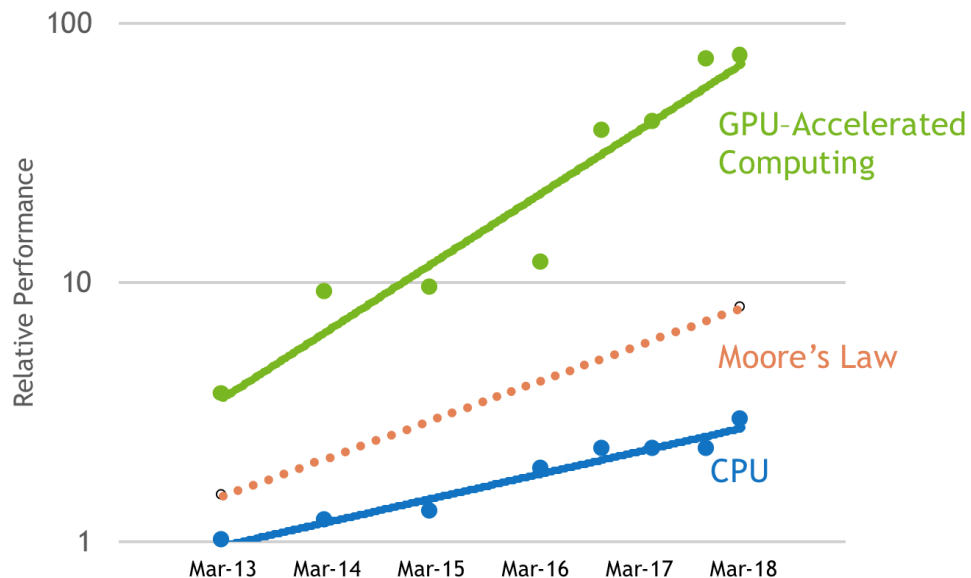
## Progress Of Stack In 5 Years

2013

cuBLAS: 5.0
cuFFT: 5.0
cuRAND: 5.0
cuSPARSE: 5.0
NPP: 5.0
Thrust: 1.5.3
CUDA: 5.0
Resource Mgr: r304
Base OS: CentOS 6.2



Accelerated Server  
With Fermi



Measured performance of Amber, CHROMA, GTC, LAMMPS, MILC, NAMD, Quantum Espresso, SPECFEM3D

2018

cuBLAS: 10.0
cuFFT: 10.0
cuRAND: 10.0
cuSOLVER: 10.0
cuSPARSE: 10.0
NPP: 10.0
Thrust: 1.9.0
CUDA: 10.0
Resource Mgr: r384
Base OS: Ubuntu 16.04



Accelerated Server  
with Volta

# TESLA UNIVERSAL ACCELERATION PLATFORM

Single Platform Drives Utilization and Productivity

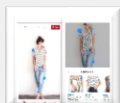
## CUSTOMER USECASES



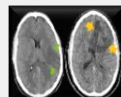
Speech



Translate



Recommender



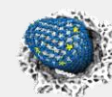
Healthcare



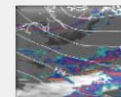
Manufacturing



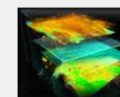
Finance



Molecular Simulations



Weather Forecasting



Seismic Mapping

CONSUMER INTERNET

INDUSTRIAL APPLICATIONS

SCIENTIFIC APPLICATIONS

## APPS & FRAMEWORKS



## NVIDIA SDK & LIBRARIES

MACHINE LEARNING/ ANALYTICS

cuDF

cuML

cuGRAPH

DEEP LEARNING

cuDNN

cuBLAS

CUTLASS

NCCL

TensorRT

HPC

CuBLAS

CuFFT

OpenACC

CUDA

## TESLA GPUs & SYSTEMS



TESLA GPU



VIRTUAL GPU



NVIDIA DGX FAMILY



NVIDIA HGX



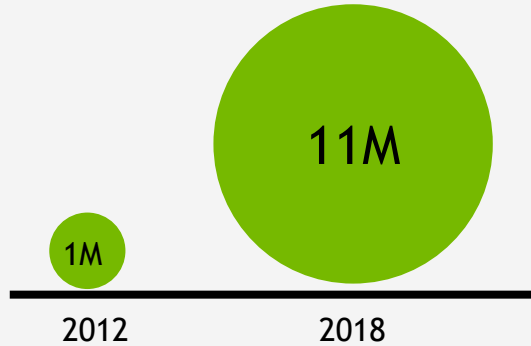
SYSTEM OEM



CLOUD

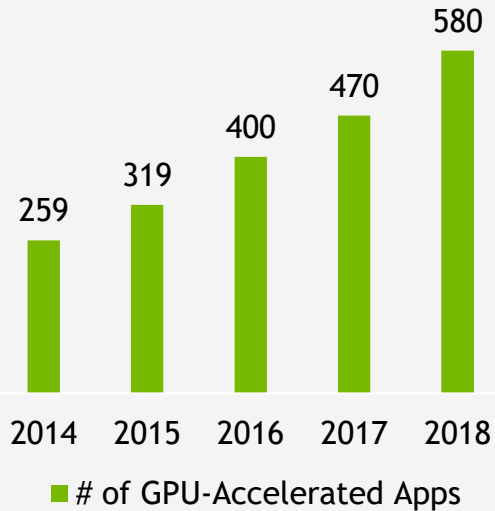
# MOST ADOPTED PLATFORM FOR ACCELERATING HPC

11M CUDA Downloads



11X CUDA DOWNLOADS

580 Applications Accelerated



ALL TOP 15 APPLICATIONS  
ACCELERATED

127 Systems on Top 500




World's #1 Summit: 144 PF  
World's #2 Sierra: 95 PF  
Europe's #1 Piz Daint: 21 PF  
Japan's #1 ABCI: 20 PF  
Industrial #1 ENI: 12 PF

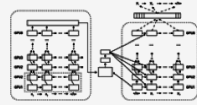
NEW HIGHS IN TOP 500 LIST

# MOST ADOPTED PLATFORM FOR ACCELERATING AI

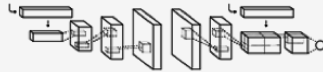
**Convolutional Networks**



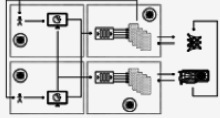
**Recurrent Networks**



**Generative Adversarial Networks**



**Reinforcement Learning**



**BROADEST ARRAY OF NETWORKS**

**Chainer**



**mxnet** **ONNX**



**PaddlePaddle** **PYTORCH**






**TensorFlow** **MATLAB**






**EVERY DEEP LEARNING  
FRAMEWORK ACCELERATED**

**Alibaba Cloud** **aws** **Google Cloud**


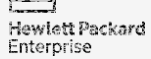



**IBM Cloud** **Microsoft Azure** **Tencent Cloud**






**Cloud Services**


**DELL** **Hewlett Packard Enterprise** **IBM**



**inspur** **Lenovo** **SUPERMICR**



**Systems**



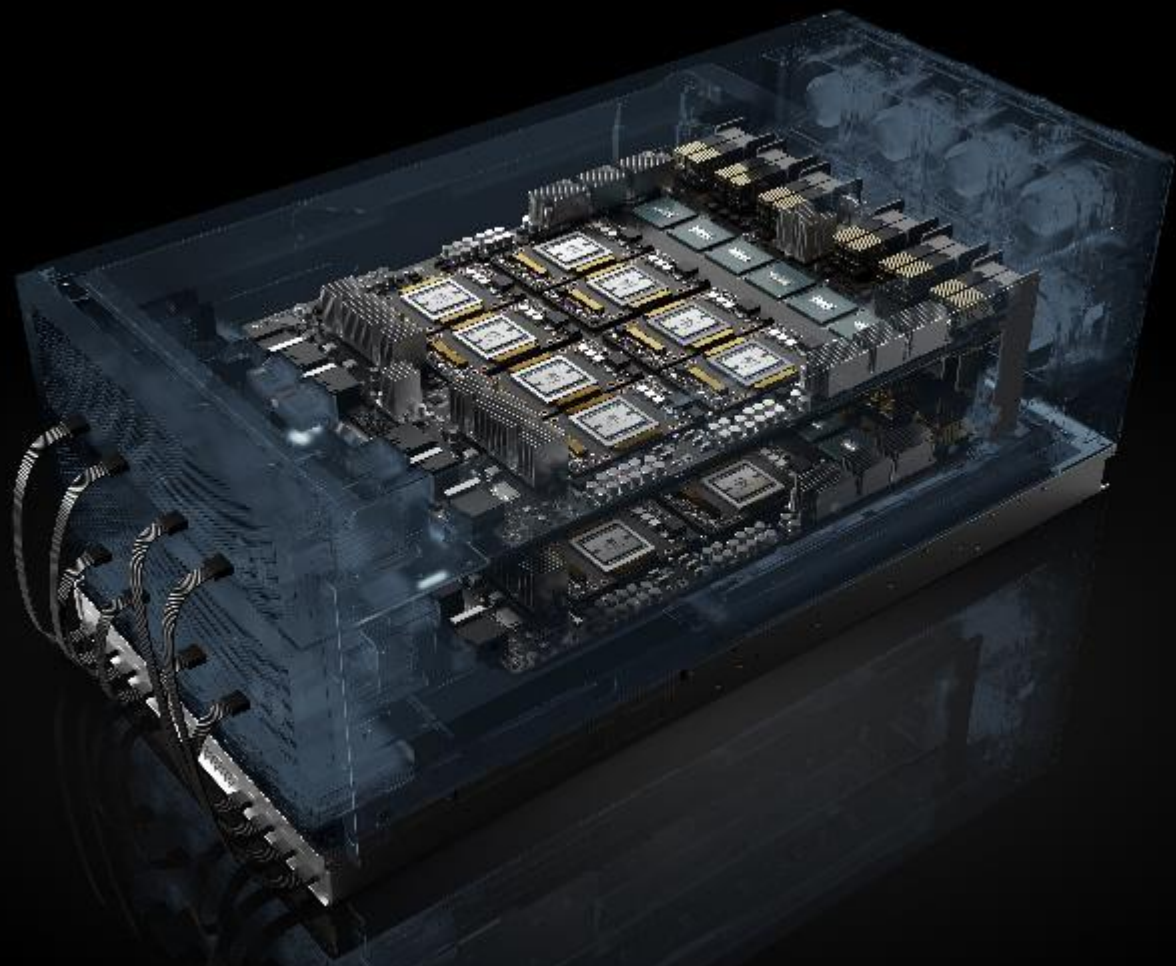
**Desktops**

**AVAILABLE EVERYWHERE**

# TESLA HGX-2

Fusing HPC and AI into  
One Unified Computing  
Architecture

Multi-precision Computing  
2 PFLOPS AI | 250 TFLOPS FP32  
| 125 TFLOPS FP64  
16 Tesla V100 GPUs |  
0.5TB Memory | 2.4 TB/s |  
16TB/s Memory Bandwidth





# TESLA T4

WORLD'S MOST ADVANCED SCALE-OUT GPU

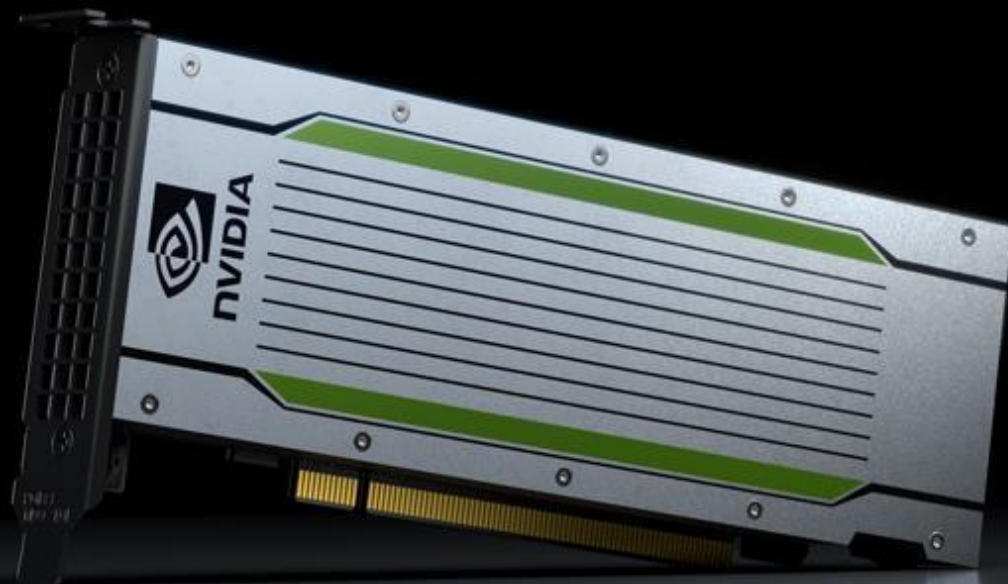
320 Turing Tensor Cores

2,560 CUDA Cores

65 FP16 TFLOPS | 130 INT8 TOPS | 260 INT4 TOPS

16GB | 320GB/s

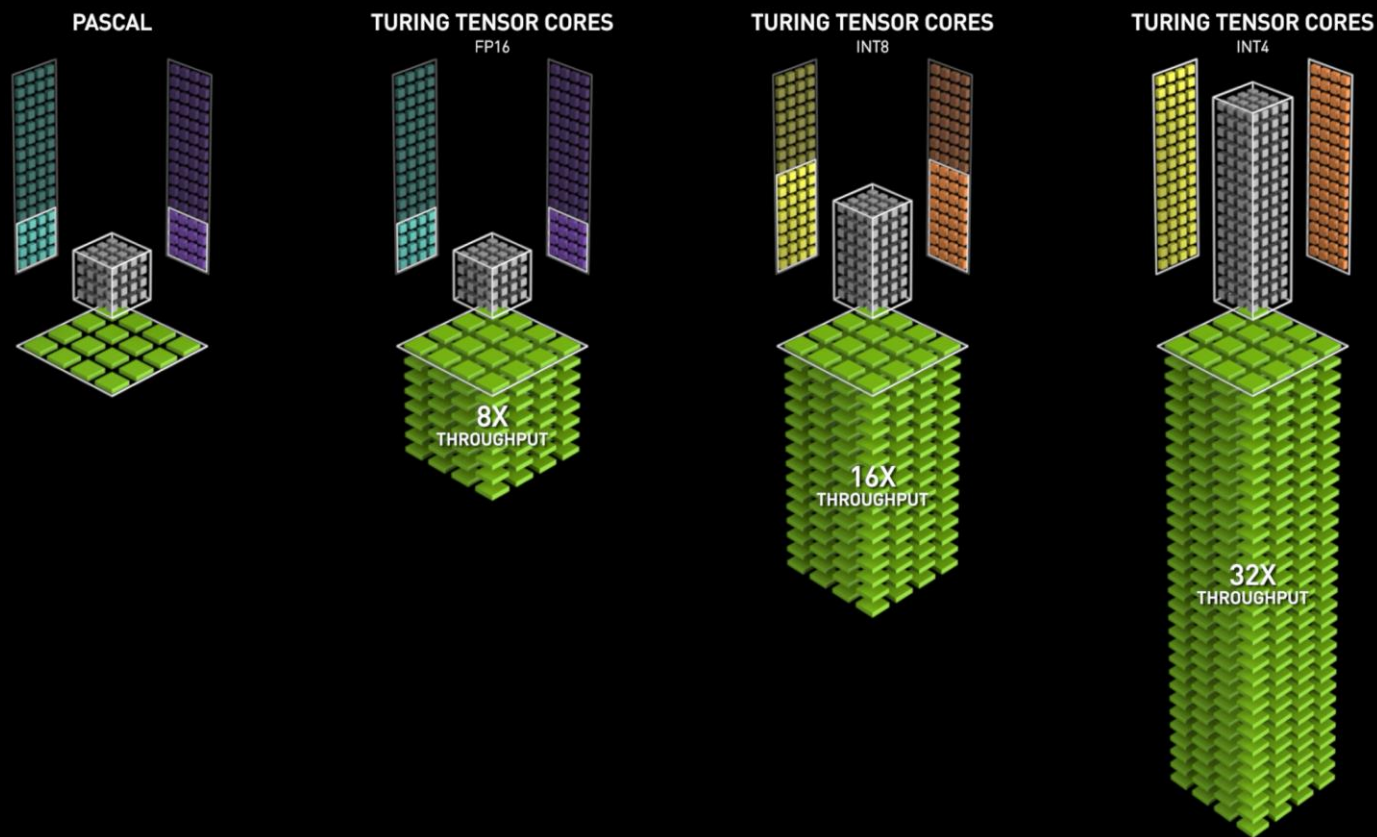
70 W



# NEW TURING TENSOR CORE

MULTI-PRECISION FOR AI INFERENCE & ENTRY LEVEL TRAINING

65 TFLOPS FP16 | 130 TeraOPS INT8 | 260 TeraOPS INT4



# ACCELERATING MACHINE LEARNING

## The RAPIDS Ecosystem

### Open Source Community



### Enterprise Data Science Platforms



### Startups



### Deep Learning Integration



# RAPIDS

### GPU Servers



### Storage Partners



# JETSON POWERING AUTONOMOUS MACHINES

## WAREHOUSE



## DELIVERY



## AGRICULTURE



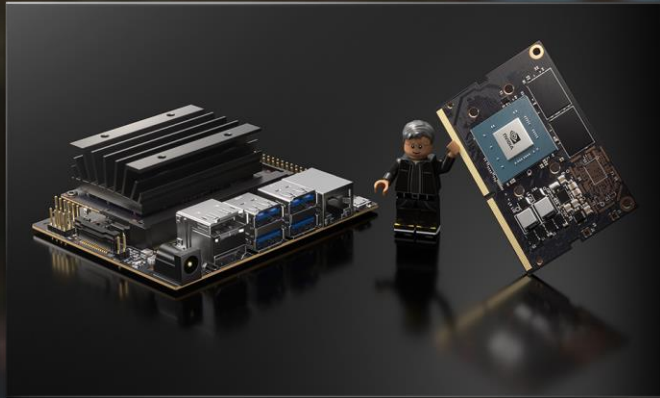
## RETAIL



## INDUSTRIAL



# JETSON NANO | ISAAC | CONSTELLATION | TOYOTA



# NVIDIA AND MATHWORKS COLLABORATION

Working Together to Accelerate the Pace of Engineering and Science

- Integrate the power of NVIDIA systems with MATLAB and Simulink, a leading platform for technical computing and system development
- Accelerate performance across the enterprise, including embedded devices, desktops and laptops, and HPC/Cloud
- Applications include deep learning, embedded vision, and autonomous systems, as well as general-purpose technical computing

# NVIDIA AND MATHWORKS COLLABORATION

Integrate the TESLA Platform with MATLAB & Simulink Across the Enterprise

## EMBEDDED SYSTEMS

OPTIMIZED CUDA GENERATION  
FROM MATLAB CODE

MATLAB code



GPU Coder



CUDA

cuDNN, cuSolver,  
cuBLAS TensorRT



## GENERAL-PURPOSE TECHNICAL COMPUTING

NVIDIA GPU SUPPORT IN  
HUNDREDS OF FUNCTIONS in:

- MATLAB
- Deep Learning Toolbox
- Image Processing Toolbox
- Statistics & Machine Learning Toolbox
- Signal Processing Toolbox
- Optimization Toolbox

No need to write CUDA code.  
Custom CUDA code can be reused.

## HPC AND CLOUD

PRE-BUILT MATLAB CONTAINERS  
FOR NVIDIA GPU CLOUD

Instantly access on-premises and  
cloud GPUs with MATLAB

- **CLOUD VENDORS:** Alibaba Cloud, AWS, Azure, Google, and Oracle
- **ON-PREM:** NVIDIA DGX



# NVIDIA AND MATHWORKS COLLABORATION

## A Deep Learning Example









