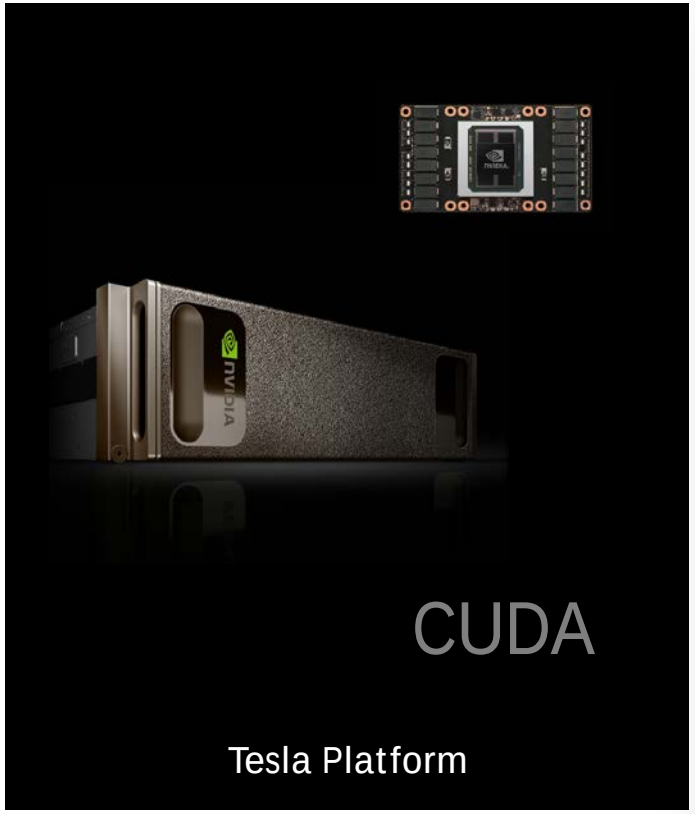
NVIDIA AT TERATEC 2018

June ‘18

Piero Altoe, HPC business developer

NVIDIA.

ONE PLATFORM BUILT FOR BOTH DATA SCIENCE & COMPUTATIONAL SCIENCE



Tesla Platform

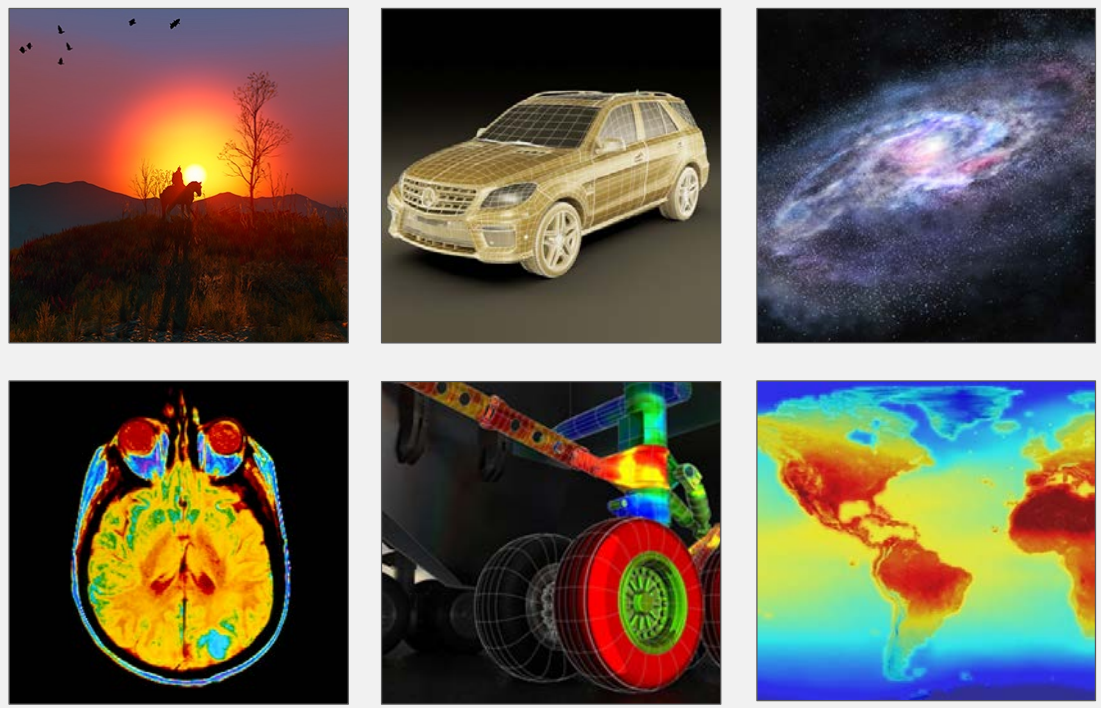
Accelerating AI

Accelerating HPC

70% OF THE WORLD'S SUPERCOMPUTING WORKLOAD ACCELERATED

VASP
AMBER
NAMD
GROMACS
Gaussian
Simulia Abaqus
WRF
OpenFOAM
ANSYS
LS-DYNA
BLAST
LAMMPS
ANSYS Fluent
Quantum Espresso
GAMESS

Top 15 HPC Applications



500+ Accelerated Applications

Intersect360 Research, Nov 2017 "HPC Application Support for GPU Computing"

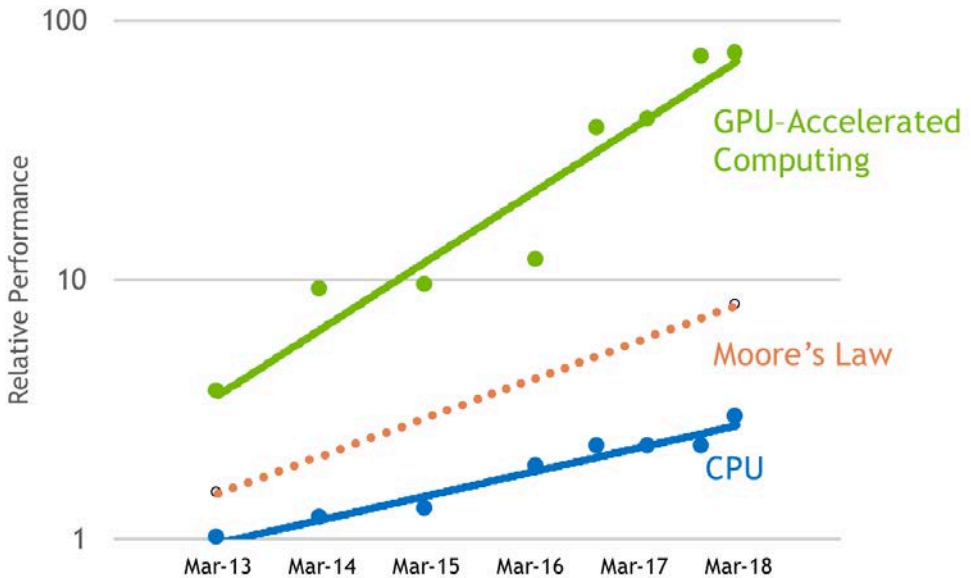
BEYOND MOORE'S LAW

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: 9.0
- cuFFT: 9.0
- cuRAND: 9.0
- cuSOLVER: 9.0
- cuSPARSE: 9.0
- NPP: 9.0
- Thrust: 1.9.0
- CUDA: 9.0
- Resource Mgr: r384
- Base OS: Ubuntu 16.04

Accelerated Server with Volta

THE WORLD'S FIRST 2 PETAFLOPS SYSTEM

INTRODUCING NVIDIA DGX-2

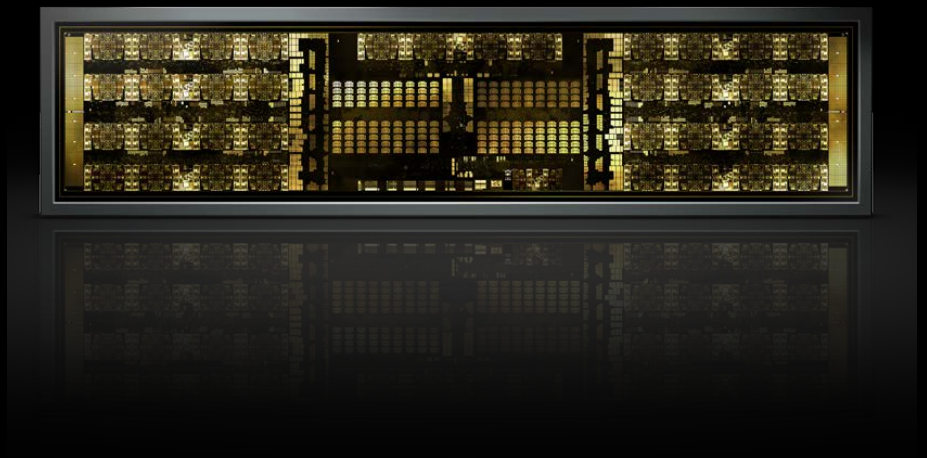
THE WORLD'S MOST POWERFUL AI SYSTEM FOR THE MOST COMPLEX AI CHALLENGES

- DGX-2 is the newest addition to the DGX family, powered by DGX software
- Deliver accelerated AI-at-scale deployment and simplified operations
- Step up to DGX-2 for unrestricted model parallelism and faster time-to-solution

NVSWITCH

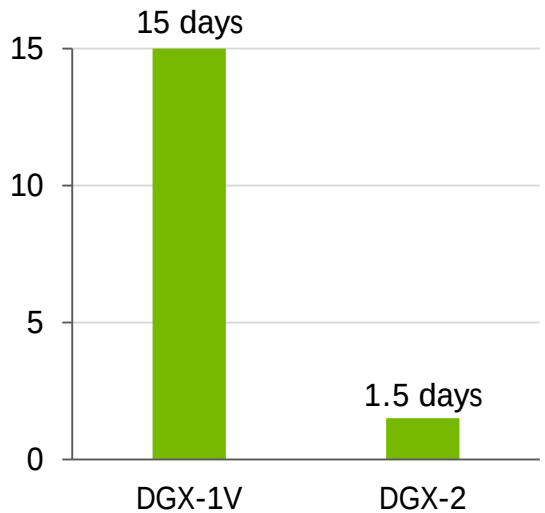
WORLD'S HIGHEST BANDWIDTH ON-NODE SWITCH

7.2 Terabits/sec or 900 GB/sec
18 NVLINK ports | 50GB/s per port bi-directional
Fully-connected crossbar
2 billion transistors | 47.5mm x 47.5mm package



10X PERFORMANCE GAIN LESS THAN A YEAR

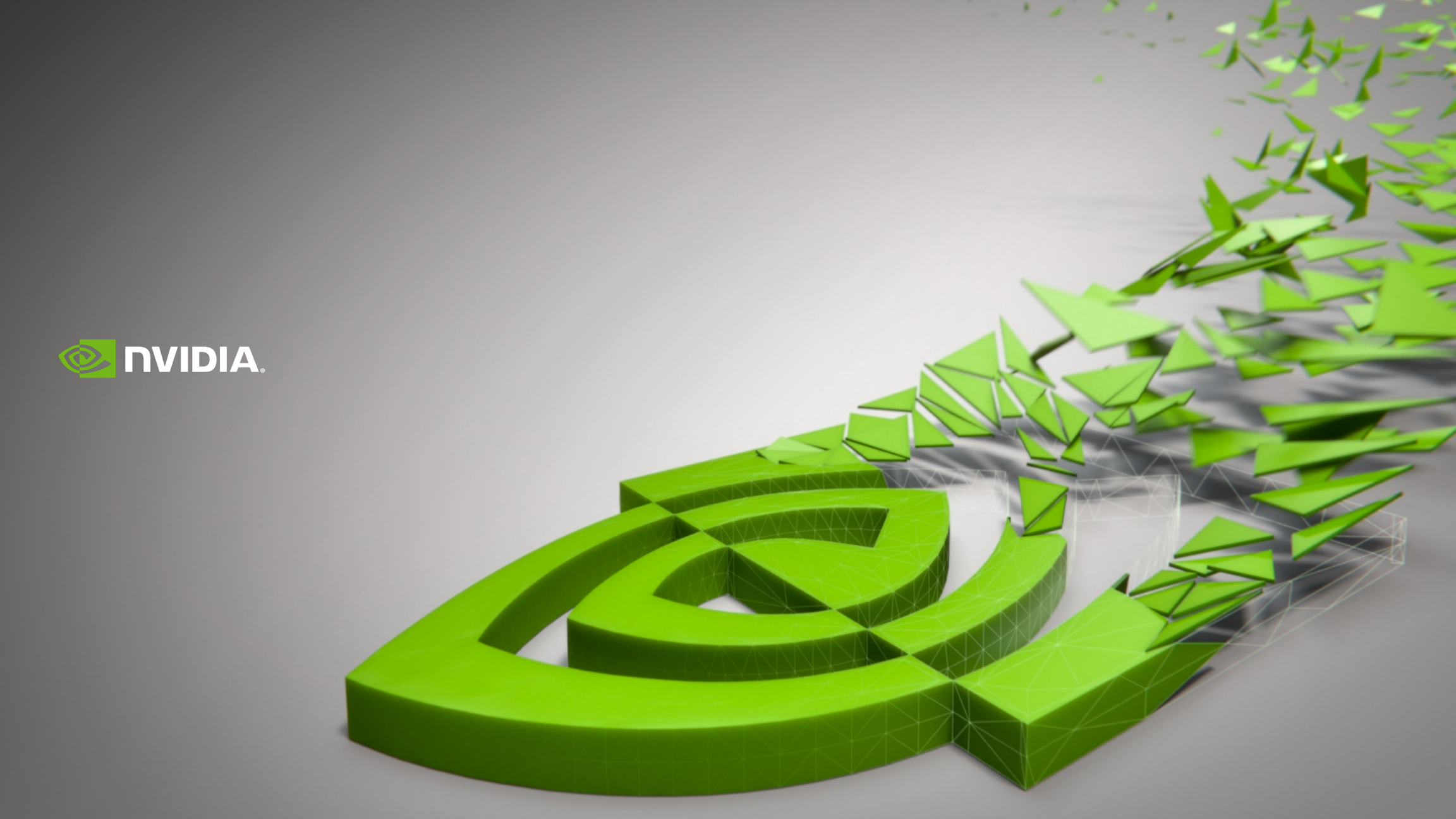
DGX-1, SEP'17



DGX-2, Q3'18

PyTorch Stack: Time to Train FAIRSEQ

software improvements across the stack including NCCL, cuDNN, etc.


NVIDIA