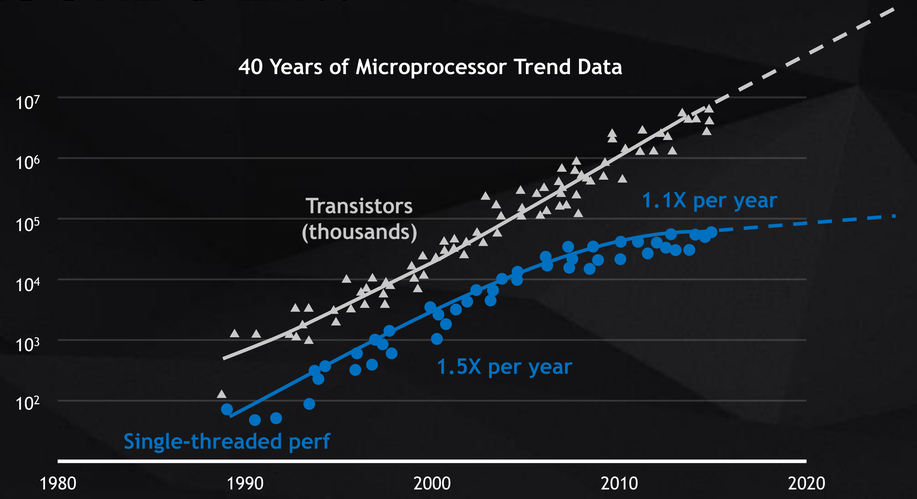
POWERING THE AI REVOLUTION

JENSEN HUANG, FOUNDER & CEO | GTC 2017

LIFE AFTER MOORE'S LAW

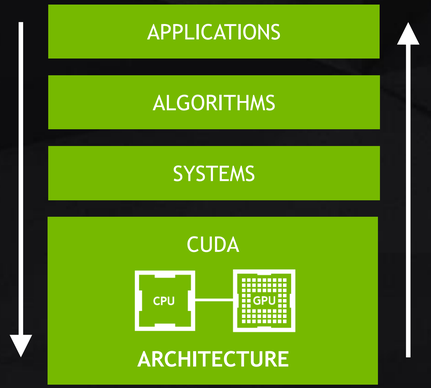
The End of Road for General Purpose Processors and the Future of Computing

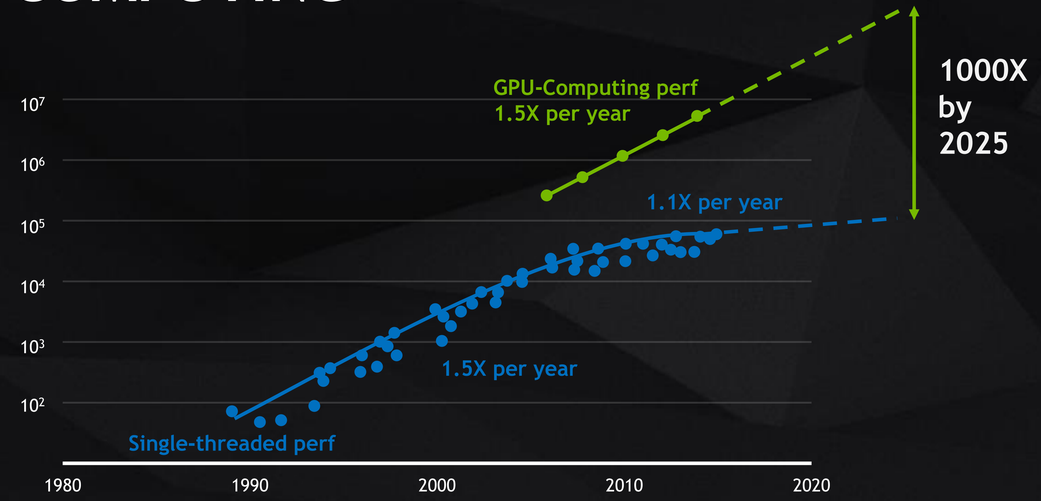
John Hennessy
Stanford University
March 2017



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

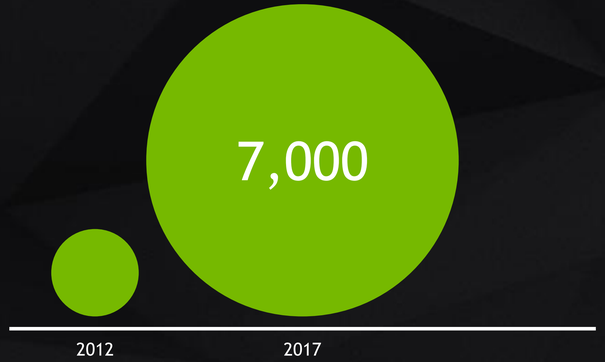
RISE OF GPU COMPUTING



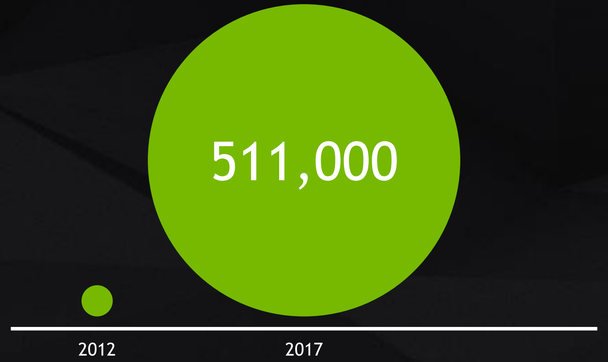


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

RISE OF GPU COMPUTING



GTC Attendees
3X in 5 Years

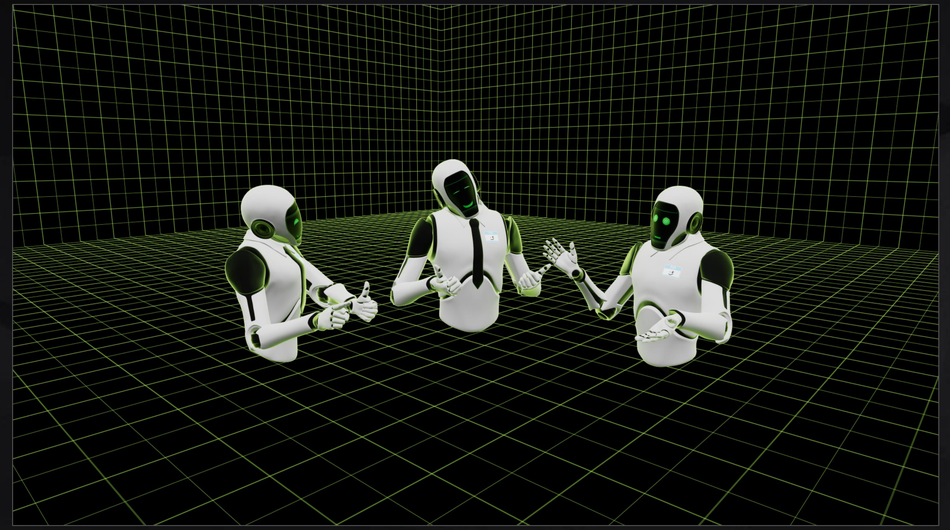


GPU Developers
11X in 5 Years

CUDA Downloads
in 2016

ANNOUNCING PROJECT HOLODECK

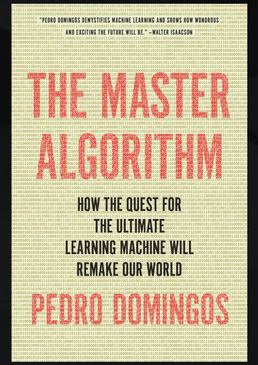
Photorealistic models
Interactive physics
Collaboration



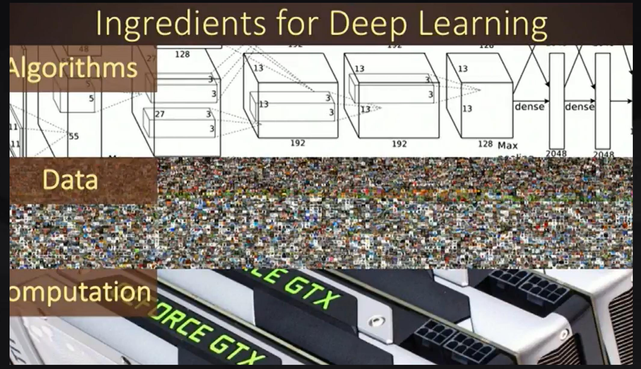
ANNOUNCING PROJECT HOLODECK

Photorealistic models
Interactive physics
Collaboration
Early access in September

ERA OF MACHINE LEARNING



"The Master Algorithm"
— Pedro Domingos



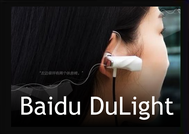
"A Quest for Intelligence"
— Fei-Fei Li

BIG BANG OF MODERN AI

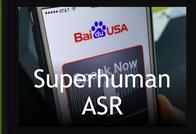
IDSIA CNN on GPU

U Toronto AlexNet on GPU

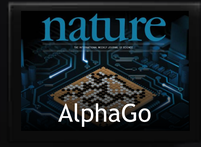
Stanford & NVIDIA Large-scale DNN on GPU



LSTM

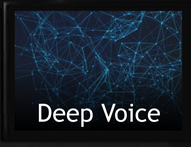


Reinforcement Learning



Auto Encoders

GAN

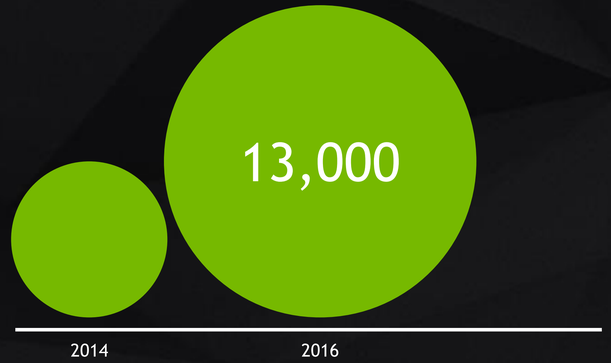


Transfer Learning

DEEP LEARNING FOR RAY TRACING

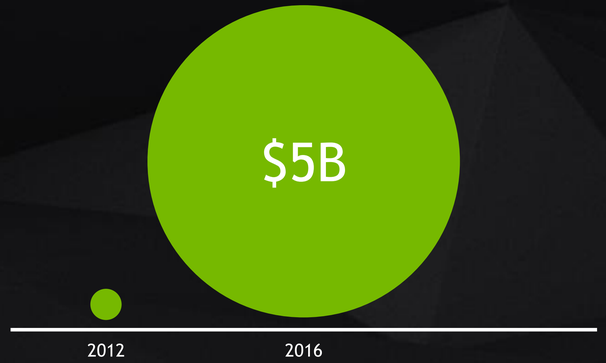
IRAY WITH DEEP LEARNING

BIG BANG OF MODERN AI



NIPS, ICML, CVPR, ICLR Attendees
2X in 2 Years

Udacity Students in AI Programs
100X in 2 Years



AI Startups Investments
9X in 4 Years

POWERING THE AI REVOLUTION

NVIDIA
DEEP LEARNING
SDK

FRAMEWORKS

Caffe2
Microsoft Cognitive Toolkit
mxnet
PYTORCH
TensorFlow
theano

GPU SYSTEMS

cisco
DELL
Hewlett Packard Enterprise
IBM
Lenovo

GPU AAS

Alibaba Cloud aliyun.com
amazon web services
Baidu Cloud
Google Cloud Platform
IBM Cloud
Tencent Cloud
Microsoft Azure

NVAIL

INTERNET SERVICES

HEALTHCARE

NVIDIA RESEARCH

INCEPTION

ENTERPRISE

NVIDIA INCEPTION — 1,300 DEEP LEARNING STARTUPS

HEALTHCARE

athelas · Atomwise (Better medicines faster.) · ARTERYS · babylon · BAYLABS · CLOUD MEDX · deep genomics · Genetesis · imagia · lumiata · PHENOMIC AI (BIOIMAGE INFORMATICS) · zebra (MEDICAL VISION)

RETAIL, ETAIL

focal (Systems) · GOFIND · GROKSTYLE · huew · MARIANA · mashgin · netra · ProductAI · shoppr

FINANCIAL

Alpaca · CAPE (ANALYTICS) · Deep Learni.ng · Orbital Insight · Quantenstein

SECURITY, IVA

AIRSYS · ENTROPIX (RESOLUTION ON DEMAND™) · herta · KUNA · umbo CV

IOT & MANUFACTURING

3DSignals · ABEJA · CLOUDBRAIN · IFM · KONUX · Preferred Networks · PREDII · Reliability Solutions. (The Art of Prediction) · VERDIGRIS · VIMOC Technologies

AUTONOMOUS MACHINES

AIMOTIVE · ARGO AI · BLUE RIVER TECHNOLOGY · dispatch (Powering delivery smart) · drive.ai · marble · nuTonomy · OPTIMUS RIDE · pilot.ai · tuSimple · ZENUITY · ZOOX

CYBER

CYLANCE · deepinstinct · UNIFYID · GreatHorn

AEC

CUBICASA · HOVER · PRENAV · SMARTVID.IO

PLATFORMs & APIs

clarifai · deep gram · Gridspace · KITT.AI · 7INVENSUN · VisionLabs (visual recognition company)

DATA MANAGEMENT

CrowdAI · DATALOGUE · understand.ai

DEVELOPMENT PLATFORMS

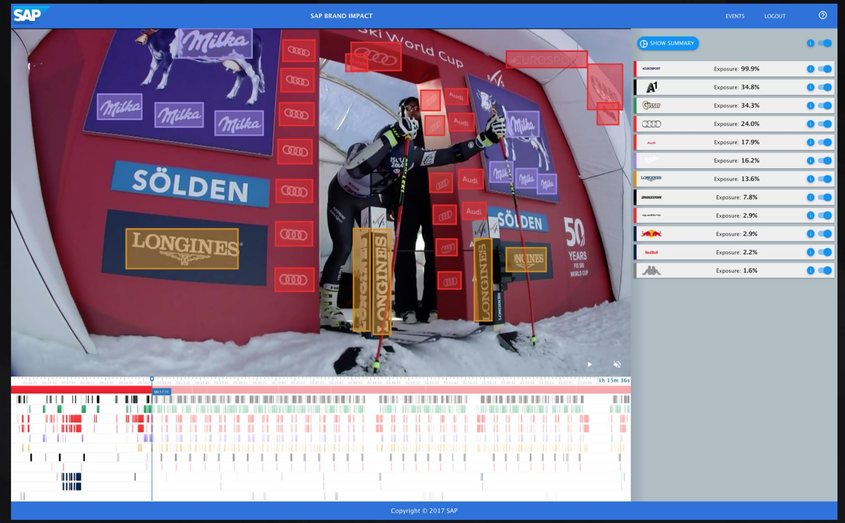
bonsai · SIGOPT · FLOYD · H_2O.ai · skymind

BUSINESS INTELLIGENCE & VISUALIZATION

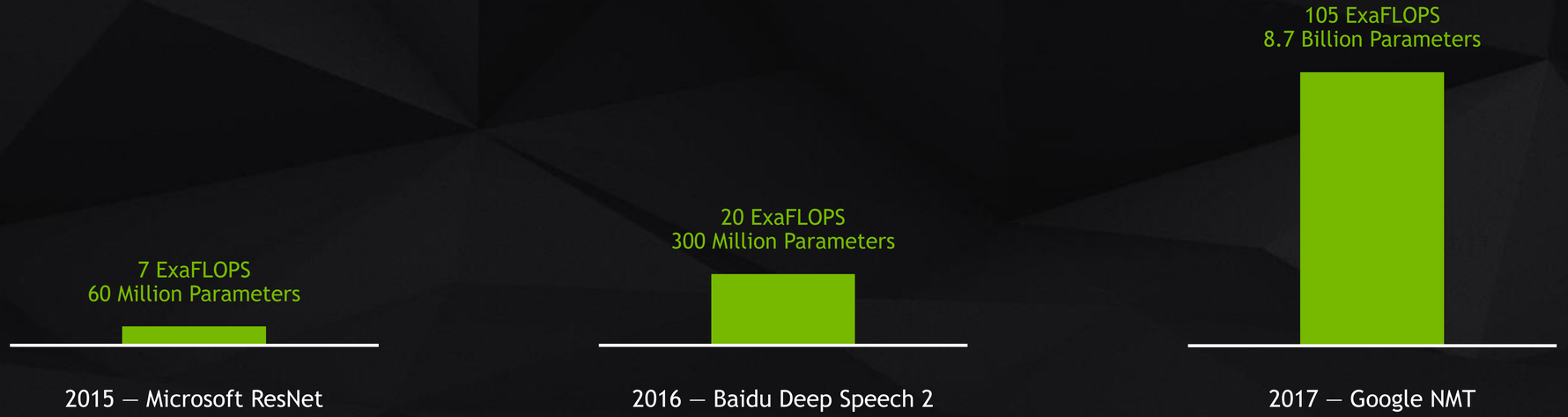
Blazing · graphistry · kinetica · MAPD · SQREAM TECHNOLOGIES

SAP AI FOR THE ENTERPRISE

First commercial AI offerings from SAP
Brand Impact, Service Ticketing, Invoice-to-Record applications
Powered by NVIDIA GPUs on DGX-1 and AWS



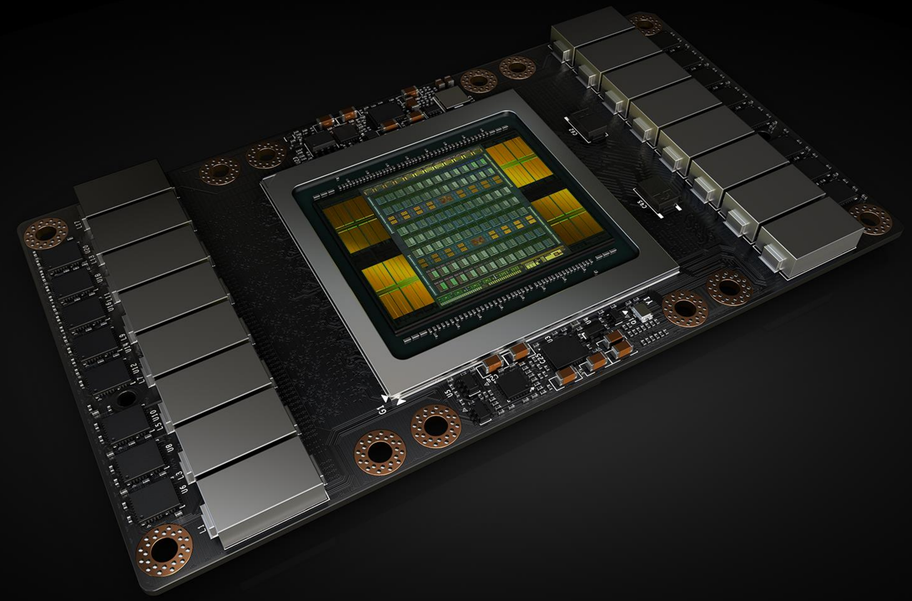
MODEL COMPLEXITY IS EXPLODING



ANNOUNCING TESLA V100

GIANT LEAP FOR AI & HPC
VOLTA WITH NEW TENSOR CORE

21B xtors | TSMC 12nm FFN | 815mm^2
5,120 CUDA cores
7.5 FP64 TFLOPS | 15 FP32 TFLOPS
NEW 120 Tensor TFLOPS
20MB SM RF | 16MB Cache | 16GB HBM2 @ 900 GB/s
300 GB/s NVLink

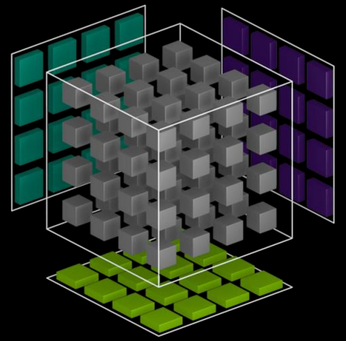


NEW TENSOR CORE

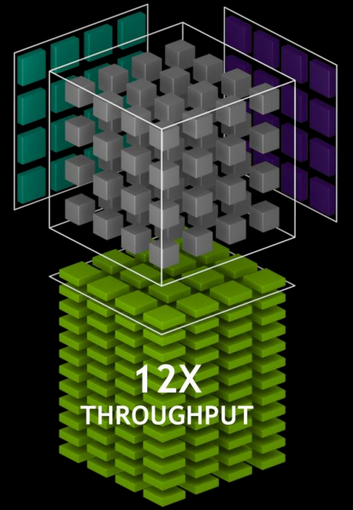
New CUDA TensorOp instructions & data formats
4x4 matrix processing array
D[FP32] = A[FP16] * B[FP16] + C[FP32]
Optimized for deep learning

Activation Inputs Weights Inputs Output Results

PASCAL



VOLTA TENSOR CORES



ANNOUNCING TESLA V100

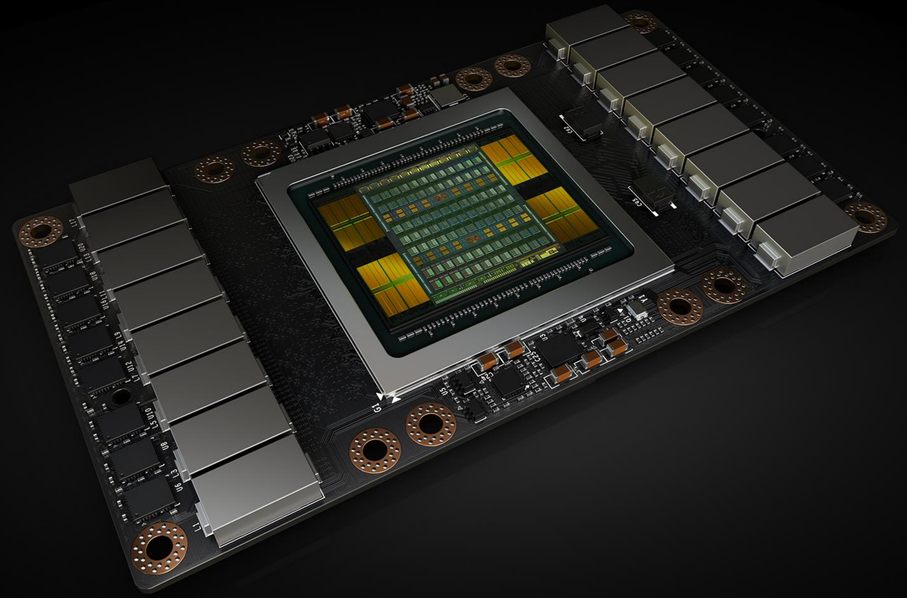
GIANT LEAP FOR AI & HPC
VOLTA WITH NEW TENSOR CORE

Compared to Pascal

1.5X General-purpose FLOPS for HPC

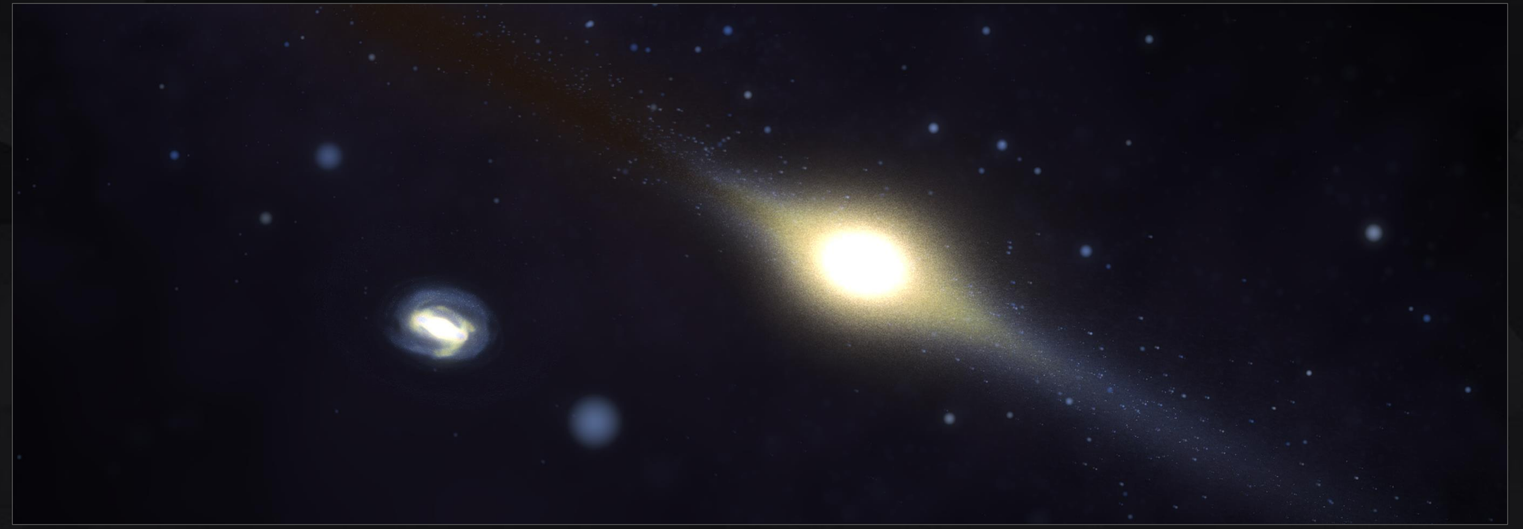
12X Tensor FLOPS for DL training

6X Tensor FLOPS for DL inferencing

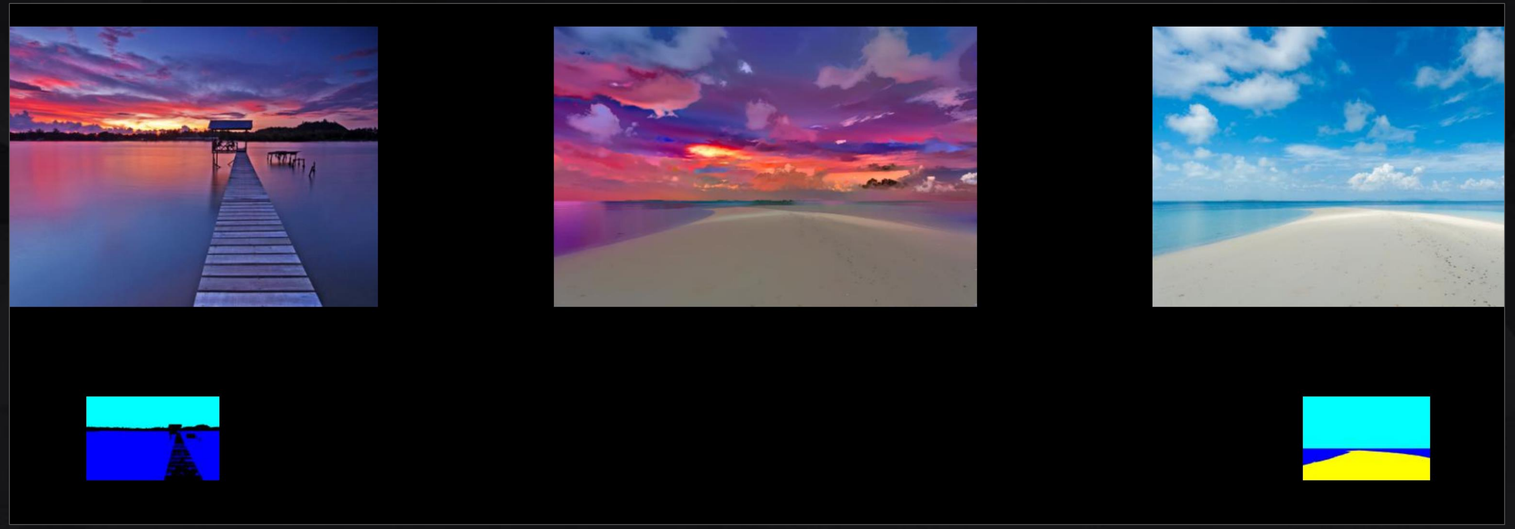



SQUARE ENIX®

GALAXY

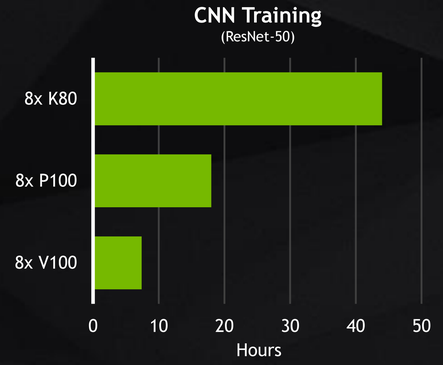


DEEP LEARNING FOR STYLE TRANSFER

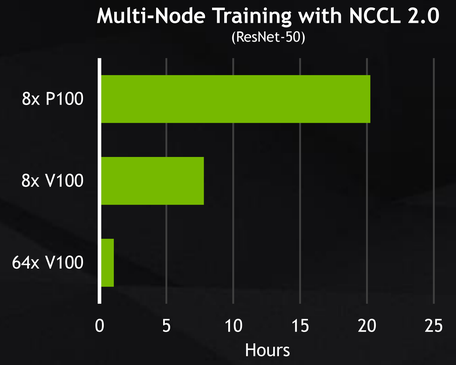


ANNOUNCING NEW FRAMEWORK RELEASES FOR VOLTA

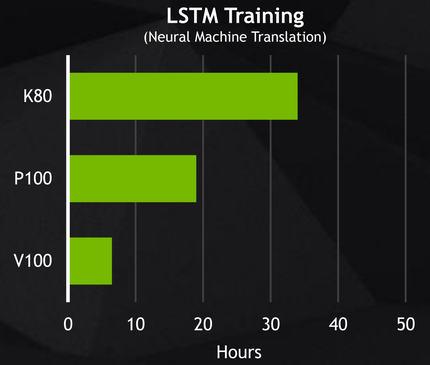
Caffe2



Microsoft Cognitive Toolkit



mxnet



MATT WOOD

General Manager
Deep Learning and AI, AWS

ANNOUNCING
NVIDIA DGX-1 WITH TESLA V100

ESSENTIAL INSTRUMENT OF AI RESEARCH

960 Tensor TFLOPS | 8x Tesla V100 | NVLink Hybrid Cube
From 8 days on TITAN X to 8 hours
400 servers in a box

ANNOUNCING
NVIDIA DGX-1 WITH TESLA V100

ESSENTIAL INSTRUMENT OF AI RESEARCH

960 Tensor TFLOPS | 8x Tesla V100 | NVLink Hybrid Cube
From 8 days on TITAN X to 8 hours
400 servers in a box

ANNOUNCING
NVIDIA DGX-1 WITH TESLA V100

ESSENTIAL INSTRUMENT OF AI RESEARCH

960 Tensor TFLOPS | 8x Tesla V100 | NVLink Hybrid Cube
From 8 days on TITAN X to 8 hours
400 servers in a box
$149,000
Order today: nvidia.com/DGX-1

ANNOUNCING
NVIDIA DGX STATION

PERSONAL DGX

480 Tensor TFLOPS | 4x Tesla V100 16GB | NVLink Fully Connected
3x DisplayPort | 1500W | Water Cooled

ANNOUNCING
NVIDIA DGX STATION

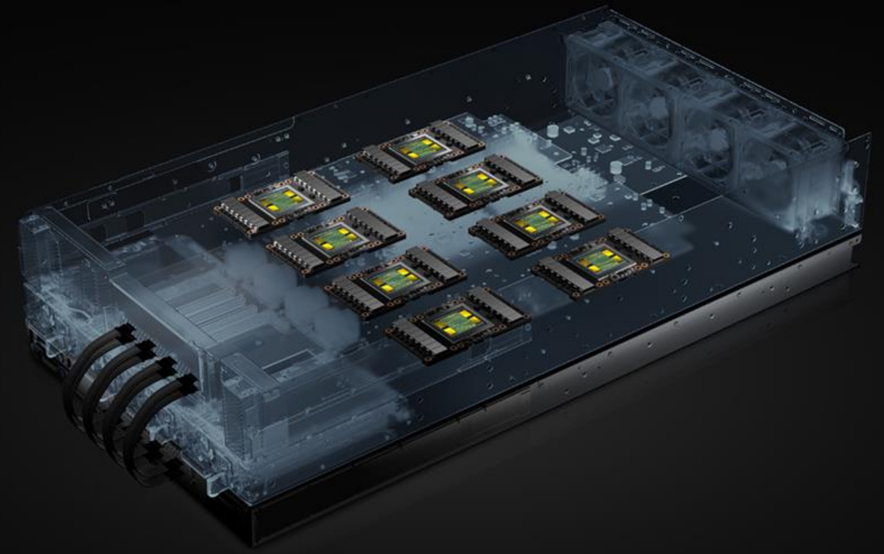
PERSONAL DGX

480 Tensor TFLOPS | 4x Tesla V100 16GB | NVLink Fully Connected
3x DisplayPort | 1500W | Water Cooled
$69,000
Order today: nvidia.com/DGX-Station

ANNOUNCING
HGX-1 WITH TESLA V100

VERSATILE GPU CLOUD COMPUTING

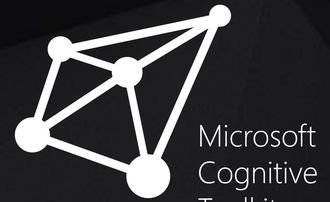
8x Tesla V100 with NVLINK Hybrid Cube
2C:8G | 2C:4G | 1C:2G Configurable
NVIDIA Deep Learning, GRID graphics, CUDA HPC stacks



Microsoft Azure

JASON ZANDER

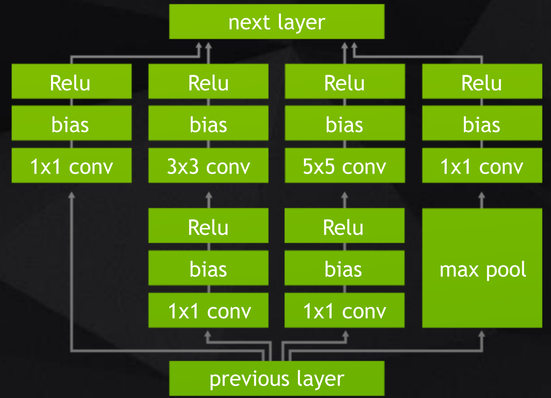
Corporate Vice President
Microsoft Azure, Microsoft



ANNOUNCING TENSORRT FOR TENSORFLOW

COMPILER FOR DEEP LEARNING INFERENCING

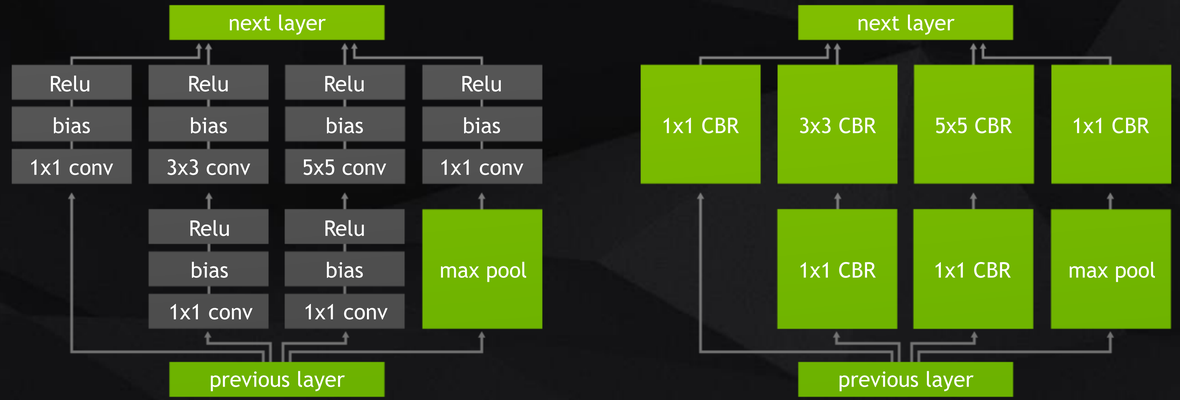
Graph optimizations for vertical and horizontal layer fusion
GPU-specific optimizations
Import models from Caffe and TensorFlow



ANNOUNCING TENSORRT FOR TENSORFLOW

COMPILER FOR DEEP LEARNING INFERENCING

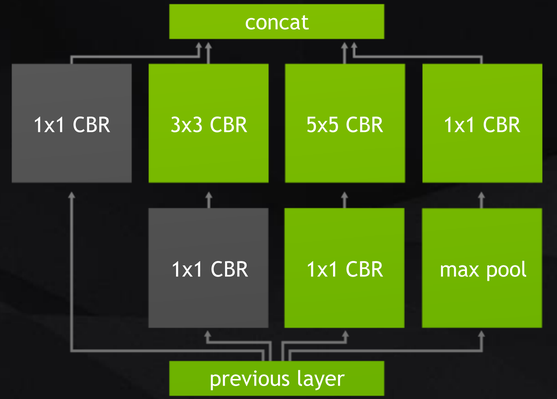
Graph optimizations for vertical and horizontal layer fusion
GPU-specific optimizations
Import models from Caffe and TensorFlow

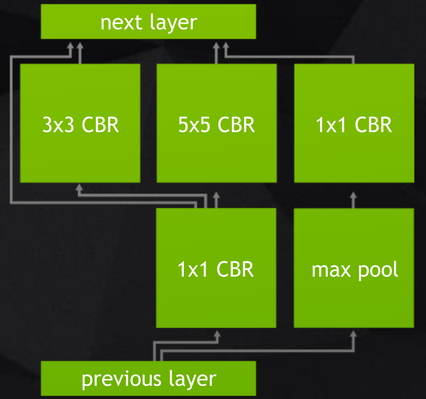


ANNOUNCING TENSORRT FOR TENSORFLOW

COMPILER FOR DEEP LEARNING INFERENCING

Graph optimizations for vertical and horizontal layer fusion
GPU-specific optimizations
Import models from Caffe and TensorFlow

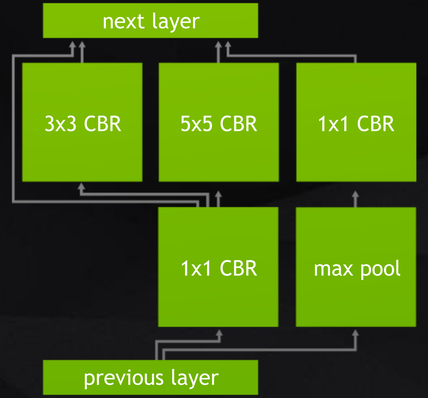


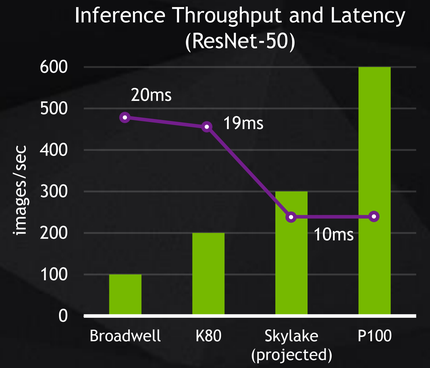


ANNOUNCING TENSORRT FOR TENSORFLOW

COMPILER FOR DEEP LEARNING INFERENCING

Graph optimizations for vertical and horizontal layer fusion
GPU-specific optimizations
Import models from Caffe and TensorFlow

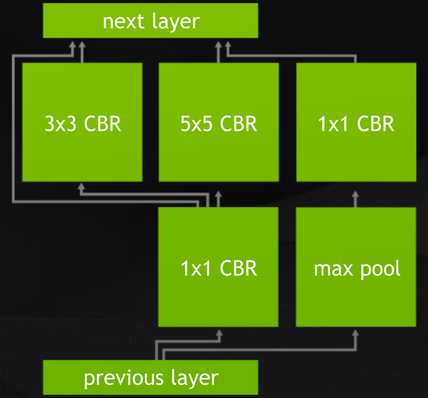


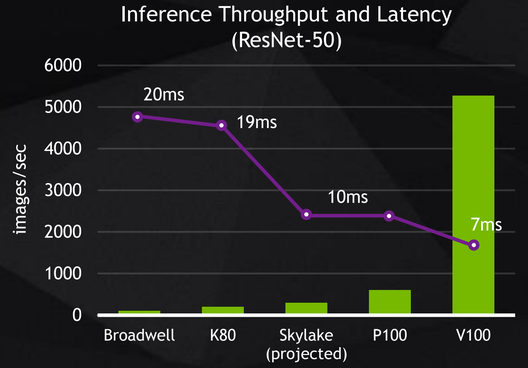


ANNOUNCING TENSORRT FOR TENSORFLOW

COMPILER FOR DEEP LEARNING INFERENCING

Graph optimizations for vertical and horizontal layer fusion
GPU-specific optimizations
Import models from Caffe and TensorFlow

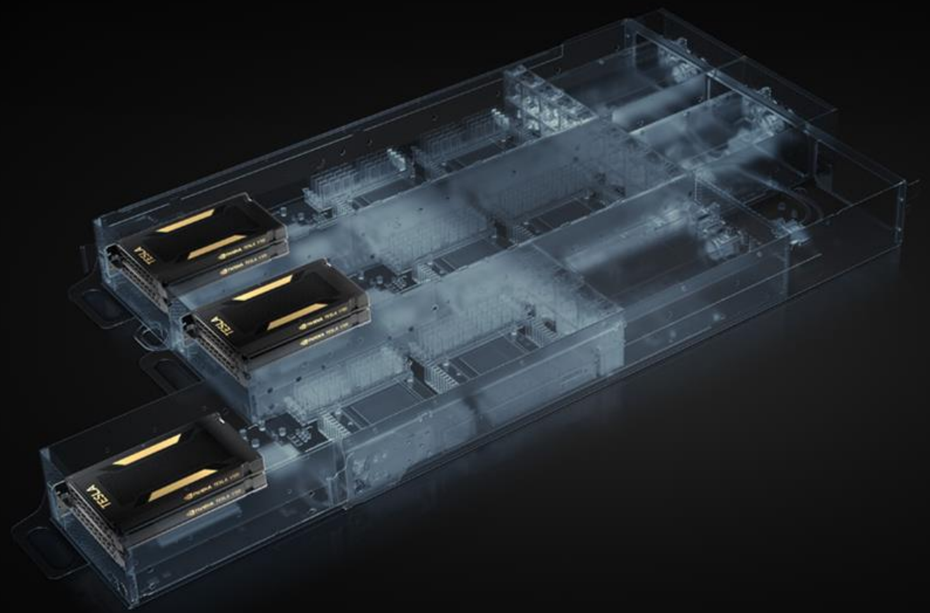




ANNOUNCING TESLA V100 FOR HYPERSCALE INFERENCE

15-25X INFERENCE SPEED-UP VS SKYLAKE

150W | FHHL PCIE



THE CASE FOR GPU ACCELERATED DATACENTERS

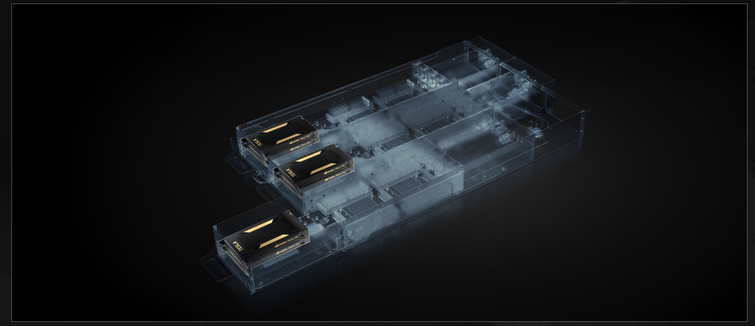
300K inf/s Datacenter

@300 inf/s > 1K CPUs

1K CPUs > 500 Nodes

@$3K = $1.5M

@500W = 250KW



500 Nodes CPU Servers → Tesla V100 Reduce 15X → 33 Nodes GPU Accelerated Server

NVIDIA
DEEP LEARNING STACK

DEEP LEARNING FRAMEWORKS

DEEP LEARNING LIBRARIES
NVIDIA cuDNN, NCCL,
cuBLAS, TensorRT

CUDA DRIVER

OPERATING SYSTEM

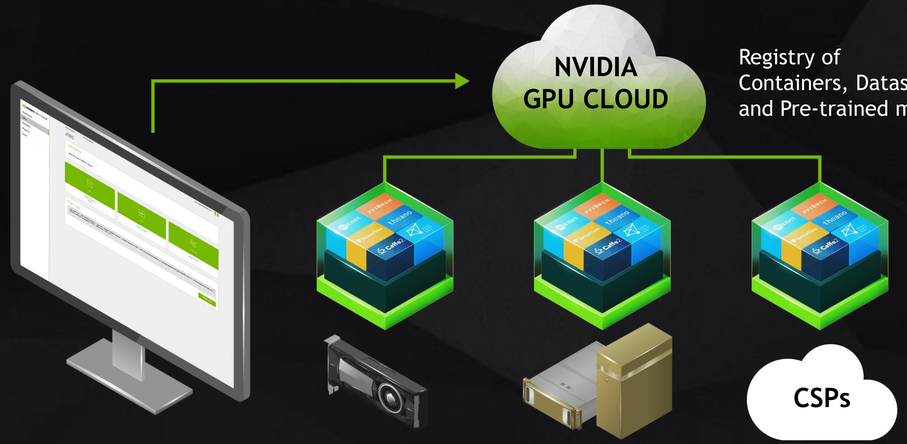
GPU

SYSTEM

ANNOUNCING
NVIDIA GPU CLOUD

GPU-ACCELERATED CLOUD PLATFORM OPTIMIZED FOR DEEP LEARNING

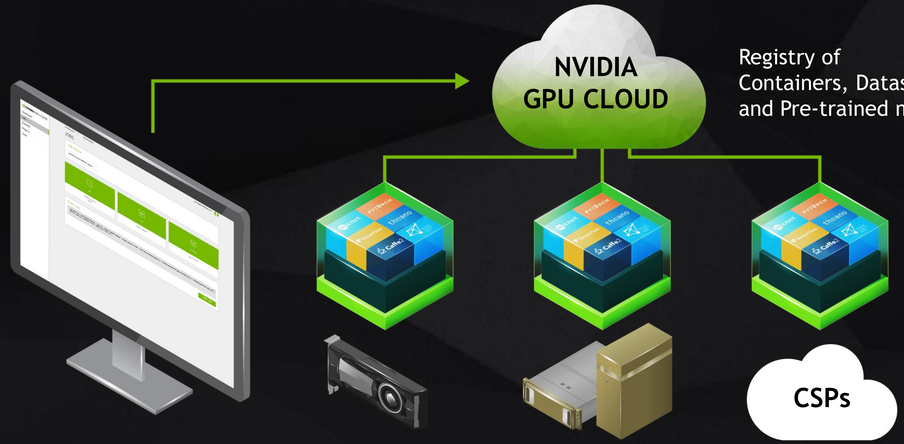
Containerized in NVDocker
Optimization across the full stack
Always up-to-date
Fully tested and maintained by NVIDIA



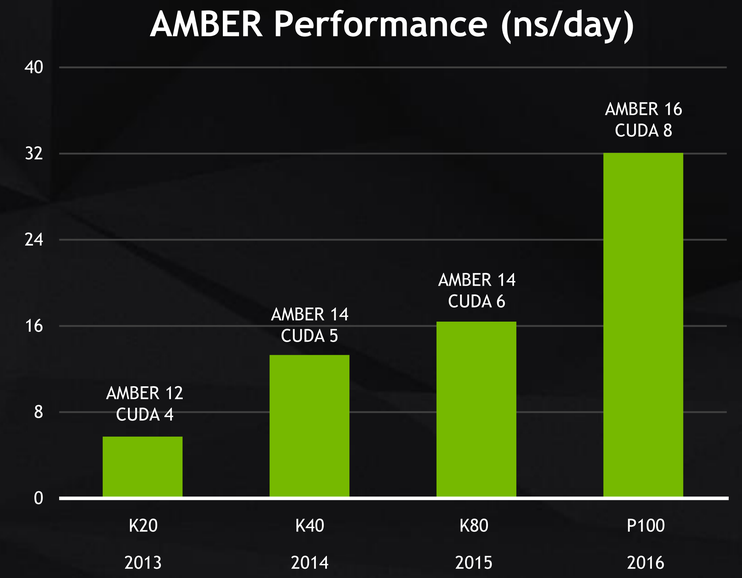
ANNOUNCING
NVIDIA GPU CLOUD

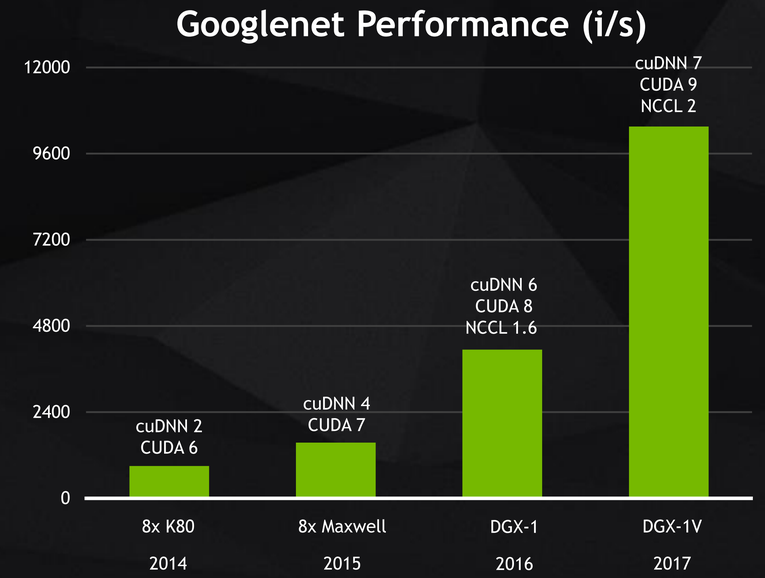
GPU-ACCELERATED CLOUD PLATFORM OPTIMIZED FOR DEEP LEARNING

Containerized in NVDocker
Optimization across the full stack
Always up-to-date
Fully tested and maintained by NVIDIA
Beta in July

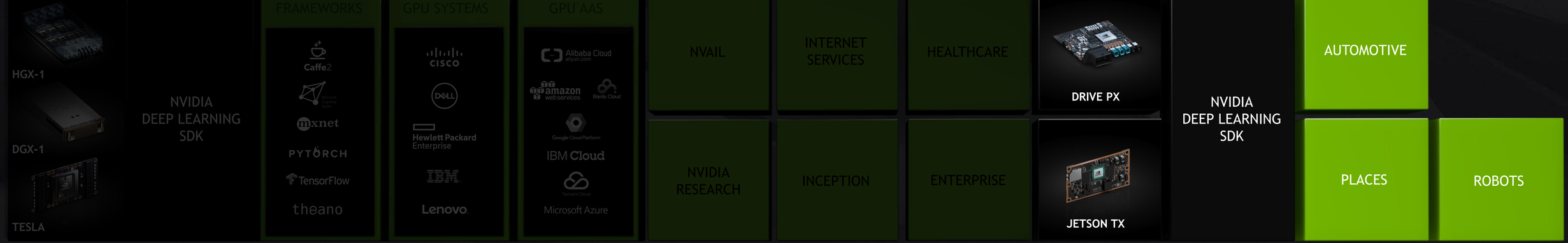


POWER OF GPU COMPUTING



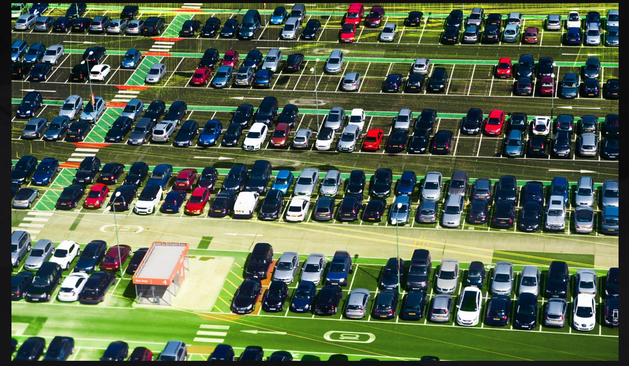


POWERING THE AI REVOLUTION



AI REVOLUTIONIZING TRANSPORTATION

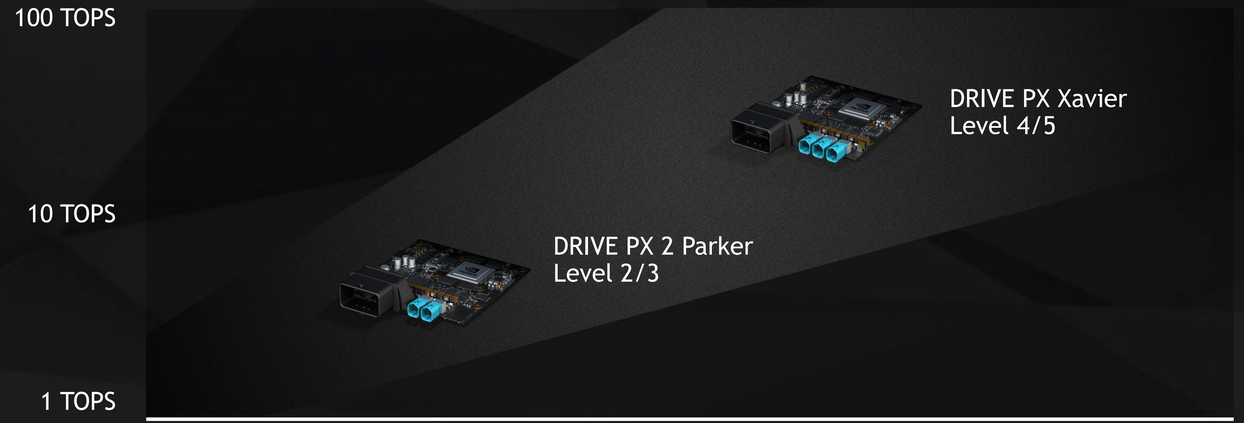
280B Miles per Year

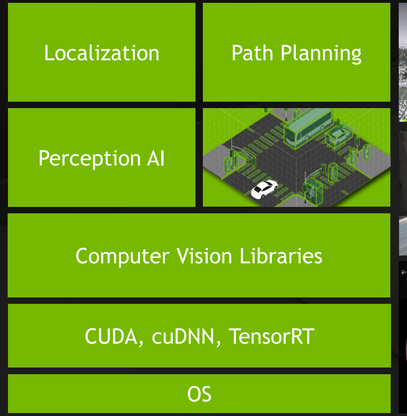


800M Parking Spots for 250M Cars in U.S.

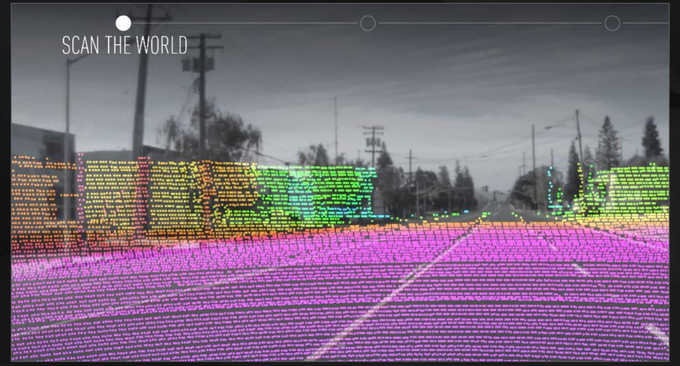
Domino’s: 1M Pizzas Delivered per Day

NVIDIA DRIVE — AI CAR PLATFORM





NVIDIA DRIVE



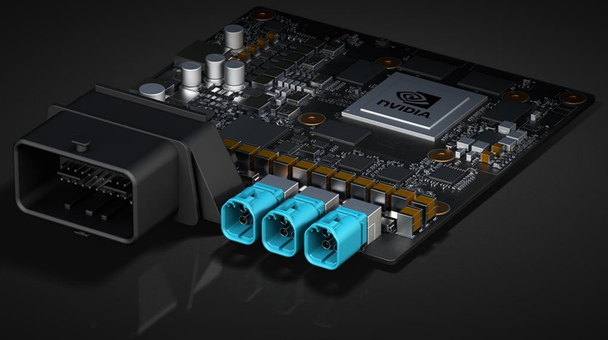
Mapping-to-Driving

Co-Pilot

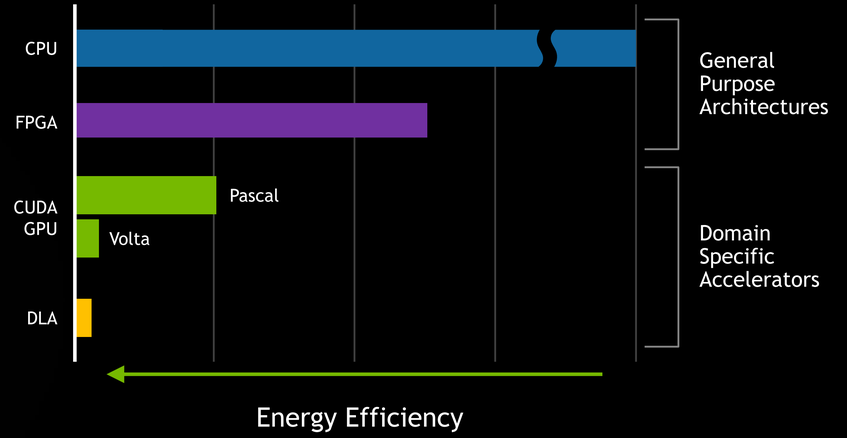
Guardian Angel

ANNOUNCING TOYOTA SELECTS NVIDIA DRIVE PX FOR AUTONOMOUS VEHICLES

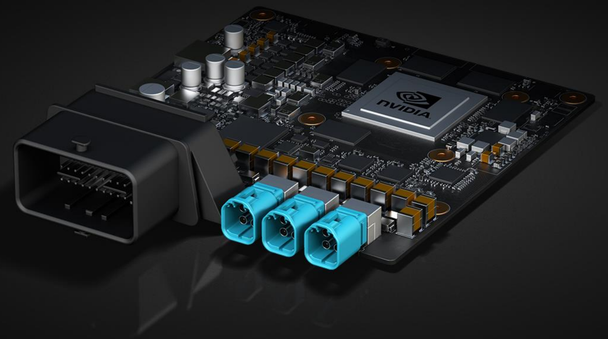
AI PROCESSOR FOR AUTONOMOUS MACHINES



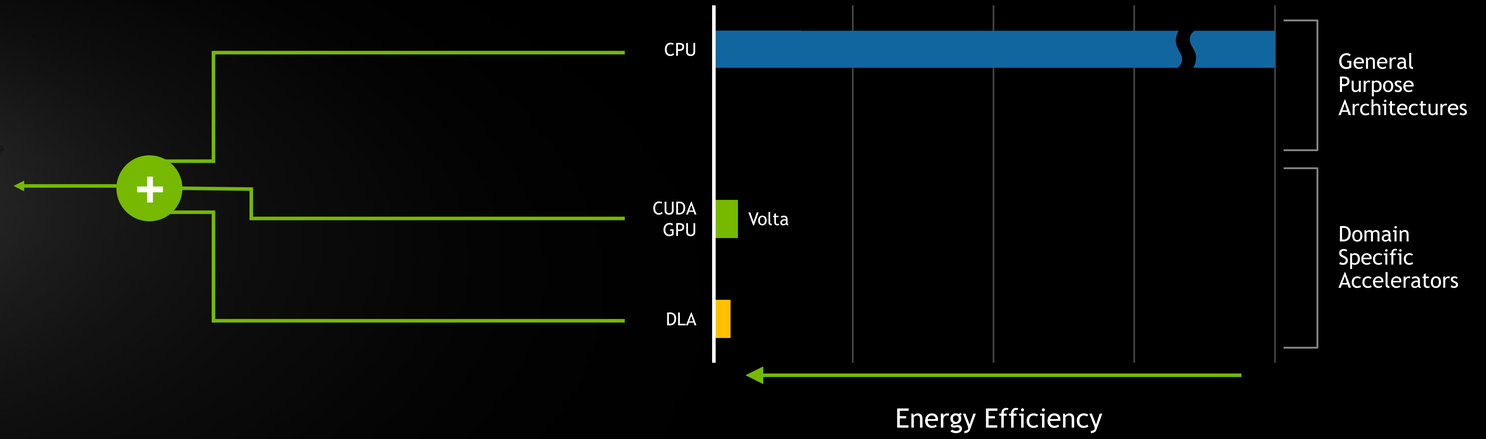
XAVIER
30 TOPS DL | 30W
Custom ARM64 CPU | 512 Core Volta GPU | 10 TOPS DL Accelerator



AI PROCESSOR FOR AUTONOMOUS MACHINES

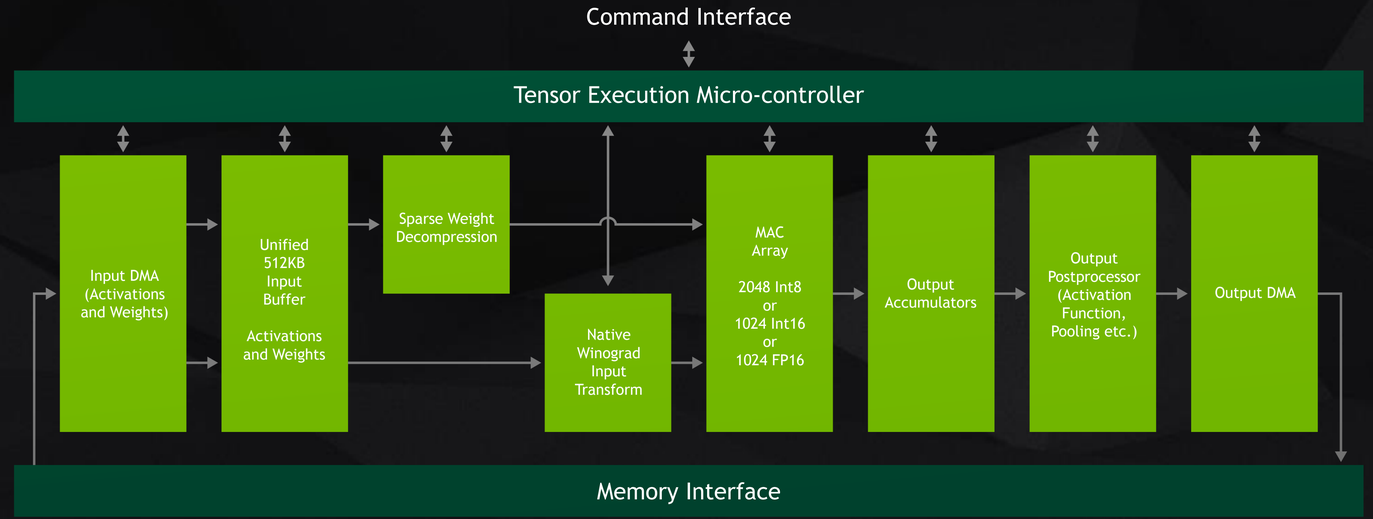


XAVIER
30 TOPS DL | 30W
Custom ARM64 CPU | 512 Core Volta GPU | 10 TOPS DL Accelerator

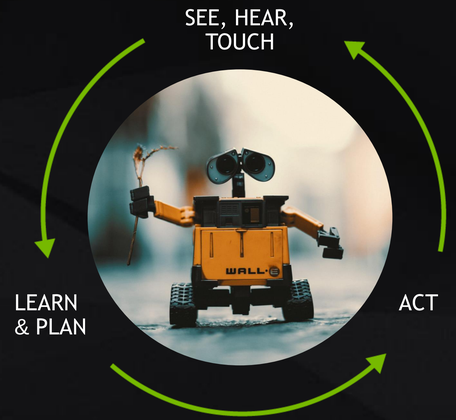


ANNOUNCING XAVIER DLA NOW OPEN SOURCE

Early Access July
General Release September



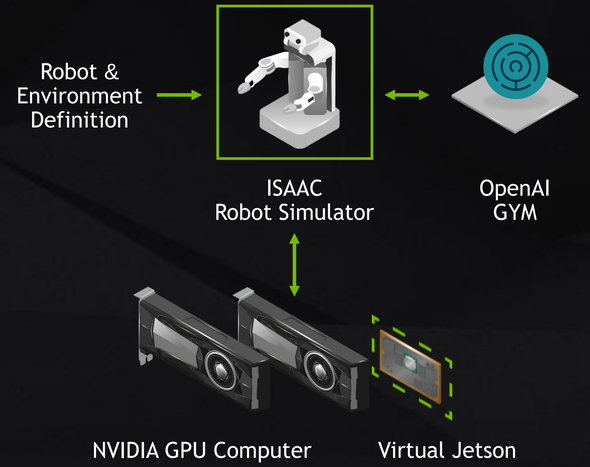
ROBOTS



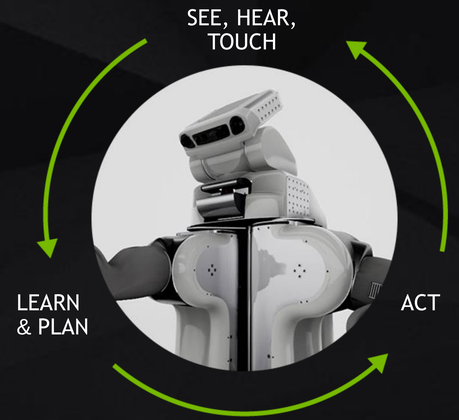
ROBOTS

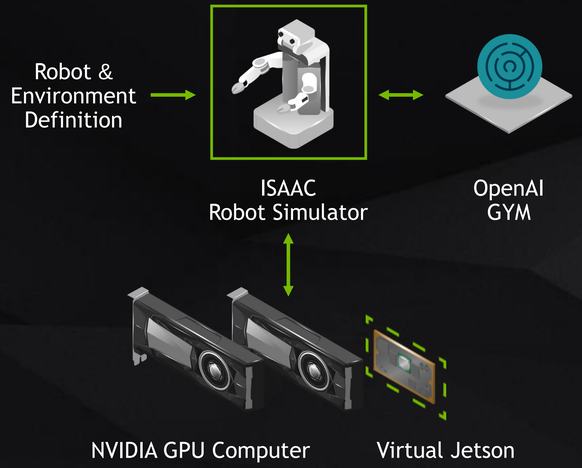
Credit: Yevgen Chebotar, Karol Hausman, Marvin Zhang, Sergey Levine

ANNOUNCING ISAAC ROBOT SIMULATOR

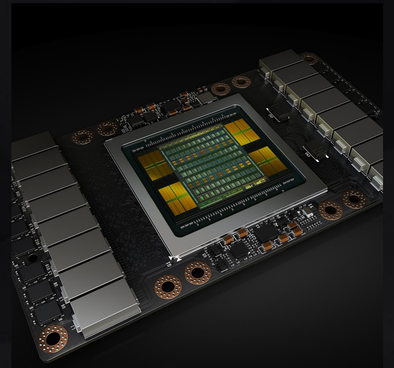


ANNOUNCING ISAAC ROBOT SIMULATOR

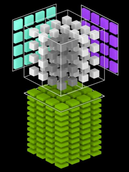




NVIDIA POWERING THE AI REVOLUTION



Tesla V100



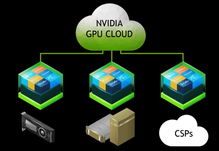
Tensor Core

TensorRT

Alibaba amazon Baidu百度 facebook

Google Microsoft Tencent

NVIDIA GPU in Every Cloud

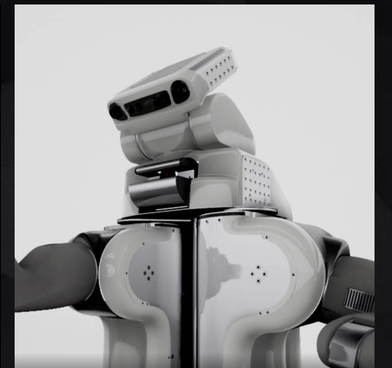


NVIDIA GPU Cloud

DGX-1 and DGX Station

Xavier DLA
Open Source

TOYOTA



ISAAC