
NVIDIA

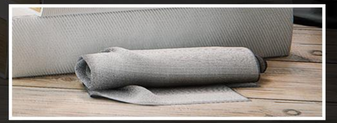
Amazing
Graphics

Amazing
Science

Amazing
AI

Amazing
Robots

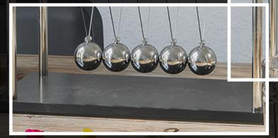
SCREEN-SPACE AMBIENT OCCLUSION
BAKED LIGHTING

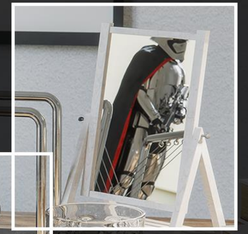


GLOBAL ILLUMINATION

SCREEN-SPACE REFLECTIONS
ENVIRONMENT MAPS

RAY TRACED REFLECTIONS



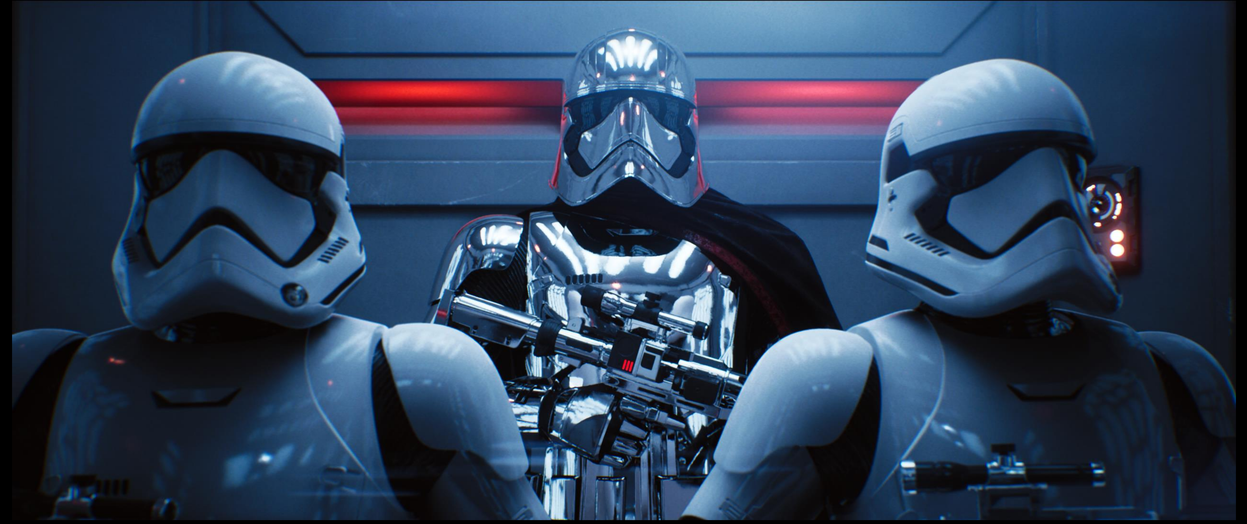


SCREEN-SPACE REFRACTION
DEPTH SORTING

CAUSTICS

SUBSURFACE SHADING APPROXIMATION

SUBSURFACE SCATTERING

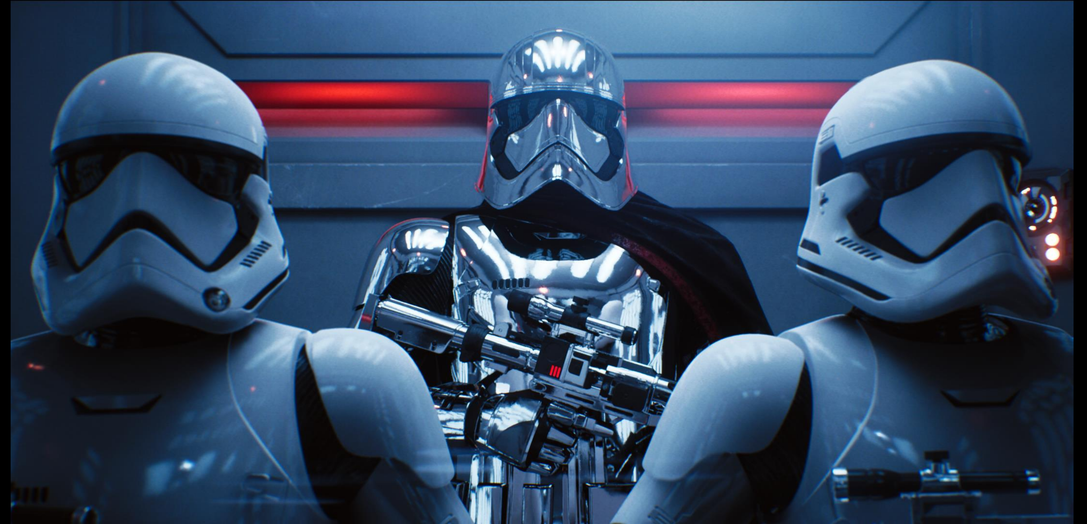


ANNOUNCING
NVIDIA RTX TECHNOLOGY

ILMxLAB
IMMERSIVE ENTERTAINMENT

UNREAL
ENGINE

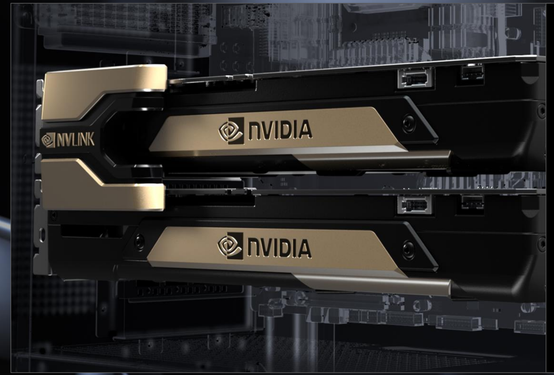
NVIDIA.



ANNOUNCING QUADRO GV100 WITH NVIDIA RTX TECHNOLOGY

GIANT LEAP FOR REAL-TIME COMPUTER GRAPHICS

2 GV100s Connected by NVLink2
64GB HBM2 Memory
10,240 CUDA Cores
236 TFLOPS Tensor Cores



NVIDIA OptiX | Vulkan | Microsoft DXR

NVIDIA RTX Technology

NVIDIA Volta GPU

ONE BILLION IMAGES RENDERED EVERY YEAR

GAMING
400 Games

MEDIA & ENTERTAINMENT
500 Movies

PRODUCT DESIGN
12M Designers

ARCHITECTURE
150,000 Architects

TRADITIONAL RENDER FARM

280 Dual-CPU Servers
168 kW

NVIDIA RTX QUADRO GV100

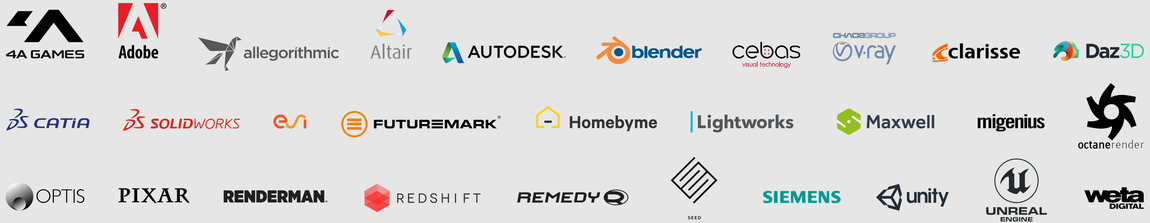
BIG SAVINGS FOR RENDERING

14 Quad-GPU Servers
24 kW

1/5 the Cost 1/7 the Space 1/7 the Power

NVIDIA RTX EXCITEMENT

TOOLS | ENGINES
GAMING | MEDIA & ENTERTAINMENT | PRODUCT DESIGN | ARCHITECTURE



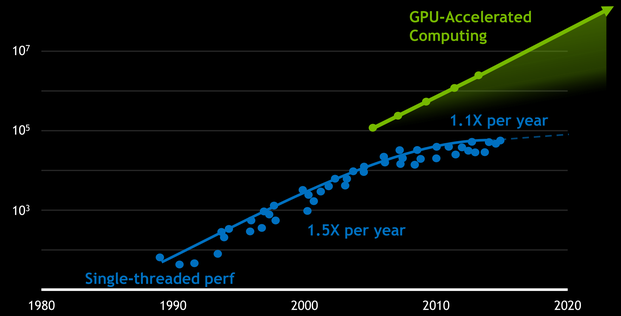
"With RTX we can now do ray tracing renders interactively. It's just fantastic!"

—Sébastien Guichou, CTO, Isotropix

"NVIDIA RTX opens the door to make real-time ray tracing a reality!"

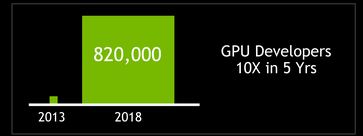
—Kim Libreri, CTO, Epic Games

RISE OF GPU COMPUTING

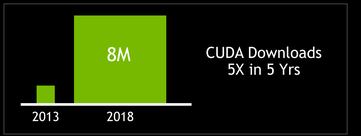


40 Years of CPU Trend Data

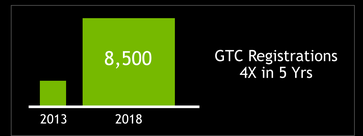
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp



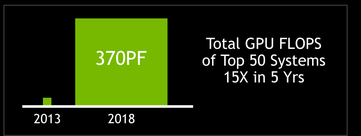
GPU Developers
10X in 5 Yrs



CUDA Downloads
5X in 5 Yrs



GTC Registrations
4X in 5 Yrs



Total GPU FLOPS
of Top 50 Systems
15X in 5 Yrs

SCIENCE NEEDS SUPERCHARGED COMPUTERS

Caltech

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

PRINCETON UNIVERSITY

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

ETH zürich

MeteoSwiss

CSCS

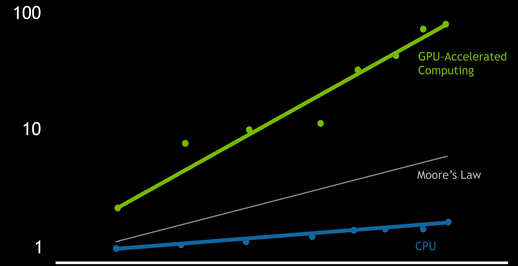
ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

NCSA

SUPERCHARGED COMPUTING

Fermi GPU Server 2013

HPC Applications	Amber	12
	NAMD	2.9
GPU Acceleration Stack	cuBLAS	5.0
	cuFFT	5.0
	NPP	5.0
	CUDA	5.0
	cuRAND	5.0
	cuSPARSE	5.0
	Res Mgr	R304
	BaseOS	CentOS 6.2



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

Volta GPU Server 2018

HPC Applications	Amber	16
	CHROMA	2018
	Gyrokinetic TC	2017
	LAMMPS	2018
	MILC	2018
	NAMD	2.13
	Quantum Esp.	6.1
	SPECFEM3D	2018
GPU Acceleration Stack	cuBLAS	9.0
	cuFFT	9.0
	NPP	9.0
	CUDA	9.0
	cuRAND	9.0
	cuSPARSE	9.0
	Res Mgr	R384
	BaseOS	Ubuntu 16.04

TRADITIONAL HPC CLUSTER

600 Dual-CPU Servers
360 kW

NVIDIA TESLA V100

BIG SAVINGS FOR HPC

30 Quad-GPU Servers
48 kW

1/5 the Cost 1/7 the Space 1/7 the Power

CLARA — MEDICAL IMAGING SUPERCOMPUTER

IMAGING & VISUALIZATION APPS

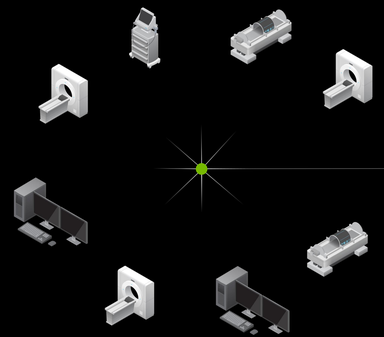
CUDA | CUDNN | TENSORRT | OGL | RTX

GPU CONTAINERS | VGPU

NVIDIA GPU SERVER

https://www.philips.com/a-w/about/news/archive/standard/news/press/2017/20170508-philips-new-ob-gyn-ultrasound-innovations-with-anatomical-intelligence-provide-lifelike-3d-images.html

CLARA — MEDICAL IMAGING SUPERCOMPUTER

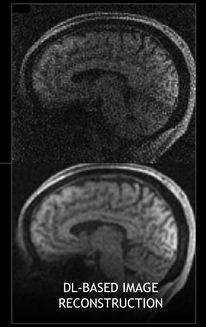


IMAGING & VISUALIZATION APPS

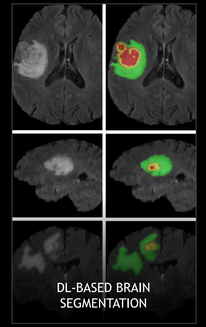
CUDA | CUDNN | TENSORRT | OGL | RTX

GPU CONTAINERS | VGPU

NVIDIA GPU SERVER



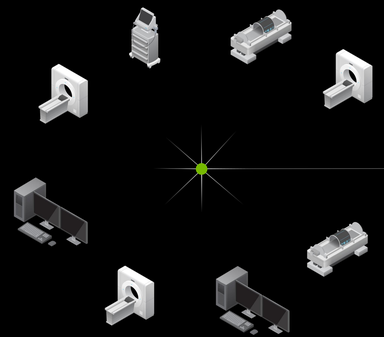
DL-BASED IMAGE RECONSTRUCTION



DL-BASED BRAIN SEGMENTATION

CINEMATIC RENDERING

CLARA — MEDICAL IMAGING SUPERCOMPUTER

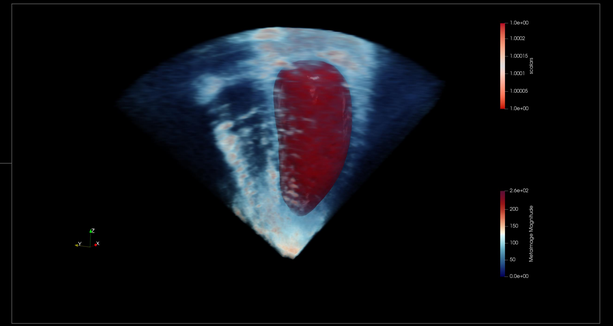


IMAGING & VISUALIZATION APPS

CUDA | CUDNN | TENSORRT | OGL | RTX

GPU CONTAINERS | VGPU

NVIDIA GPU SERVER

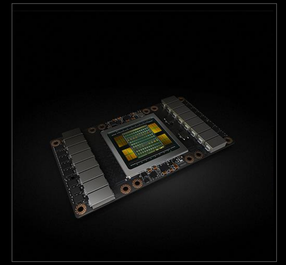


IMAGING DEVELOPMENT PARTNERS

ULTRASOUND | MRI | CT | X-RAY | MAMMO | PET

HEALTHCARE PROVIDERS | STARTUPS | IMAGING COMPANIES

NVIDIA AI PLATFORM

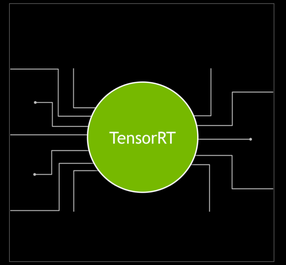


Tesla V100

DGX-1 and DGX Station

Every Cloud
Every Computer Maker

NVIDIA GPU Cloud



NVIDIA AI Inference

TITAN V

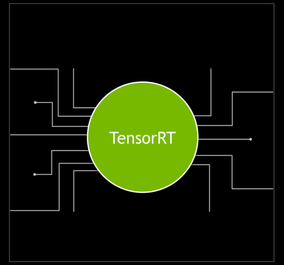
NVIDIA AI PLATFORM

Tesla V100

DGX-1 and DGX Station

Every Cloud
Every Computer Maker

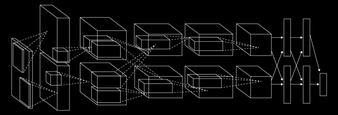
NVIDIA GPU Cloud



NVIDIA AI Inference

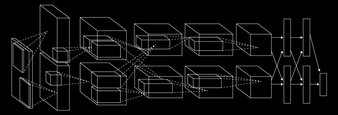
TITAN V

AlexNet



CAMBRIAN EXPLOSION

Convolutional Networks

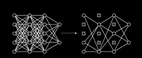


Encoder/Decoder

ReLu

BatchNorm

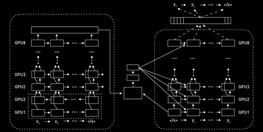
Concat



Dropout

Pooling

Recurrent Networks



LSTM

GRU

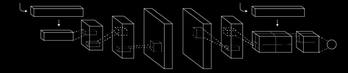
Beam Search

WaveNet

CTC

Attention

Generative Adversarial Networks



3D-GAN

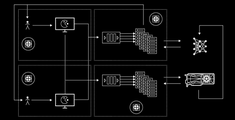
MedGAN

Conditional GAN

Coupled GAN

Speech Enhancement GAN

Reinforcement Learning

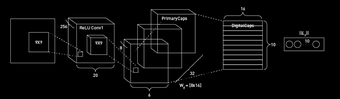


DQN

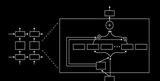
Simulation

DDPG

New Species



Capsule Nets



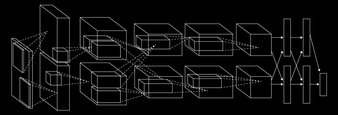
Mixture of Experts

Neural Collaborative Filtering

Block Sparse LSTM

CAMBRIAN EXPLOSION

Convolutional Networks

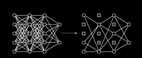


Encoder/Decoder

ReLu

BatchNorm

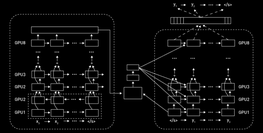
Concat



Dropout

Pooling

Recurrent Networks



LSTM

GRU

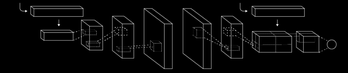
Beam Search

WaveNet

CTC

Attention

Generative Adversarial Networks



3D-GAN

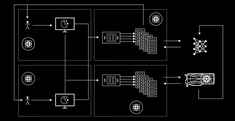
MedGAN

Conditional GAN

Coupled GAN

Speech Enhancement GAN

Reinforcement Learning

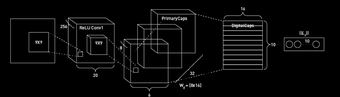


DQN

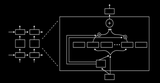
Simulation

DDPG

New Species



Capsule Nets



Mixture of Experts

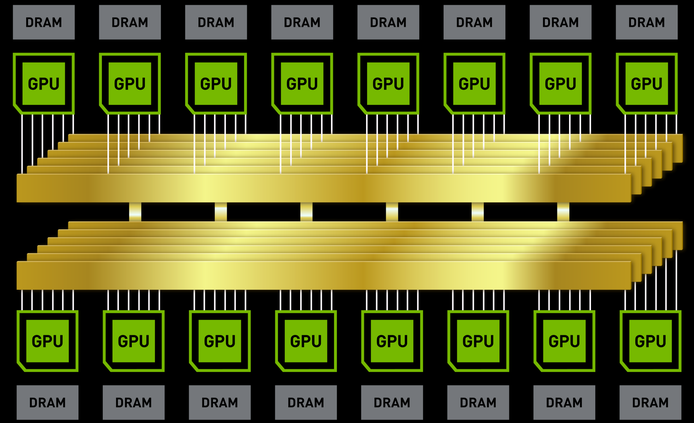
Neural Collaborative Filtering

Block Sparse LSTM

“THE WORLD WANTS A GIGANTIC GPU”

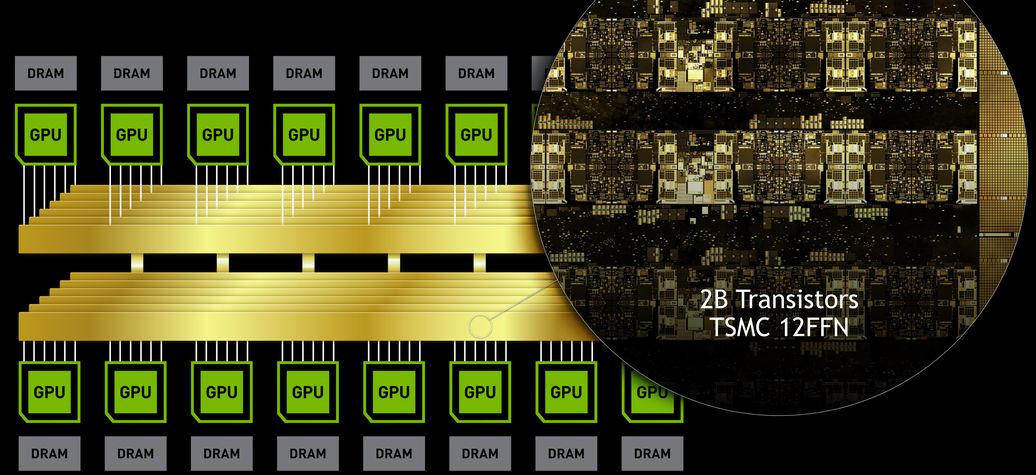
“THE WORLD’S LARGEST GPU”

16 Tesla V100 32GB Connected by NVSwitch
On-chip Memory Fabric Semantic Extended Across All GPUs
512GB HBM2 and 14.4TB/sec Aggregate
81,920 CUDA Cores
2,000 TFLOPS Tensor Cores



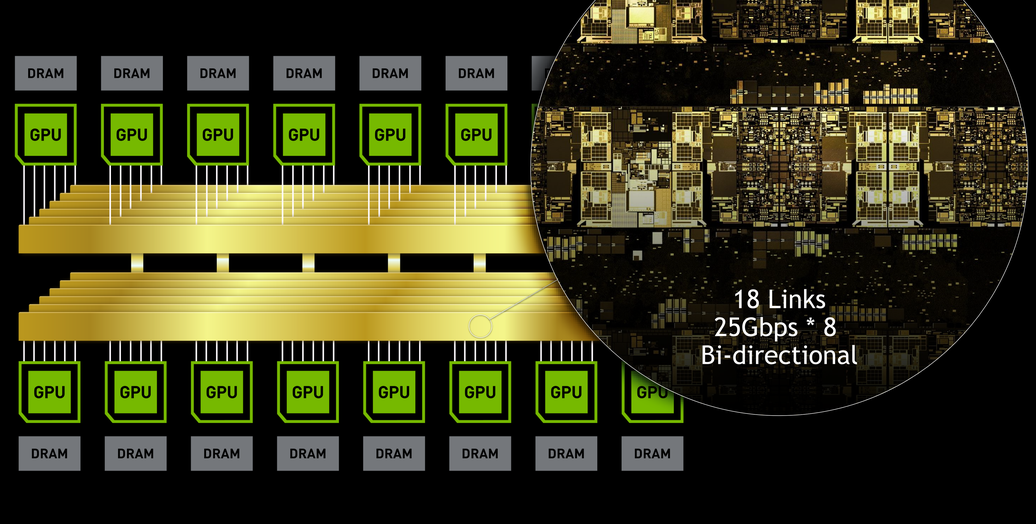
“THE WORLD’S LARGEST GPU”

16 Tesla V100 32GB Connected by NVSwitch
On-chip Memory Fabric Semantic Extended Across All GPUs
512GB HBM2 and 14.4TB/sec Aggregate
81,920 CUDA Cores
2,000 TFLOPS Tensor Cores



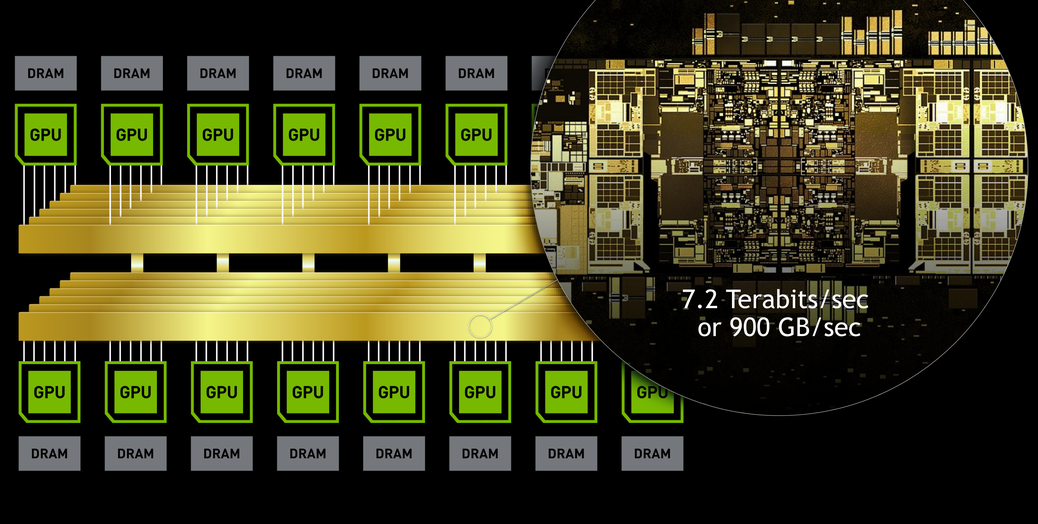
“THE WORLD’S LARGEST GPU”

16 Tesla V100 32GB Connected by NVSwitch
On-chip Memory Fabric Semantic Extended Across All GPUs
512GB HBM2 and 14.4TB/sec Aggregate
81,920 CUDA Cores
2,000 TFLOPS Tensor Cores



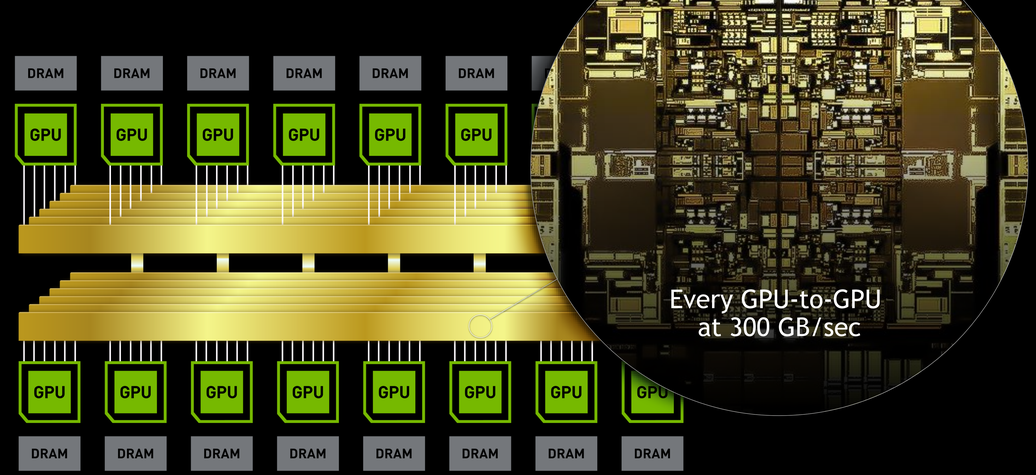
“THE WORLD’S LARGEST GPU”

16 Tesla V100 32GB Connected by NVSwitch
On-chip Memory Fabric Semantic Extended Across All GPUs
512GB HBM2 and 14.4TB/sec Aggregate
81,920 CUDA Cores
2,000 TFLOPS Tensor Cores



“THE WORLD’S LARGEST GPU”

16 Tesla V100 32GB Connected by NVSwitch
On-chip Memory Fabric Semantic Extended Across All GPUs
512GB HBM2 and 14.4TB/sec Aggregate
81,920 CUDA Cores
2,000 TFLOPS Tensor Cores



ANNOUNCING NVIDIA DGX-2

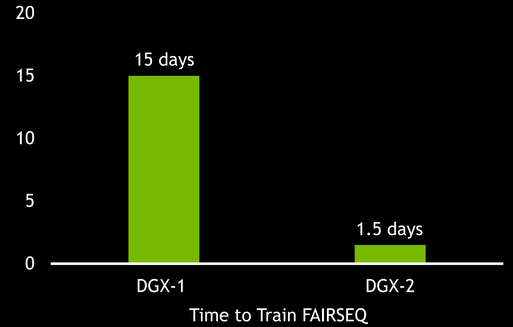
THE LARGEST GPU EVER CREATED

2 PFLOPS
512GB HBM2
10 kW
350 lbs

10X IN 6 MONTHS

DGX-1 V100 16GB — SEPT ‘17

Framework	pyTorch	0.2
	TensorFlow	1.3
	MXNet	0.11
	Caffe2	0.8.1
	CNTK	2.0
	Python	2.7
System Software Stack	NCCL	2.0.2
	cuDNN	7.0.2
	cuBLAS	9.0
	cuFFT	9.0
	NPP	9.0
	CUDA	9.0
	Res Mgr	R384
	BaseOS	2.0



DGX-2 V100 32GB — MAR ‘18

Framework	pyTorch	0.3
	TensorFlow	1.7
	MXNet	1.0
	Caffe2	0.8.1
	CNTK	2.3
	Python	2.7 or 3.6
System Software Stack	NCCL	2.2
	cuDNN	7.1
	cuBLAS	9.2
	cuFFT	9.2
	NPP	9.2
	CUDA	9.2
	Res Mgr	R396
	BaseOS	3.1.2

Fairseq is a neural machine translation network, published by Facebook in May'17
Fairseq is trained with WMT'14 English-French dataset in 55 epochs

ANNOUNCING NVIDIA DGX-2

$1.5M
Available in Q3

ANNOUNCING NVIDIA DGX-2

~~$1M~~ $399K
Available in Q3

TRADITIONAL HYPERSCALE CLUSTER

300 Dual-CPU Servers
$3M
180 kW

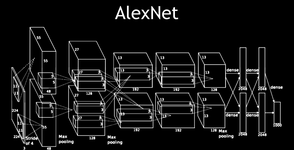
NVIDIA DGX-2
FOR DEEP LEARNING

1 DGX-2
Big Savings for Deep Learning
10 kW

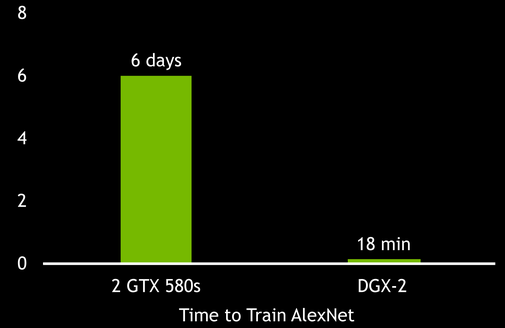
1/8 the Cost 1/60 the Space 1/18 the Power

“500X” IN 5 YEARS

2 GTX 580s — DEC ‘12



Framework	cuda-convnet	
System Software Stack	NCCL	N/A
	cuDNN	N/A
	cuBLAS	5.0
	cuFFT	5.0
	NPP	5.0
	CUDA	5.0
	Res Mgr	R304



DGX-2 — MAR ‘18

Framework	NV Caffe	0.17
System Software Stack	NCCL	2.2
	cuDNN	7.1
	cuBLAS	9.2
	cuFFT	9.2
	NPP	9.2
	CUDA	9.2
	Res Mgr	R396

NVIDIA GPU CLOUD

OPTIMIZED STACKS FOR EVERY CLOUD

20,000+ Registered Organizations
30 Containers

NOW on AWS, GCP, AliCloud,
Oracle Cloud, DGX

GPU CONTAINERS

DEEP LEARNING
Caffe
Caffe2
Chainer
PaddlePaddle
CNTK
CUDA
Digits
H2O.ai
MXNet
PyTorch
TensorFlow
TensorRT
Theano
Torch

HPC
GAMESS
Gromacs
LAMMPS
NAMD
RELION
CHROMA
MILC
CANDLE
Lattice Microbes

HPC Viz
ParaView Holodeck
ParaView IndeX
ParaView OptiX
IndeX
VMD

ANALYTICS
MapD
Kinetica

PLASTER

NVIDIA AI INFERENCE

ASR
RNN++

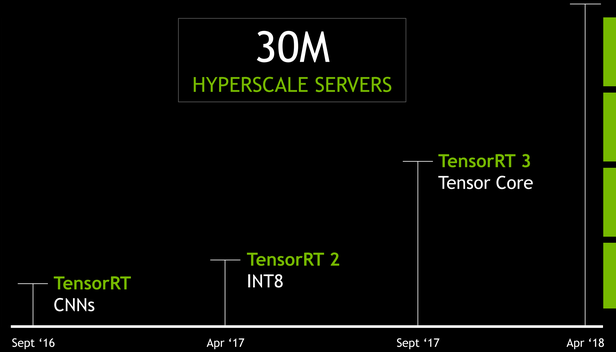
SPEECH SYNTH
DGN, S2S

RECOMMENDER
MLP-NCF

NLP
RNN

IMAGE / VIDEO
CNN

30M
HYPERSCALE SERVERS



TensorRT 4

TensorFlow Integration

Kaldi Optimization

ONNX

WinML

190X
IMAGE / VIDEO
ResNet-50 with TensorFlow Integration

50X
NLP
GNMT

45X
RECOMMENDER
Neural Collaborative Filtering

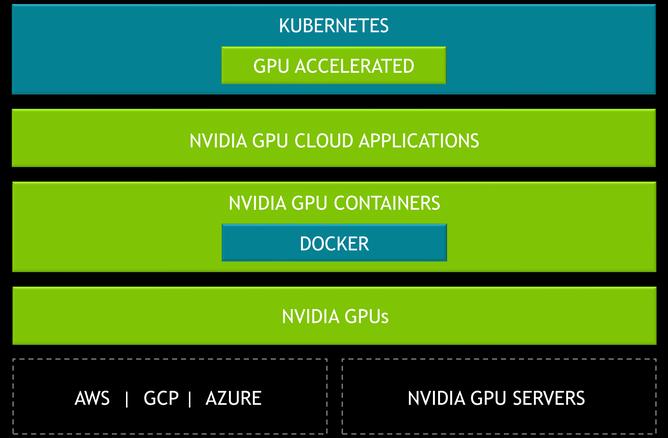
36X
SPEECH SYNTH
WaveNet

60X
ASR
DeepSpeech 2 DNN

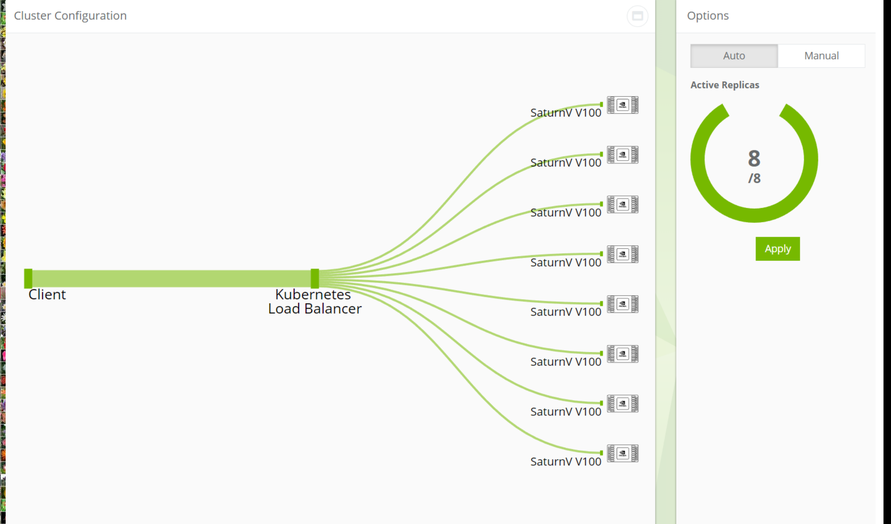
All speed-ups are chip-to-chip CPU to GV100.

ANNOUNCING KUBERNETES ON NVIDIA GPUs

Scale-up Thousands of GPUs Instantly
Multi-region, Self-healing Cluster Orchestration
GPU Optimized Out-of-the-Box



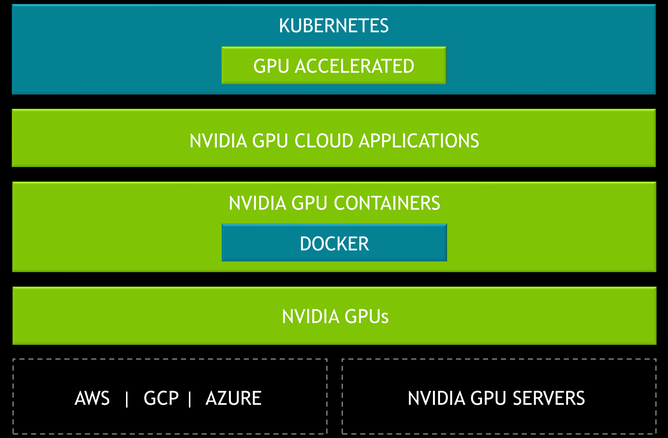

Kubernetes on NVIDIA GPUs
Scale Out Infrastrucure for the Accelerated Datacenter
Images Per Sec: 5269


Cluster Configuration
SaturnV V100
SaturnV V100
SaturnV V100
SaturnV V100
SaturnV V100
SaturnV V100
SaturnV V100
SaturnV V100
Client
Kubernetes
Load Balancer

Options
Auto
Manual
Active Replicas
8
/8
Apply

ANNOUNCING KUBERNETES ON NVIDIA GPUs

Scale-up Thousands of GPUs Instantly
Multi-region, Self-healing Cluster Orchestration
GPU Optimized Out-of-the-Box



NVIDIA AI INFERENCE

CSPs | VIDEO ANALYTICS | SPEECH | RECOMMENDATION SYSTEMS | MAPPING
AUTOMOTIVE | ROBOTICS | SMART CITIES | ETAIL | HEALTHCARE | MANUFACTURING

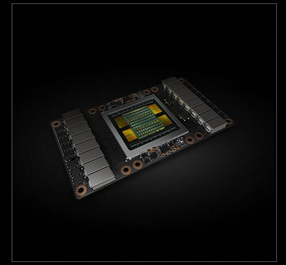
"NVIDIA's inference platform made it possible to derive real-time understanding of live videos."

—Nicolas Koumchatzky, Head of Cortex, Twitter

"We believe TensorRT could dramatically improve productivity for our enterprise customers."

—Markus Noga, Head of Machine Learning, SAP

NVIDIA AI PLATFORM

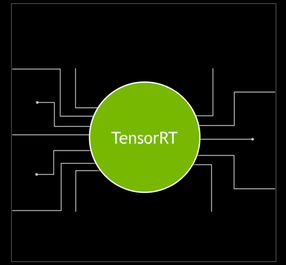


Tesla V100
NEW 32GB

DGX Systems
NEW with V100 32GB
NEW DGX-2

Every Cloud
NGC Now on AWS, GCP, AliCloud, Oracle

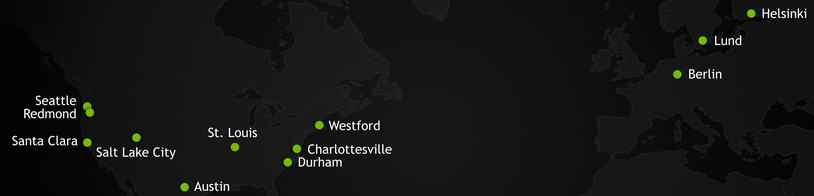
NVIDIA GPU Cloud
30 GPU-Optimized Containers



NVIDIA AI Inference
NEW TensorRT 4, TensorFlow
Kaldi, ONNX, WinML

TITAN V
Out of stock!

NVIDIA RESEARCH

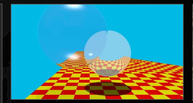


Bill Dally
NVIDIA Chief Scientist

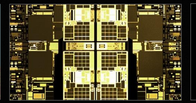
200 RESEARCHERS

- Graphics
- Deep Learning
- Robotics
- Computer Vision
- Parallel Architectures
- Programming Systems
- Circuits
- VLSI
- Networks

RECENT WORKS



RTX



NVSwitch

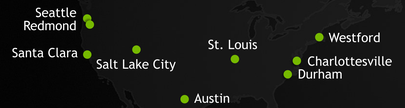
CuDNN

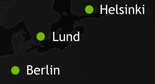
CNN Image Inpainting

Noise-to-Noise Denoising

Progressive GAN

NVIDIA RESEARCH



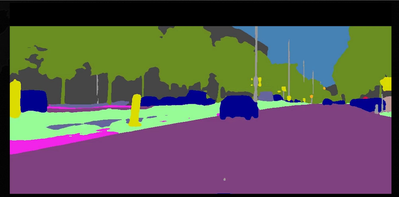


Bill Dally
NVIDIA Chief Scientist

200 RESEARCHERS

- Graphics
- Deep Learning
- Robotics
- Computer Vision
- Parallel Architectures
- Programming Systems
- Circuits
- VLSI
- Networks

CONDITIONAL GANs



EVERYTHING THAT MOVES WILL BE AUTONOMOUS

Cars

Robotaxis

Trucks

Delivery Vans

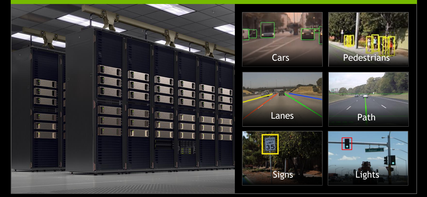
Buses

Tractors

NVIDIA DRIVE END-TO-END PLATFORM

COLLECT DATA

TRAIN MODELS



SIMULATE

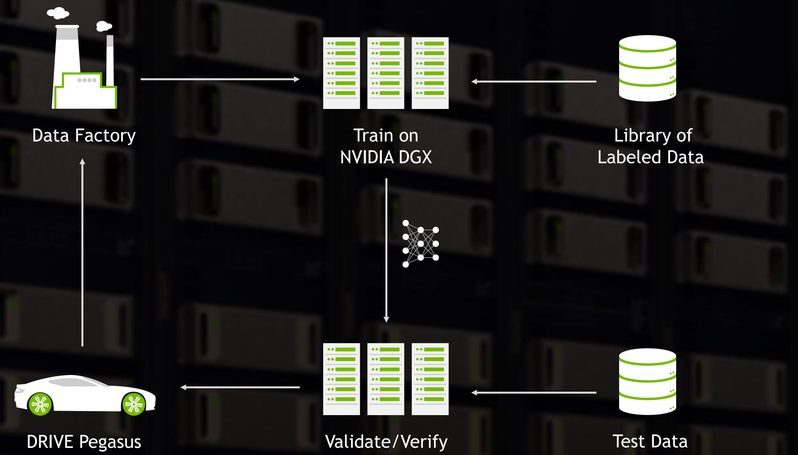
DRIVE

NVIDIA PERCEPTION INFRASTRUCTURE

LARGE-SCALE DEEP LEARNING MODEL DEVELOPMENT

Workflow, Tools, Supercomputing Infrastructure
Data Ingest, Labeling, Training, Validation, Adaptation
Automation, Best Model Discovery, Traceability, Reproducibility
Purpose-built for Safety Standards of Automotive

"Data is the new source code"

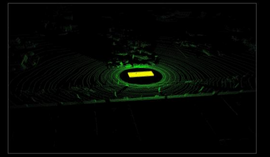


Perception

Free Space Perception

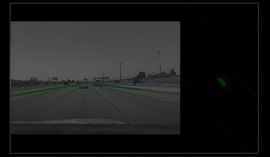
Distance Perception

Weather

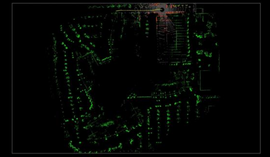


LIDAR Perception

Camera-based Mapping



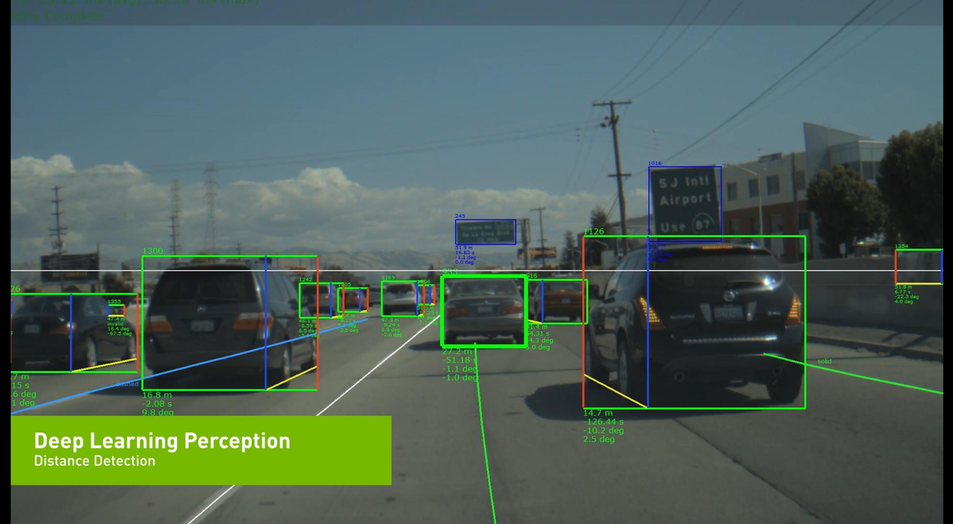
Camera Localization to HD Map



LIDAR Localization to HD Map

Path Perception

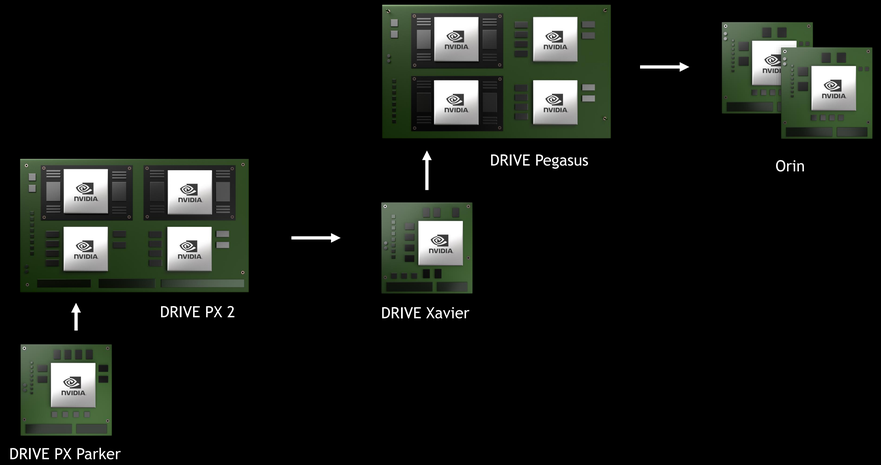
Scene Perception


S J Intl
Airport
Use 87
Deep Learning Perception
Distance Detection

NVIDIA DRIVE ROADMAP

ONE ARCHITECTURE

Auto-Grade
Super Energy-Efficient
ASIL-D Functional Safety



SIMULATION — THE PATH TO BILLIONS OF MILES

World drives trillions of miles each year.

U.S. has 770 accidents per billion miles.

A fleet of 20 test cars cover 1 million miles per year.

ANNOUNCING NVIDIA DRIVE SIM AND CONSTELLATION

AV VALIDATION SYSTEM

Virtual Reality AV Simulator

Same Architecture as DRIVE Computer

Simulate Rare and Difficult Conditions, Recreate Scenarios, Run Regression Tests, Drive Billions of Virtual Miles

10,000 Constellations Drive 3B Miles per Year

ANNOUNCING NVIDIA DRIVE SIM AND CONSTELLATION

AV VALIDATION SYSTEM

Virtual Reality AV Simulator

Same Architecture as DRIVE Computer

Simulate Rare and Difficult Conditions, Recreate Scenarios, Run Regression Tests, Drive Billions of Virtual Miles

10,000 Constellations Drive 3B Miles per Year

ANNOUNCING NVIDIA DRIVE SIM AND CONSTELLATION

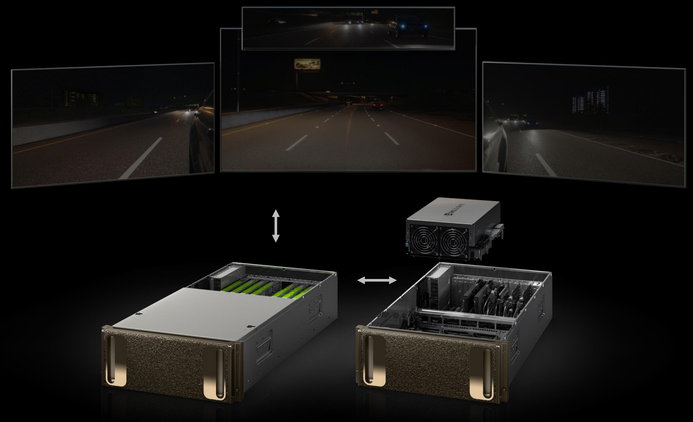
AV VALIDATION SYSTEM

Virtual Reality AV Simulator

Same Architecture as DRIVE Computer

Simulate Rare and Difficult Conditions, Recreate Scenarios, Run Regression Tests, Drive Billions of Virtual Miles

10,000 Constellations Drive 3B Miles per Year



ANNOUNCING NVIDIA DRIVE SIM AND CONSTELLATION

AV VALIDATION SYSTEM

Virtual Reality AV Simulator

Same Architecture as DRIVE Computer

Simulate Rare and Difficult Conditions, Recreate Scenarios, Run Regression Tests, Drive Billions of Virtual Miles

10,000 Constellations Drive 3B Miles per Year

ANNOUNCING NVIDIA DRIVE SIM AND CONSTELLATION

AV VALIDATION SYSTEM

Virtual Reality AV Simulator

Same Architecture as DRIVE Computer

Simulate Rare and Difficult Conditions, Recreate Scenarios, Run Regression Tests, Drive Billions of Virtual Miles

10,000 Constellations Drive 3B Miles per Year

370 PARTNERS DEVELOPING ON NVIDIA DRIVE

CARS | TRUCKS | MOBILITY SERVICES | SUPPLIERS | MAPPING | LIDAR | CAMERA/RADAR | STARTUPS

ROBOTICS BOOSTS EVERY INDUSTRY

Delivery

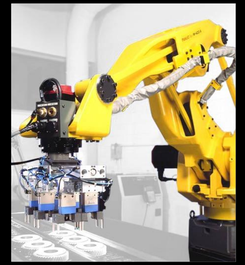
Consumer

Healthcare

Agriculture

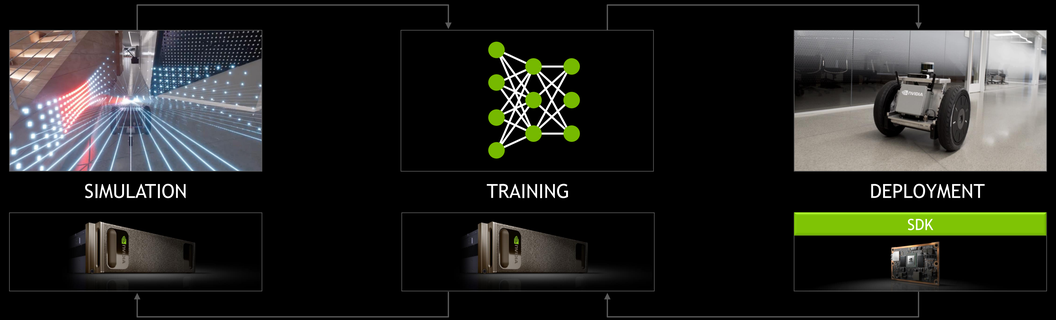
Retail

Logistics

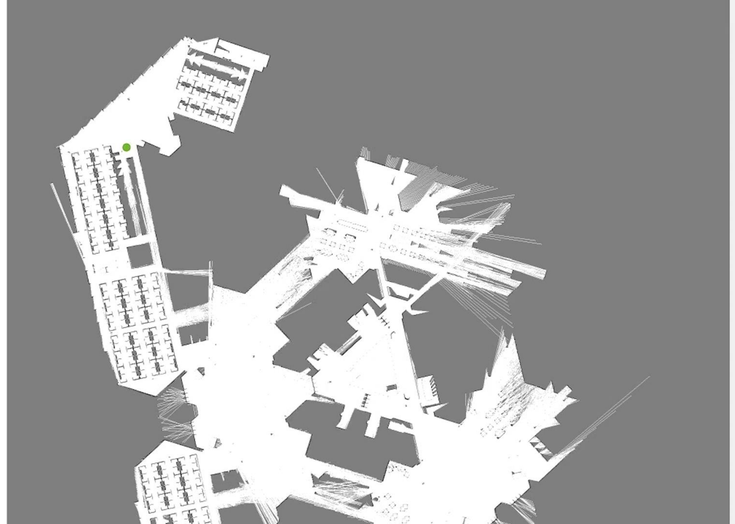


Manufacturing

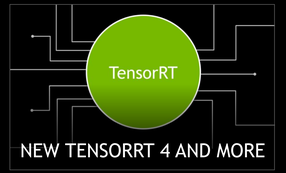
NVIDIA ISAAC ROBOTICS PLATFORM



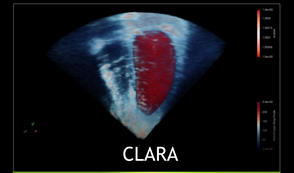
https://developer.nvidia.com/isaac-sdk



THE GPU COMPUTING REVOLUTION CONTINUES



Kubernetes
On
NVIDIA
GPUs



GRAPHICS

AI

AUTO

NEW PLATFORMS


NVIDIA