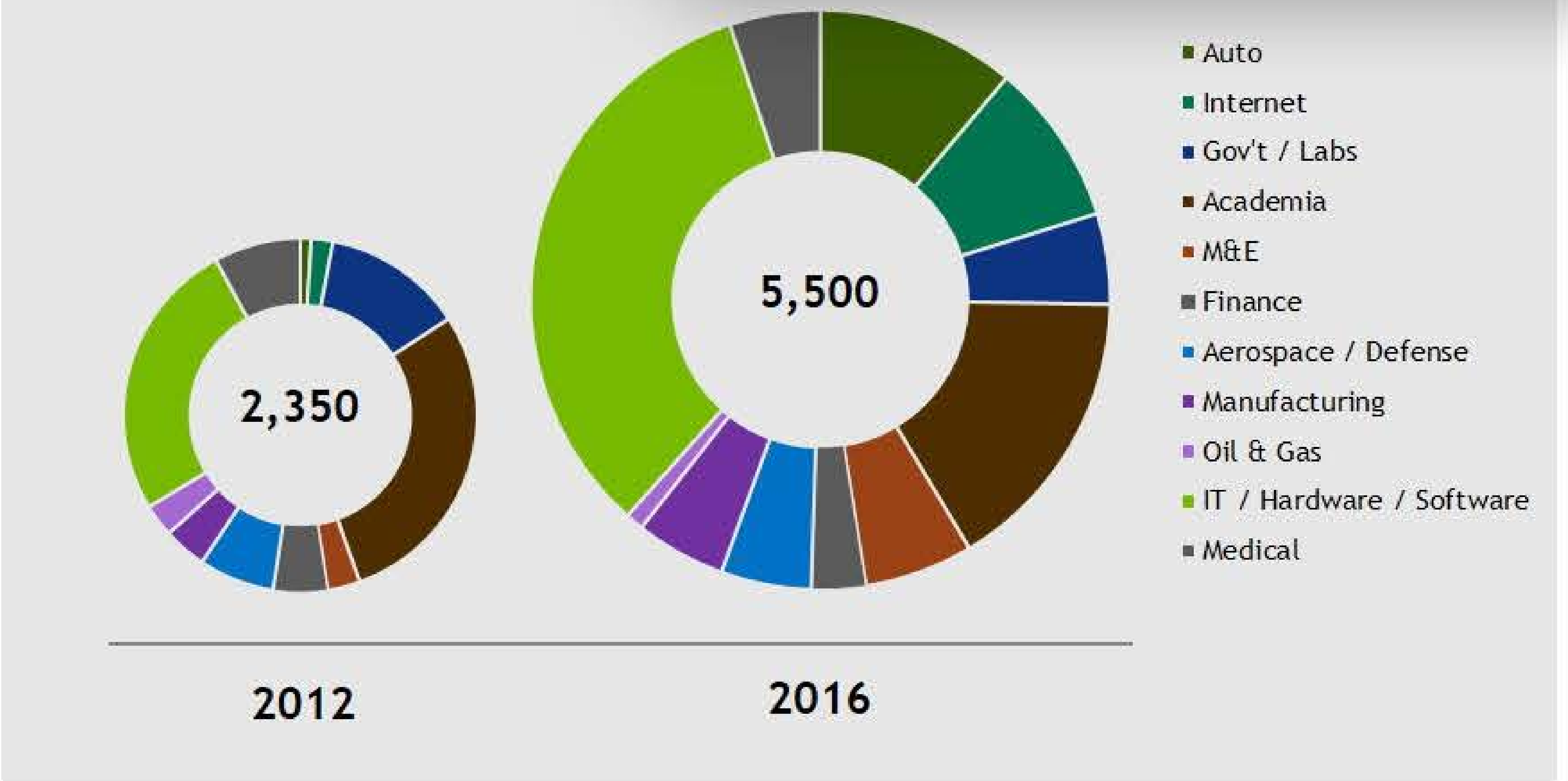
GPU TECHNOLOGY CONFERENCE

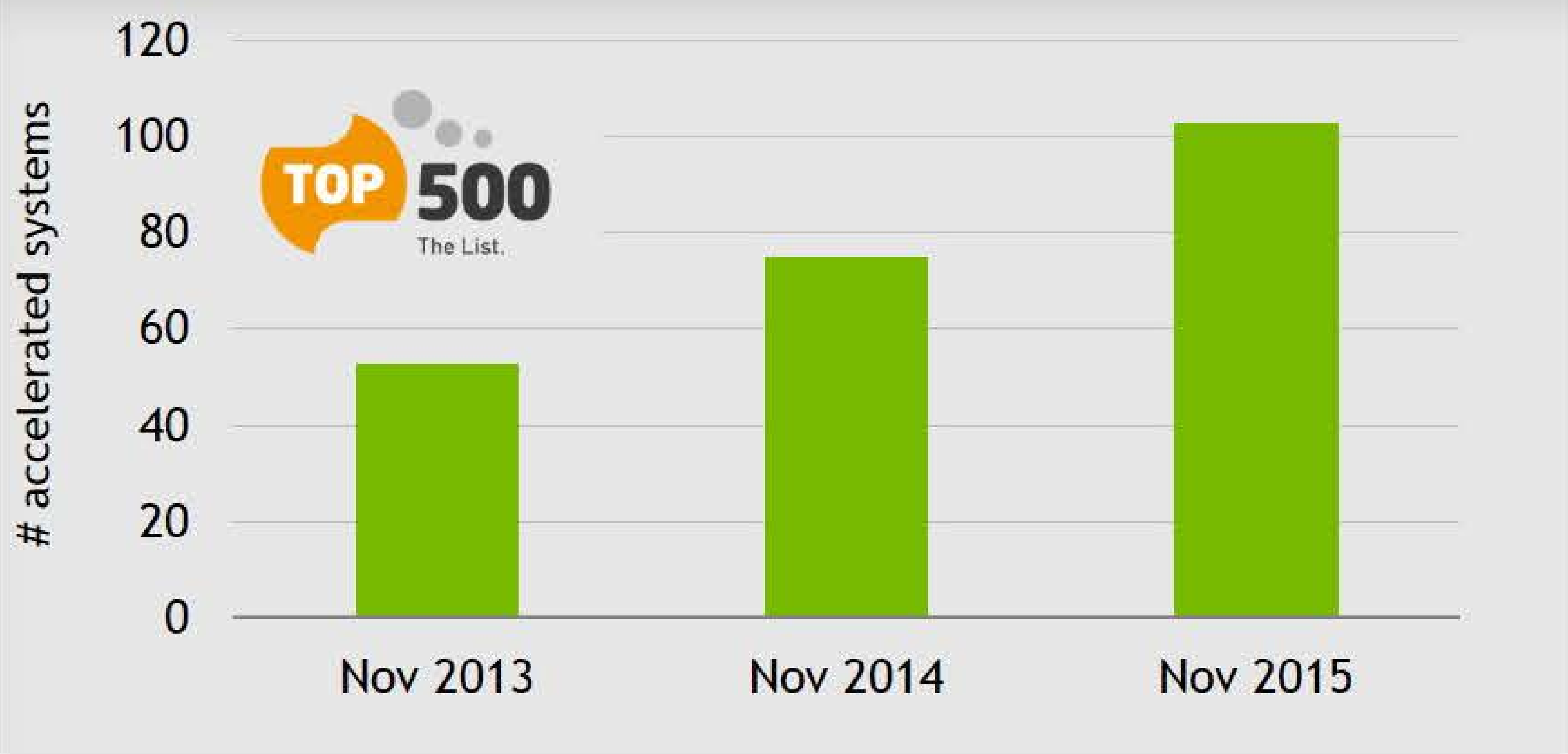
A NEW COMPUTING MODEL

JEN-HSUN HUANG, CO-FOUNDER & CEO | GTC 2016

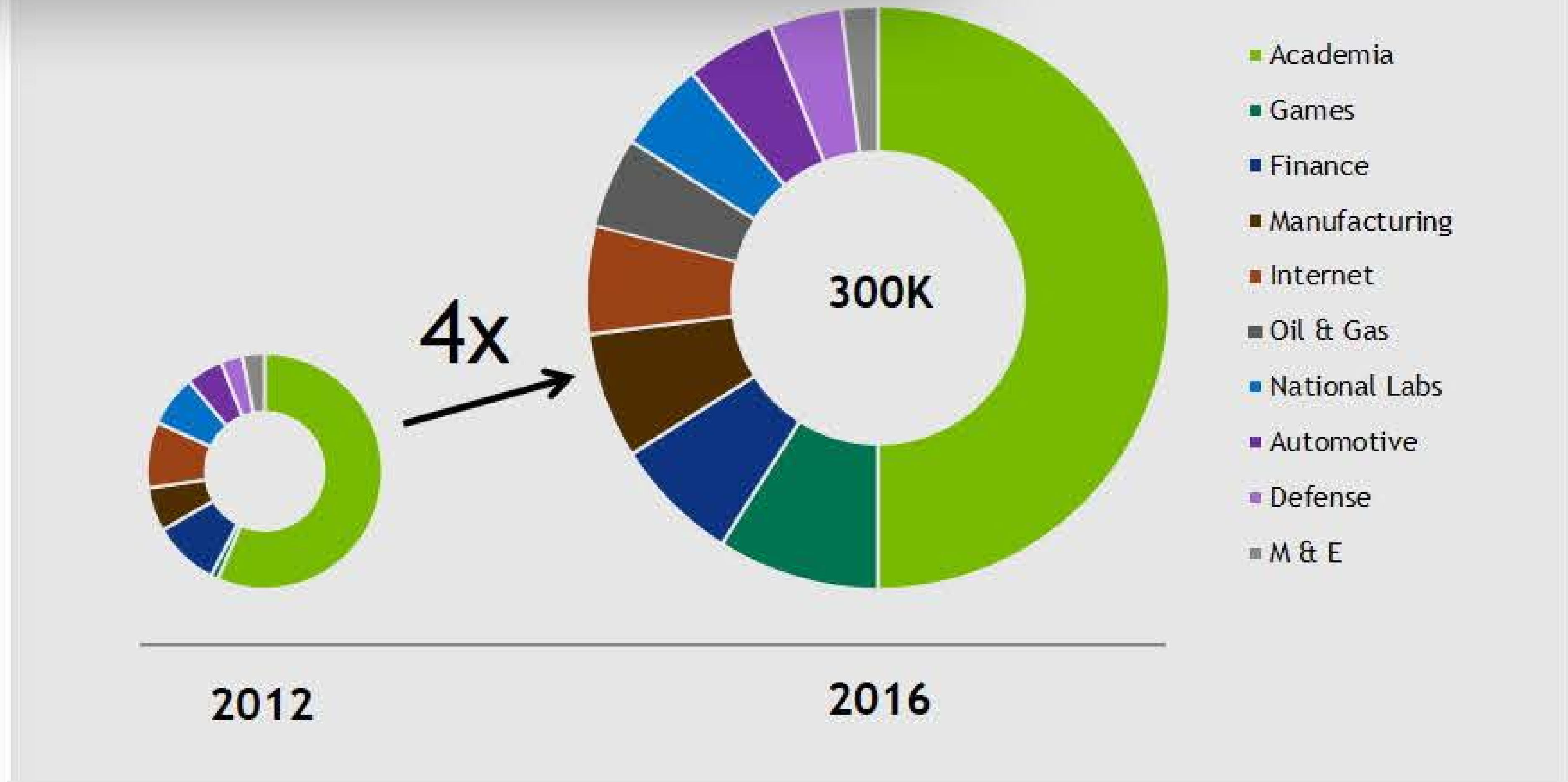
LEAPS IN ADOPTION



2X GTC Attendees



2X Accelerated Systems,
96% of New Systems on NVIDIA



4X CUDA Developers,
10X in Hyperscale + Auto

5 THINGS

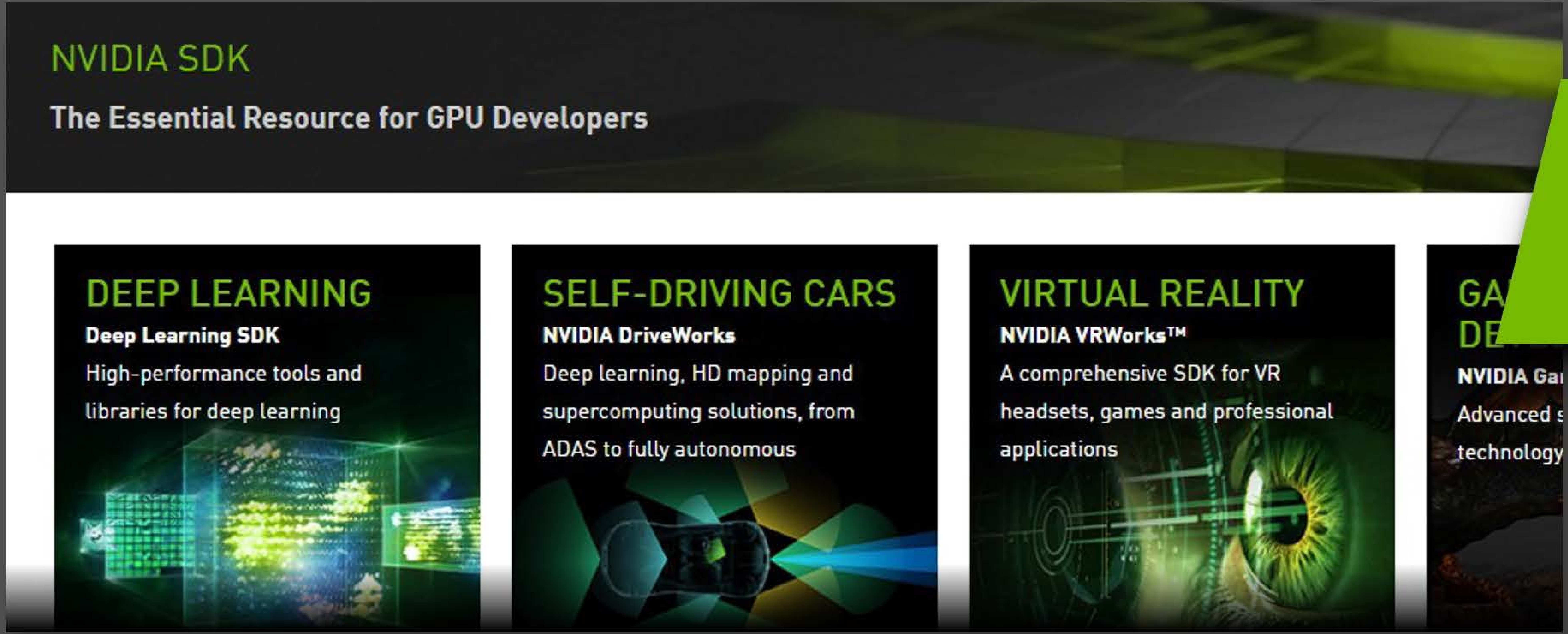
A Toolbox

A Deep Learning Chip

A Deep Learning Box

VR

A Deep Learning Car



NVIDIA SDK

The Essential Resource for GPU Developers

DEEP LEARNING

Deep Learning SDK

High-performance tools and libraries for deep learning

SELF-DRIVING CARS

NVIDIA DriveWorks

Deep learning, HD mapping and supercomputing solutions, from ADAS to fully autonomous

VIRTUAL REALITY

NVIDIA VRWorks™

A comprehensive SDK for VR headsets, games and professional applications

GA

DE

NVIDIA Ga

Advanced s

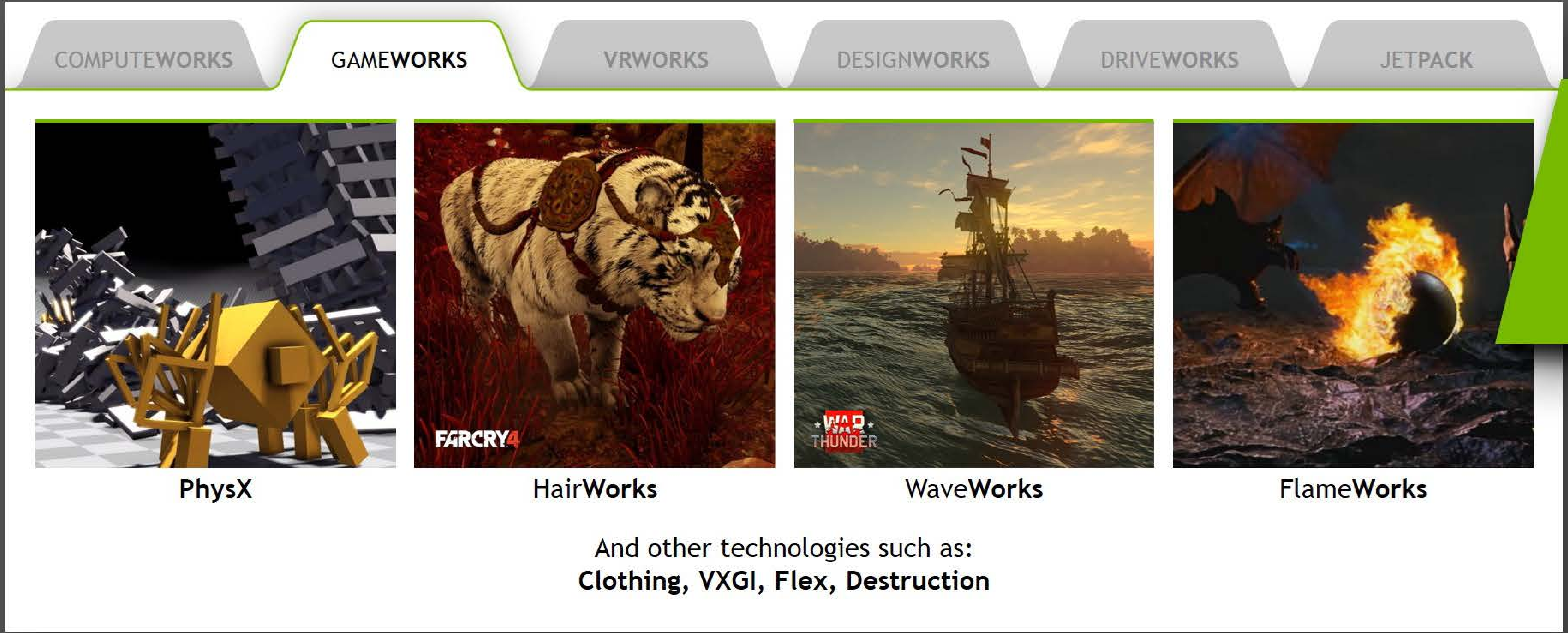
technology

NVIDIA SDK

THE ESSENTIAL RESOURCE FOR GPU COMPUTING

developer.nvidia.com

Available Now



PhysX

HairWorks

WaveWorks

FlameWorks

And other technologies such as:
Clothing, VXGI, Flex, Destruction

NVIDIA GAMEWORKS

Volumetric Lighting

Voxel Accelerated Ambient Occlusion

Hybrid Frustum Traced Shadows

Available Now

COMPUTEWORKS GAMEWORKS VRWORKS DESIGNWORKS DRIVEWORKS JETPACK

Iray

MDL

OptiX

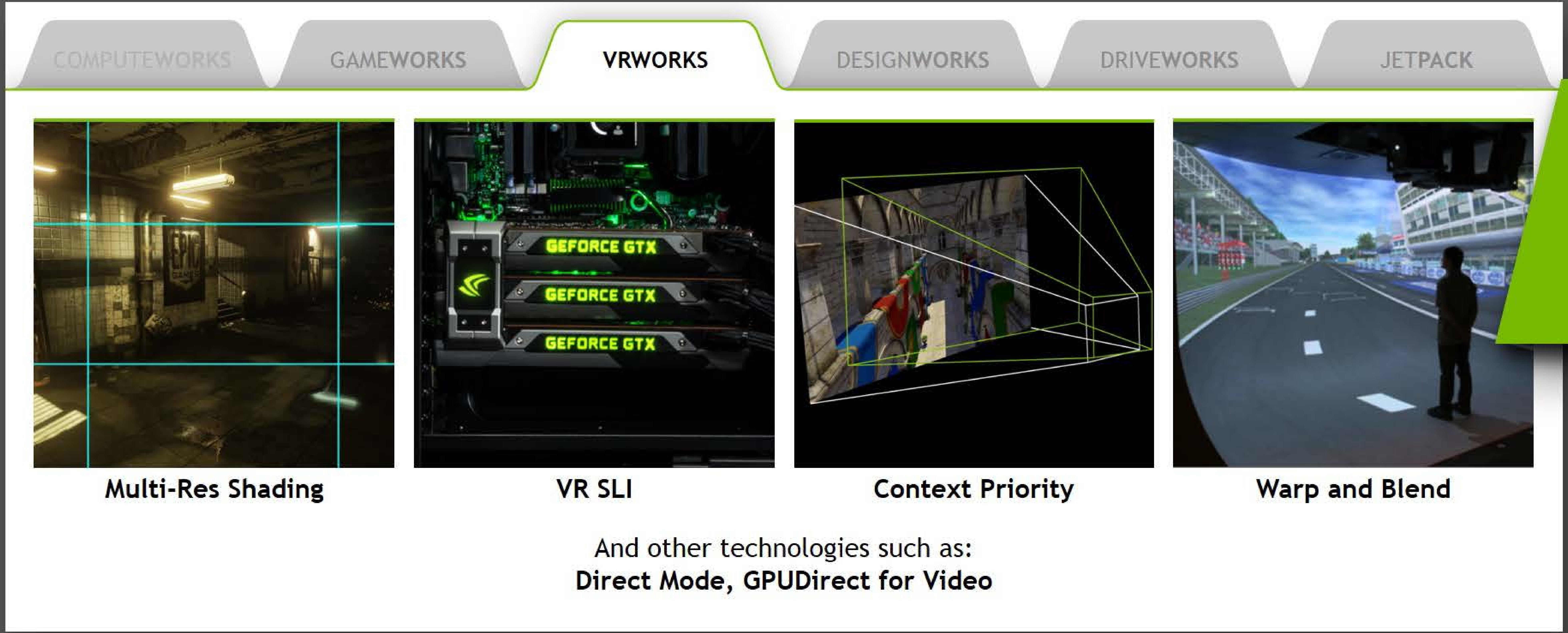
Path Rendering

and other technologies such as:
GL Extensions, GRID, GPU Direct for Video, Mosaic, VXGI, Warp and Blend

NVIDIA DESIGNWORKS

Adobe support of MDL

Siemens NX adopts Iray



Multi-Res Shading

VR SLI

Context Priority

Warp and Blend

And other technologies such as:
Direct Mode, GPUDirect for Video

NVIDIA VRWORKS

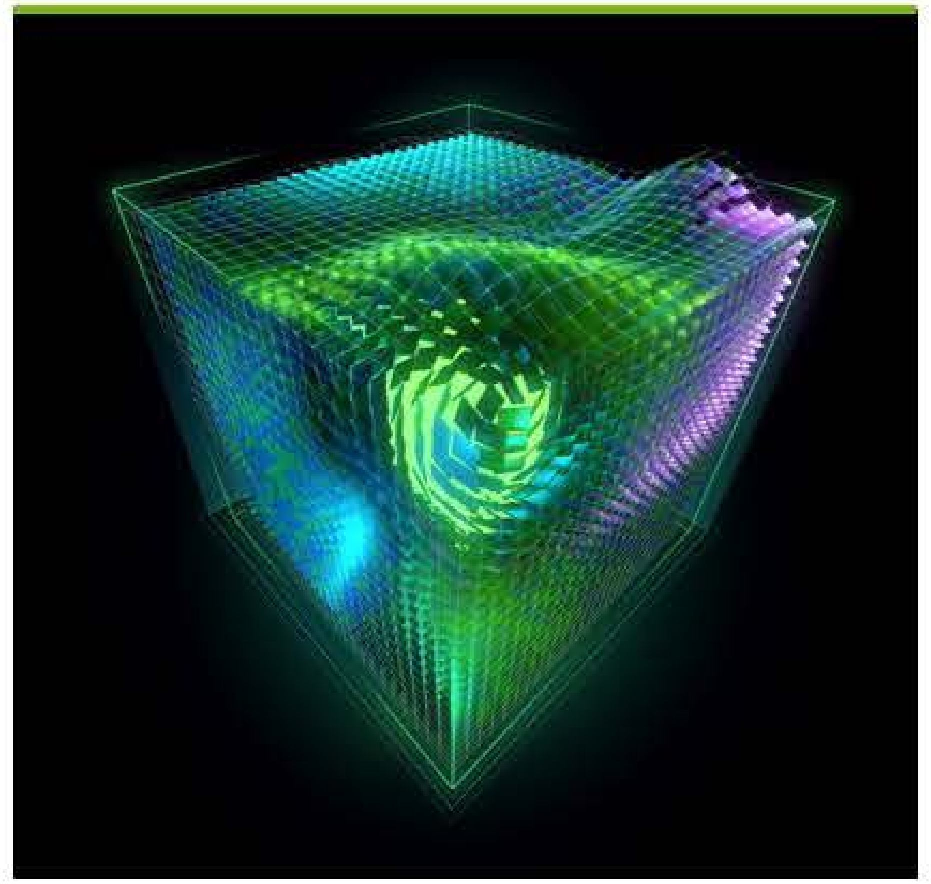
Oculus Rift and HTC Vive integration

Epic, Max Play and Unity game engines

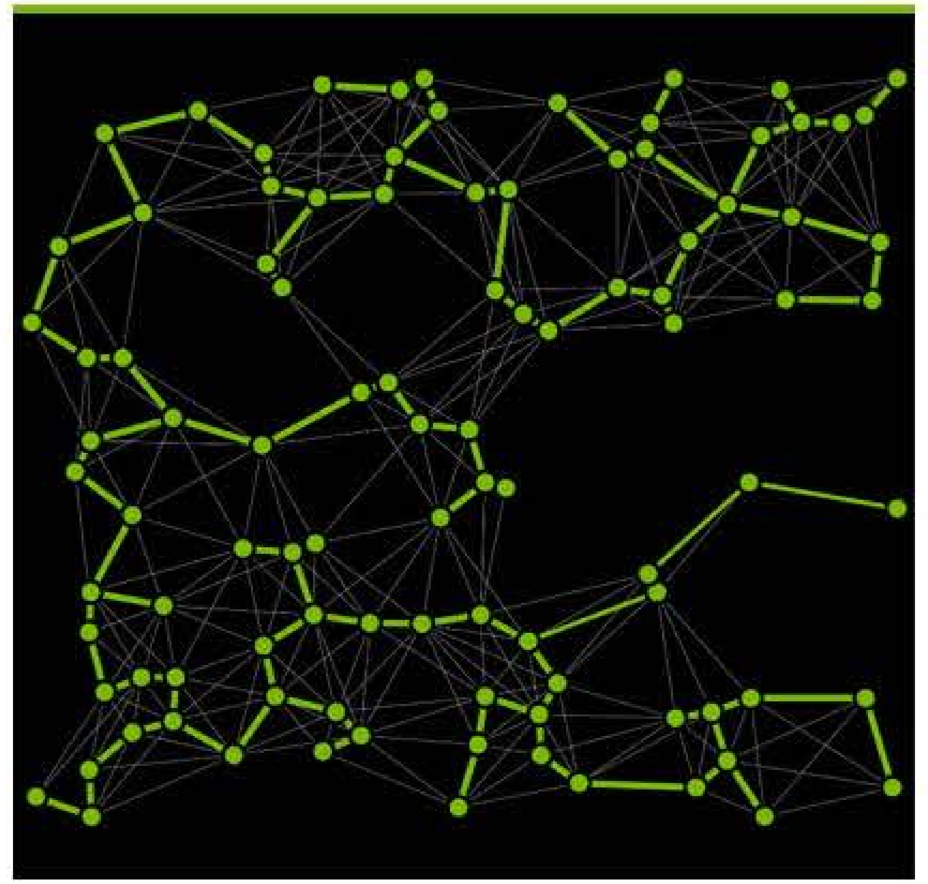
Available Now

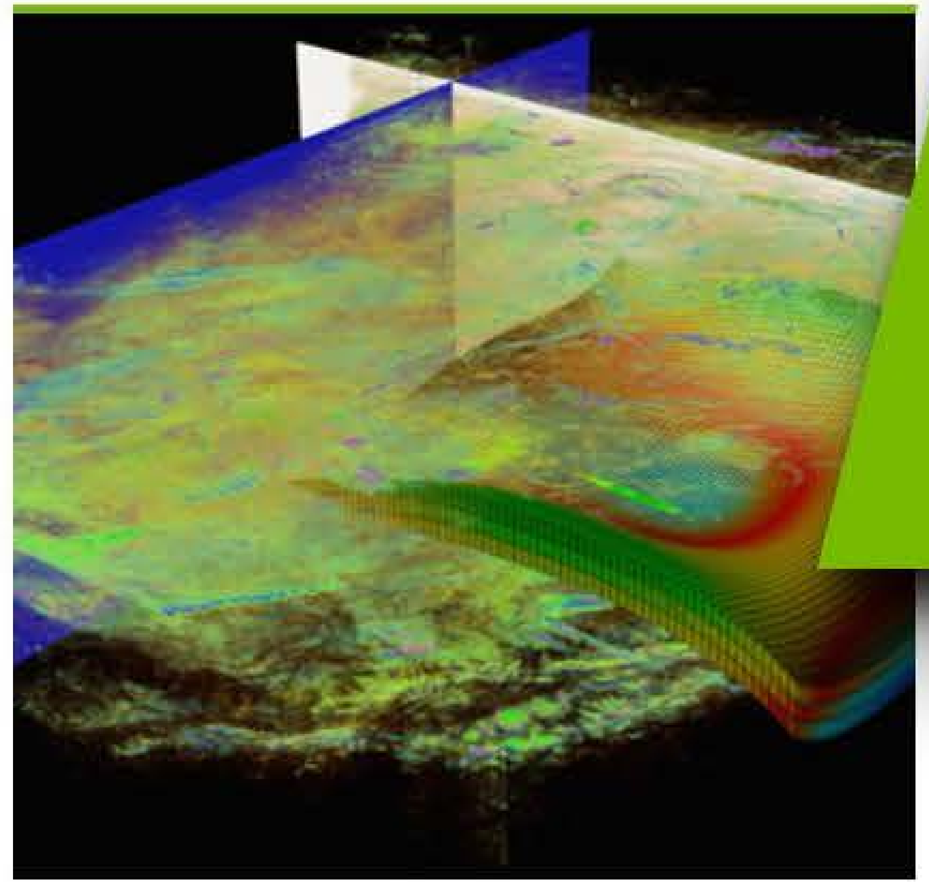
NVIDIA COMPUTEWORKS

COMPUTEWORKS | GAMEWORKS | VRWORKS | DESIGNWORKS | DRIVEWORKS | JETPACK


CUDA

cuDNN


nvGRAPH


IndeX

And other technologies such as:
AMGx, cuSOLVER, cuSPARSE, OpenACC, NSIGHT, THRUST

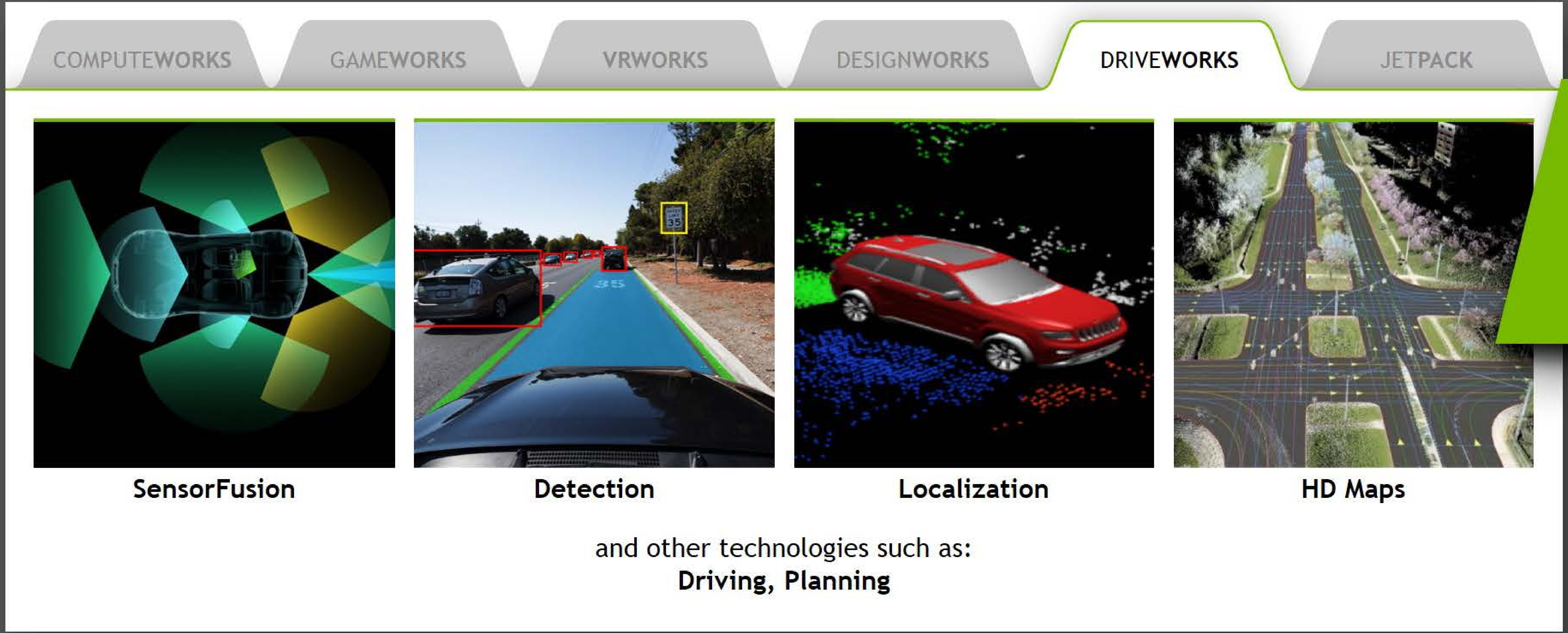
CUDA 8 — Available June

cuDNN 5 — Available April

nvGRAPH — Available June

IndeX plug-in for ParaView — Available May

NVIDIA DRIVEWORKS



SensorFusion

Detection

Localization

HD Maps

and other technologies such as:
Driving, Planning

UNDER CONSTRUCTION

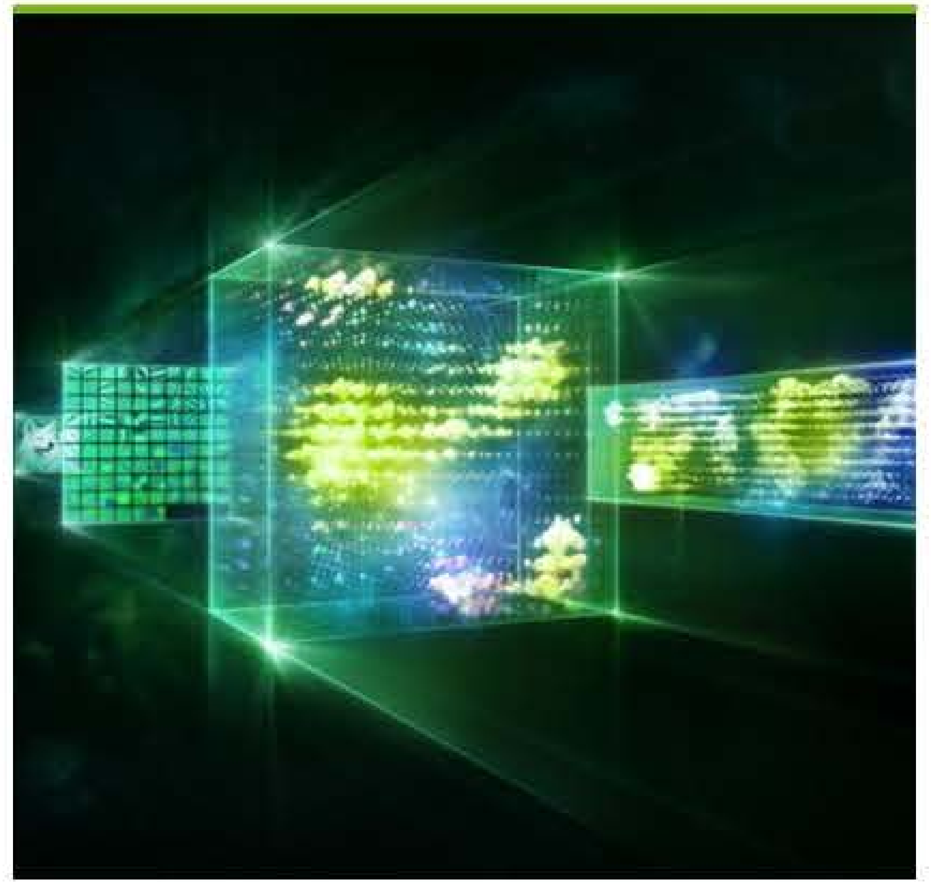
JPL — Available Now

EAP — Available Q2’16

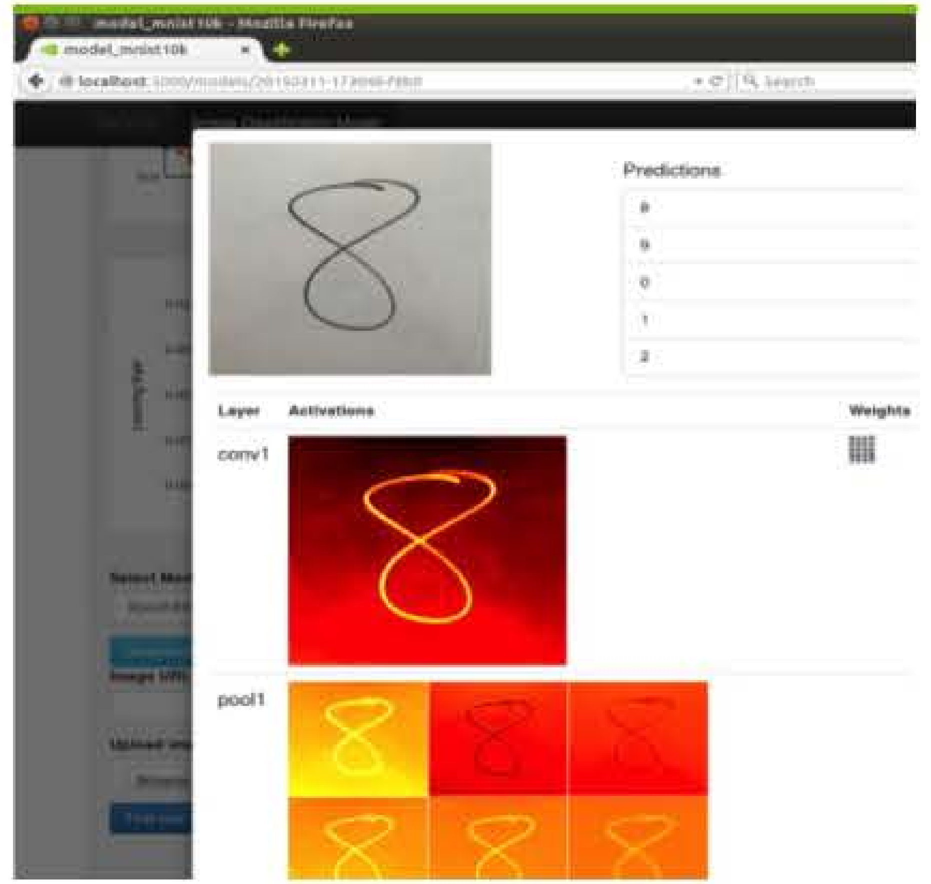
General release — Available Q1’17

NVIDIA JETPACK

COMPUTEWORKS GAMEWORKS VRWORKS DESIGNWORKS DRIVEWORKS JETPACK



Deep Learning SDK



DIGITS Workflow

VisionWorks

Jetson Media SDK

and other technologies such as:

Linux4Tegra, NSIGHT EE, OpenCV4Tegra, OpenGL, System Trace, Visual Profiler, Vulkan

GIE - GPU Inference Engine — Available May

Jetson TX1: 24 images/s/W

5 THINGS

NVIDIA DEVELOPER

NVIDIA SDK

The Essential Resource for GPU Developers

DEEP LEARNING
Deep Learning SDK
High-performance tools and libraries for deep learning

SELF-DRIVING CARS
NVIDIA DriveWorks
Deep learning, HD mapping and supercomputing solutions, from ADAS to fully autonomous

VIRTUAL REALITY
NVIDIA VRWorks™
A comprehensive SDK for VR headsets, games and professional applications

GAME DEVELOPME
NVIDIA GameWorks™
Advanced simulation technology for game

NVIDIA SDK

A Deep Learning Chip

A Deep Learning Box

VR

A Deep Learning Car

A START OF A NEW PLATFORM

Samsung, Oculus, HTC release headsets

Google announces Jump VR camera platform

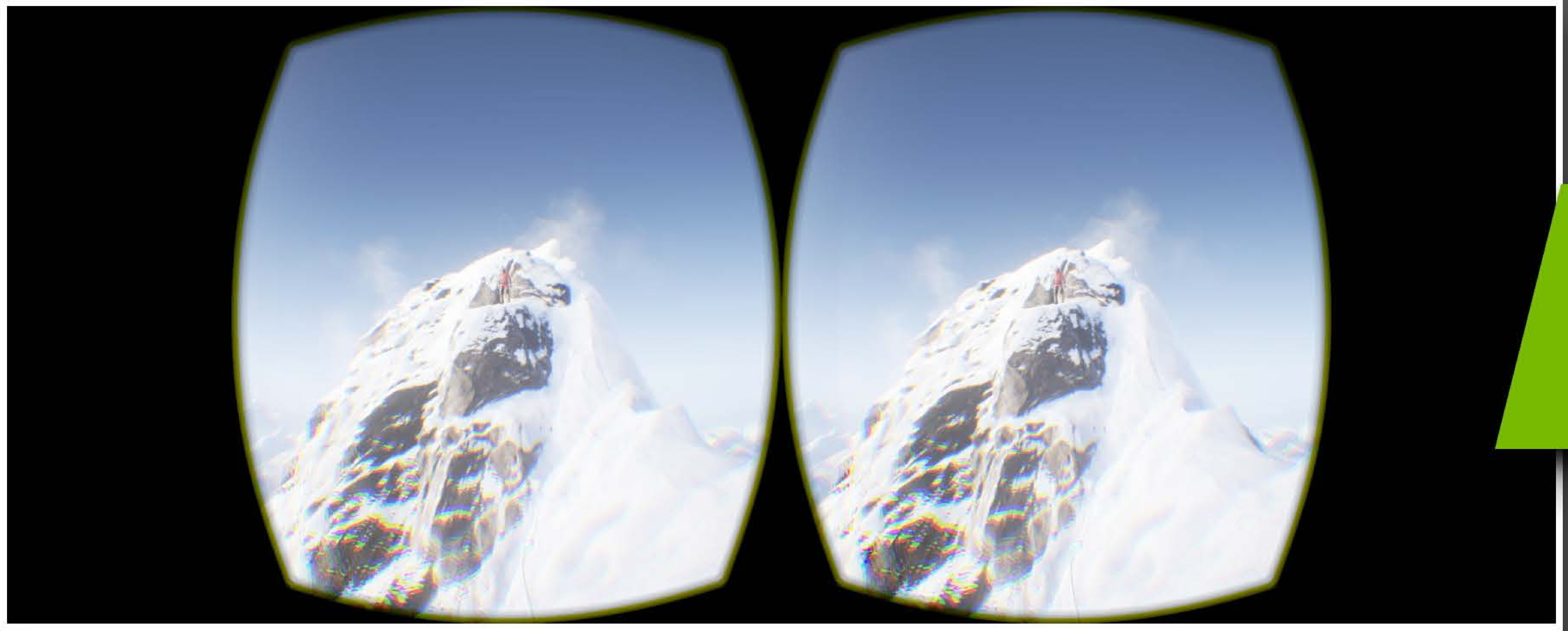
New York Times ships Cardboard to subscribers

Microsoft demonstrates Holoportation

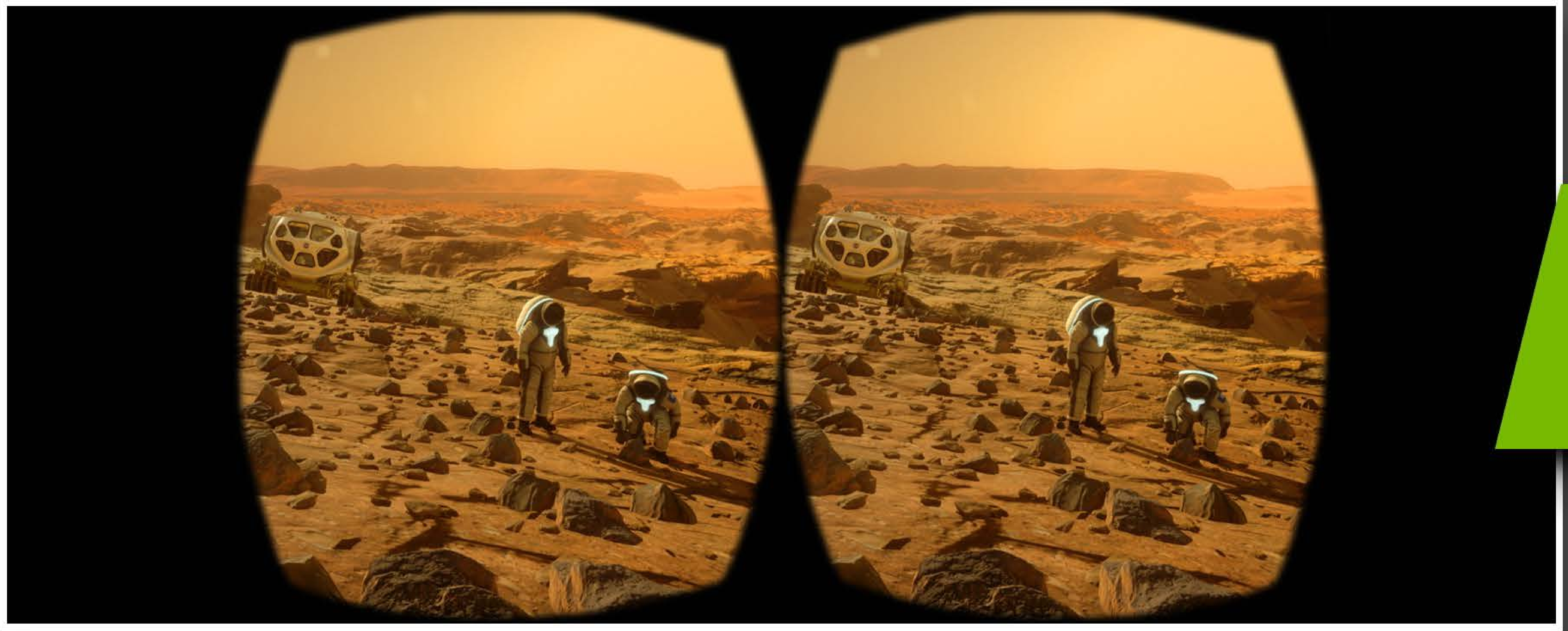
NEXTVR JAUNT

magic leap

VR Startups Raise $1.5B in funding



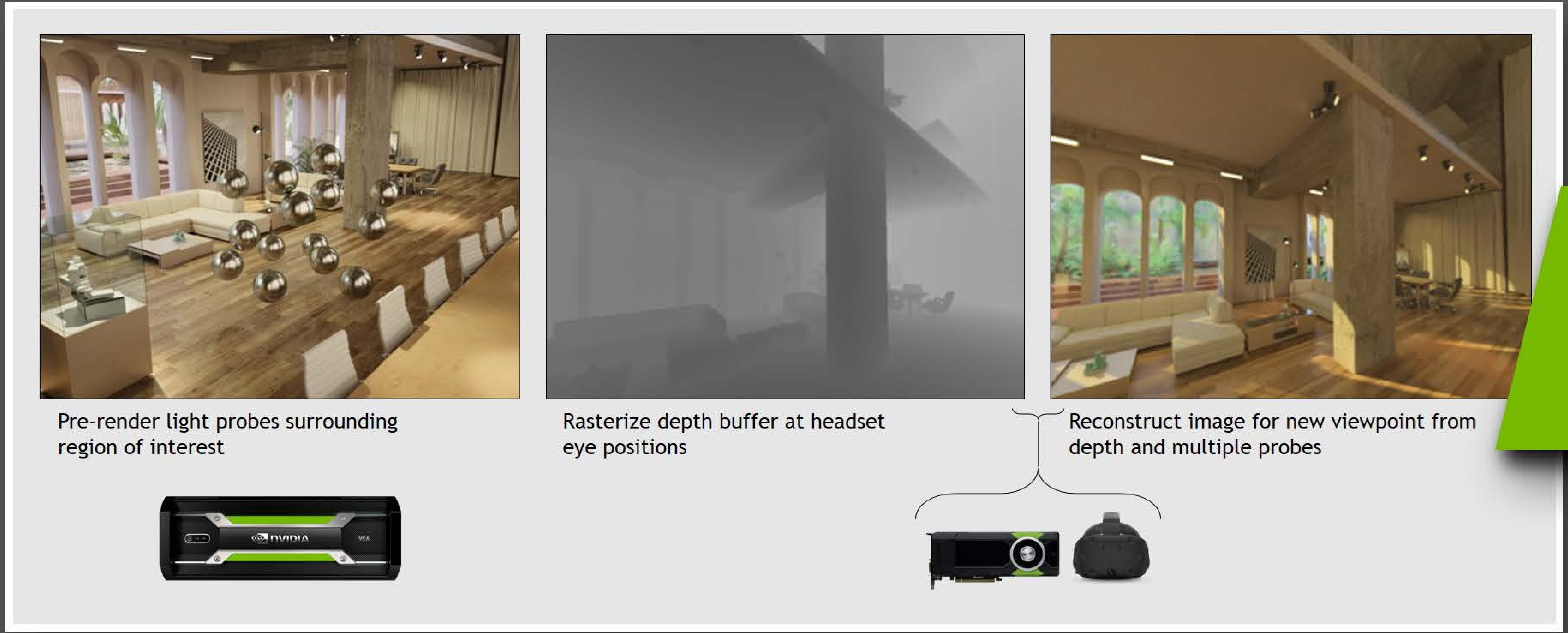
EVEREST VR



MARS 2030

IRAY VR

BREAKTHROUGH PHOTOREAL VR

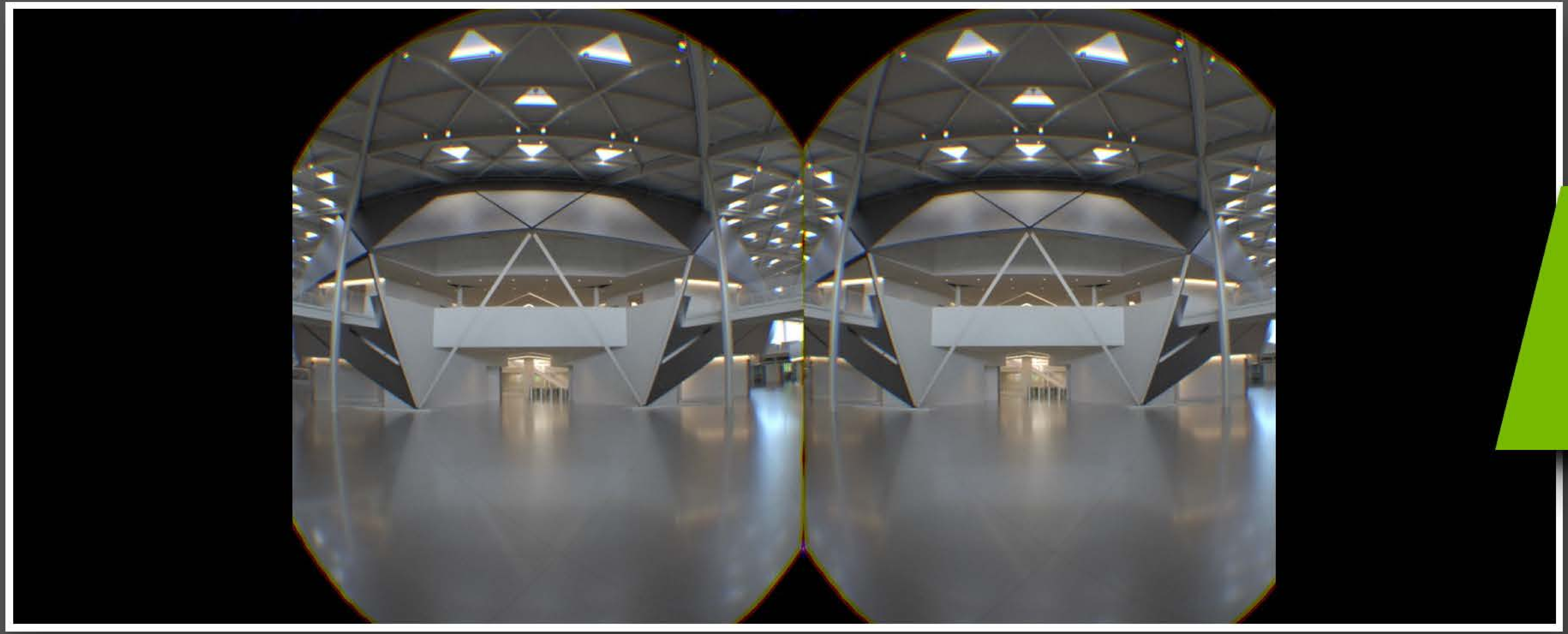


Pre-render light probes surrounding region of interest

Rasterize depth buffer at headset eye positions

Reconstruct image for new viewpoint from depth and multiple probes

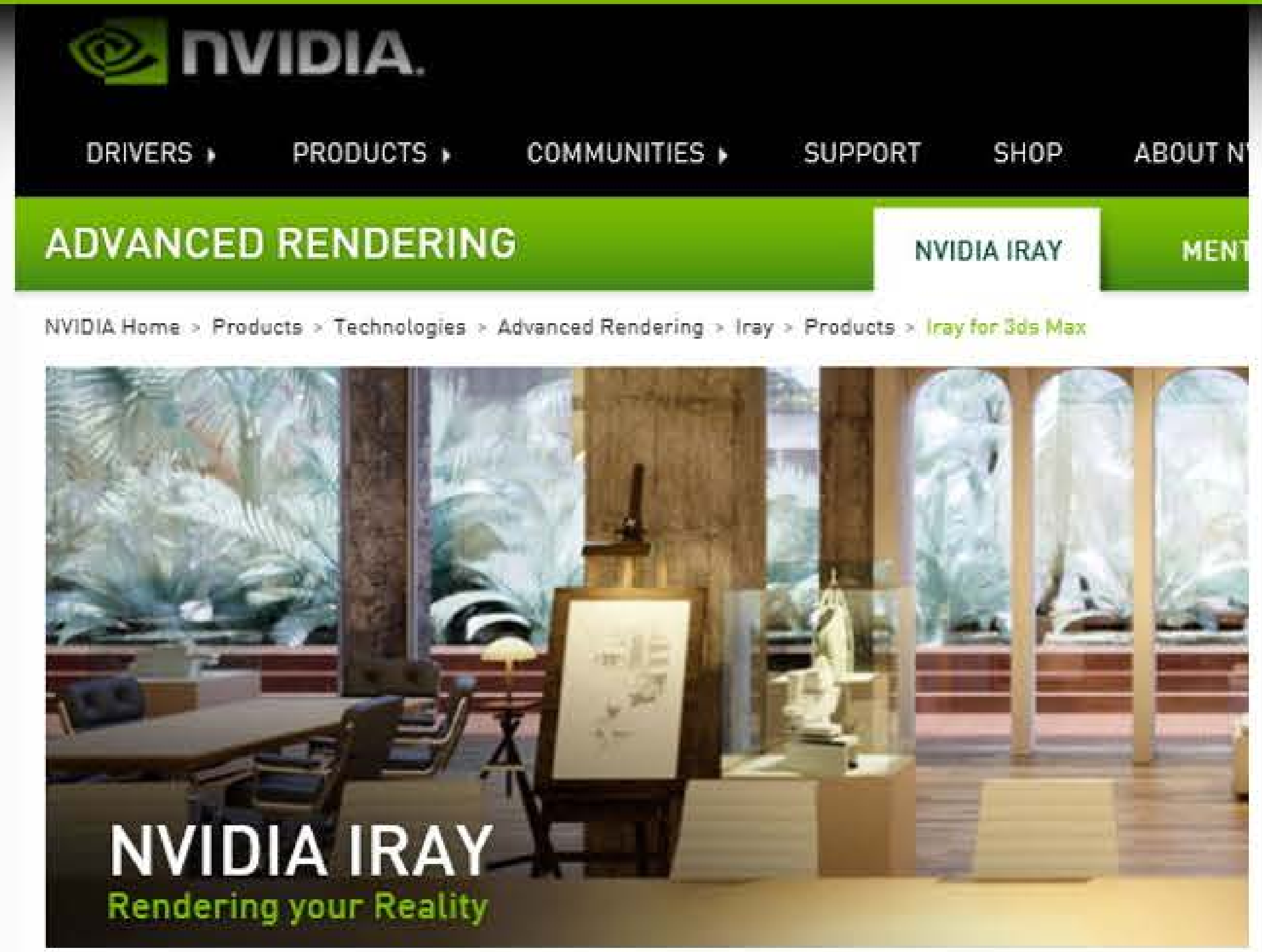
Available starting in June



IRAY VR

IRAY VR LITE

1. Design in 3ds Max



2. Download Iray for 3ds Max Plug-in

3. Download Android Viewer

4. Get VR HMD

Available in June

5 THINGS

NVIDIA DEVELOPER

NVIDIA SDK

The Essential Resource for GPU Developers

DEEP LEARNING
Deep Learning SDK
High-performance tools and libraries for deep learning

SELF-DRIVING CARS
NVIDIA DriveWorks
Deep learning, HD mapping and supercomputing solutions, from ADAS to fully autonomous

VIRTUAL REALITY
NVIDIA VRWorks™
A comprehensive SDK for VR headsets, games and professional applications

GAME DEVELOPME
NVIDIA GameWorks™
Advanced simulation technology for game

NVIDIA SDK

IRAY VR

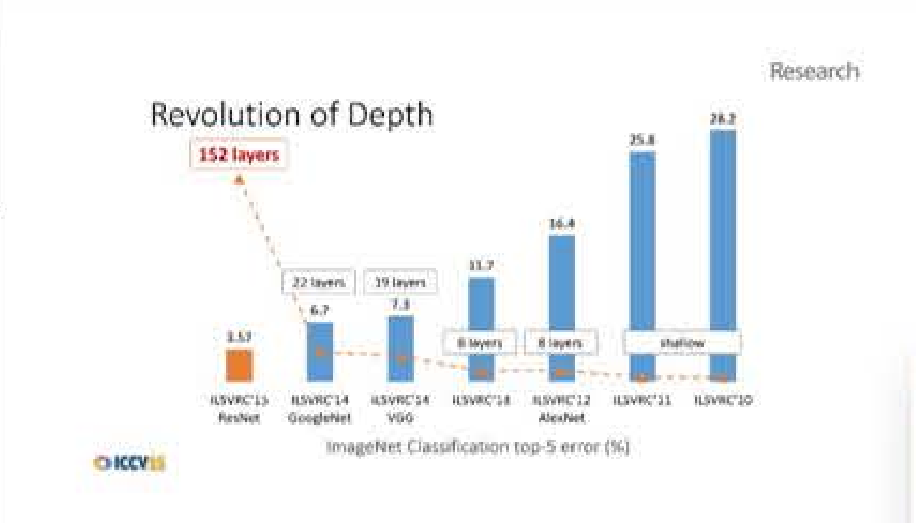
A Deep Learning Chip

A Deep Learning Box

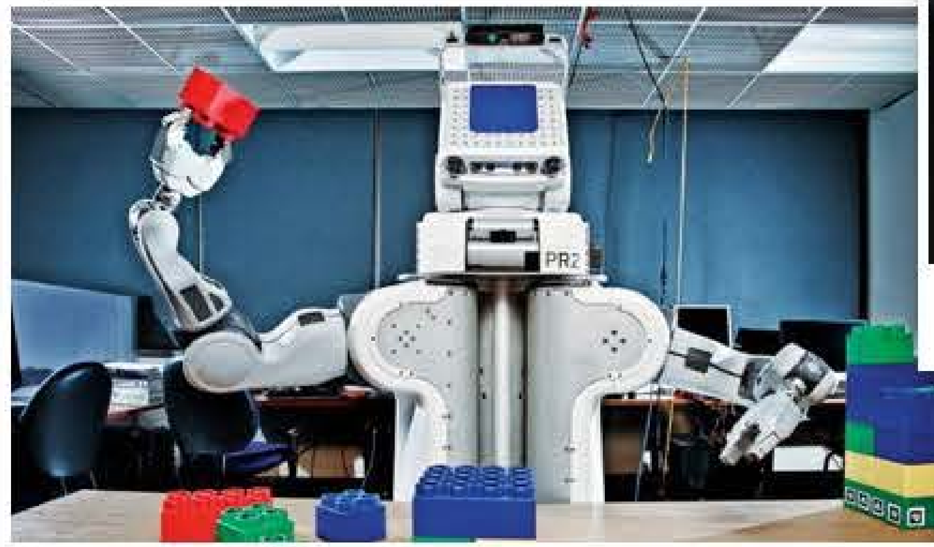
A Deep Learning Car

AN AMAZING YEAR IN AI

Microsoft & Google
“Superhuman” Image Recognition



Microsoft
“Super Deep Network”



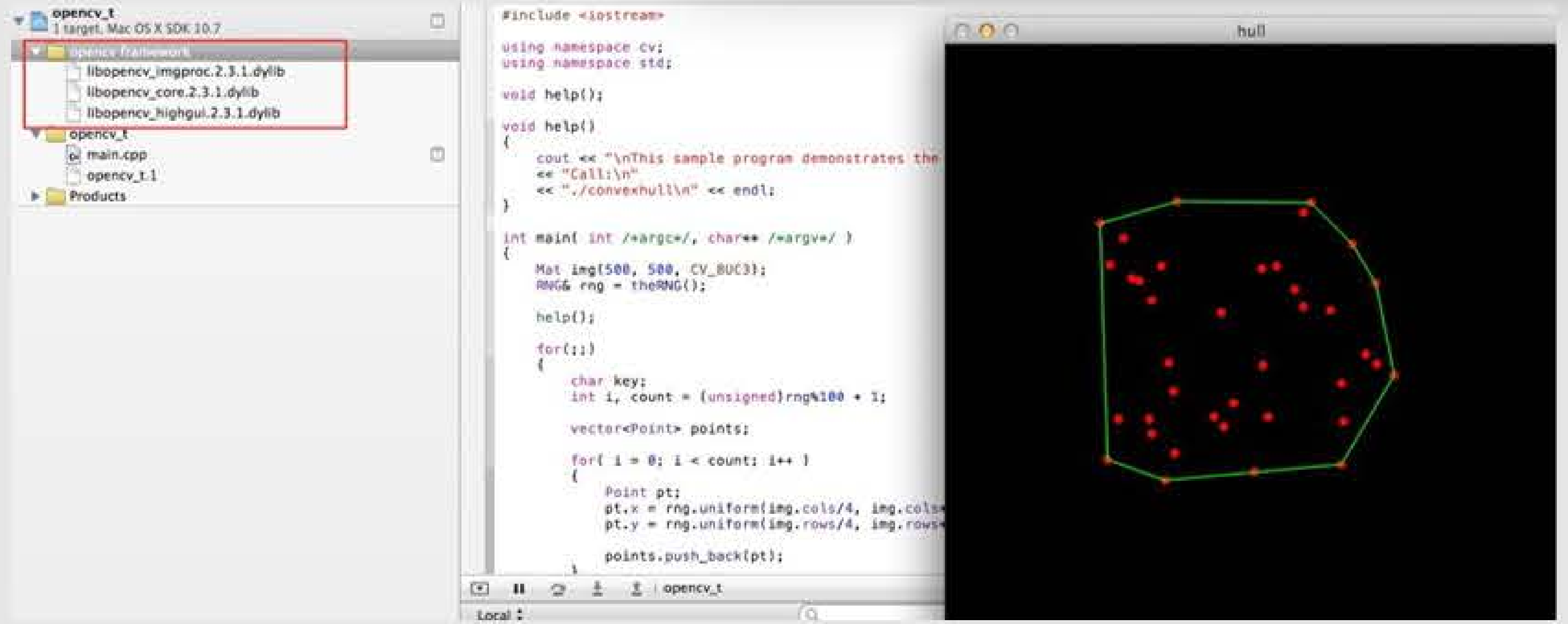
Berkeley’s Brett
One network, everything robotics

Deep Speech 2
One network, 2 languages

A New Computing Model
Hits Pop Culture

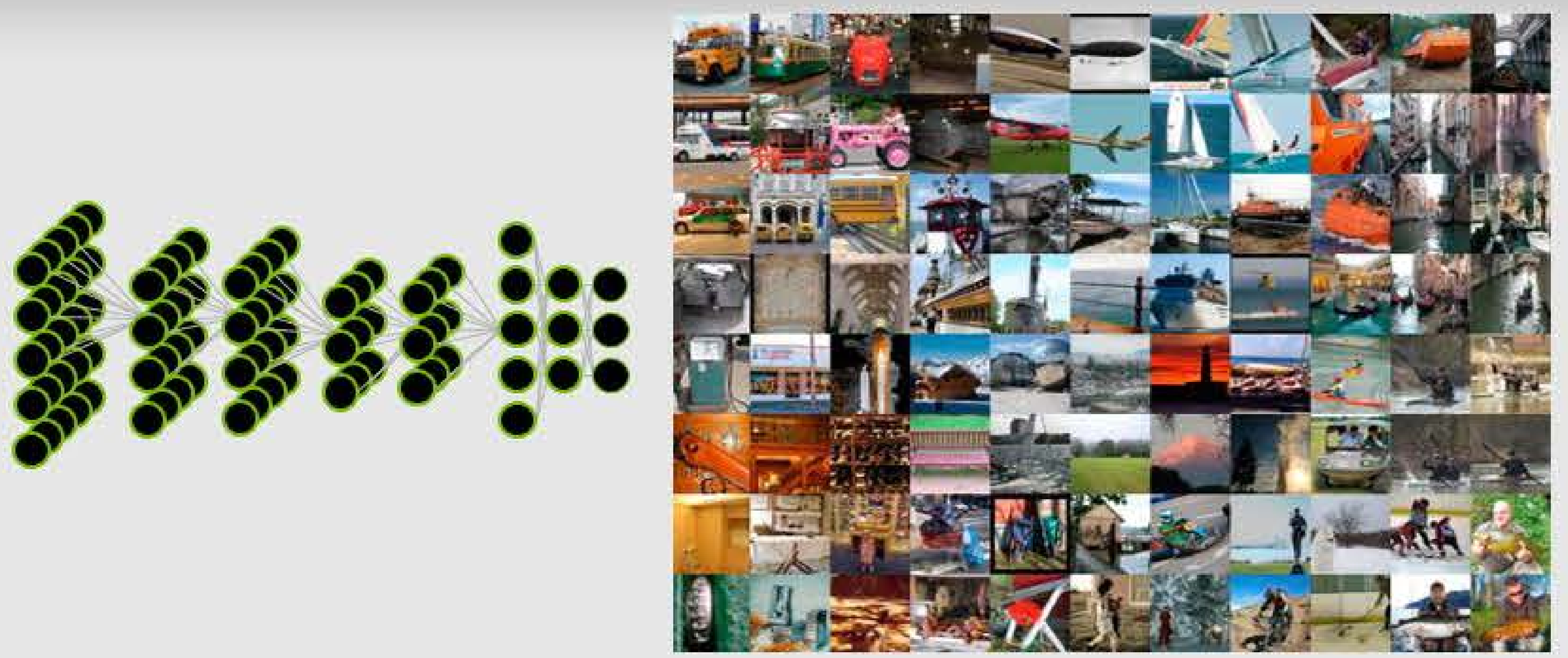
AlphaGo
Rivals a World Champion

A NEW COMPUTING MODEL

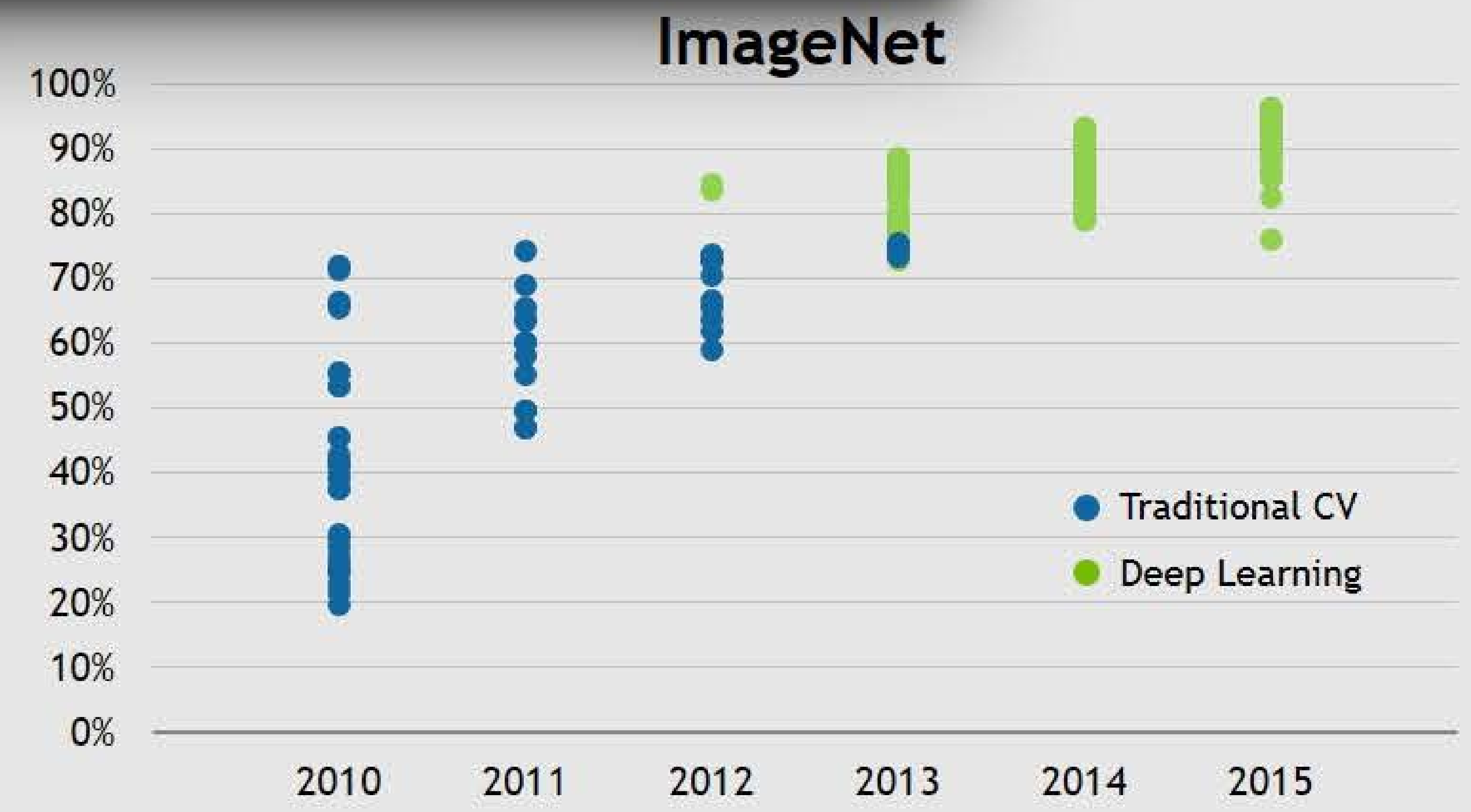

opencv_t
1 target, Mac OS X SDK 10.7
opencv framework
libopencv_imgproc.2.3.1.dylib
libopencv_core.2.3.1.dylib
libopencv_highgui.2.3.1.dylib
opencv_t
main.cpp
opencv_t.1
Products
opencv_t
Local
hull

```
#include <iostream>

using namespace cv;
using namespace std;

void help();

void help()
{
    cout << "\nThis sample program demonstrates the
    << "Call:\n"
    << "./convexhull\n" << endl;
}

int main( int /*argc*/, char** /*argv*/ )
{
    Mat img(500, 500, CV_8UC3);
    RNG& rng = theRNG();

    help();

    for(;;)
    {
        char key;
        int i, count = (unsigned)rng%100 + 1;

        vector<Point> points;

        for( i = 0; i < count; i++ )
        {
            Point pt;
            pt.x = rng.uniform(img.cols/4, img.cols
            pt.y = rng.uniform(img.rows/4, img.rows

            points.push_back(pt);
        }
```

Traditional Computer Vision
Experts + Time



Deep Learning Object Detection
DNN + Data + HPC



Deep Learning Achieves “Superhuman” Results

THE EXPANDING UNIVERSE OF MODERN AI

" THE BIG BANG "

Big Data
GPU
Algorithms

RESEARCH

Berkeley UNIVERSITY OF CALIFORNIA
Carnegie Mellon University
DEEPMIND
MIT Massachusetts Institute of Technology
NYU
OpenAI
Université de Montréal
UNIVERSITY OF OXFORD
UNIVERSITY OF TORONTO

CORE TECHNOLOGY / FRAMEWORKS

Preferred Networks
Chainer
facebook. torch
Université de Montréal
theano
Google
TensorFlow
Berkeley UNIVERSITY OF CALIFORNIA
Caffe
Microsoft
CNTK
UNIVERSITY OF OXFORD
MatConvNet
NVIDIA. cuDNN

AI-as-a-PLATFORM

amazon web services
IBM Watson
Google
Microsoft Azure

START-UPS

api.ai
Personal Assistants
conversational interface

BLUE RIVER TECHNOLOGY
Agriculture
crop-yield optimization

clarifai
Tech
visual recognition platform

drive.ai
Automotive
computer vision

MetaMind
eCommerce & Medical
recommendation engines

deep genomics
Genomics
genetic interpretation

Morpho
Tech
computer vision

Orbital Insight
Geospatial
predictions from images

nervana
Tech
AI-as-a-service

SADAKO
Waste Management
sorting robots

SocialEyes®
Medical
diabetic retinopathy

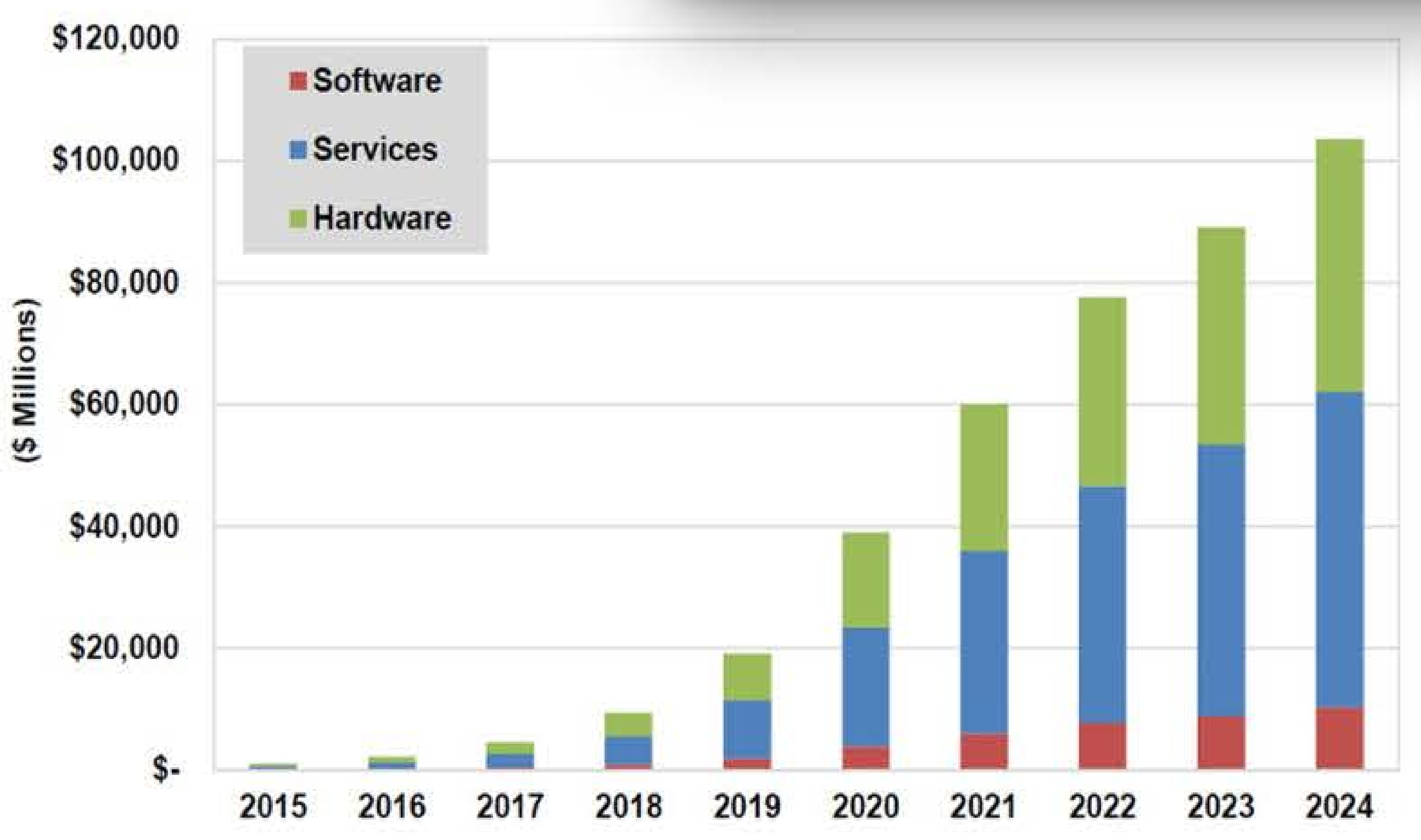
HOW ARE YOU
Education
teaching robots

1,000+ AI START-UPS
$5B IN FUNDING
Source: Venture Scanner

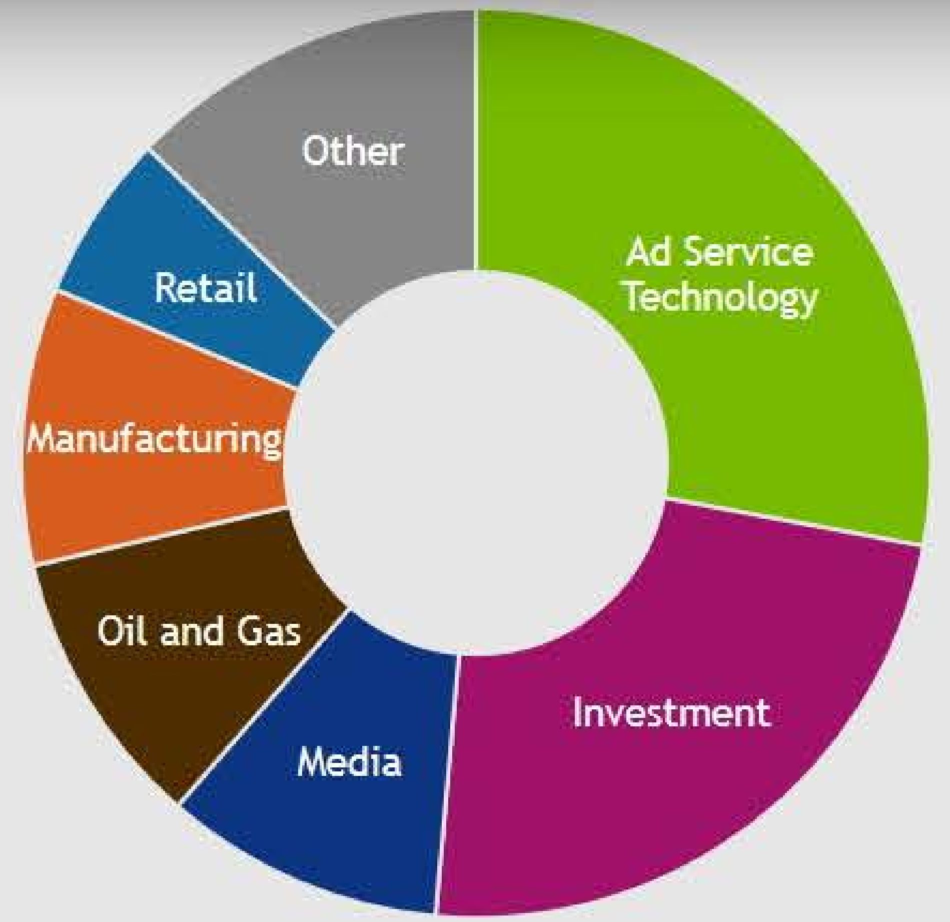
INDUSTRY LEADERS

Alibaba.com
AstraZeneca
Audi
Bai du 百度
Bloomberg
B M W
charles SCHWAB
CISCO
ebay
FANUC ROBOTICS
Ford
GE
gsk
THE HOME DEPOT
MGH MASSACHUSETTS GENERAL HOSPITAL
Mercedes-Benz
MERCK
Pinterest
Schlumberger
SIEME
TARG
TESL
TOY
UBE
VOLV
Walma
YAHO
Yand
yelp

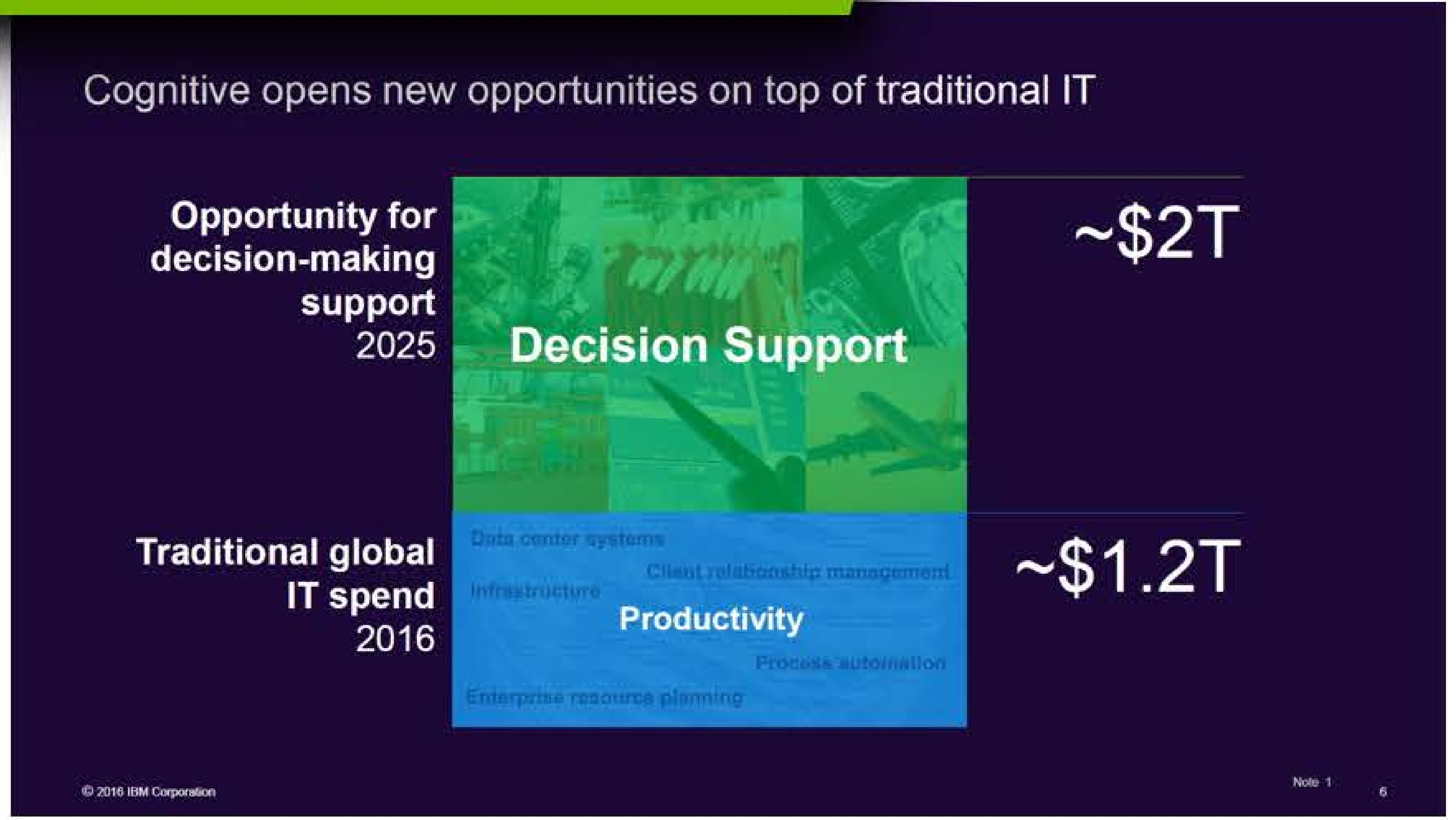
$500B OPPORTUNITY OVER 10 YRS



Deep Learning Total Revenue by Segment



Deep Learning Software Revenue by Industry



Cognitive opens new opportunities on top of traditional IT

Opportunity for decision-making support
2025

Decision Support

~$2T

Traditional global IT spend
2016

Data center systems
Client relationship management
Infrastructure
Productivity
Process automation
Enterprise resource planning

~$1.2T

© 2016 IBM Corporation

Note 1

6

IBM: “Cognitive business represents a $2T opportunity”

SOURCE: “Deep Learning for Enterprise Applications,” 4Q 2015, Tractica

NVIDIA GPU FOR HYPERSCALE

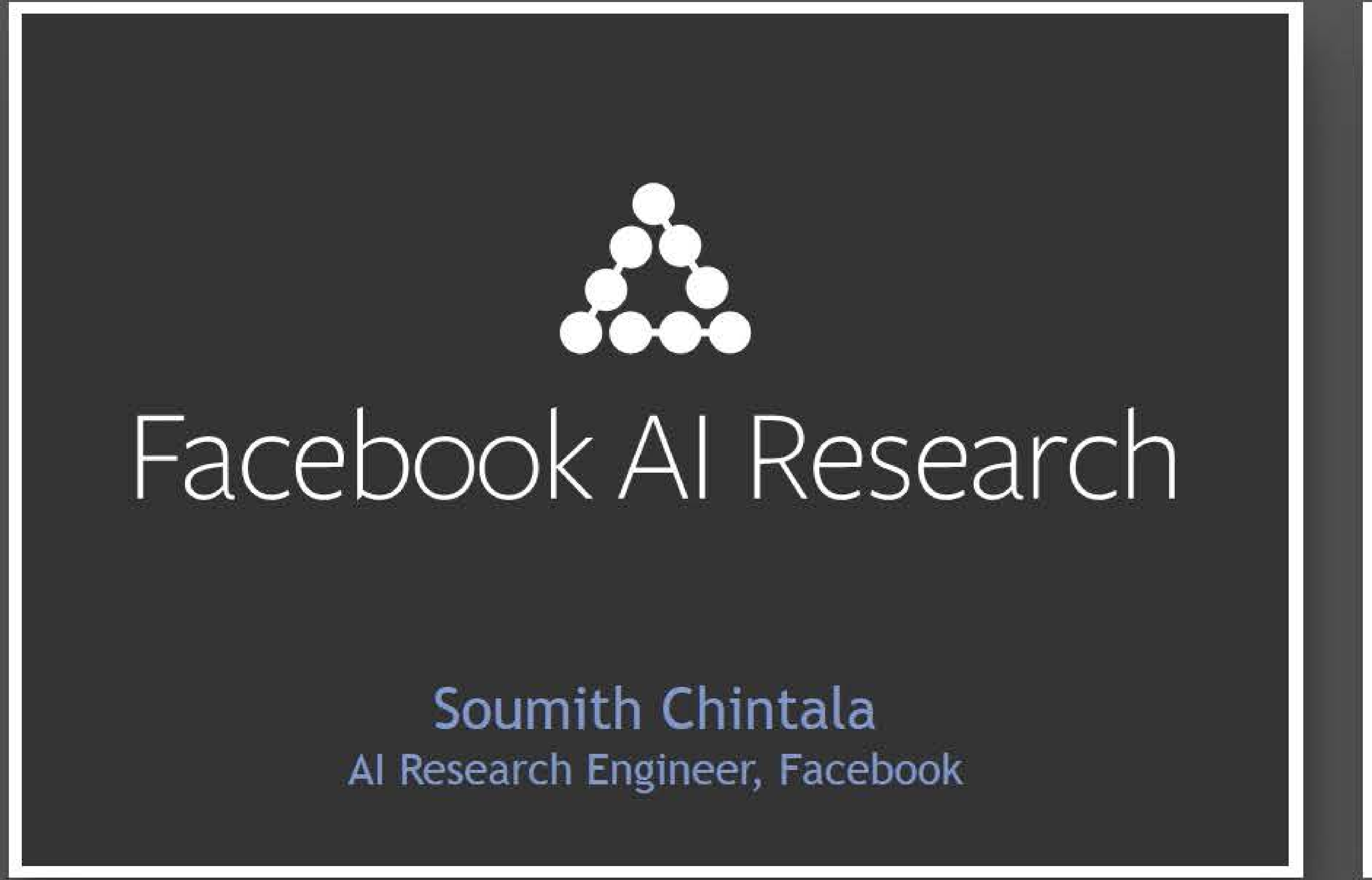
TESLA M40 + TESLA M4

10X Speed up | 20 images/s/W

Alibaba/Aliyun
Amazon
Baidu
eBay
Facebook
Flickr
Google
iFLYTEK
iQIYI
JD.com

Orange
Periscope
Pinterest
Qihoo 360
Shazam
Skype
Sogou
Twitter
Yahoo Supermarket
Yandex
Yelp

Cloud Services Powered by AI



Facebook AI Research

Soumith Chintala
AI Research Engineer, Facebook

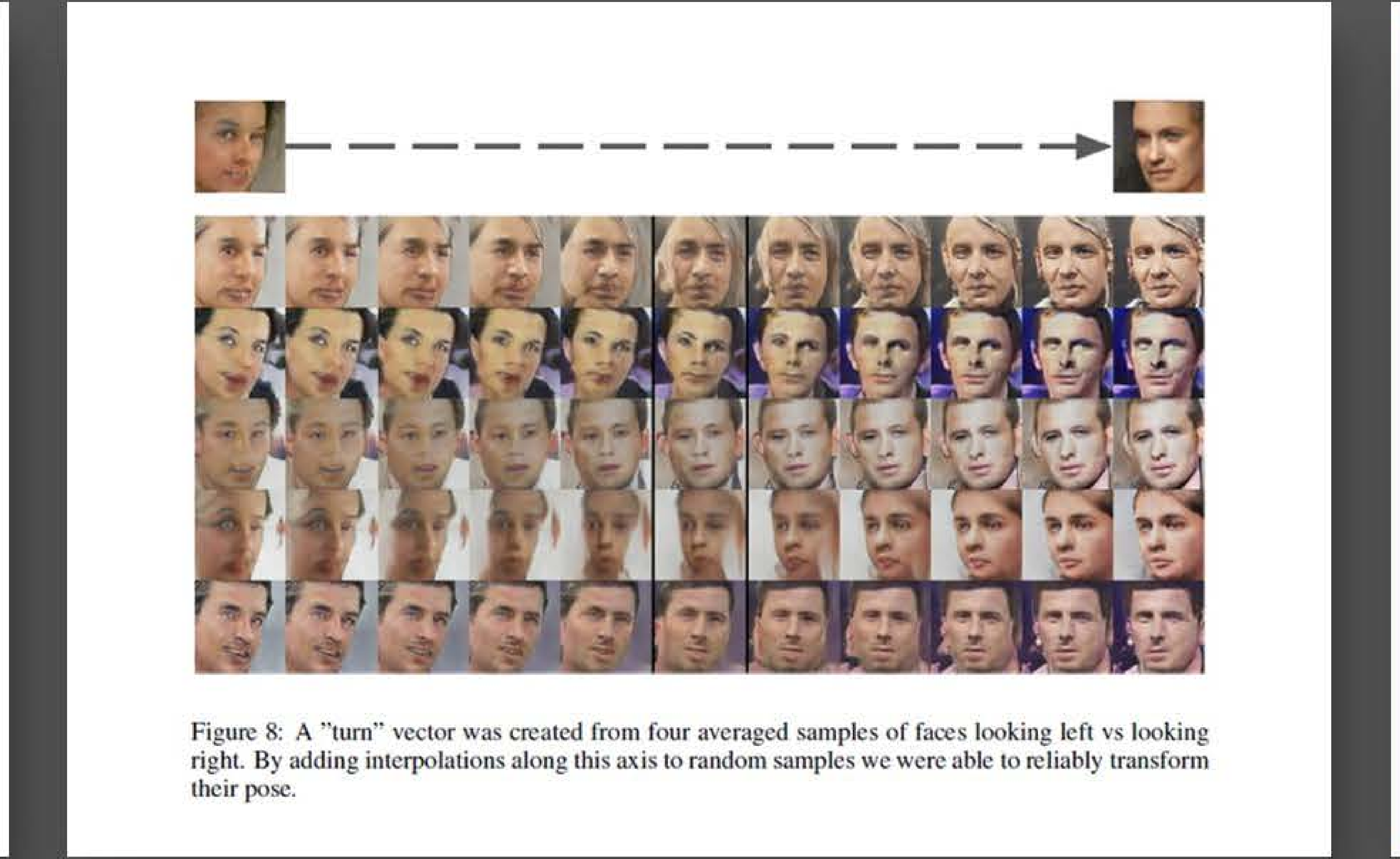
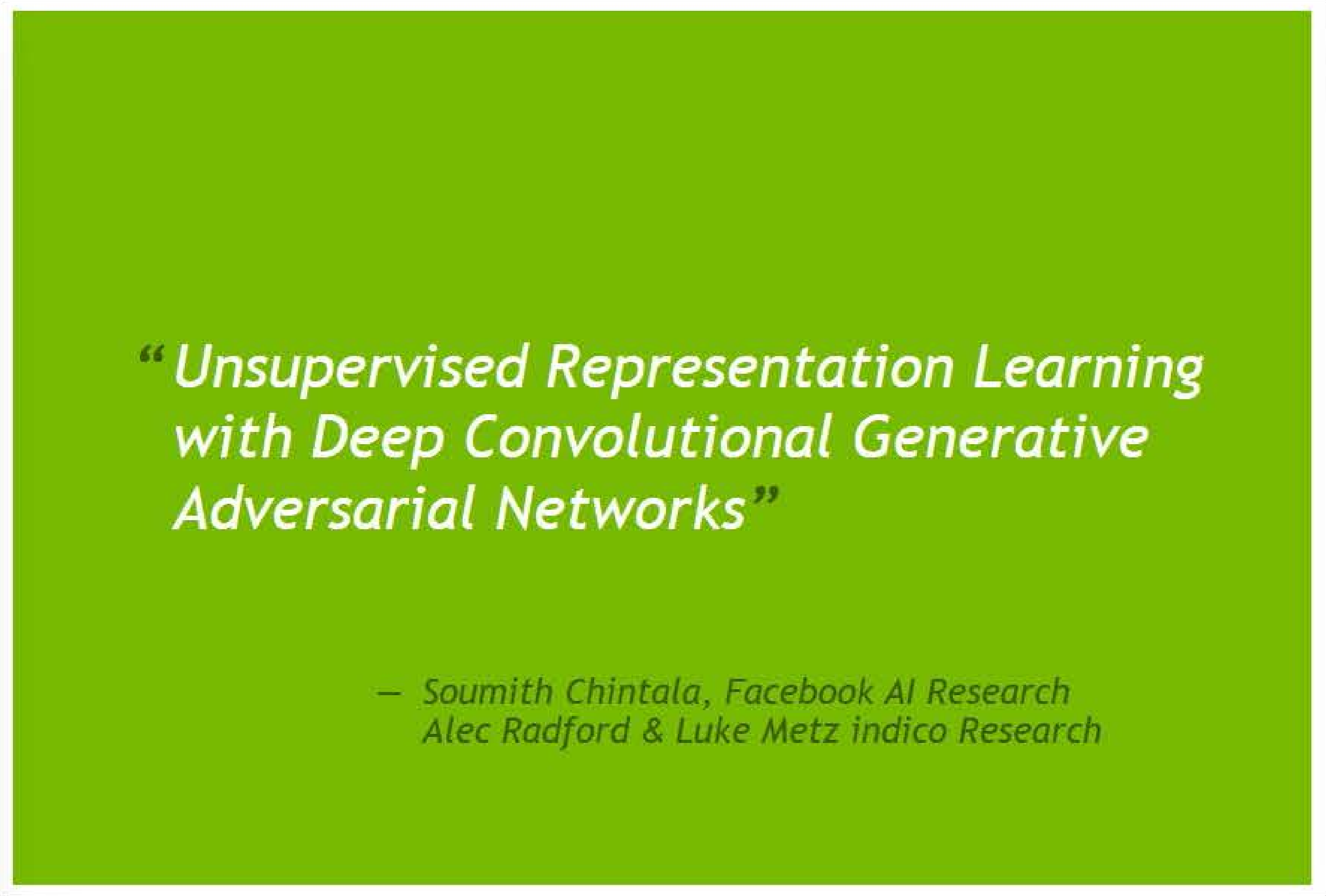


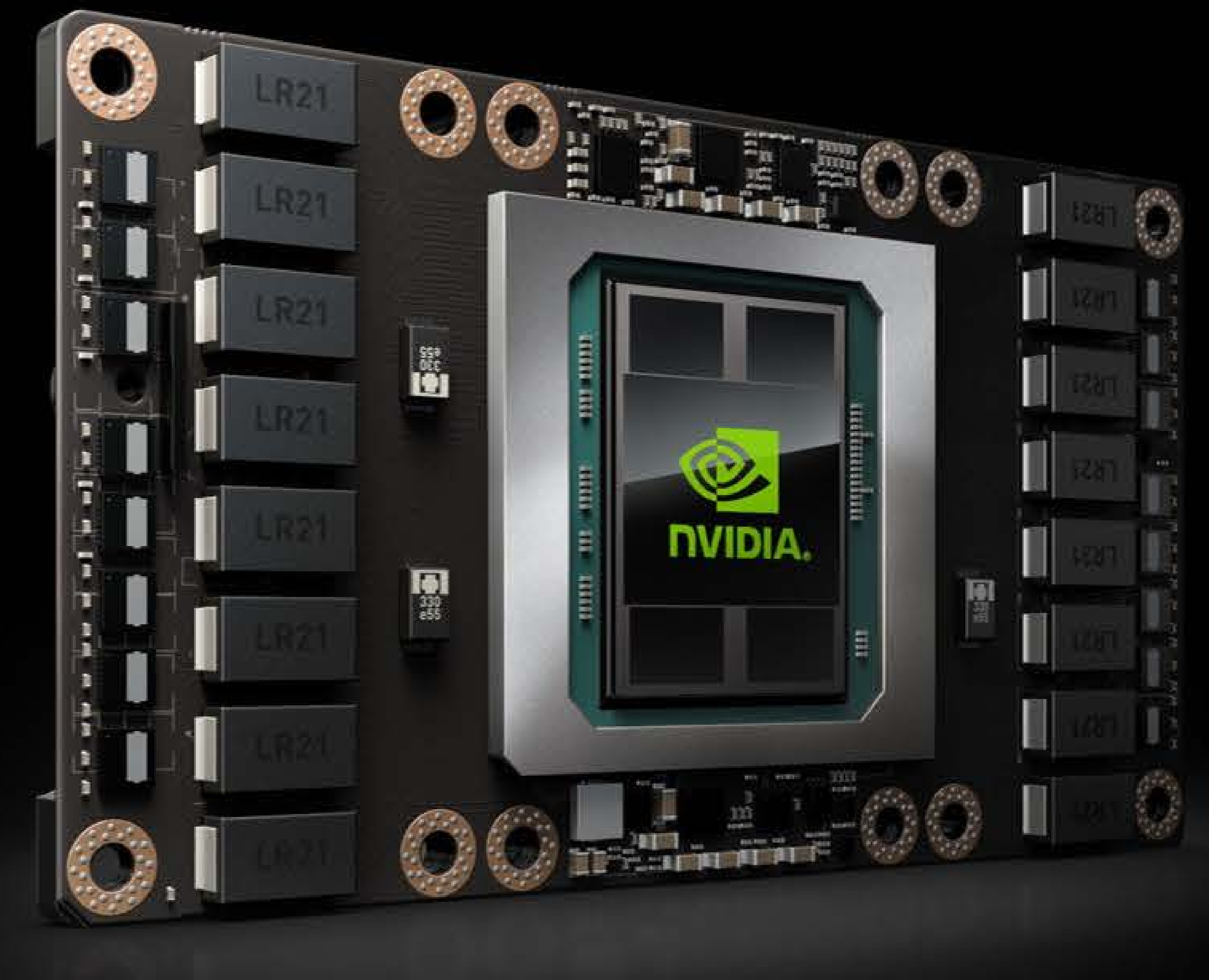
Figure 8: A ”turn” vector was created from four averaged samples of faces looking left vs looking right. By adding interpolations along this axis to random samples we were able to reliably transform their pose.



> *“Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”*
>
> *— Soumith Chintala, Facebook AI Research*
> *Alec Radford & Luke Metz indico Research*

Facebook AI Research
training
generation
generation
generation
generation
generation
generation
generation

UNSUPERVISED LEARNING



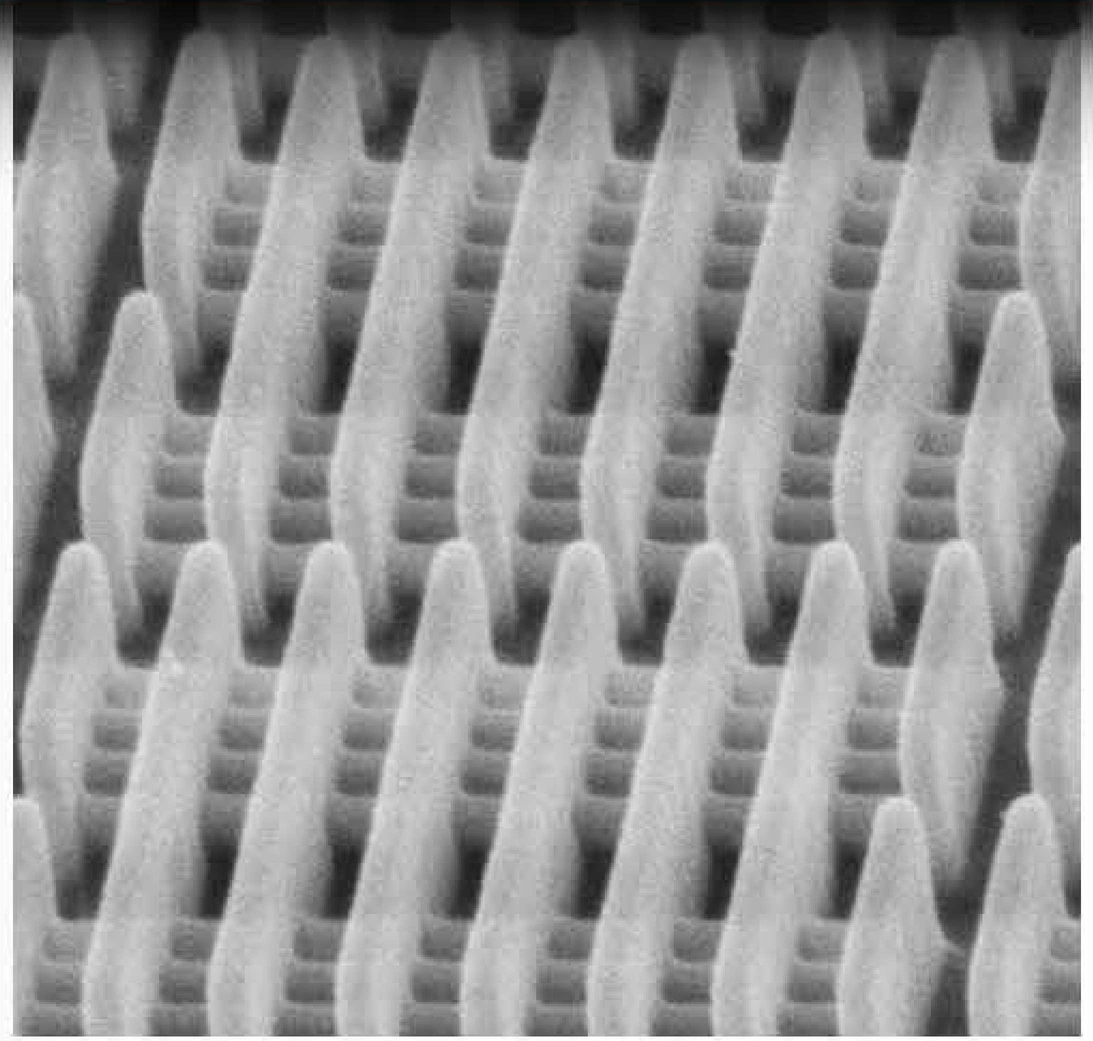
TESLA P100

THE MOST ADVANCED HYPERSCALE DATACENTER GPU EVER BUILT

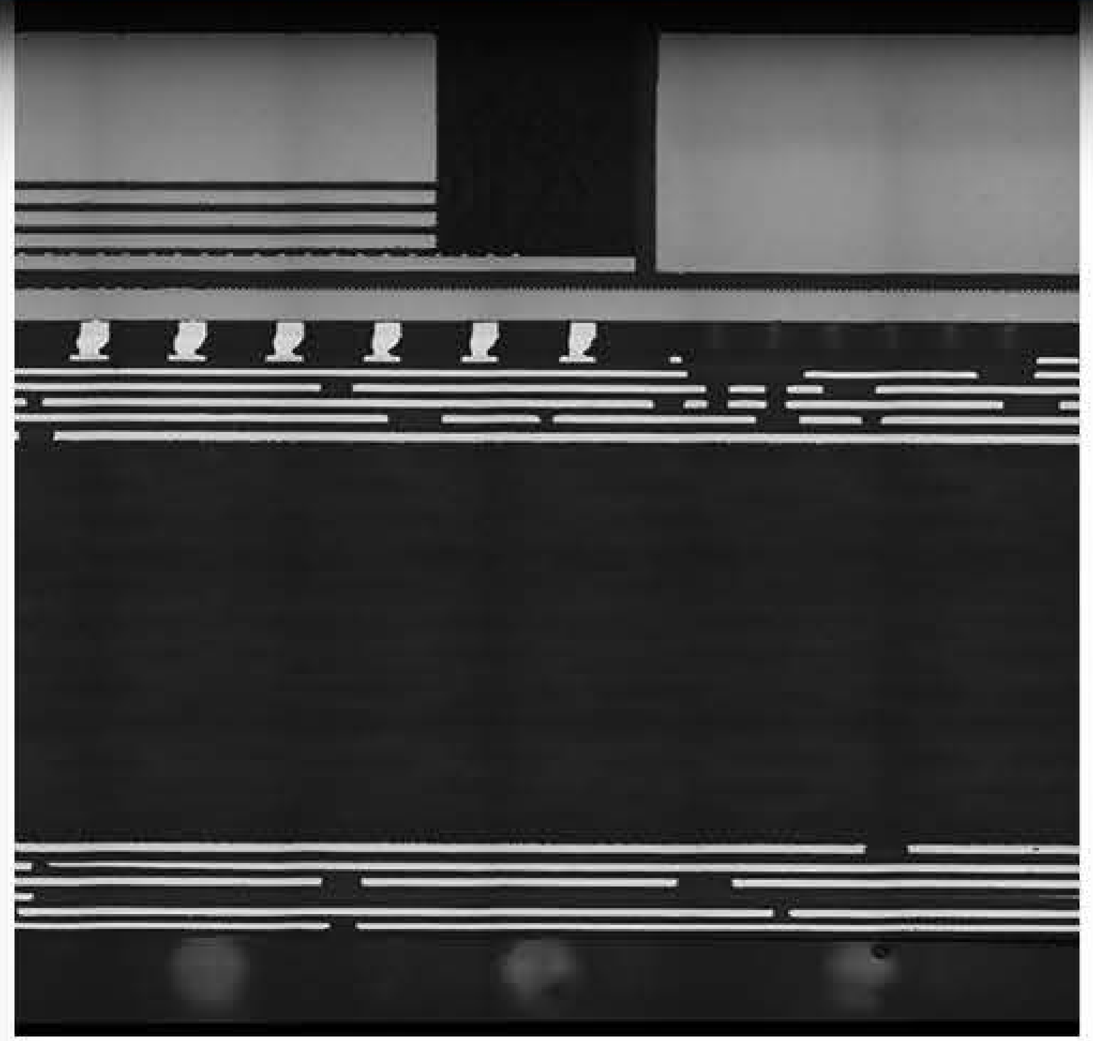
150B XTORS | 5.3TF FP64 | 10.6TF FP32 | 21.2TF FP16 | 14MB SM RF | 4MB L2 Cache

“FIVE MIRACLES”

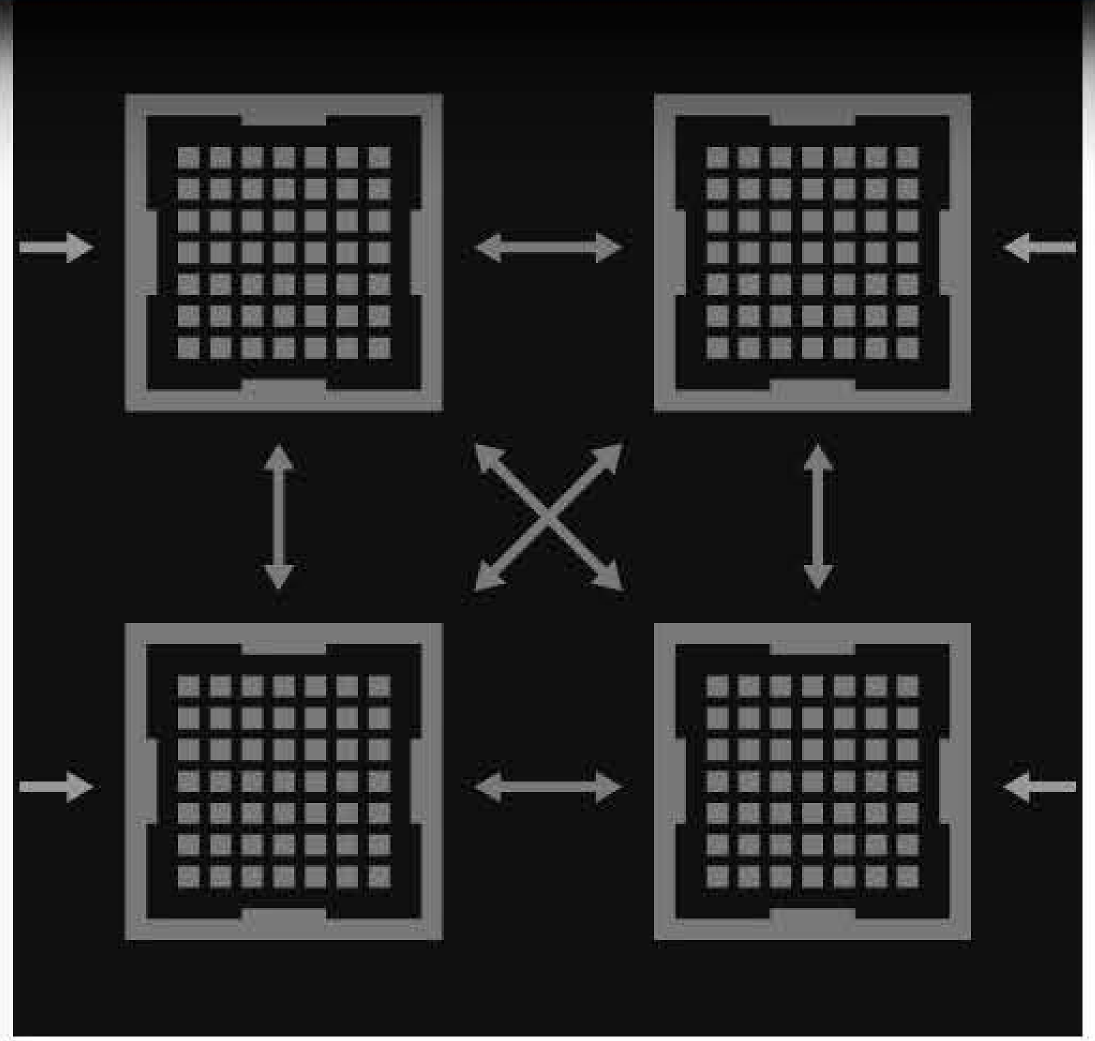
Pascal Architecture



16nm FinFET



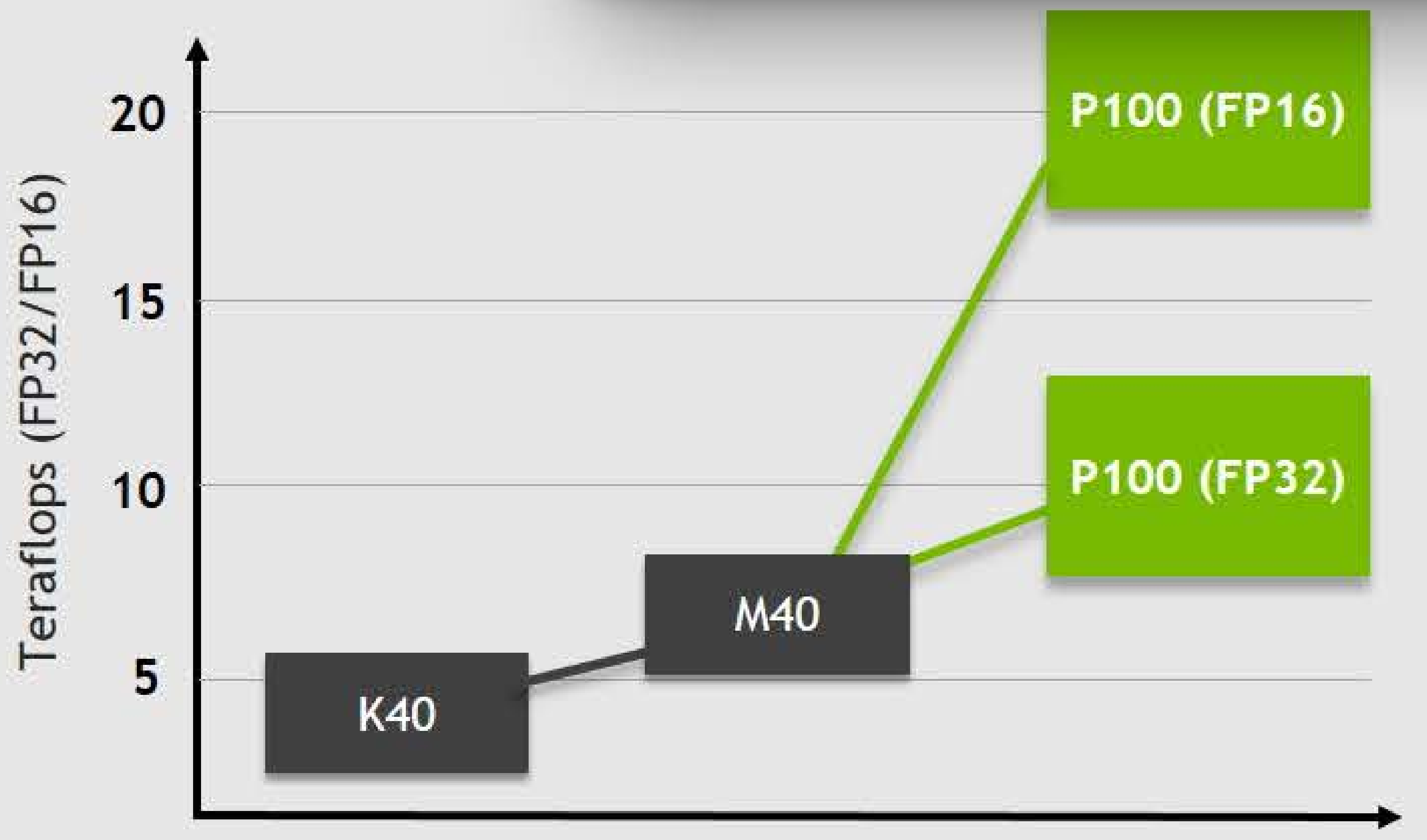
CoWoS with HBM2



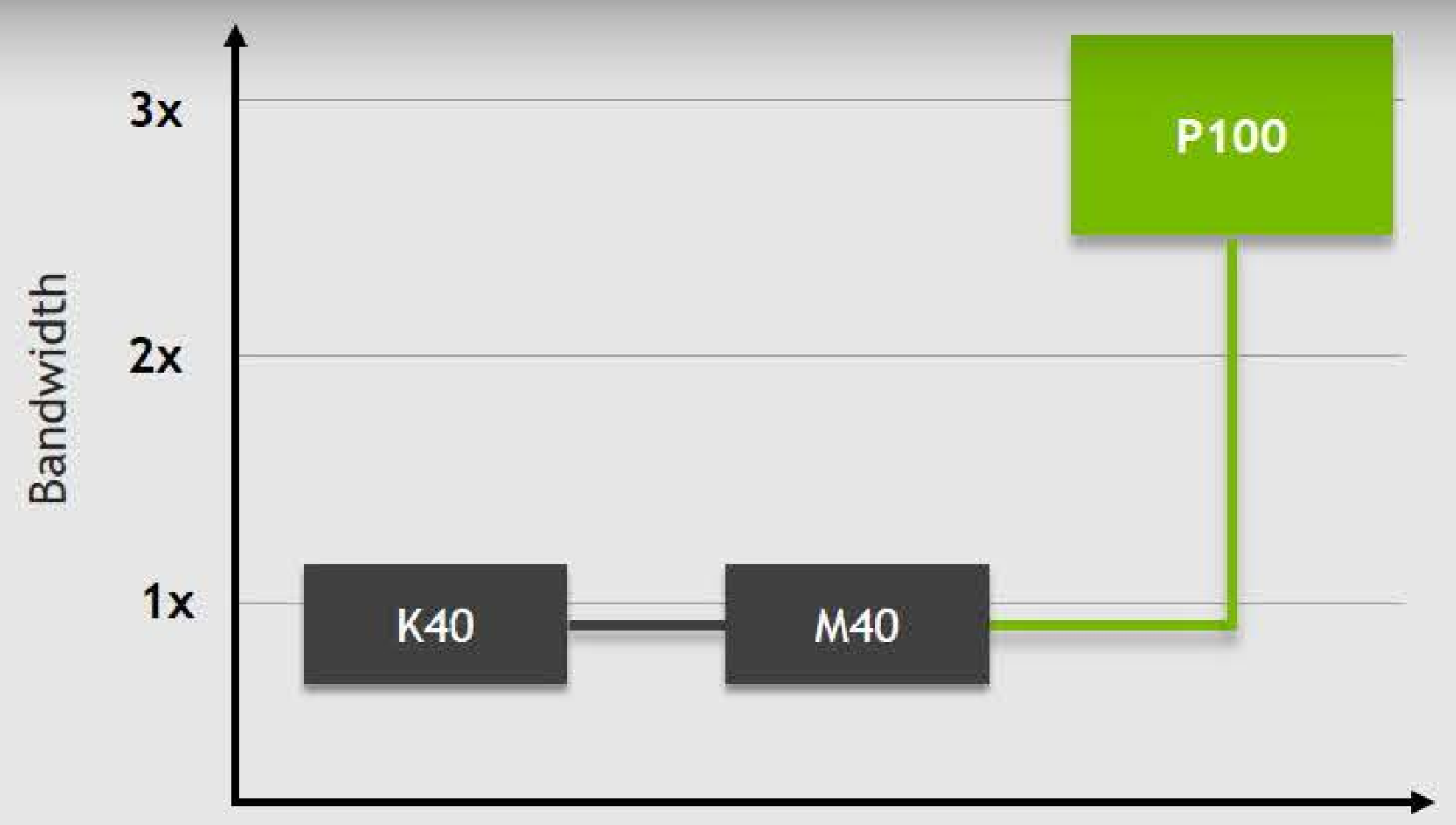
NVLink

New AI Algorithms

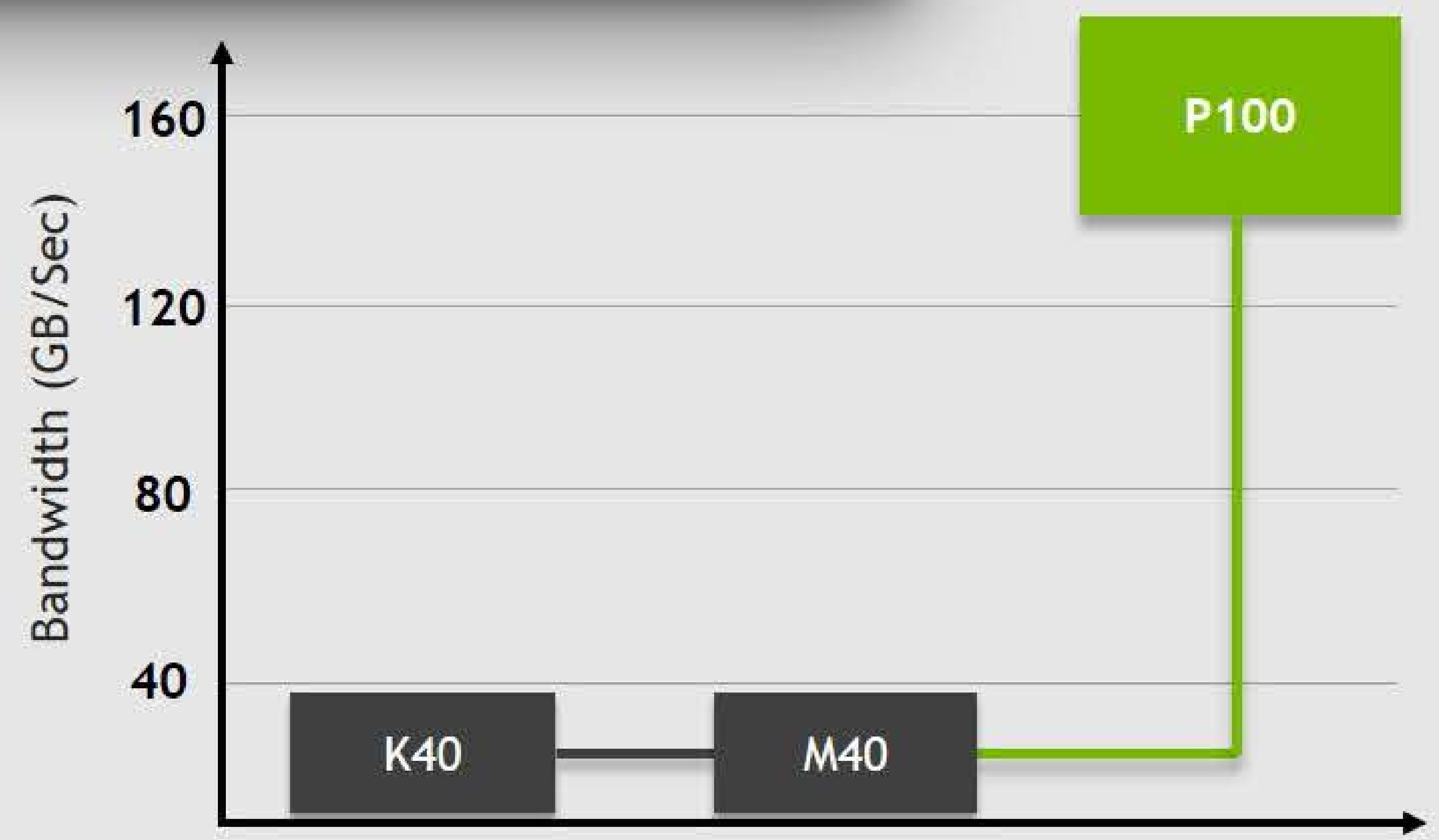
GIANT LEAPS IN EVERYTHING



3x Compute



3x GPU Mem BW



5x GPU-GPU BW

"NVIDIA GPU is accelerating progress in AI. As neural nets become larger and larger, we not only need faster GPUs with larger and faster memory, but also much faster GPU-to-GPU communication, as well as hardware that can take advantage of reduced-precision arithmetic. This is precisely what Pascal delivers."

Yann LeCun, Director of AI Research, Facebook

"This is a new era of computing. New approaches to the underlying technologies will be required for AI and cognitive. The combination of NVIDIA Pascal GPUs and IBM POWER accelerates Watson's learning of new skills. Together, IBM and NVIDIA will advance the artificial intelligence industry."

Dr. John Kelly III, SVP, Cognitive Solutions & IBM Research

"AI computers are like space rockets: The bigger the better. Pascal's throughput and interconnect will make the biggest rocket we've seen yet."

Andrew Ng, Chief Scientist, Baidu

"Microsoft is developing super deep neural networks that are more than 1000 layers. NVIDIA Tesla P100's impressive horsepower will enable Microsoft's CNTK to accelerate AI breakthroughs."

Xuedong Huang, Chief Speech Scientist, Microsoft Research

TESLA P100 SERVERS — COMING IN Q1‘17

IBM

Hewlett Packard
Enterprise

DELL

CRAY

5 THINGS

NVIDIA DEVELOPER

NVIDIA SDK

The Essential Resource for GPU Developers

DEEP LEARNING
Deep Learning SDK
High-performance tools and libraries for deep learning

SELF-DRIVING CARS
NVIDIA DriveWorks
Deep learning, HD mapping and supercomputing solutions, from ADAS to fully autonomous

VIRTUAL REALITY
NVIDIA VRWorks™
A comprehensive SDK for VR headsets, games and professional applications

GAME DEVELOPME
NVIDIA GameWorks
Advanced simulation technology for game

NVIDIA SDK

TESLA P100

A Deep Learning Box

IRAY VR

A Deep Learning Car

GPU-ACCELERATED DL FOR EVERY MARKET

阿里云 aliyun.com

amazon web services™

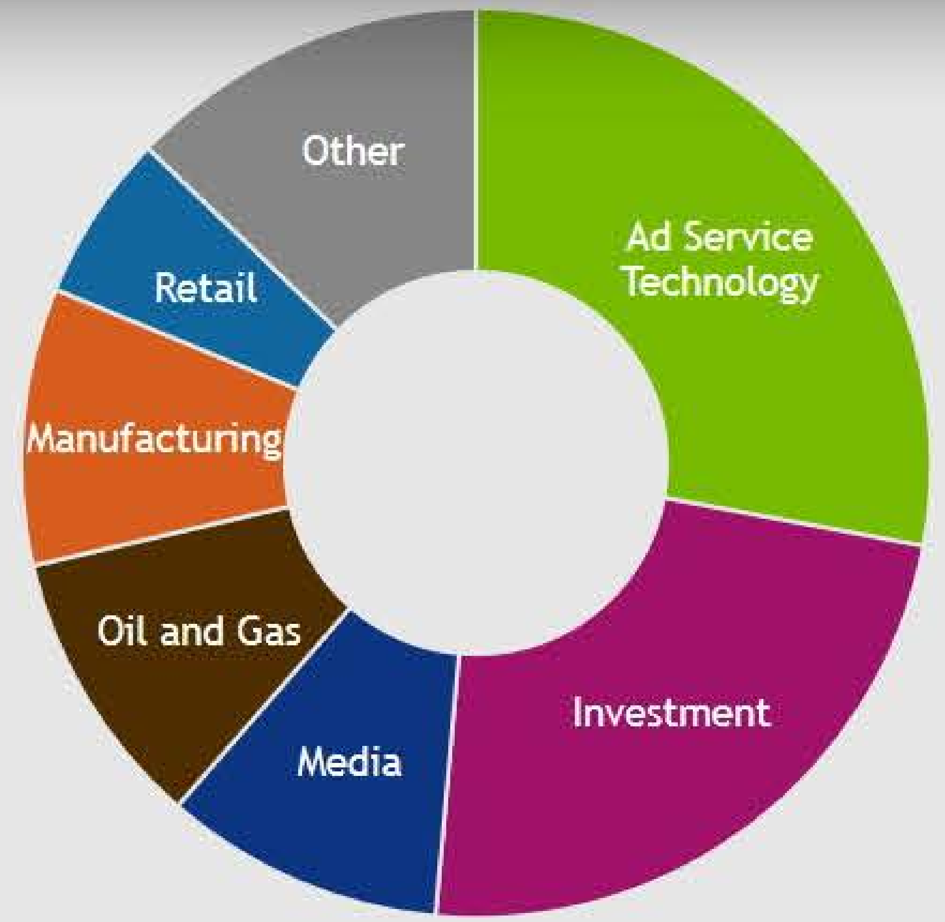
Baidu 百度

Google

Microsoft Azure

IBM Cloud

Deep Learning in the Cloud



IBM: "Cognitive business represents a $2T opportunity"

IBM

Hewlett Packard Enterprise

DELL

CRAY

Deep Learning for Enterprise

SOURCE: "Deep Learning for Enterprise Applications," 4Q 2015, Tractica

NVIDIA DGX-1

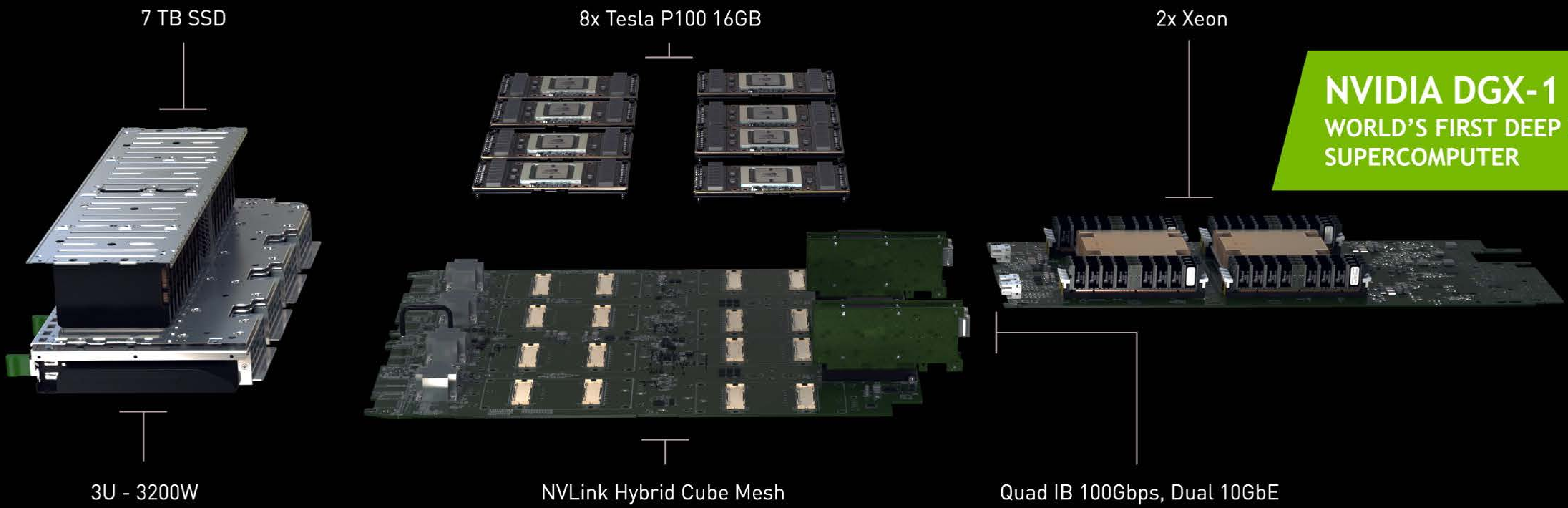
WORLD'S FIRST DEEP LEARNING SUPERCOMPUTER

Engineered for deep learning | 170TF FP16 | 8x Tesla P100 | NVLink hybrid cube mesh | Accelerates major AI frameworks

NVIDIA DGX-1

WORLD'S FIRST DEEP LEARNING SUPERCOMPUTER

170 TFLOPS

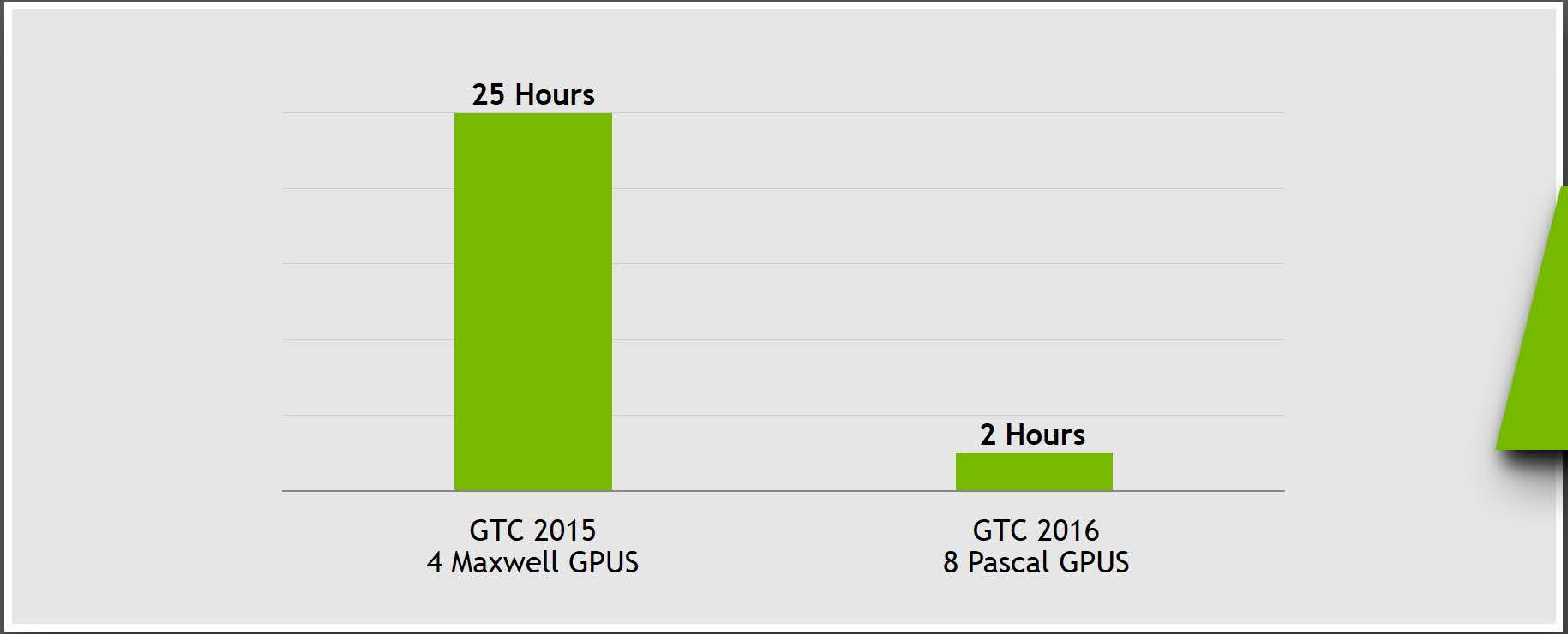


	DUAL XEON	DGX-1
FLOPS (CPU + GPU)	3 TF	170 TF
AGGREGATE NODE BW	76 GB/s	768 GB/s
ALEXNET TRAIN TIME	150 HOURS	2 HOURS
TRAIN IN 2 HOURS	>250 NODES*	1 NODE

“250 SERVERS
IN-A-BOX”

* Caffe Training on Multi-node Distributed-memory Systems Based on Intel® Xeon® Processor E5 Family (extrapolated)
Gennady Fedorov (Intel)'s picture Submitted by Gennady Fedorov (Intel), Vadim P. (Intel) on October 29, 2015

https://software.intel.com/en-us/articles/caffe-training-on-multi-node-distributed-memory-systems-based-on-intel-xeon-processor-e5



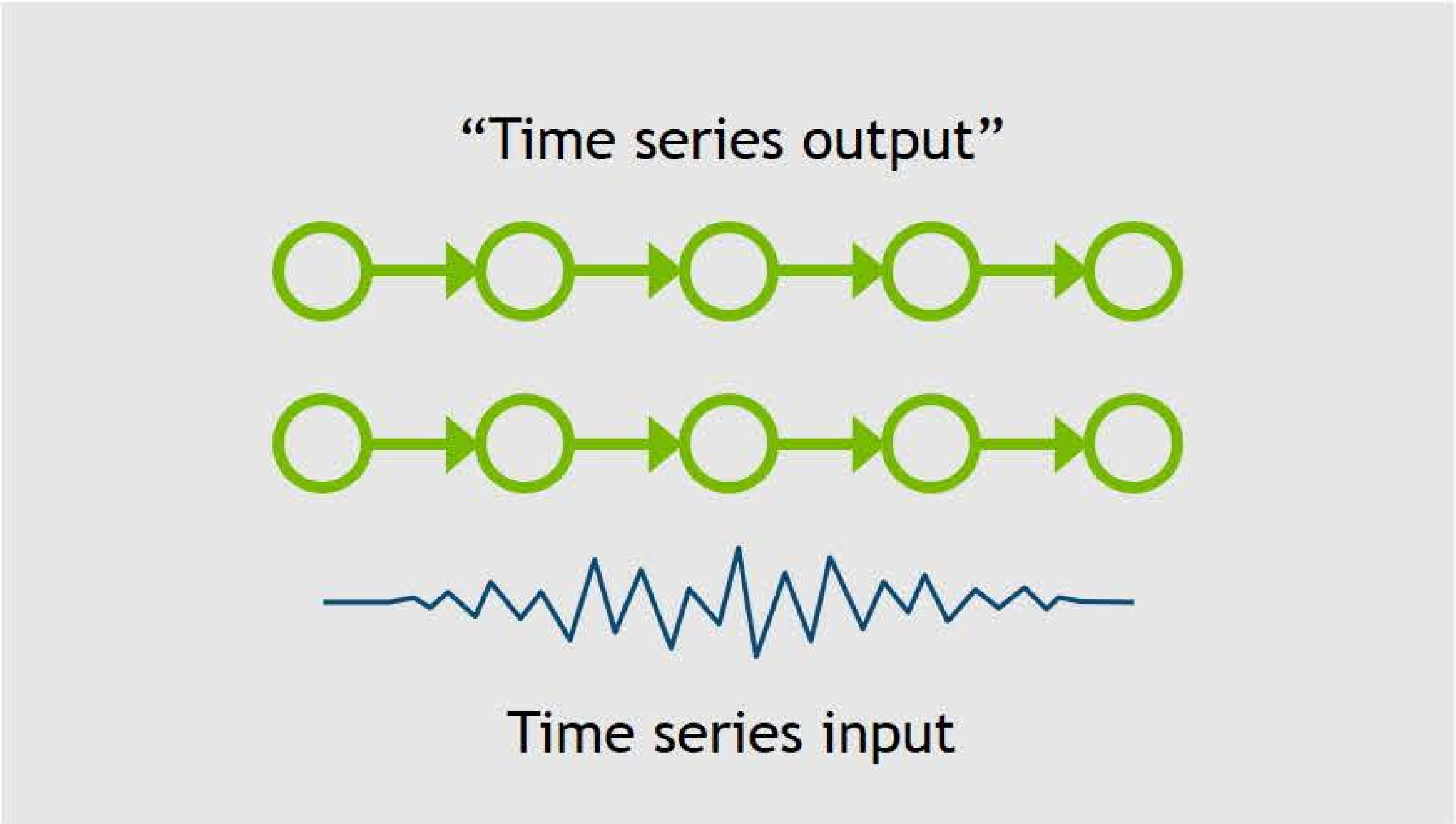
12X SPEED-UP IN ONE YEAR

1.33 billion images/day

Baidu 百度

Bryan Catanzaro

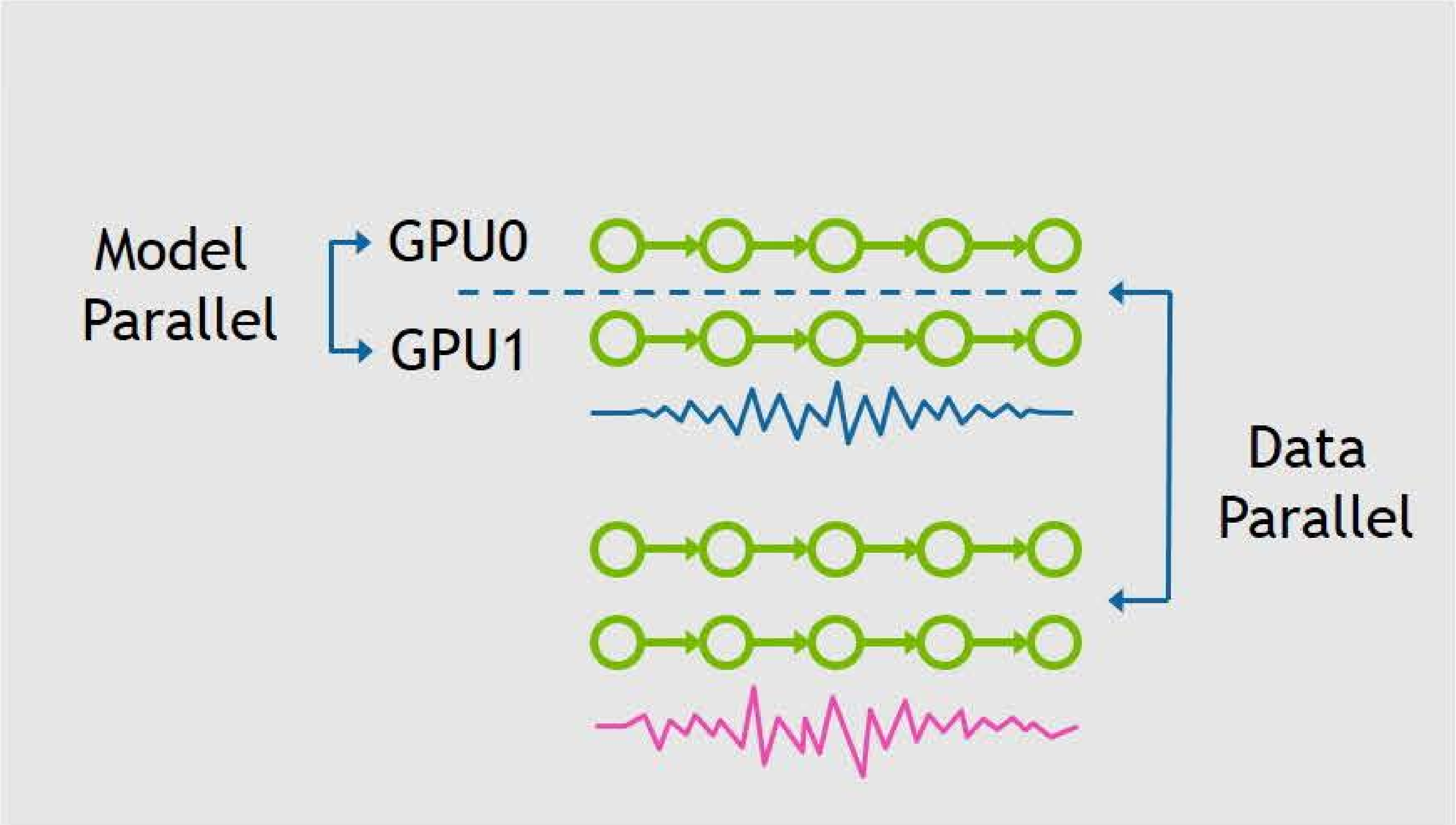
Senior Researcher, Baidu



“Time series output”

Time series input

Recurrent Neural Nets



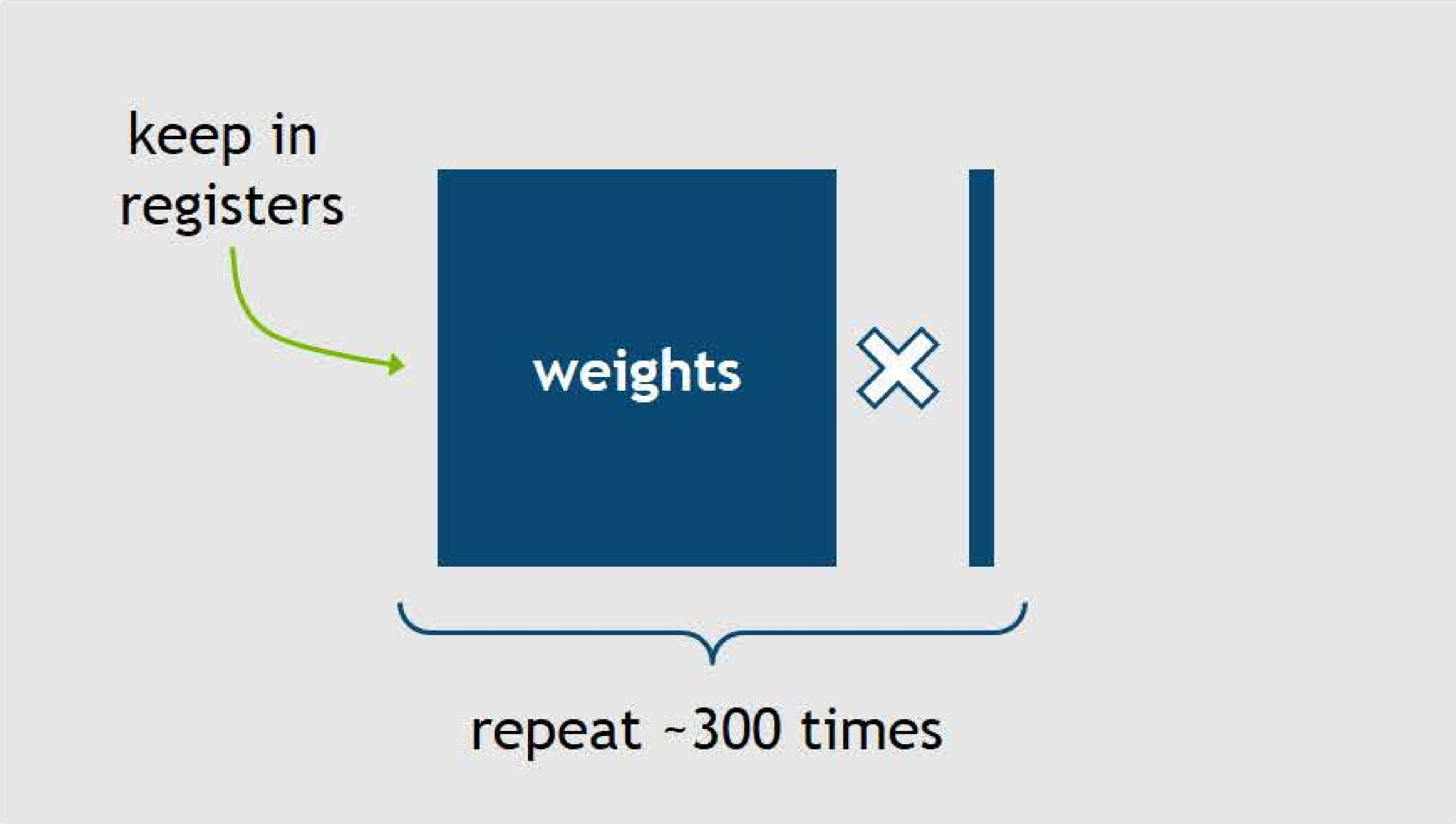
Model Parallel

GPU0

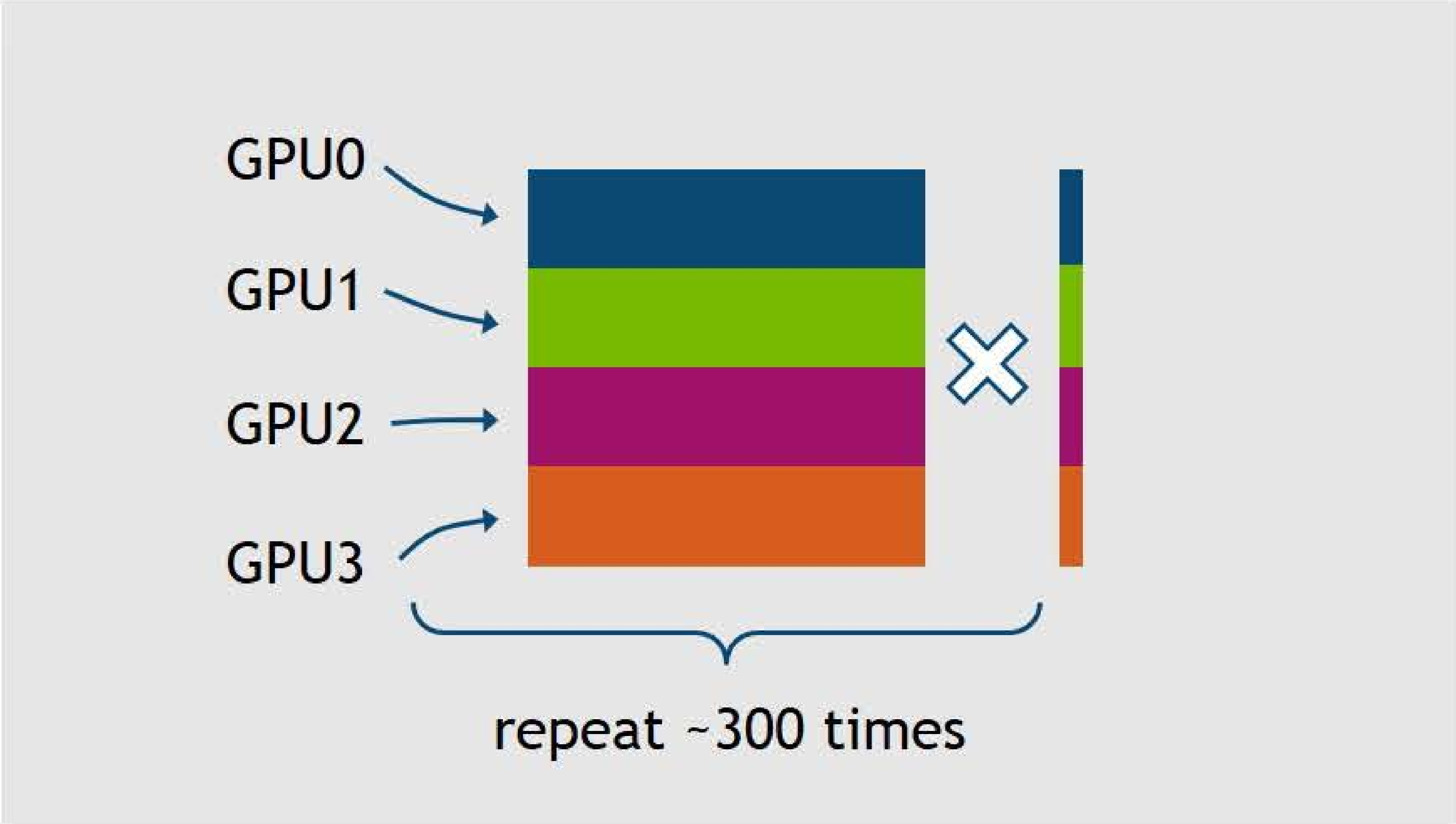
GPU1

Data Parallel

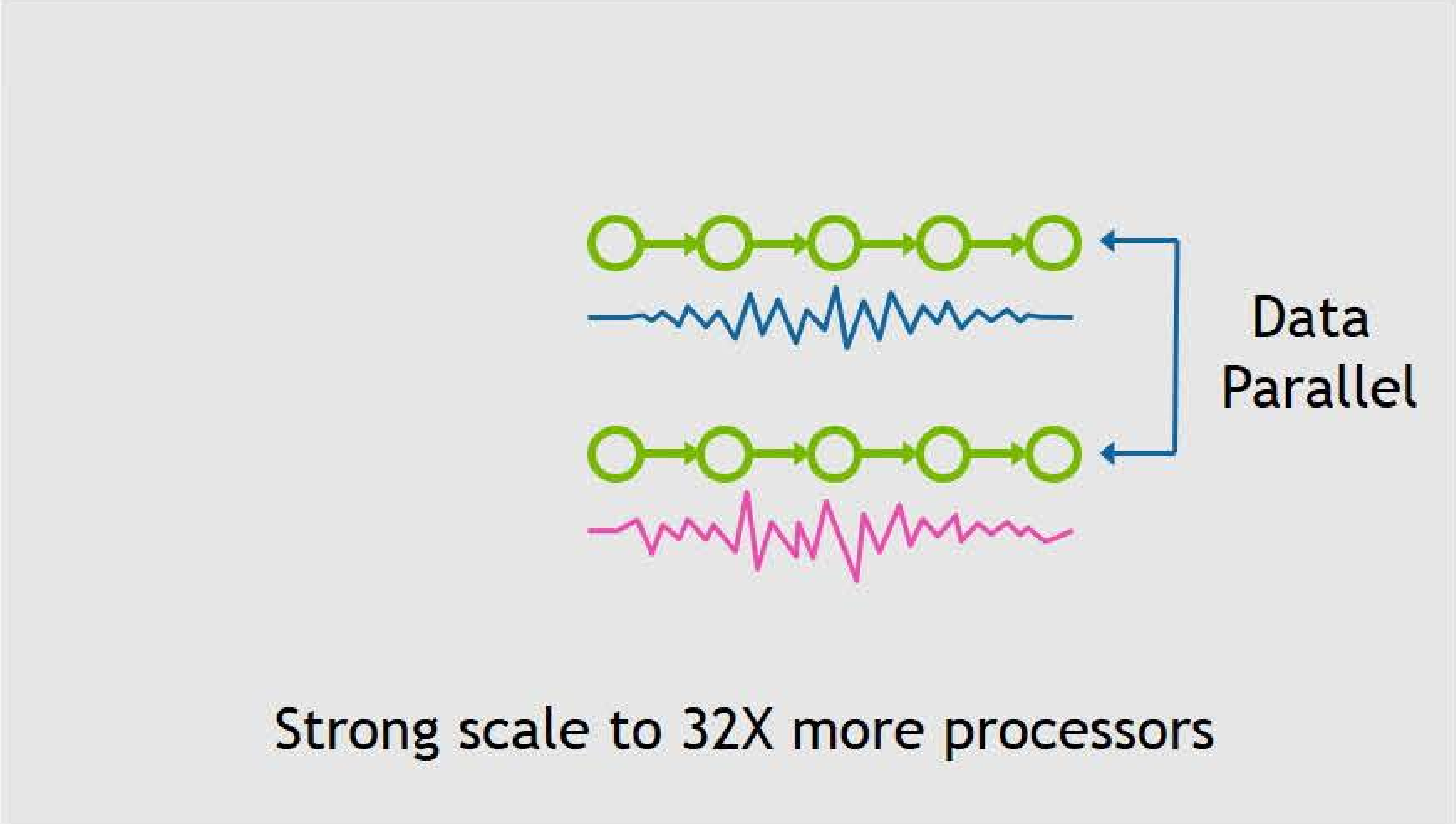
Model + Data Parallelism



Persistent RNNs: Peak FLOPs at batch of 8



Add Model Parallelism over NVLINK

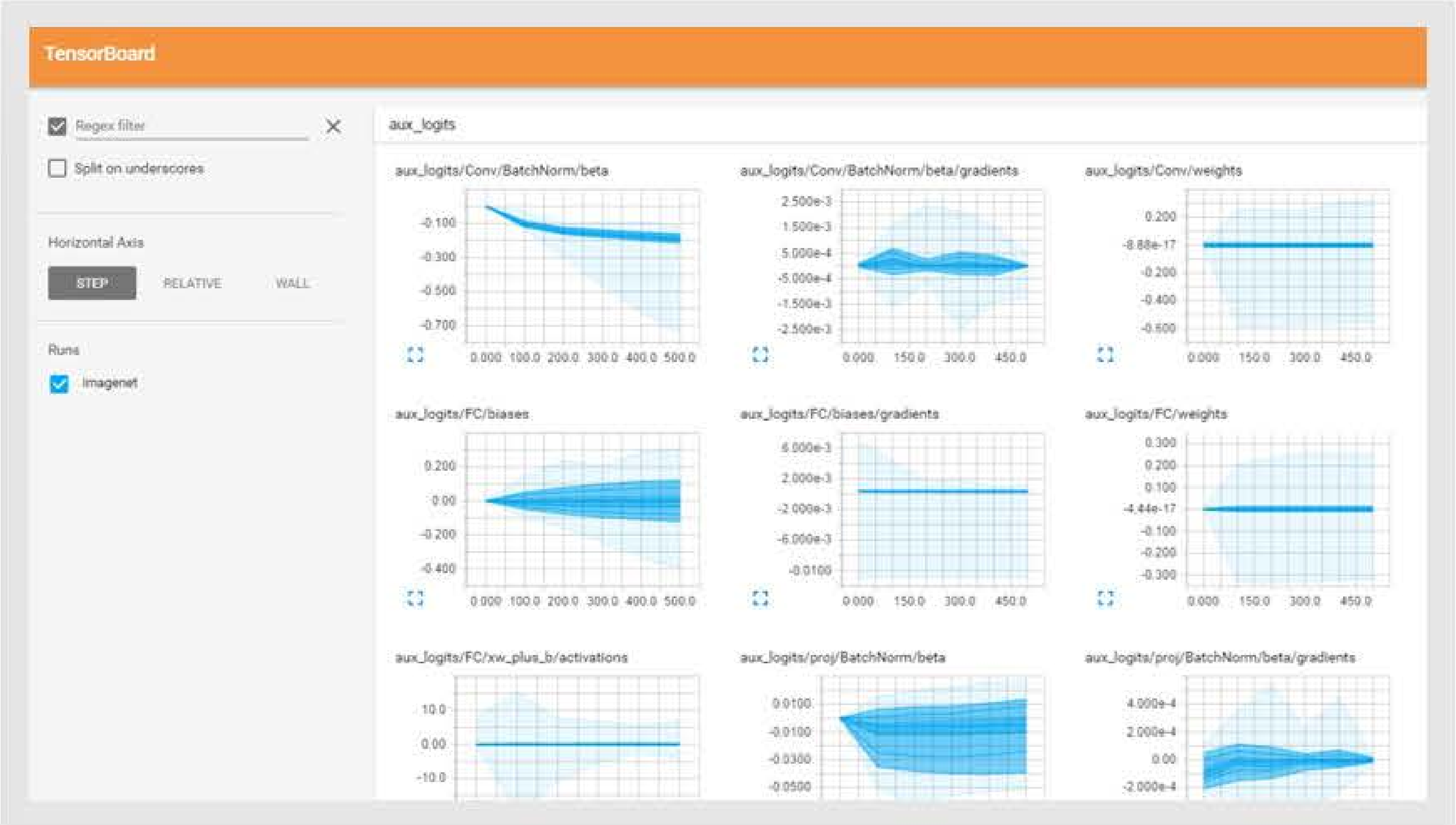


Compose with Data Parallelism

TensorFlow

Rajat Monga

TensorFlow Technical Lead & Manager, Google



TensorBoard

☑ Regex filter ✕

□ Split on underscores

Horizontal Axis: STEP RELATIVE WALL

Runs

☑ imagenet

aux_logits

NVIDIA DGX-1

WORLD'S FIRST DEEP LEARNING SUPERCOMPUTER

170TF | "250 servers in-a-box" | nvidia.com/dgx1

$129,000

PIONEERS IN AI RESEARCH

Berkeley
UNIVERSITY OF CALIFORNIA

Carnegie Mellon University

香港中文大學
The Chinese University of Hong Kong

Massachusetts Institute of Technology

NYU

Stanford University

Université de Montréal

UNIVERSITY OF OXFORD

USI/SUPSI
IDSIA

UNIVERSITY OF TORONTO

Frameworks for Multi-GPU Pascal

Large-scale Deep Learning

Reinforcement Learning

Unsupervised and Transfer Learning

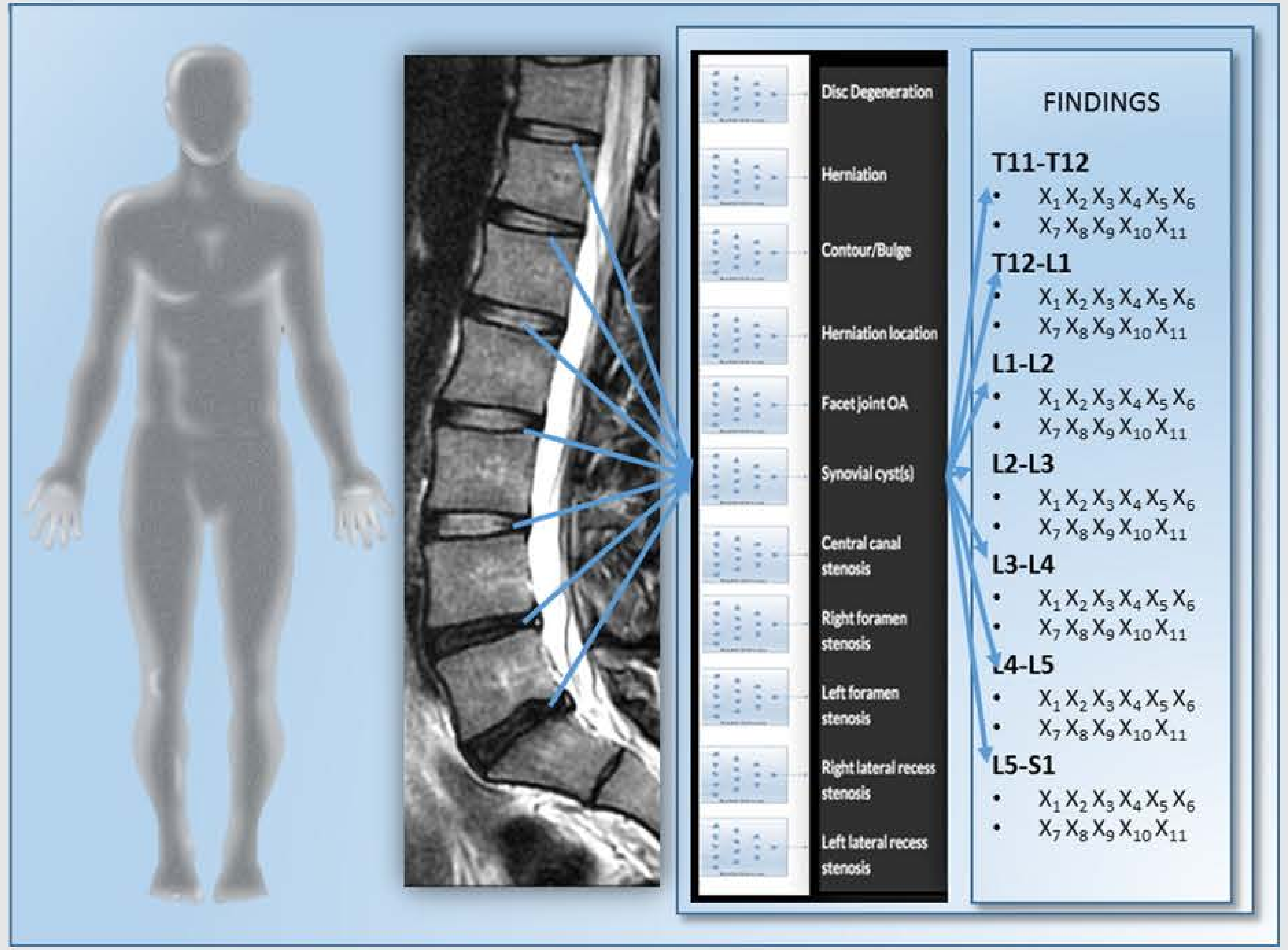
Natural Language Understanding

Autonomous Driving

Medical Applications

NVIDIA

MGH 1811 MASSACHUSETTS GENERAL HOSPITAL



Disc Degeneration

Herniation

Contour/Bulge

Herniation location

Facet joint OA

Synovial cyst(s)

Central canal stenosis

Right foramen stenosis

Left foramen stenosis

Right lateral recess stenosis

Left lateral recess stenosis

FINDINGS

T11-T12
- $X_1 X_2 X_3 X_4 X_5 X_6$
- $X_7 X_8 X_9 X_{10} X_{11}$

T12-L1
- $X_1 X_2 X_3 X_4 X_5 X_6$
- $X_7 X_8 X_9 X_{10} X_{11}$

L1-L2
- $X_1 X_2 X_3 X_4 X_5 X_6$
- $X_7 X_8 X_9 X_{10} X_{11}$

L2-L3
- $X_1 X_2 X_3 X_4 X_5 X_6$
- $X_7 X_8 X_9 X_{10} X_{11}$

L3-L4
- $X_1 X_2 X_3 X_4 X_5 X_6$
- $X_7 X_8 X_9 X_{10} X_{11}$

L4-L5
- $X_1 X_2 X_3 X_4 X_5 X_6$
- $X_7 X_8 X_9 X_{10} X_{11}$

L5-S1
- $X_1 X_2 X_3 X_4 X_5 X_6$
- $X_7 X_8 X_9 X_{10} X_{11}$

DEEP LEARNING FOR MEDICINE

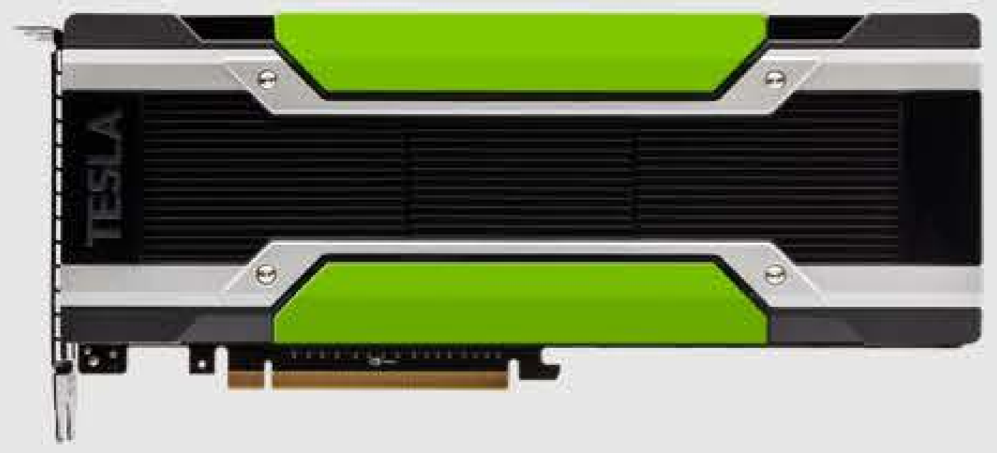
NVIDIA Founding Technology Partner of MGH Center of Clinical Data Science

10B Medical images on DGX-1 to advance radiology, pathology, genomics

TESLA FAMILY

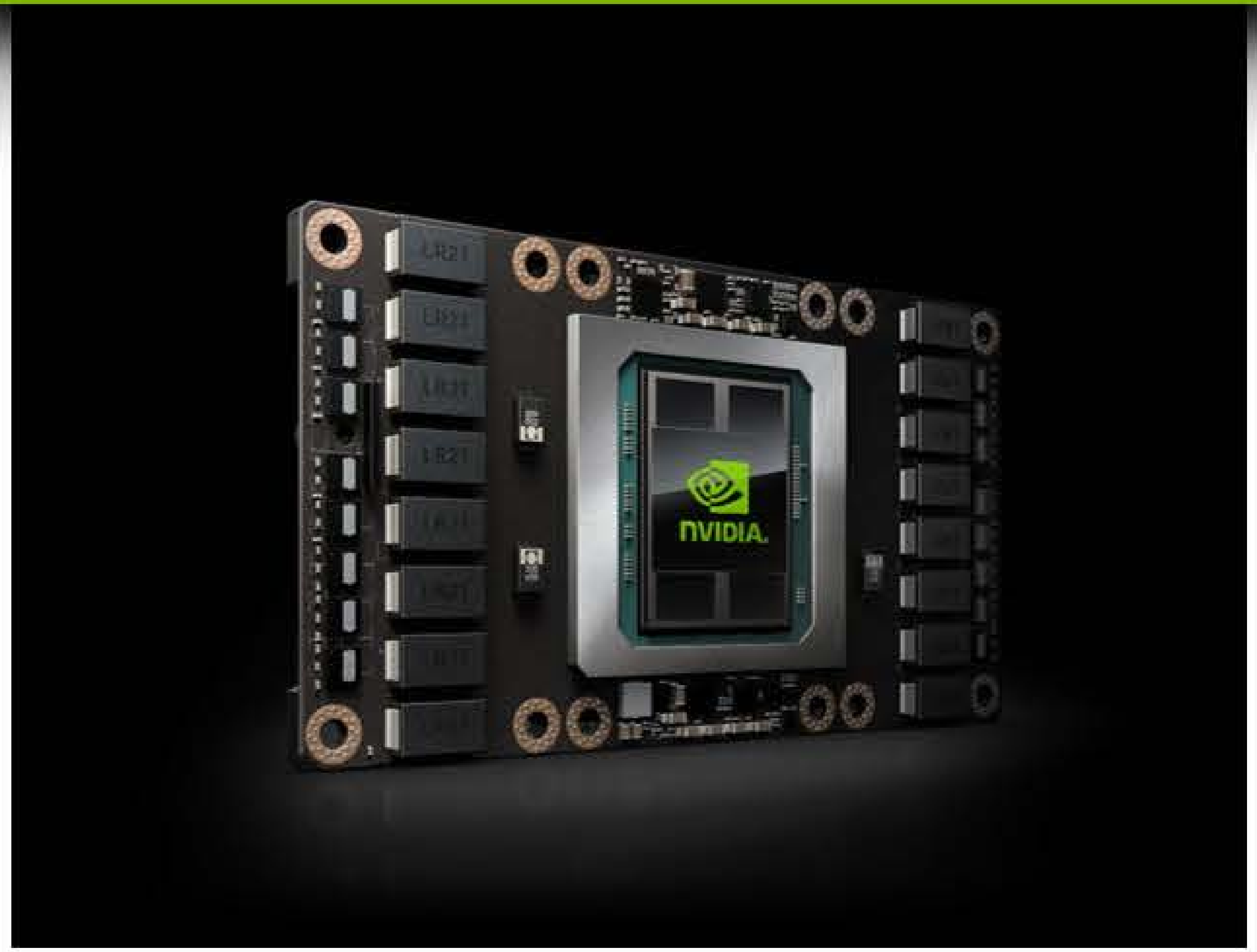
M40 + M4

Hyperscale HPC



K80

Multi-App HPC



Strong-Scale HPC

Researchers / Early Adopters

5 THINGS

NVIDIA DEVELOPER

NVIDIA SDK

The Essential Resource for GPU Developers

DEEP LEARNING
Deep Learning SDK
High-performance tools and libraries for deep learning

SELF-DRIVING CARS
NVIDIA DriveWorks
Deep learning, HD mapping and supercomputing solutions, from ADAS to fully autonomous

VIRTUAL REALITY
NVIDIA VRWorks™
A comprehensive SDK for VR headsets, games and professional applications

GAME DEVELOPME
NVIDIA GameWorks™
Advanced simulation technology for game

NVIDIA SDK

TESLA P100

NVIDIA DGX-1

IRAY VR

A Deep Learning Car

AN AMAZING YEAR FOR SELF-DRIVING CARS

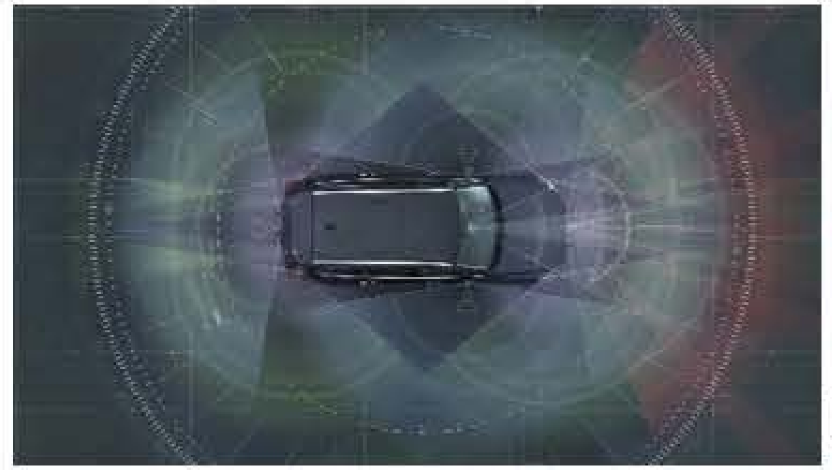
Audi, BMW, Daimler Buy HERE

Uber Enters the Race

Tesla Model S Auto-pilot

Toyota Invests $1B in AI Lab

Baidu Enters the Race



Volvo Drive Me on Public Roads in 2017

Honda, Nissan, Toyota Team Up

NHTSA: Computer Counts as Driver

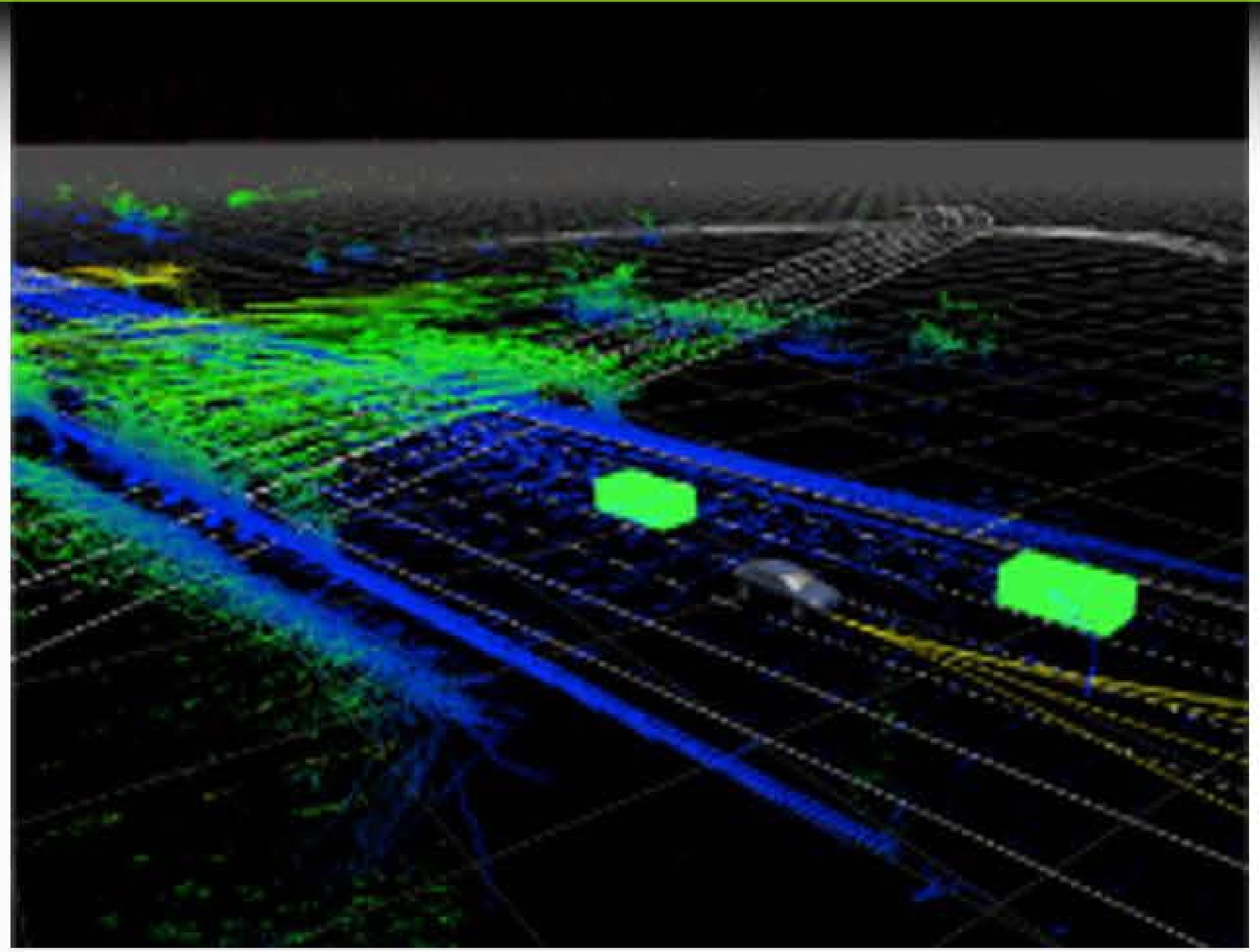
cruise
+
GENERAL MOTORS
JOIN THE DRIVERLESS REVOLUTION

GM Buys Cruise

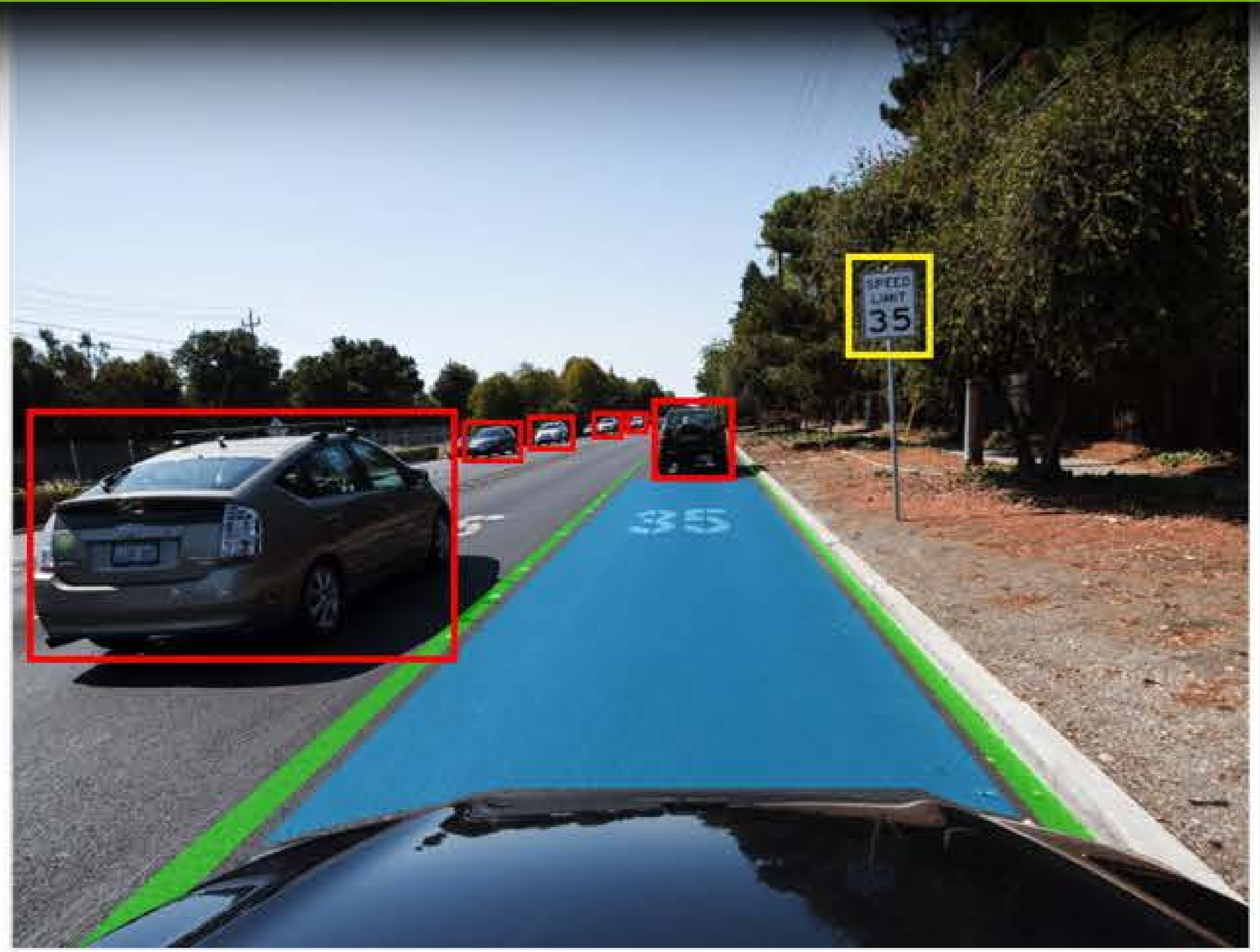
Tesla Model 3: 300K pre-orders

SELF-DRIVING LOOPS

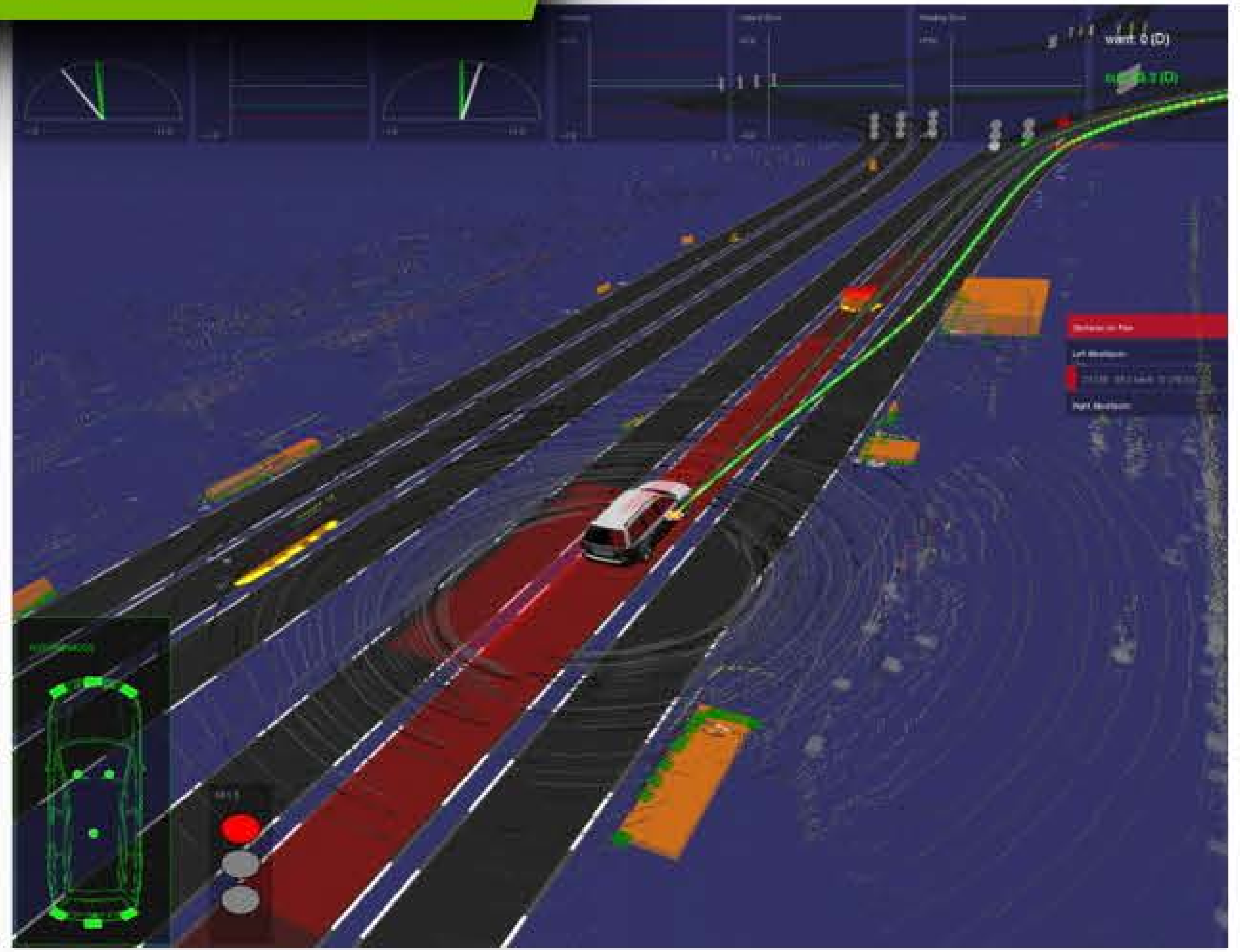
MAP



LOCALIZE

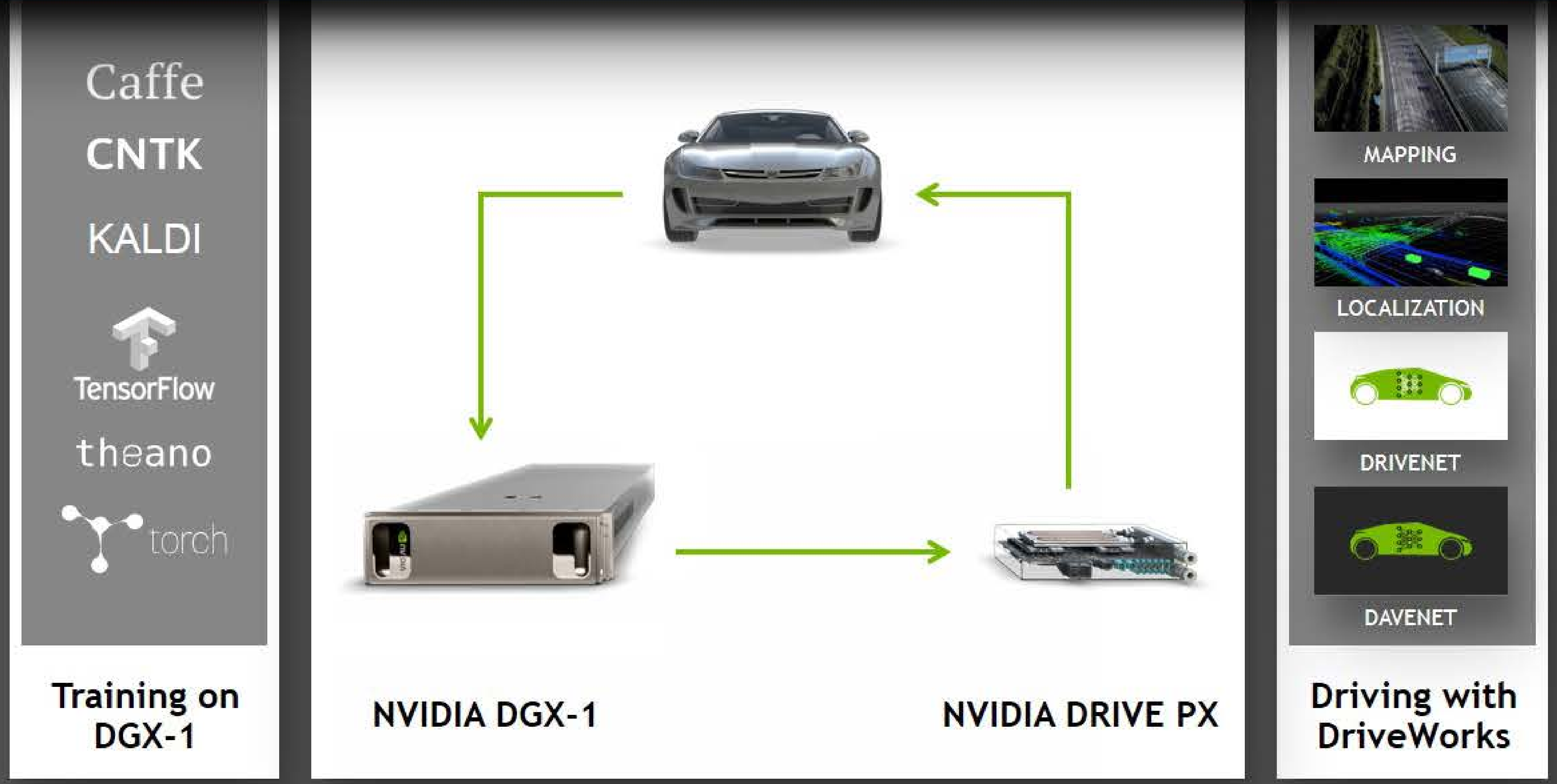


SEE



DRIVE

NVIDIA DRIVE PX
AI CAR COMPUTER



Caffe
CNTK
KALDI
TensorFlow
theano
torch

Training on DGX-1

NVIDIA DGX-1

NVIDIA DRIVE PX

MAPPING
LOCALIZATION
DRIVENET
DAVENET

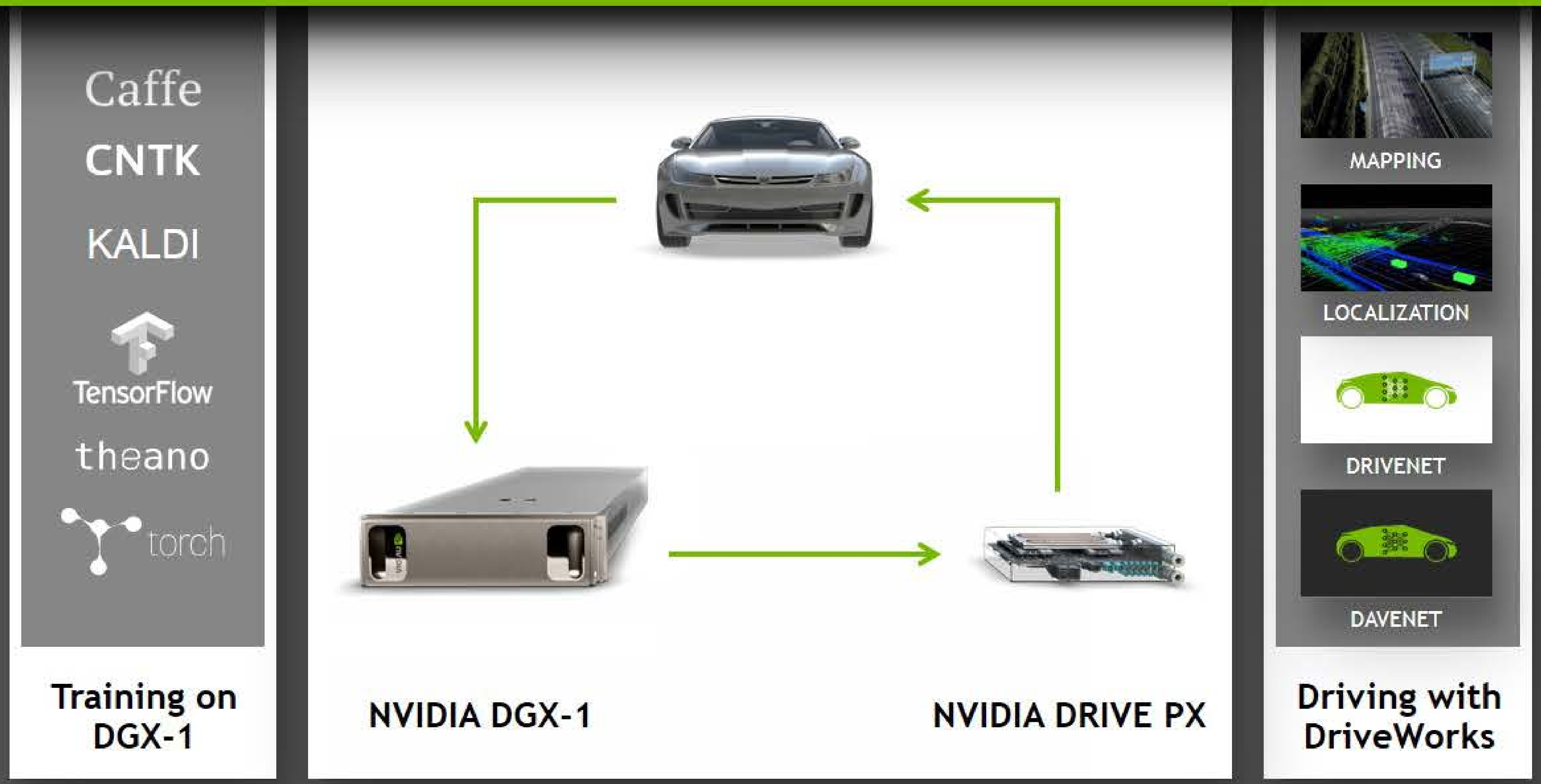
Driving with DriveWorks

World’s first DL-powered car computing platform

One scalable architecture — from DNN training to cluster, infotainment, ADAS, autonomous driving, and mapping

Open platform

NVIDIA DRIVE PX PERCEPTION



Caffe
CNTK
KALDI
TensorFlow
theano
torch

Training on DGX-1

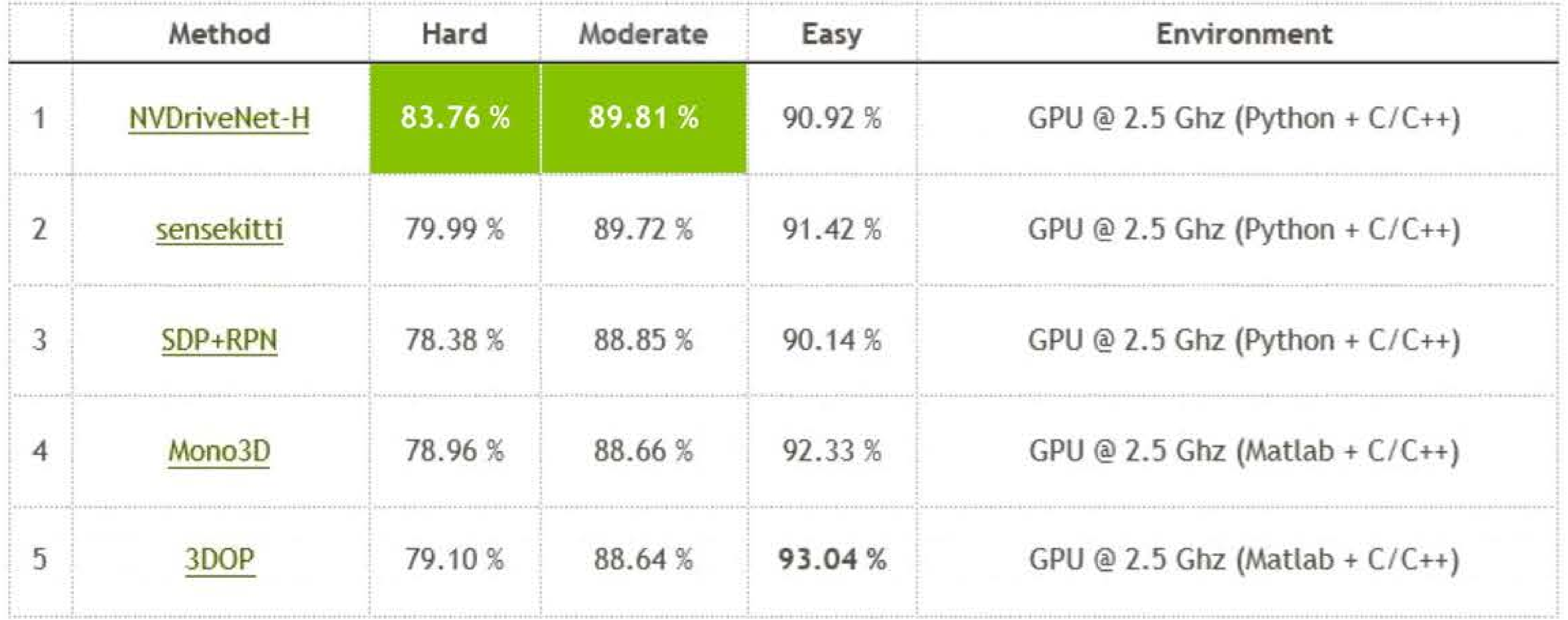
NVIDIA DGX-1

NVIDIA DRIVE PX

MAPPING
LOCALIZATION
DRIVENET
DAVENET

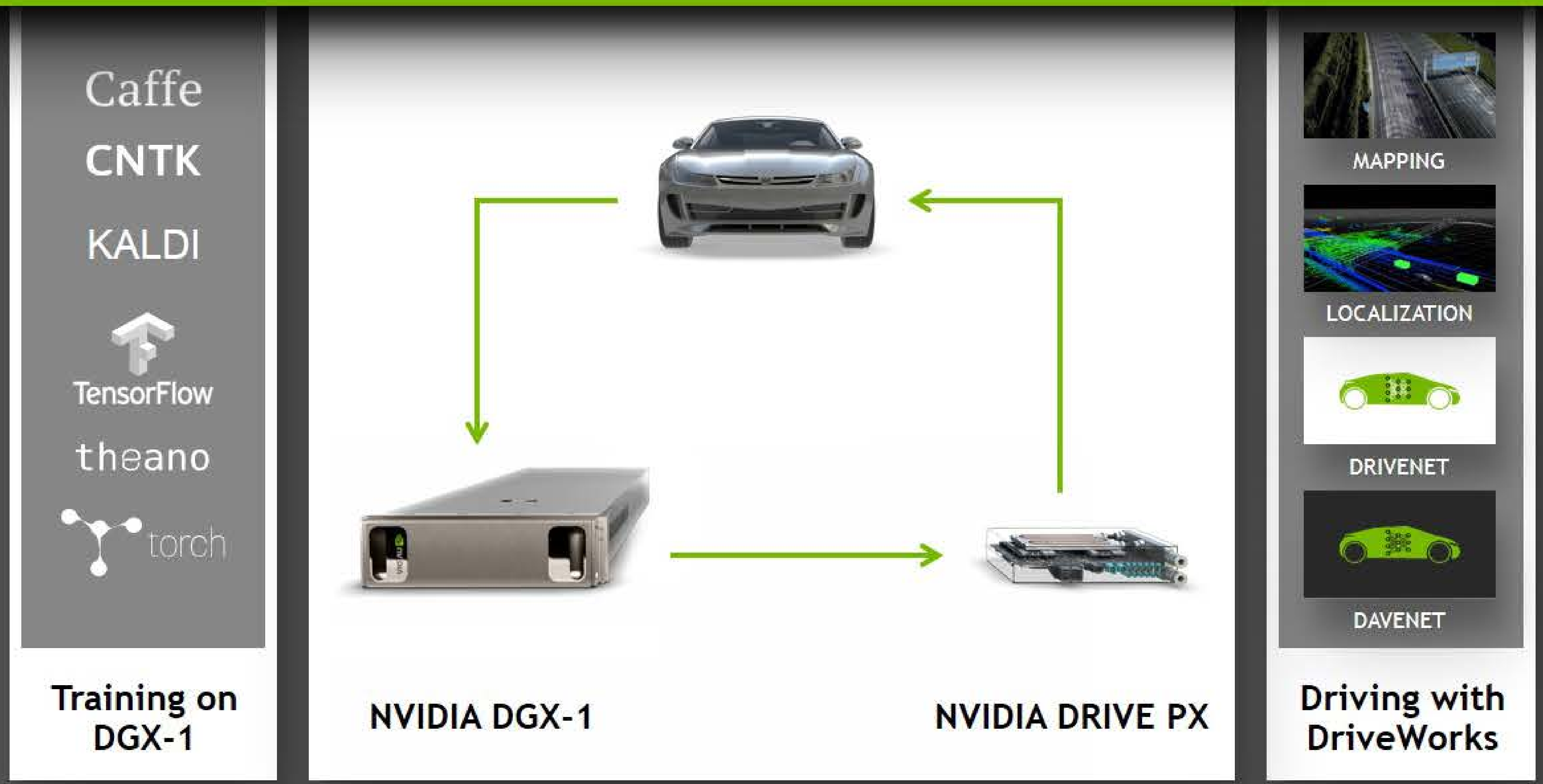
Driving with DriveWorks

NVIDIA DRIVENET
#1 accuracy score for KITTI car detection



	Method	Hard	Moderate	Easy	Environment
1	NVDriveNet-H	83.76 %	89.81 %	90.92 %	GPU @ 2.5 Ghz (Python + C/C++)
2	sensekitti	79.99 %	89.72 %	91.42 %	GPU @ 2.5 Ghz (Python + C/C++)
3	SDP+RPN	78.38 %	88.85 %	90.14 %	GPU @ 2.5 Ghz (Python + C/C++)
4	Mono3D	78.96 %	88.66 %	92.33 %	GPU @ 2.5 Ghz (Matlab + C/C++)
5	3DOP	79.10 %	88.64 %	**93.04 %**	GPU @ 2.5 Ghz (Matlab + C/C++)

NVIDIA DRIVE PX PERCEPTION



Caffe
CNTK
KALDI
TensorFlow
theano
torch

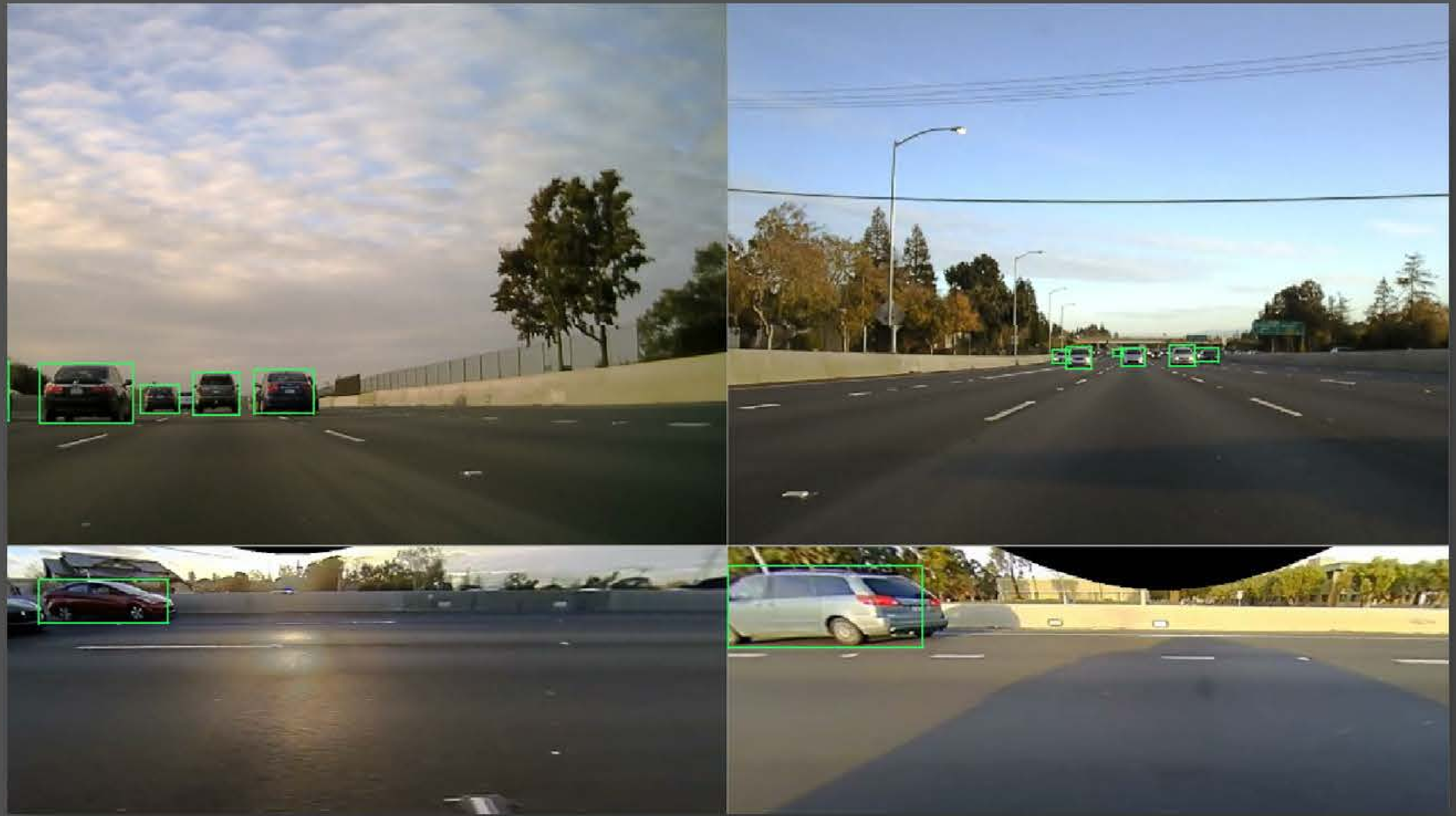
Training on DGX-1

NVIDIA DGX-1

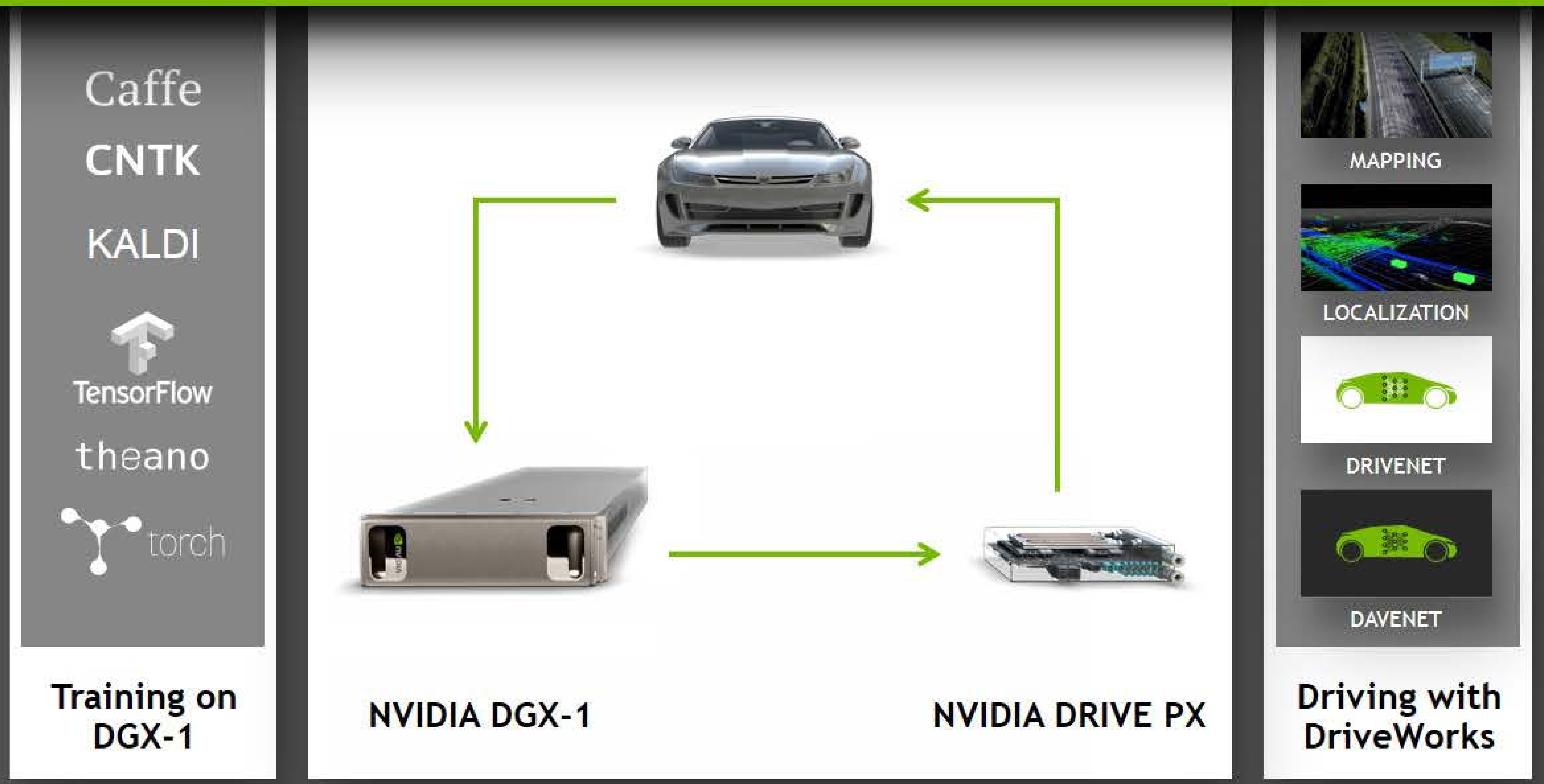
NVIDIA DRIVE PX

MAPPING
LOCALIZATION
DRIVENET
DAVENET

Driving with DriveWorks



NEW
END-TO-END HD MAPPING



Caffe

CNTK

KALDI

TensorFlow

theano

torch

Training on DGX-1

NVIDIA DGX-1

NVIDIA DRIVE PX

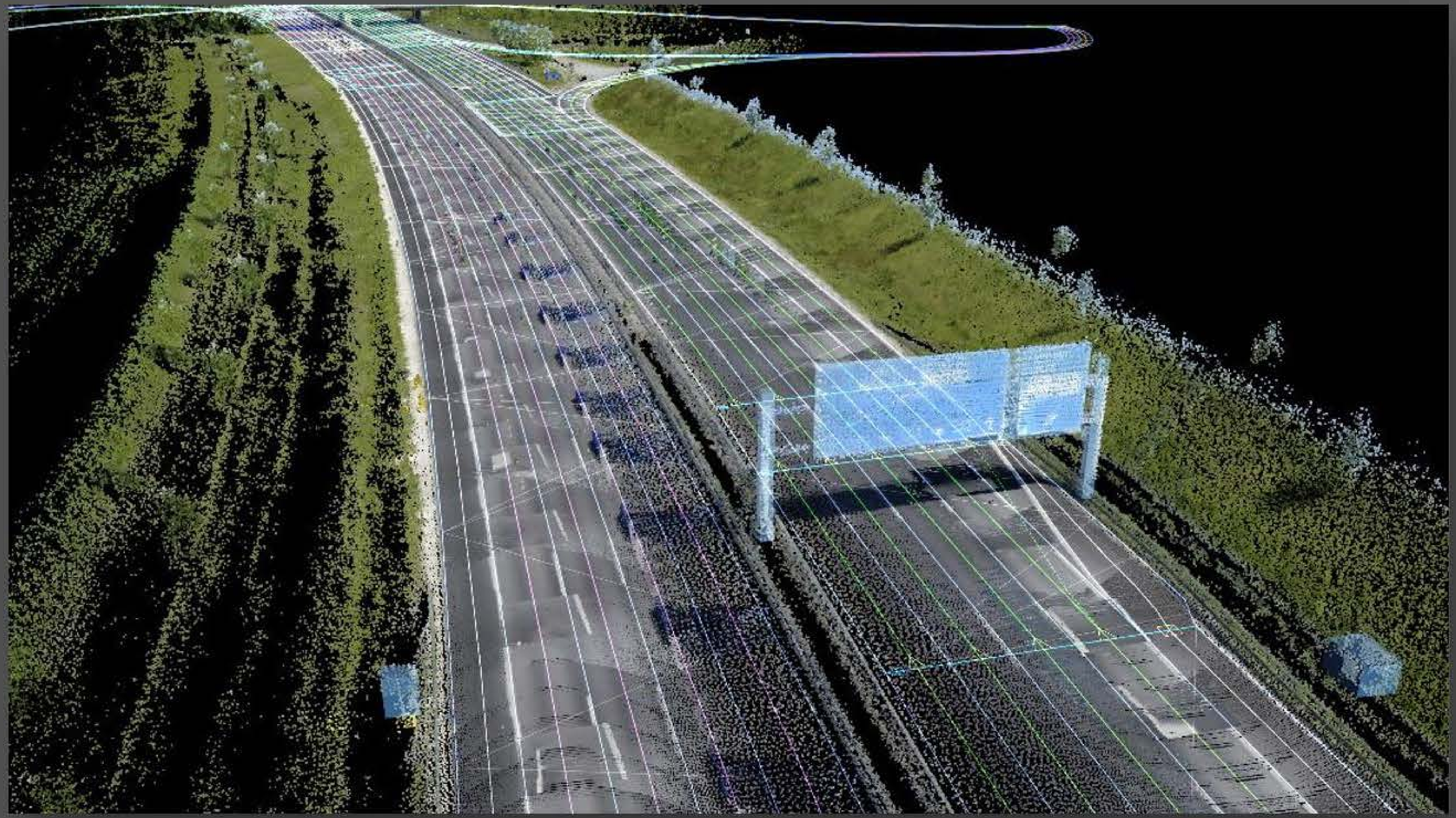
MAPPING

LOCALIZATION

DRIVENET

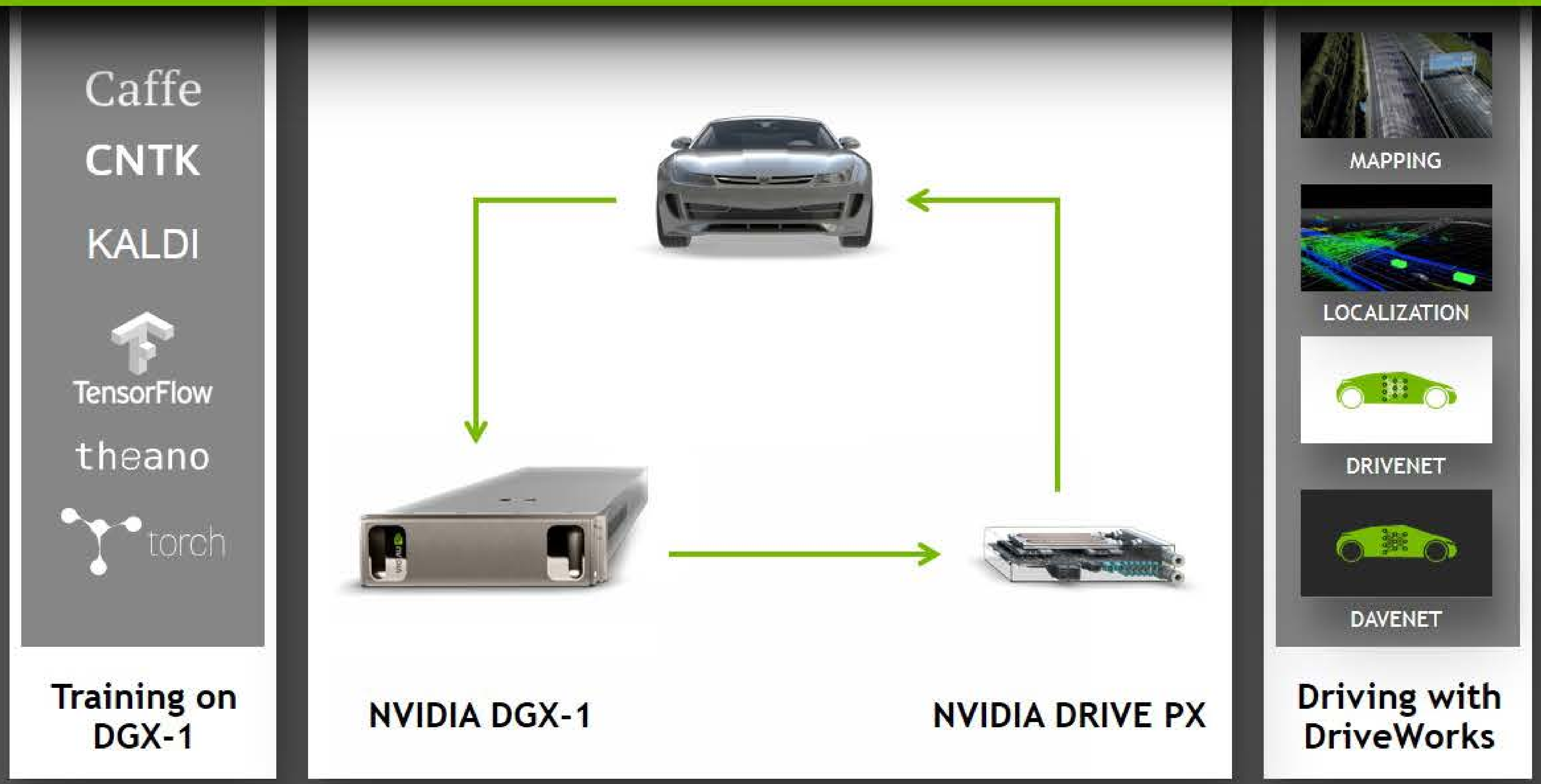
DAVENET

Driving with DriveWorks



BAIDU SELF-DRIVING CAR COMPUTER

NEW
END-TO-END HD MAPPING



Caffe

CNTK

KALDI

TensorFlow

theano

torch

Training on DGX-1

NVIDIA DGX-1

NVIDIA DRIVE PX

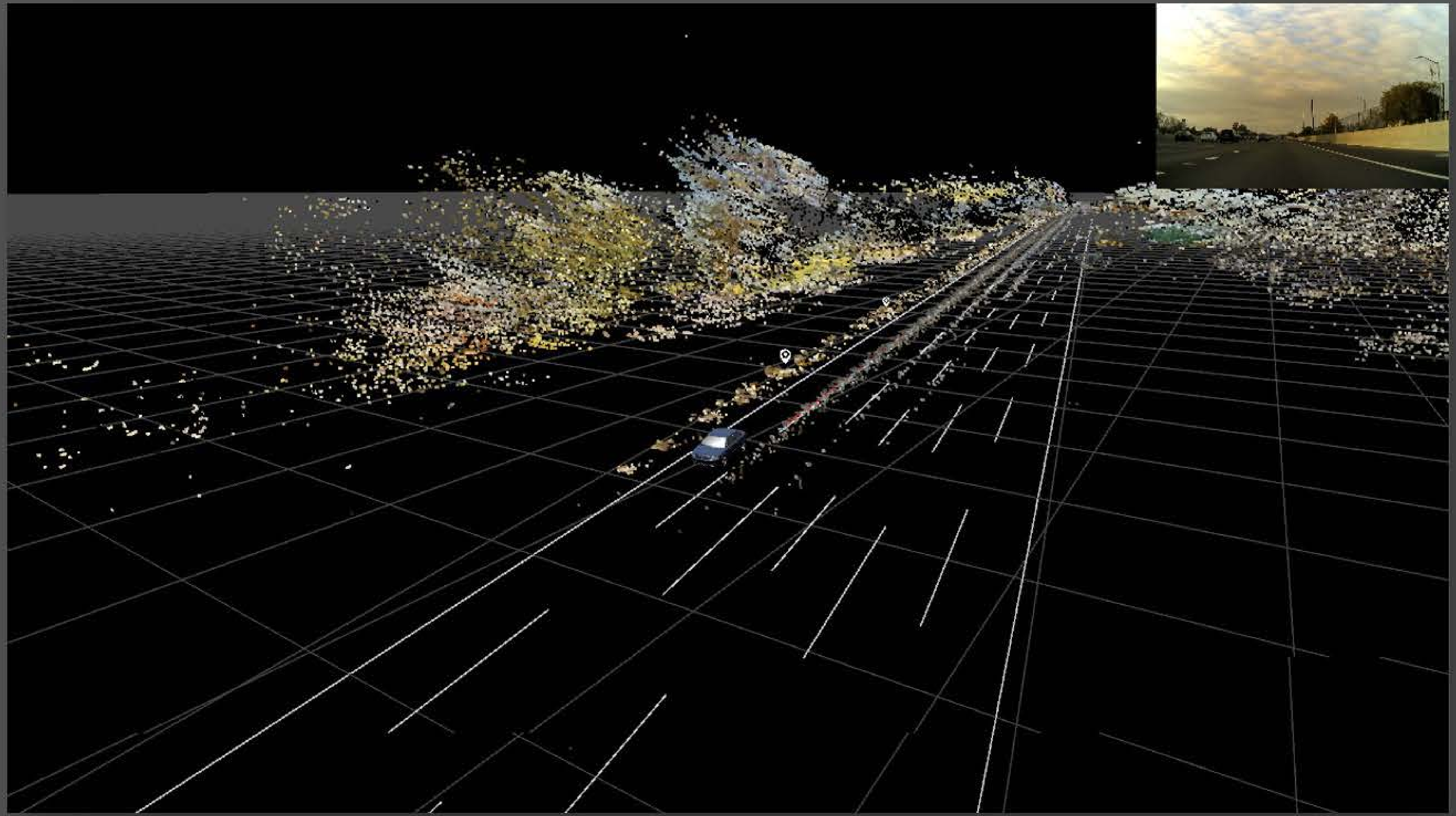
MAPPING

LOCALIZATION

DRIVENET

DAVENET

Driving with DriveWorks



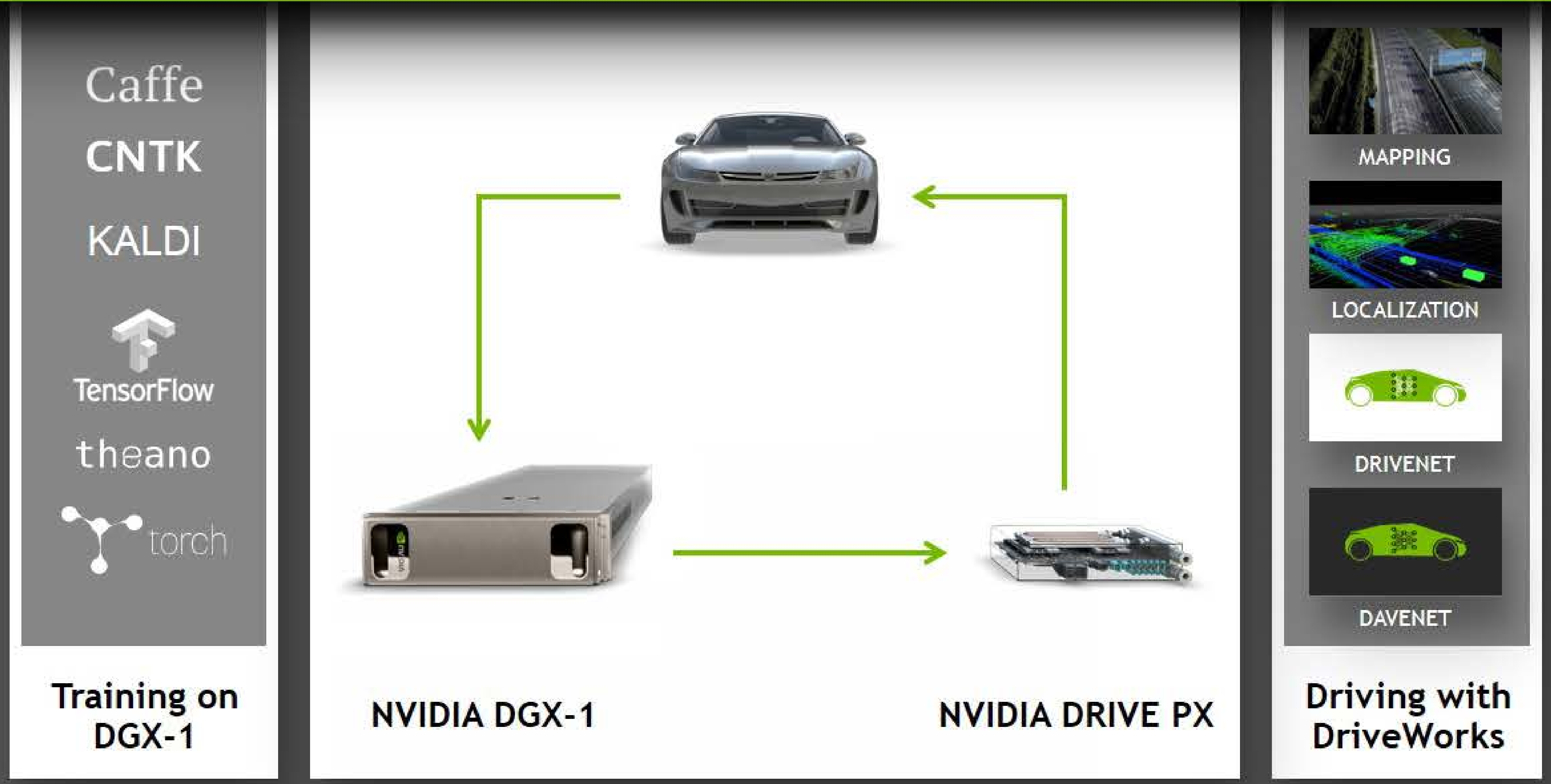
PLATFORM FOR MAPPING THE WORLD

here

TOMTOM®

ZENRIN

NEW AI DRIVING



Caffe

CNTK

KALDI

TensorFlow

theano

torch

Training on DGX-1

NVIDIA DGX-1

NVIDIA DRIVE PX

MAPPING

LOCALIZATION

DRIVENET

DAVENET

Driving with DriveWorks

NVIDIA
roborace
roborace

WORLD’S FIRST AUTONOMOUS RACE CAR

Designed by Daniel Simon

2,200 lbs

Blazing fast

WORLD'S FIRST AUTONOMOUS CAR RACE

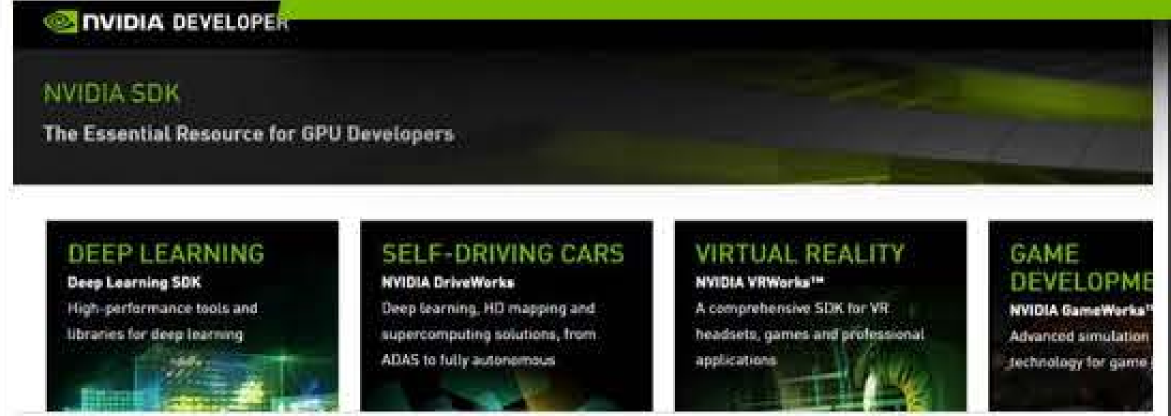
10 teams, 20 identical cars

DRIVE PX 2: The “brain” of every car

2016/17 Formula E season

ROBORACE

VR, AI, SELF-DRIVING CARS



NVIDIA DEVELOPER

NVIDIA SDK

The Essential Resource for GPU Developers

DEEP LEARNING
Deep Learning SDK
High-performance tools and libraries for deep learning

SELF-DRIVING CARS
NVIDIA DriveWorks
Deep learning, HD mapping and supercomputing solutions, from ADAS to fully autonomous

VIRTUAL REALITY
NVIDIA VRWorks™
A comprehensive SDK for VR headsets, games and professional applications

GAME DEVELOPME
NVIDIA GameWorks™
Advanced simulation technology for game

NVIDIA SDK

TESLA P100

NVIDIA DGX-1

IRAY VR

HD MAPPING, AI DRIVING

GPU TECHNOLOGY CONFERENCE