ALISON B LOWNDES

AI DevRel | EMEA

@alisonblowndes

August 2018

NVIDIA

The day job

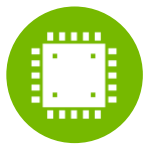
AUTOMOTIVE
Auto sensors reporting location, problems

COMMUNICATIONS
Location-based advertising

CONSUMER PACKAGED GOODS
Sentiment analysis of what's hot, problems

FINANCIAL SERVICES
Risk & portfolio analysis
New products

EDUCATION & RESEARCH
Experiment sensor analysis



HIGH TECHNOLOGY / INDUSTRIAL MFG.
Mfg. quality
Warranty analysis

LIFE SCIENCES

MEDIA/ENTERTAINMENT
Viewers / advertising effectiveness

ON-LINE SERVICES / SOCIAL MEDIA
People & career matching

HEALTH CARE
Patient sensors, monitoring, EHRs

OIL & GAS
Drilling exploration sensor analysis

RETAIL
Consumer sentiment

TRAVEL & TRANSPORTATION
Sensor analysis for optimal traffic flows

UTILITIES
Smart Meter analysis for network capacity,

LAW ENFORCEMENT & DEFENSE
Threat analysis - social media monitoring, photo analysis

www.FrontierDevelopmentLab.org

NASA
FRONTIER
DEVELOPMENT LAB

SETI INSTITUTE

NASA

esa

NVIDIA

IBM

LOCKHEED MARTIN

FDL EUROPE

A MISSION CONTROL
FOR SPACESHIP EARTH

esa

EARTH OBSERVATION

KICK OFF WORKSHOP
ESA ESRIN, ROME
25th - 29th June ‘18

RESEARCH SPRINT
30th June - 17th August

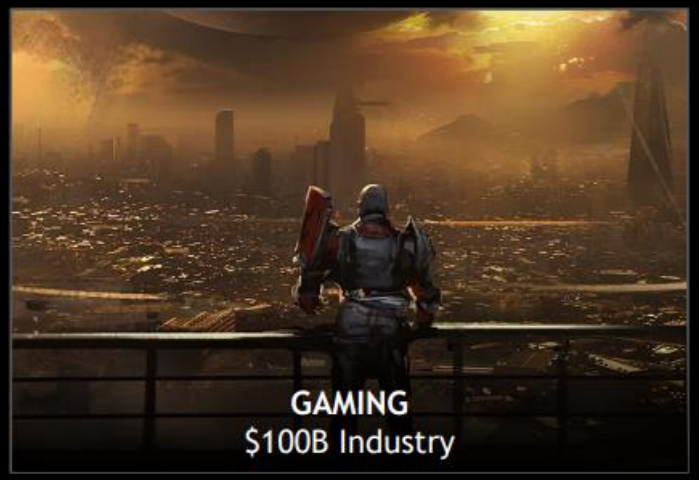
IBM

NVIDIA

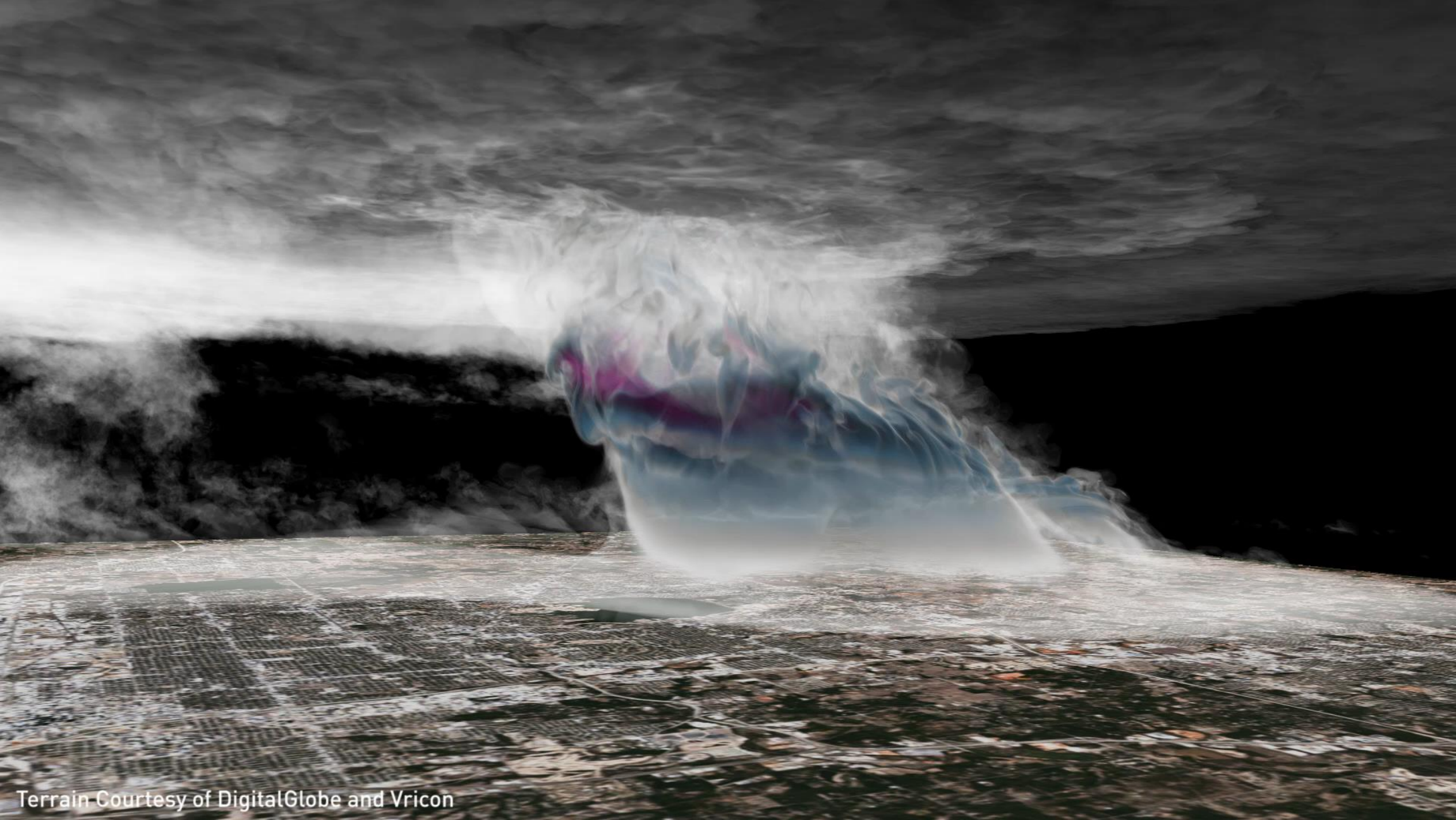
UNIVERSITY OF OXFORD

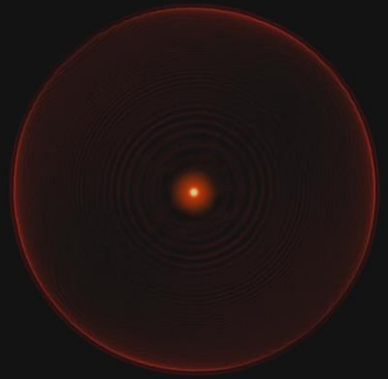
CATAPULT
Satellite Applications

THE MOST EXCITING TIME IN TECH HISTORY



NVIDIA GPU

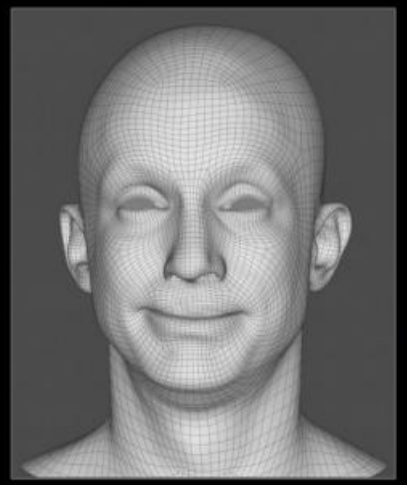

Terrain Courtesy of DigitalGlobe and Vricon




1
450

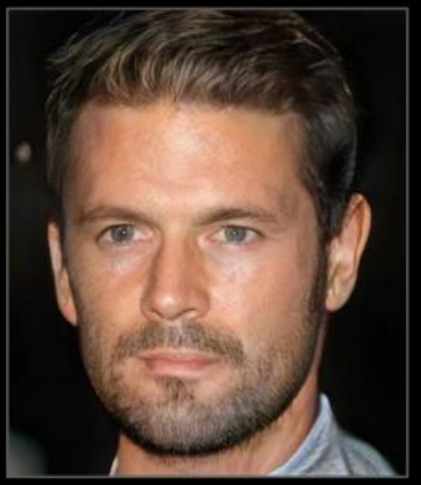
NVIDIA RESEARCH

NVIDIA Research
AI Autoencoder



NVIDIA Research / Remedy
Audio-driven Facial Animation

NVIDIA Research
Semantic Manipulation with GANs



NVIDIA Research
Progressive GAN

NVIDIA Research / AIVA
RNNs for Music

CUTLASS

http://github.com/NVIDIA/cutlass

https://devblogs.nvidia.com/cutlass-linear-algebra-cuda/

CUTLASS v1.0 for CUDA 9.2

CUDA Templates for Linear Algebra Subroutines

Optimised implementation of dense matrix products (GEMM)

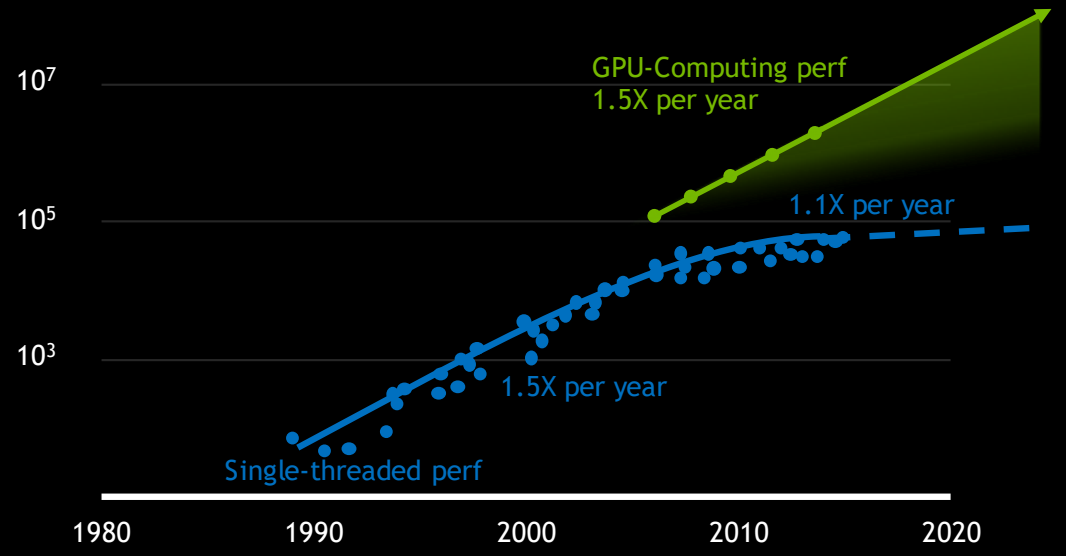
Hierarchical decomposition into modular classes

Fast kernels for deep learning

THE RISE OF GPU COMPUTING

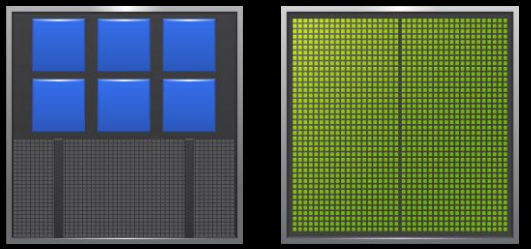
Big Data Needs Algorithms and Compute That Scales

END OF MOORE'S LAW

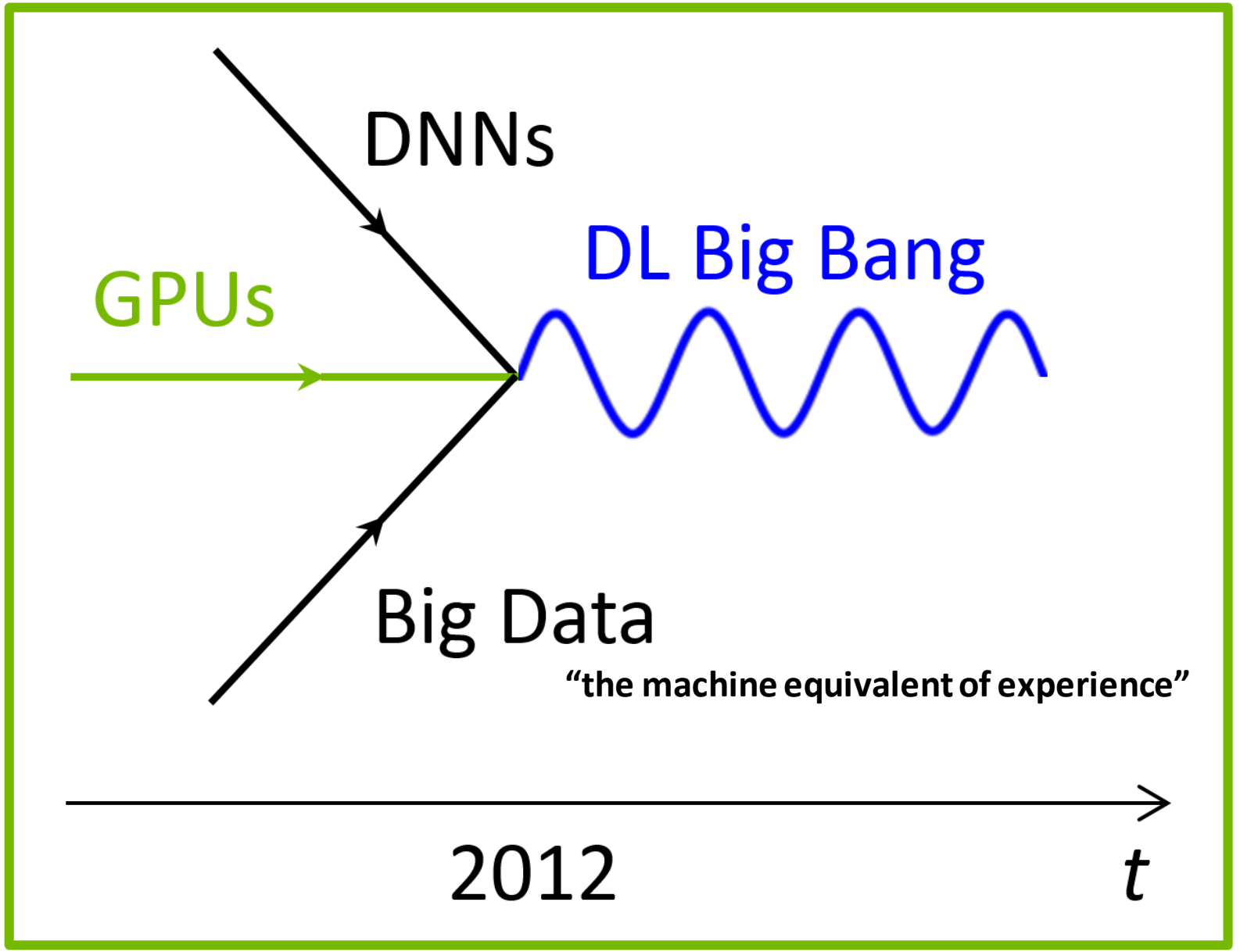


Original data up to the year 2010 collected and plotted by M. Horowitz,
F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten Newplot and data collected for 2010-2015 by K. Rupp

CPU vs. GPU

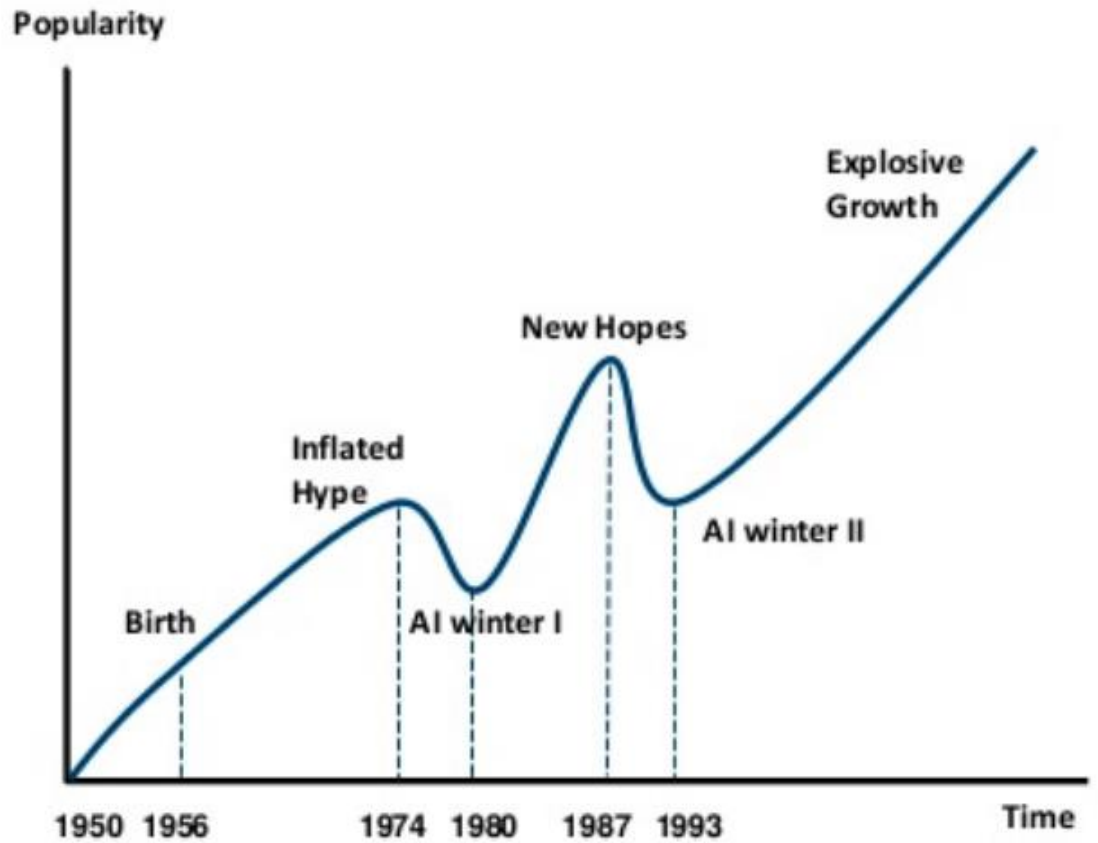


NVIDIA

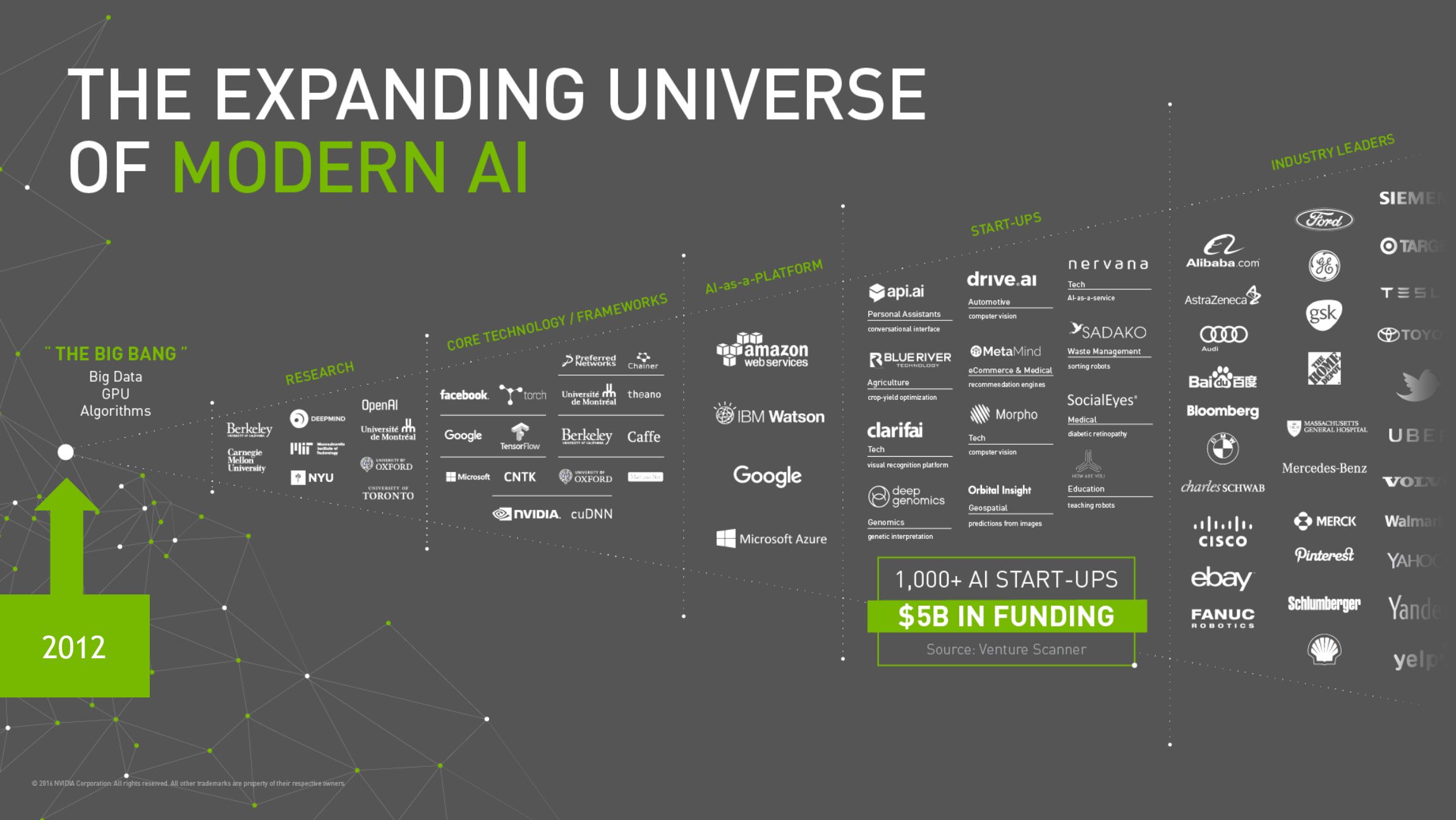


Feynman Diagram of DL, © NVIDIA 2016, author Branislav Kisačanin

History of Artificial Intelligence Hype



THE EXPANDING UNIVERSE OF MODERN AI



NVIDIA INCEPTION PROGRAM

Accelerates AI startups with a boost of GPU tools, tech and deep learning expertise

www.nvidia.com/inception

Startup Qualifications

Driving advances in the field of AI

Business plan

Incorporated

Web presence

Technology

DL startup kit*

- Pascal Titan X
- Deep Learning Institute (DLI) credit
- Connect with a DL tech expert

DGX-1 ISV discount*

Software release notification

Live webinar and office hours

**By application*

Marketing

Inclusion in NVIDIA marketing efforts

GPU Technology Conference (GTC) discount

Emerging Company Summit (ECS) participation+

Marketing kit

- One-page story template
- eBook template
- Inception web badge and banners
- Social promotion request form
- Event opportunities list

Promotion at industry events

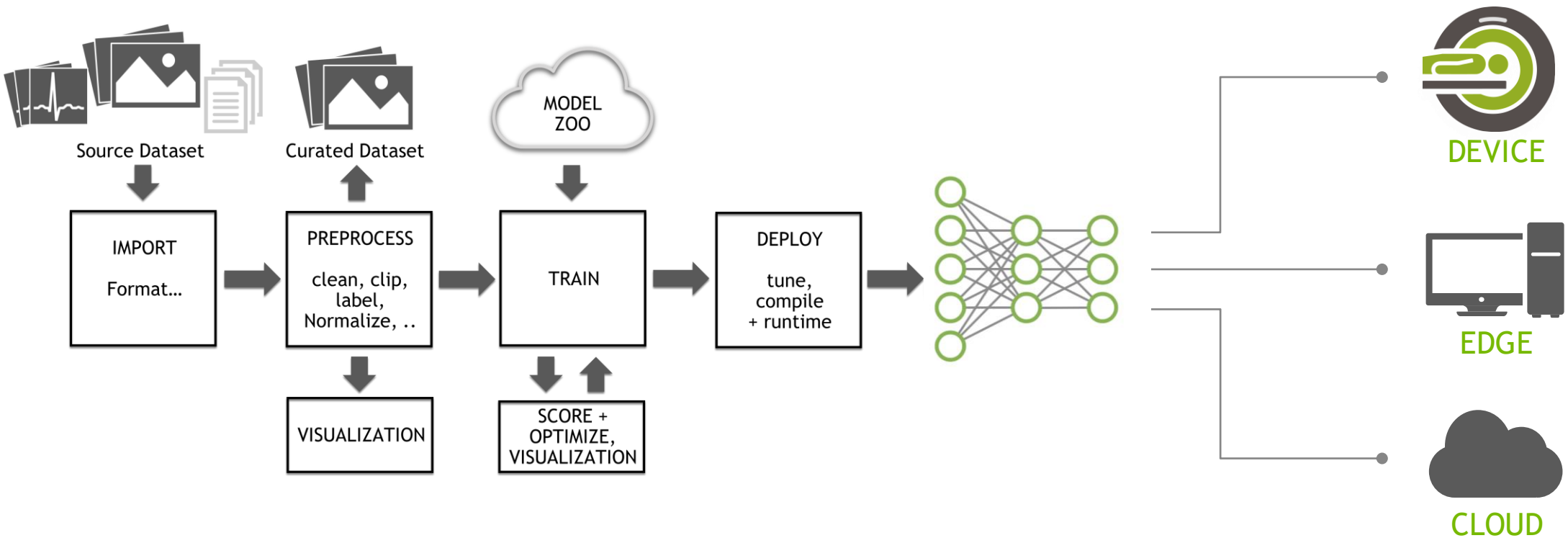
GPU ventures+

+By invitation

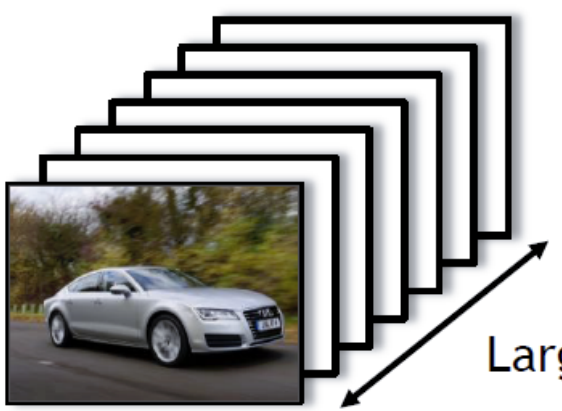
NVIDIA

DEEP LEARNING

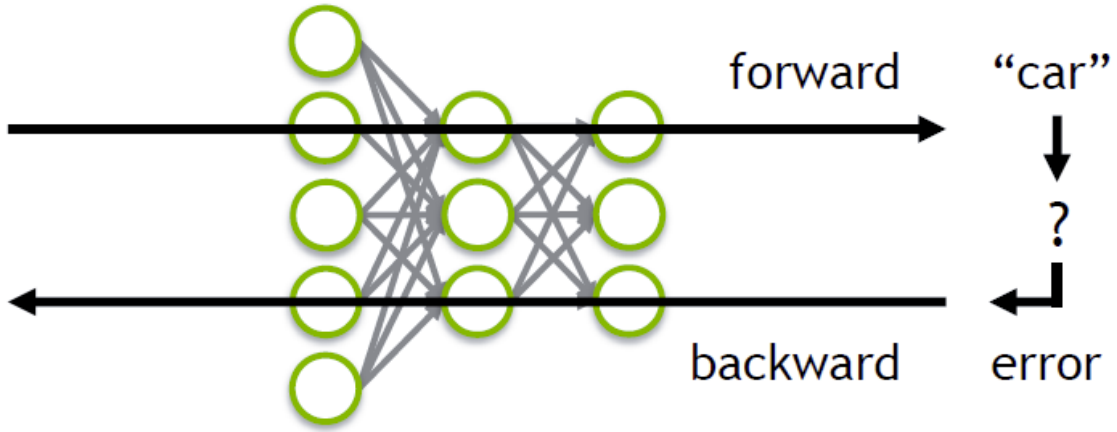
NEW PROGRAMMING MODEL



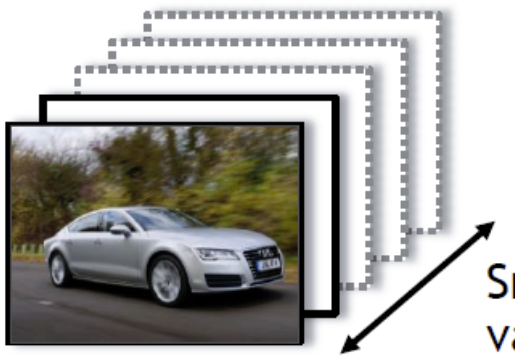
TRAINING VS INFERENCE



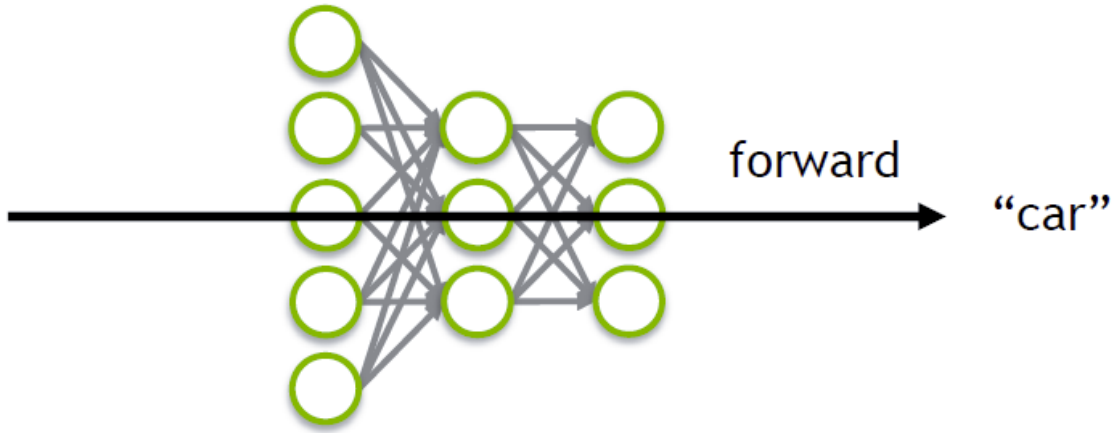
Large N



TRAINING

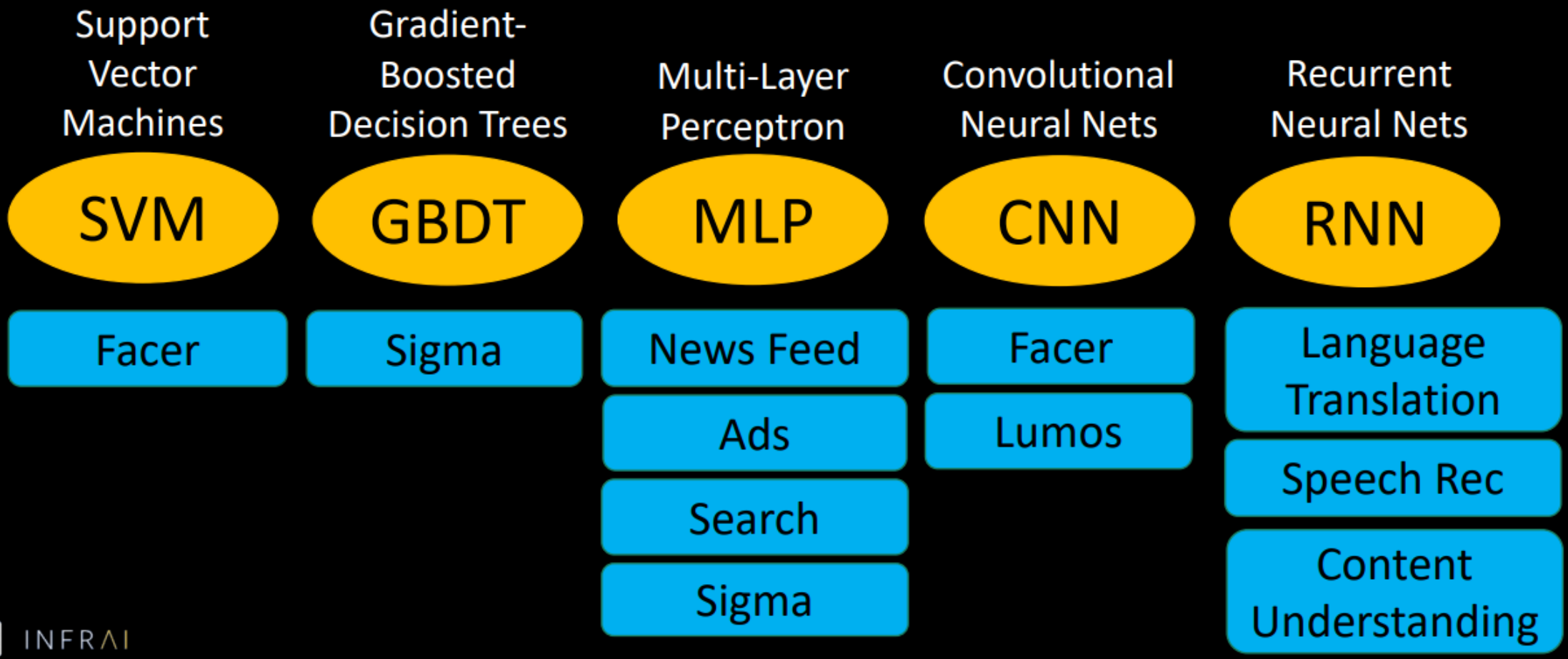


Smaller,
varied N



INFERENCE

What ML Models Do We Leverage?

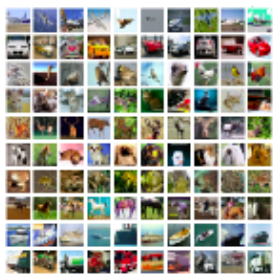


Yandex

https://catboost.yandex

Learn data

Images



CNN

Sequence

Text, DNA

RNN

Ordered features

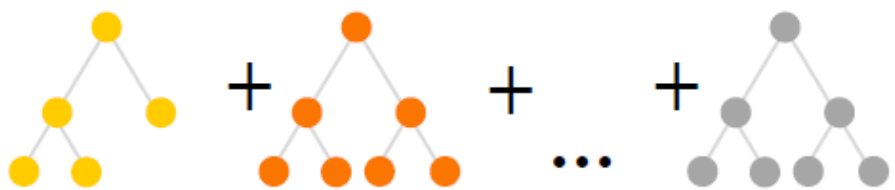
〉 Music album release year

1960 < 1970 < 1980

Gradient boosted decision trees

Categorical features

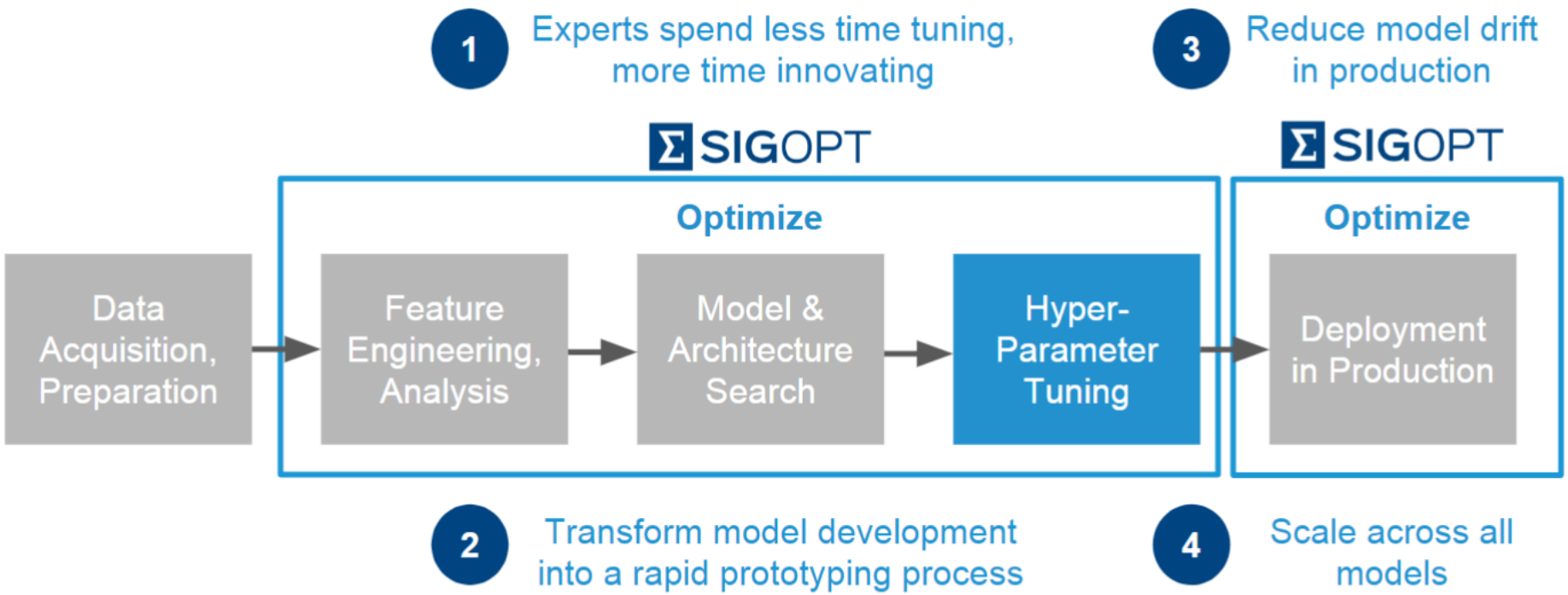
CatBoost: Categorical + Boosting



SOME KEY DECISIONS TO MAKE

FACTOR	DESCRIPTION
DL Challenge	Supervised or unsupervised, classification or regression, # of labels?
Architecture	What is the simplest architecture I can use?
Training Model	How am I going to tune my neural net? Kinds of non-linearity, loss function and weight initialization? Best training framework?
Data Quantity	How much data will be sufficient to train my model? How do I go about finding that data and is it evenly balanced?
Data Quality	Is my data directly relevant to the problem & real world data.
Data Labels	Is training data is labeled same as raw data sets, how do I 'featurize'?
Data Similarity	Is data same length vectors or does it require pre-processing?
Data Storage & Access	Where is it stored, locally and on network Data pipeline? How do I plan to extract, transform and load the data (ETL)?
Infrastructure	Cloud, On-premise, Hybrid. GPUs, CPUs or both? Single or distributed systems? Integration with languages, ent. apps/ databases.

Accelerate model development

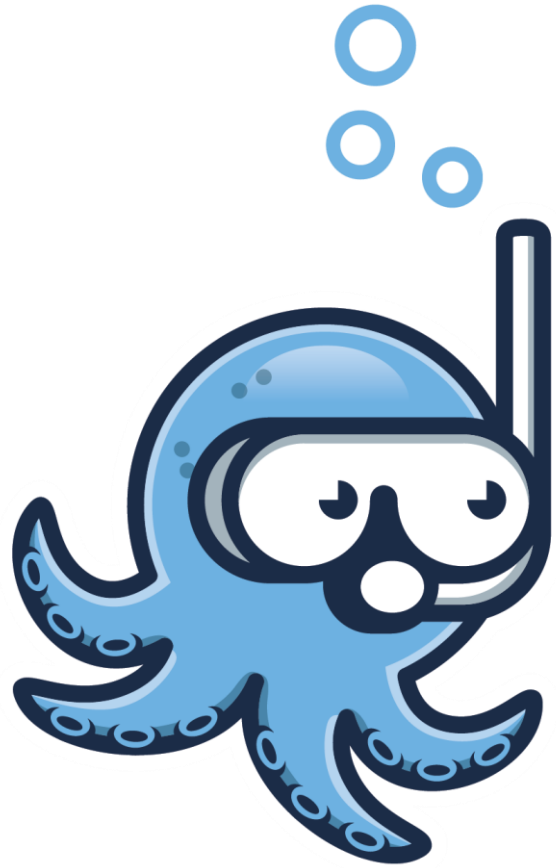


SIGOPT

Hyperparameter tuning as a service: https://sigopt.com/

Snorkel

https://github.com/HazyResearch/snorkel

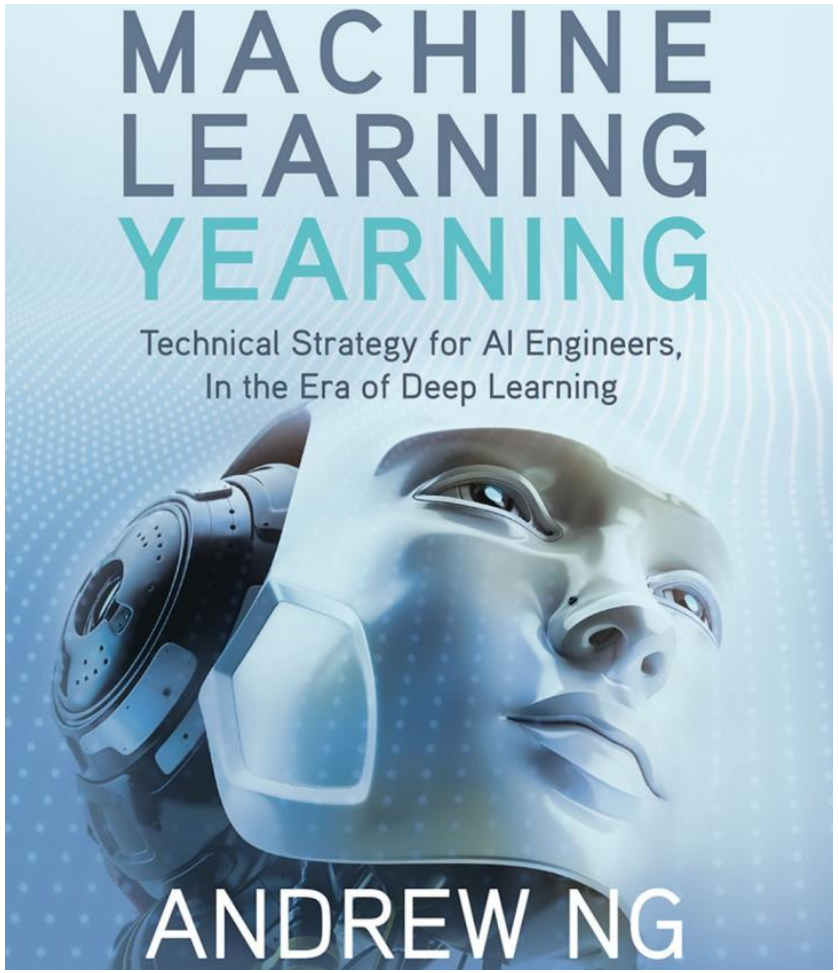


Automation of labelling data
A system for rapidly **creating, modeling, and managing training data**, For domains in which large labelled training sets are not available or easy to obtain.
Learning, essentially, which labelling functions are more accurate than others—and then using this to train a DNN
A general framework for many *weak supervision* techniques.

snorkel

“After finishing this book, you will have a deep understanding of how to set technical direction for a machine learning project.”

https://bit.ly/2wNFYlY



<title>code ninja</title>

PLASTER

Programmability, Latency, Accuracy, Size of Model,
Throughput, Energy Efficiency, Rate of Learning

https://images.nvidia.com/content/pdf/plaster-deep-learning-framework.pdf

https://devblogs.nvidia.com/parallelforall/inside-volta/

TESLA V100 32GB

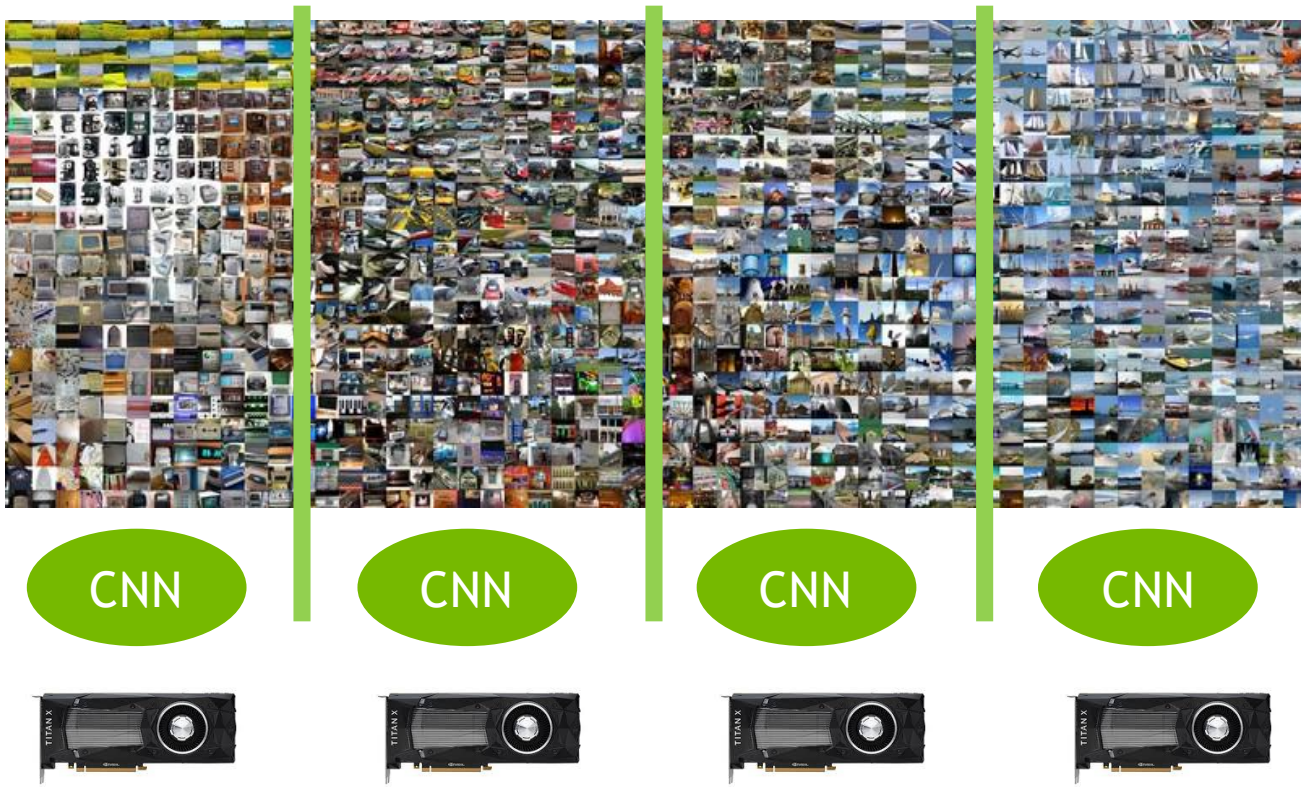
THE MOST ADVANCED DATA CENTER GPU EVER BUILT

5,120 CUDA cores
640 NEW Tensor cores
7.5 FP64 TFLOPS | 15 FP32 TFLOPS

120 Tensor TFLOP

20MB SM RF | 16MB Cache | 32GB HBM2 @ 900 GB/s
300 GB/s NVLink

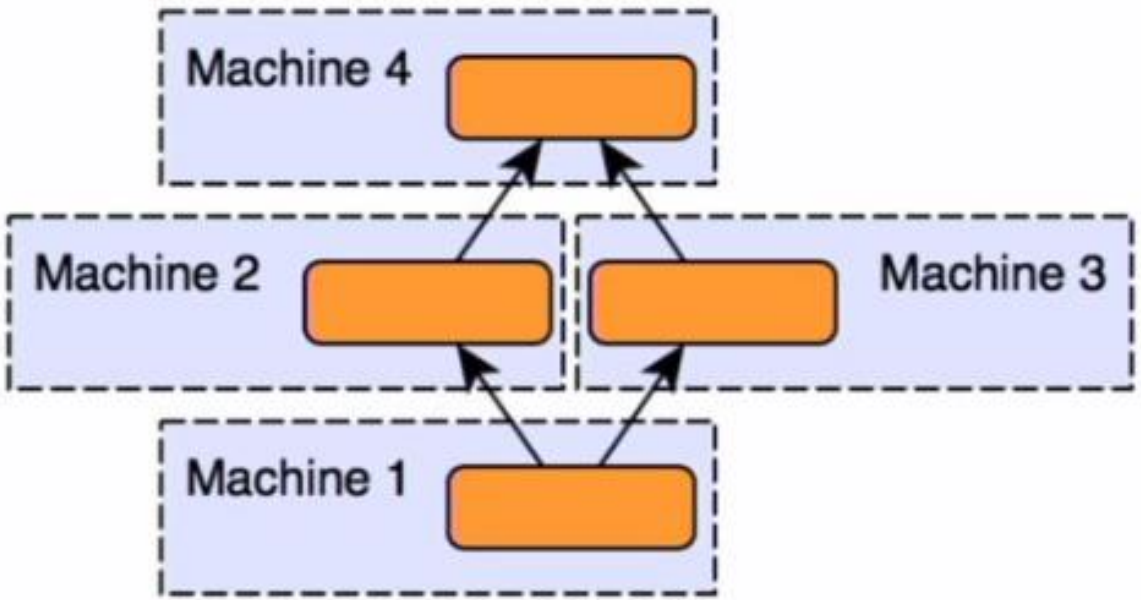
Data Parallelism



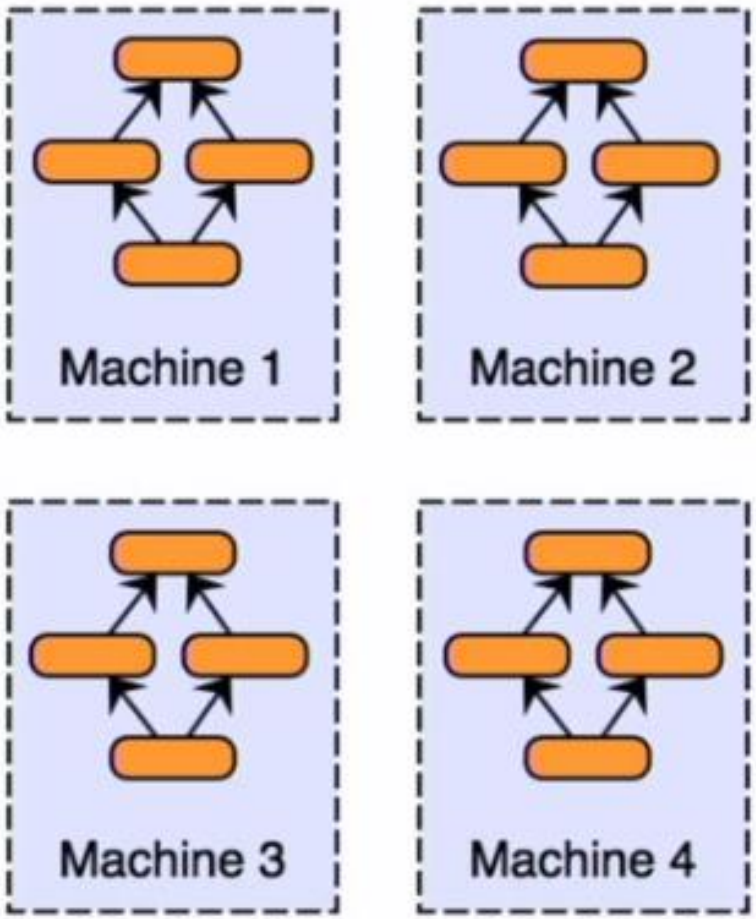
Model is replicated on ‘P’ GPUs, which requires communication.

1. Broadcast model.
2. Forward-Backward pass on each GPU with 1/P-th of batch.
3. Reduce model (sum weights).

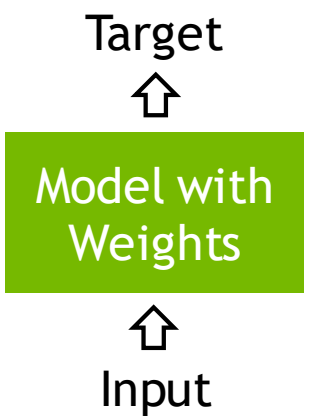
Model Parallelism



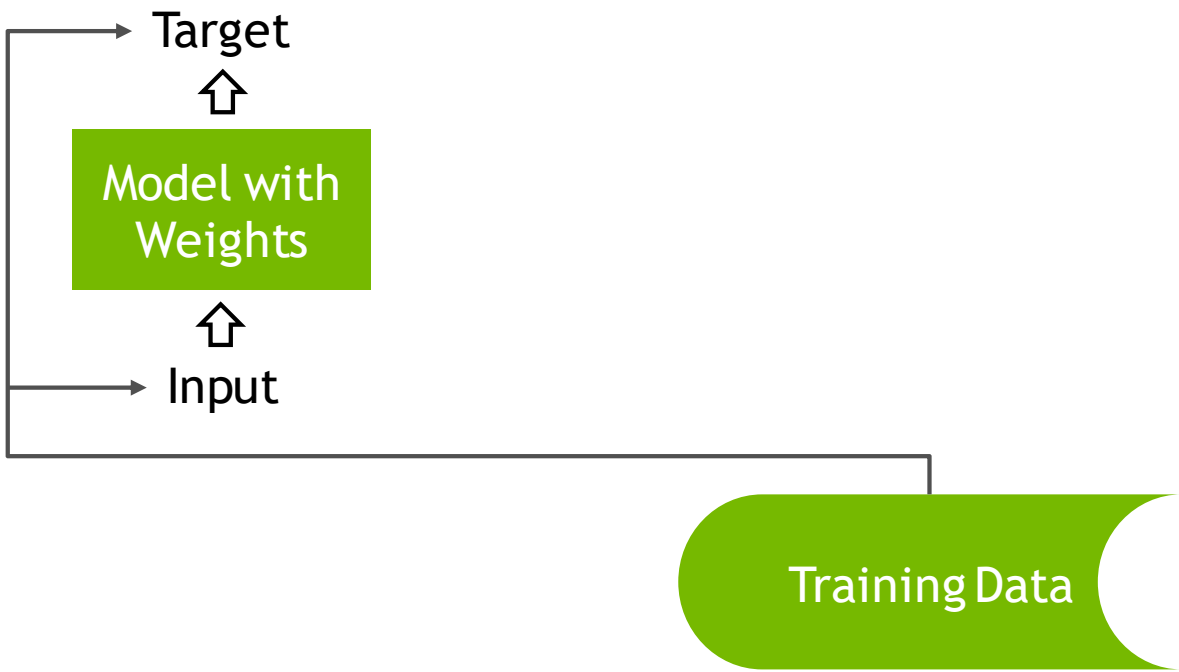
Data Parallelism



DEEP LEARNING TRAINING

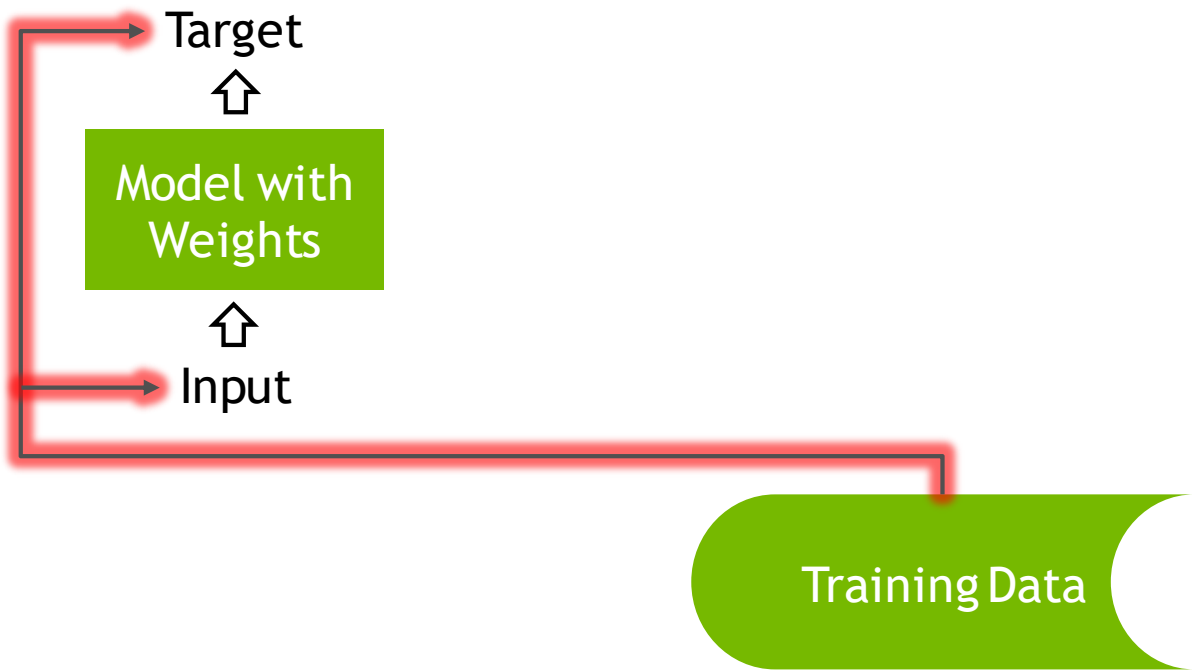


DEEP LEARNING TRAINING



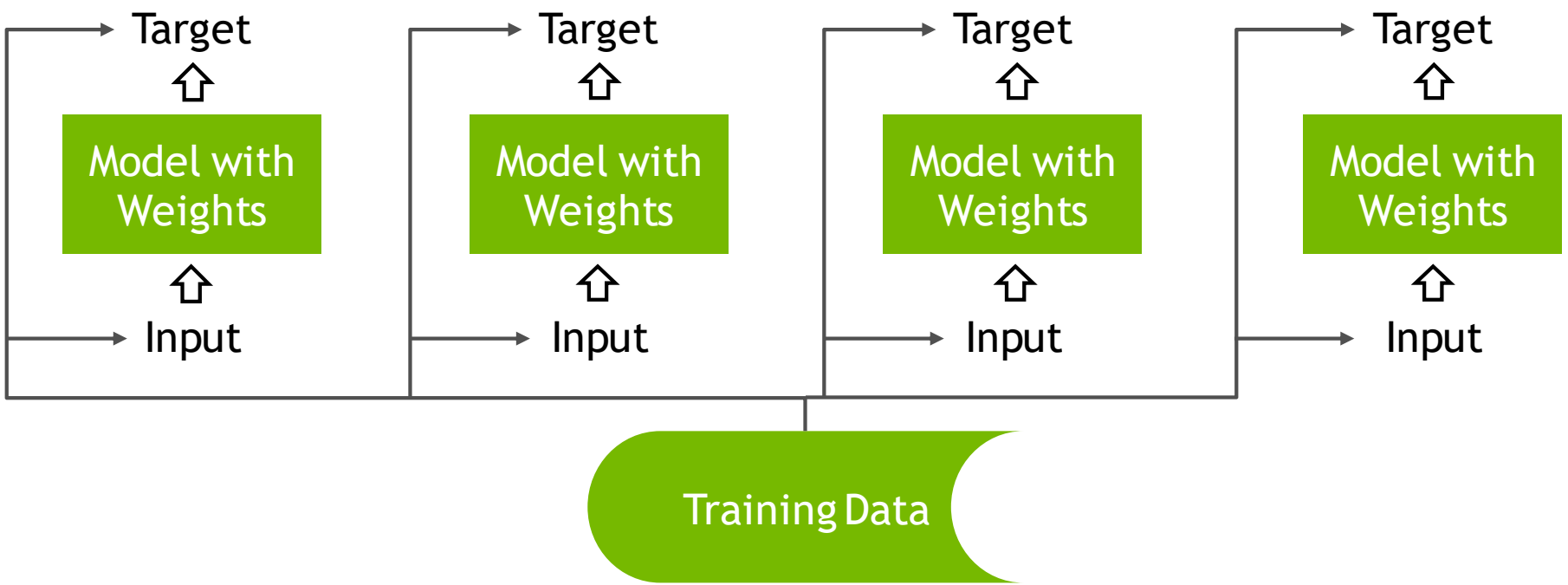
Need to load a minibatch (= some examples) of data

DEEP LEARNING TRAINING



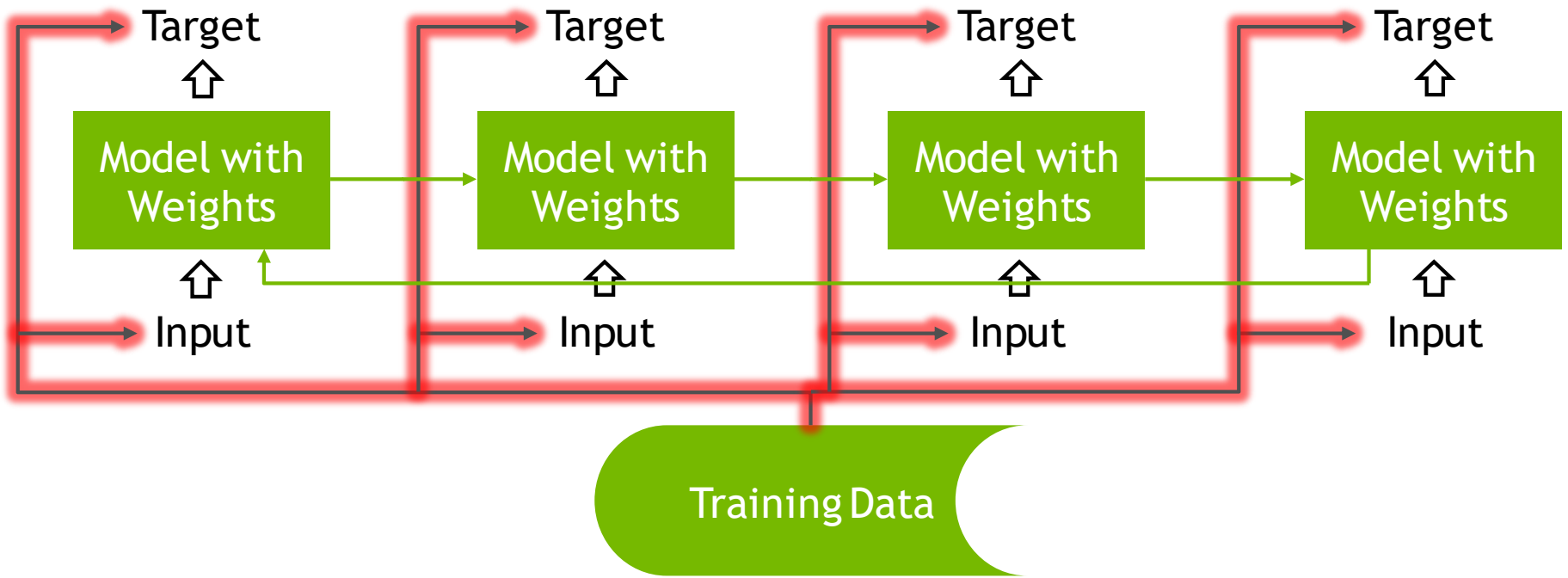
Single GPU: Load data. Weight update local to GPU.

MULTI GPU DATA PARALLEL DL TRAINING



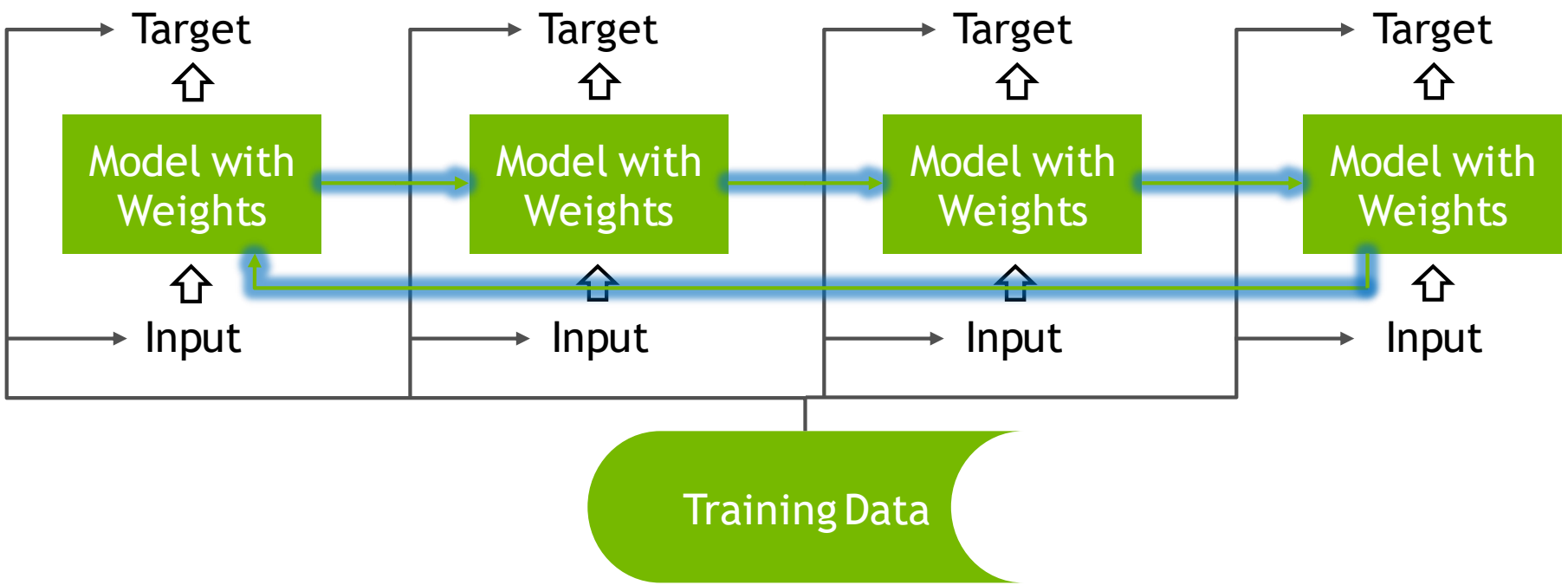
Mini-batch split between different GPUs

MULTI GPU DATA PARALLEL DL TRAINING



First load training data (different to each GPU)

MULTI GPU DATA PARALLEL DL TRAINING



Secondly, synchronize weights between the GPUs

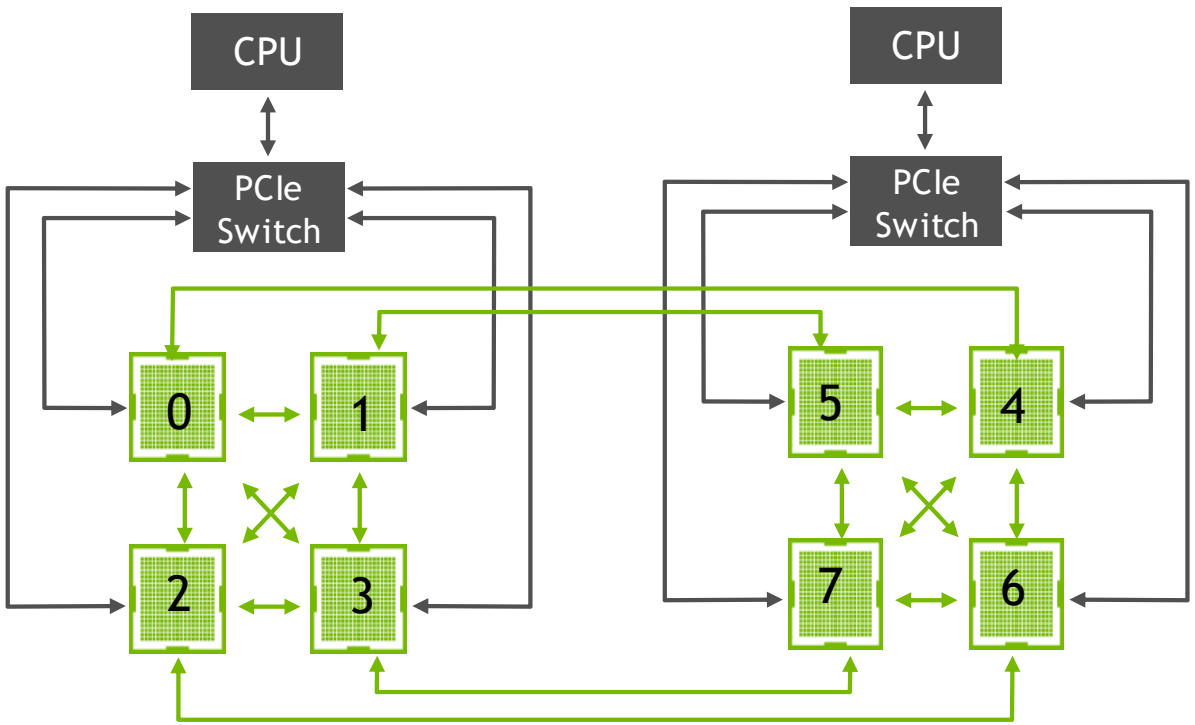
HOROVOD

“Making distributed Deep Learning fast and easy to use”

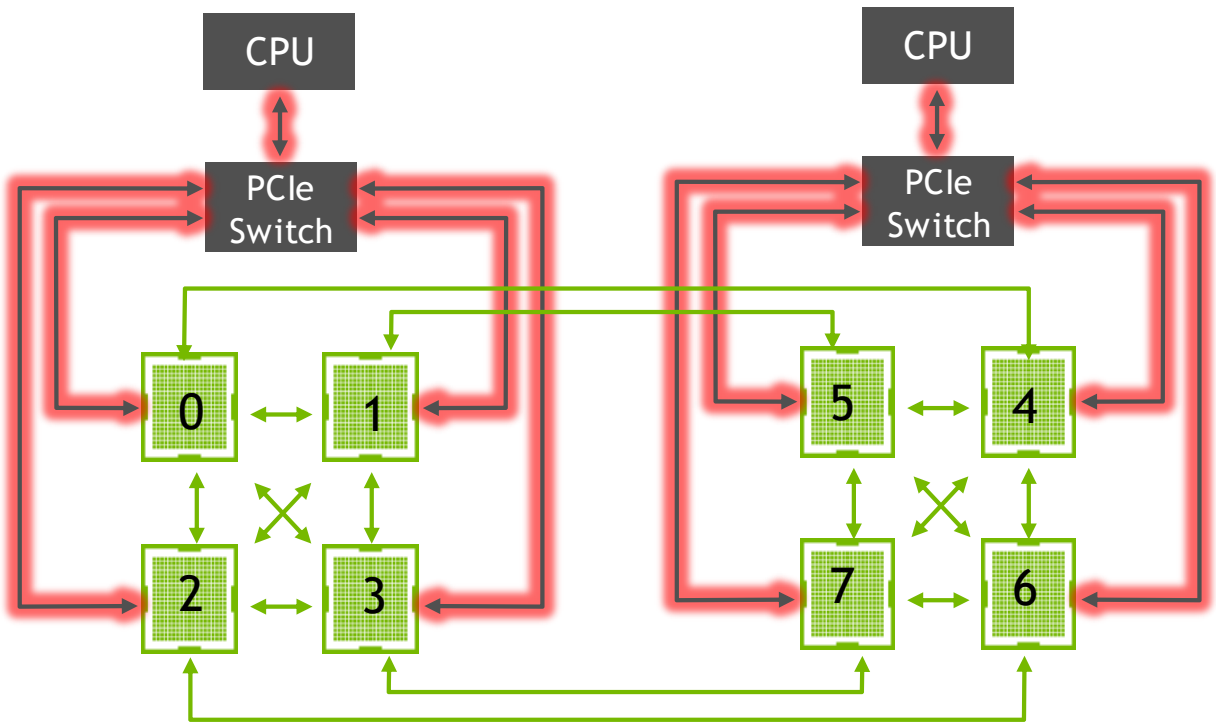
- Leverages TensorFlow + MPI + NCCL 2 to simplify development of synchronous multigpu/multinode TensorFlow
- Instead of Parameter Server architecture leverages MPI and NCCL based all reduce
- Owing to NCCL it leverages features such as:
  - NVLINK, RDMA, GPUDirectRDMA
  - Automatically detects communication topology
  - Can fall back to PCIe and TCP/IP communication

  Available in NGC TensorFlow container starting from version 18.01

DL DATA PARALLELISM - NVLINK

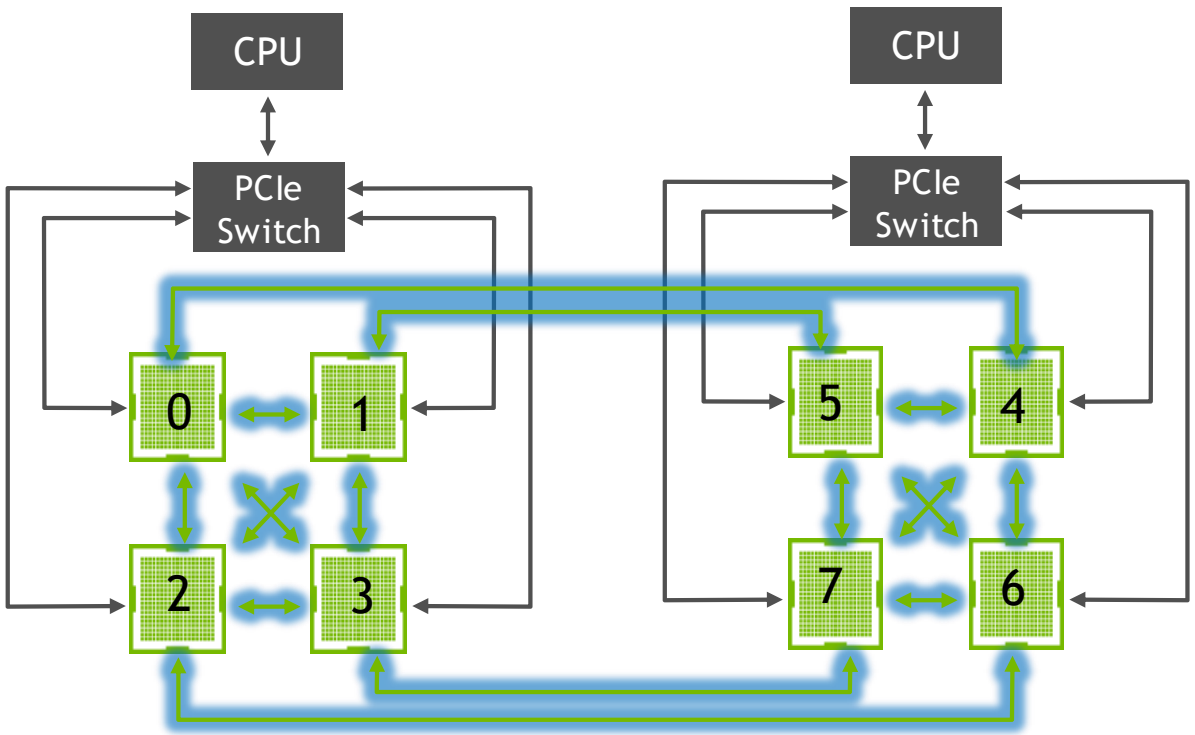


DL DATA PARALLELISM - NVLINK



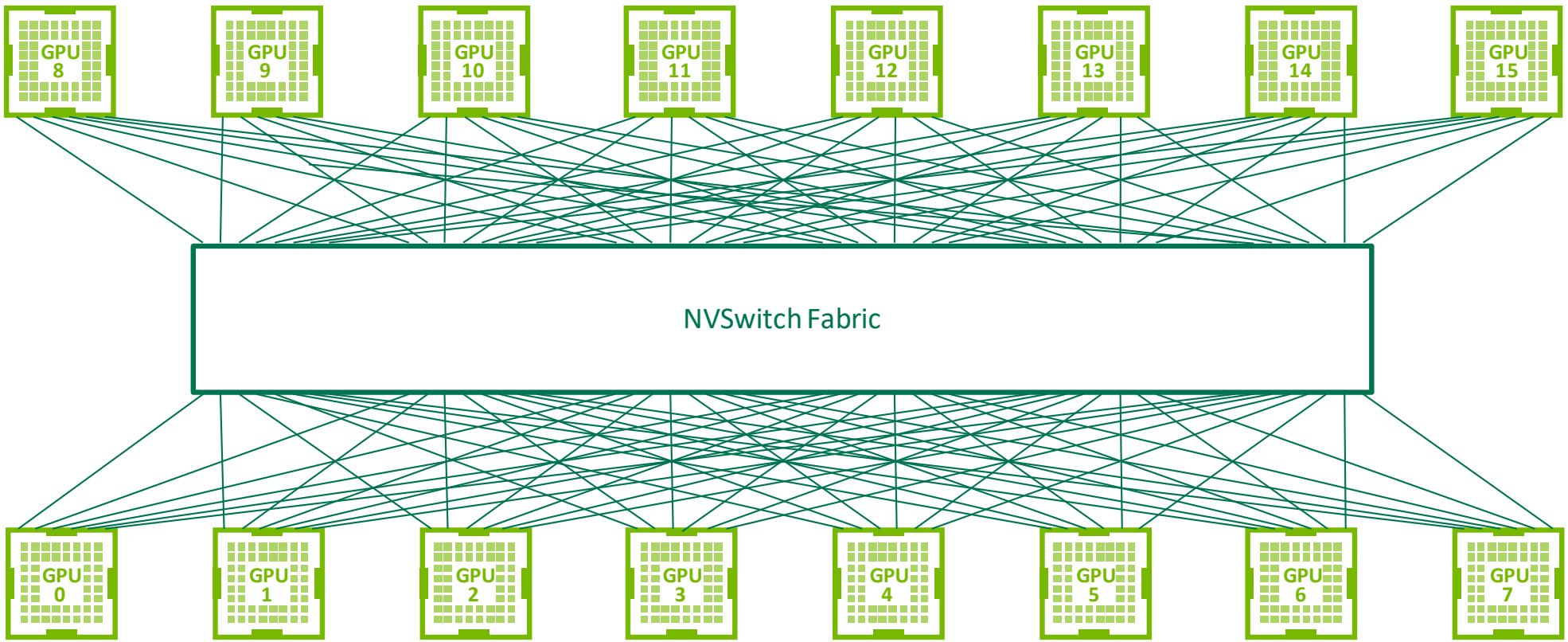
Data loading over PCIe

DL DATA PARALLELISM - NVLINK

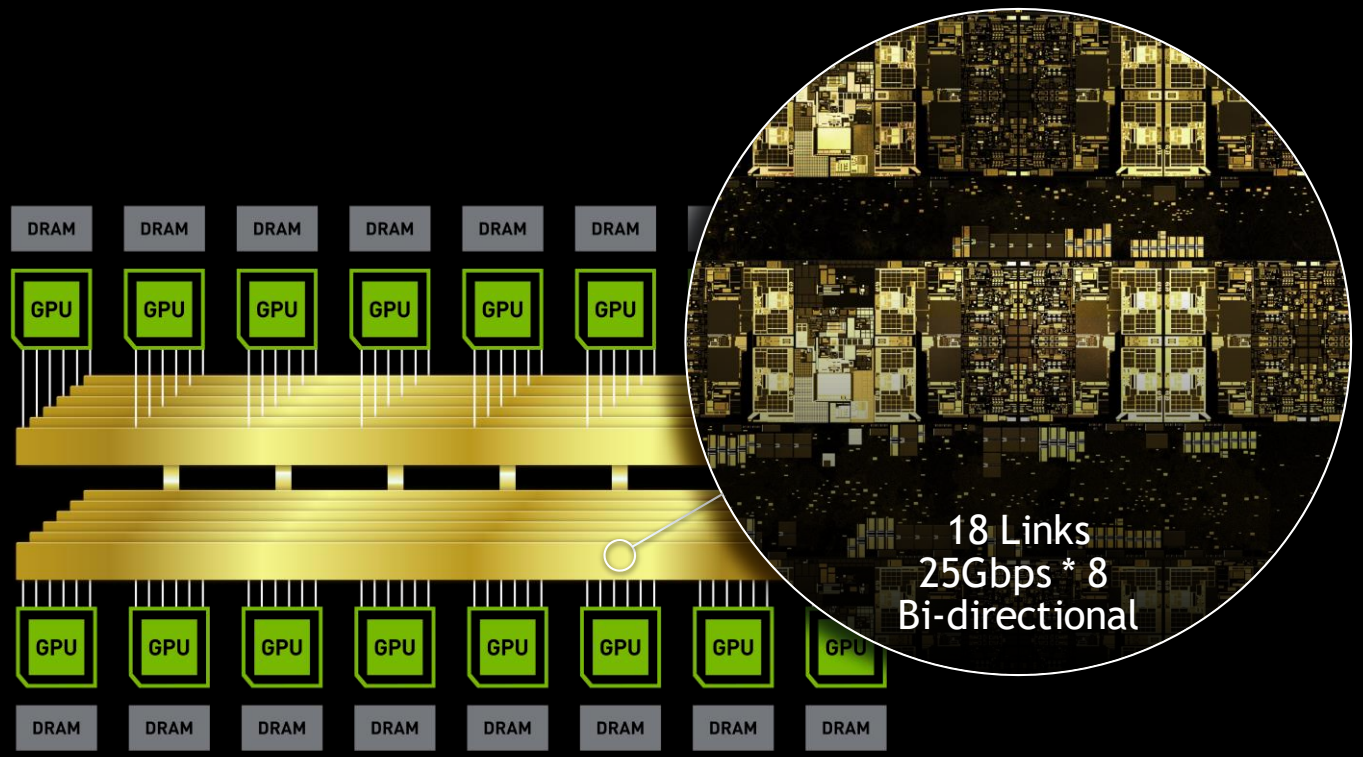


Gradient averaging over NVLink

NVSWITCH: ALL-TO-ALL CONNECTIVITY

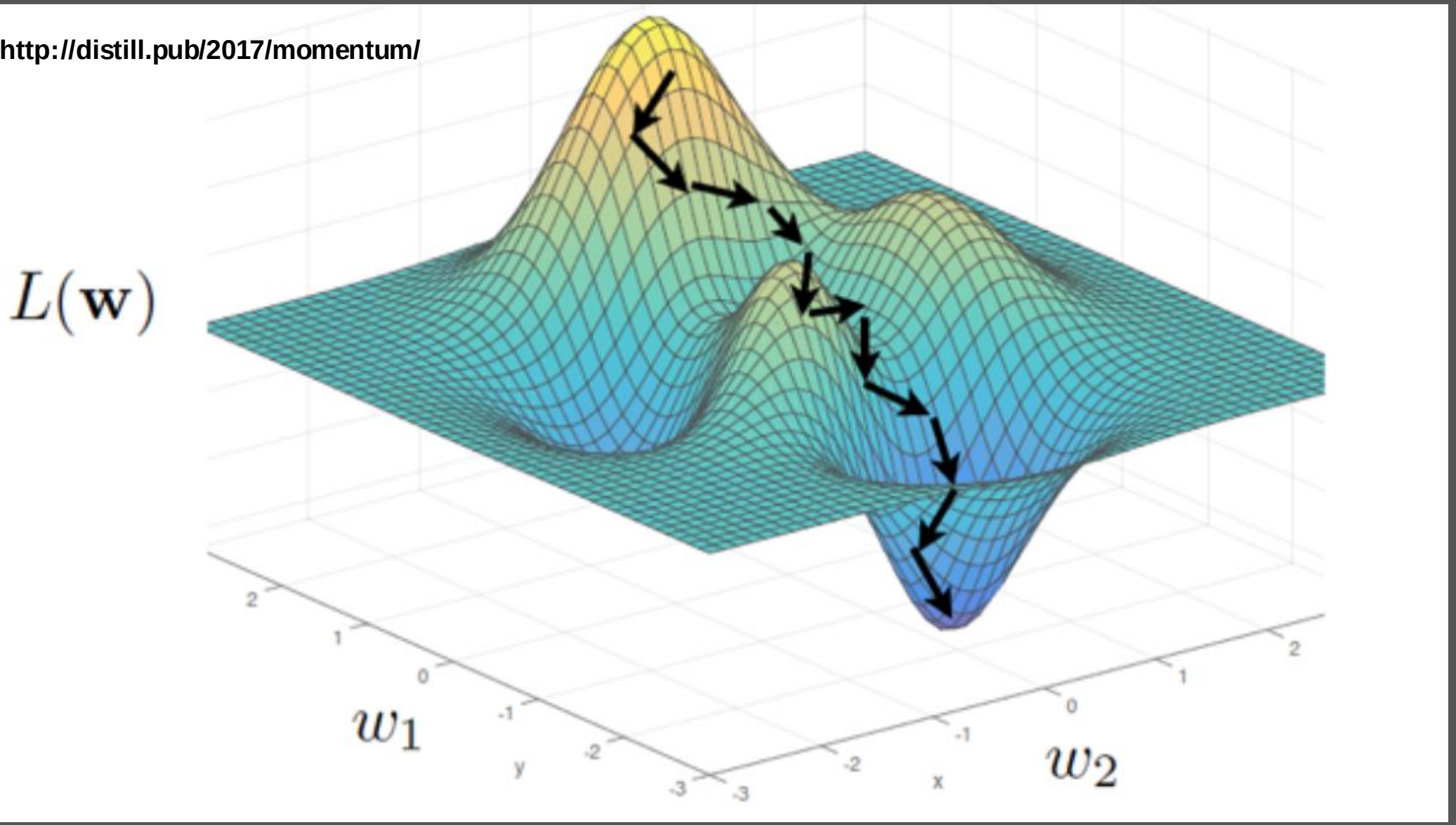


NV SWITCH



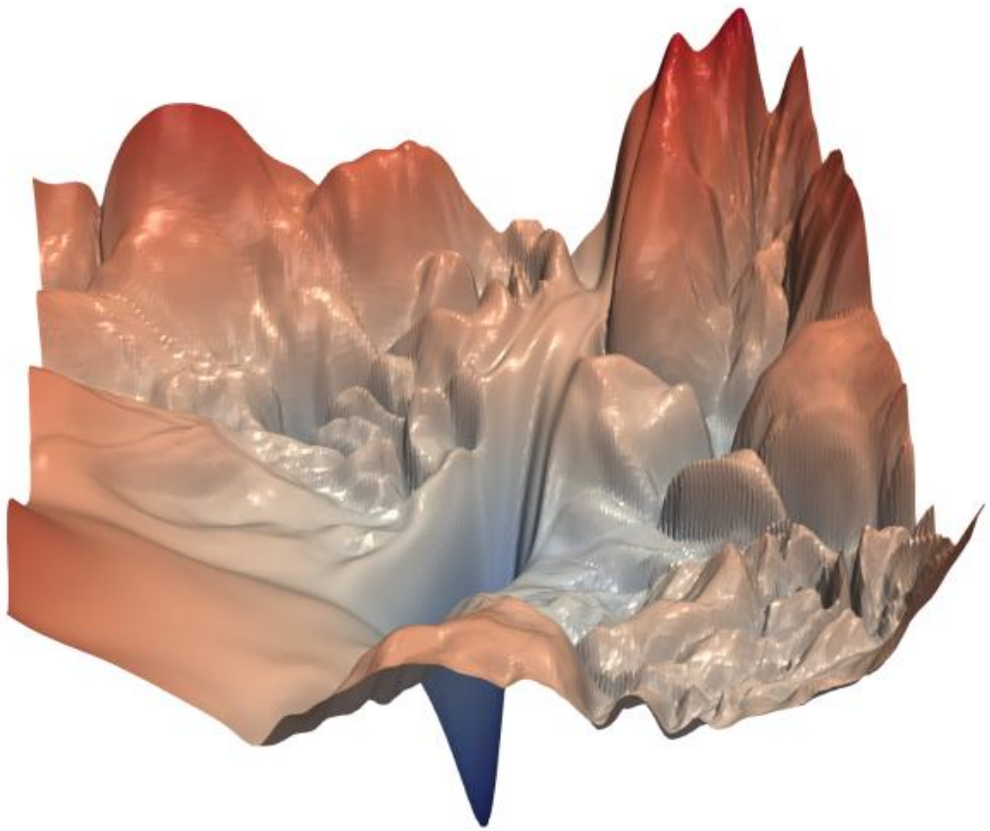
16 Tesla V100 32GB Connected by NVSwitch | On-chip Memory Fabric Semantic Extended Across All GPUs
512GB HBM2 and 14.4TB/sec Aggregate | 81,920 CUDA Cores | 2,000 TFLOPS Tensor Cores

http://distill.pub/2017/momentum/

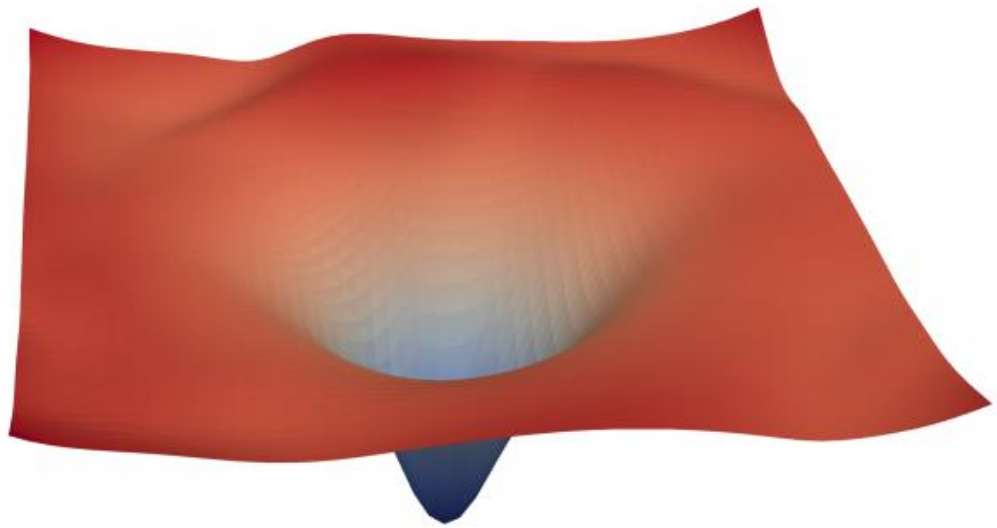


Li et al, University of Maryland & US Naval Academy

https://arxiv.org/pdf/1712.09913.pdf



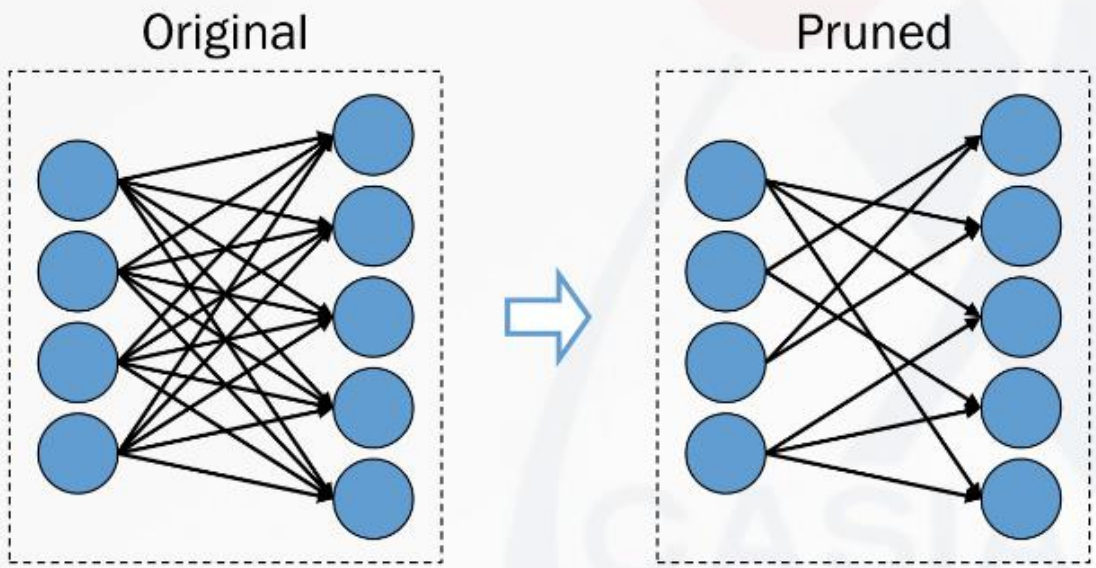
(a) without skip connections



(b) with skip connections

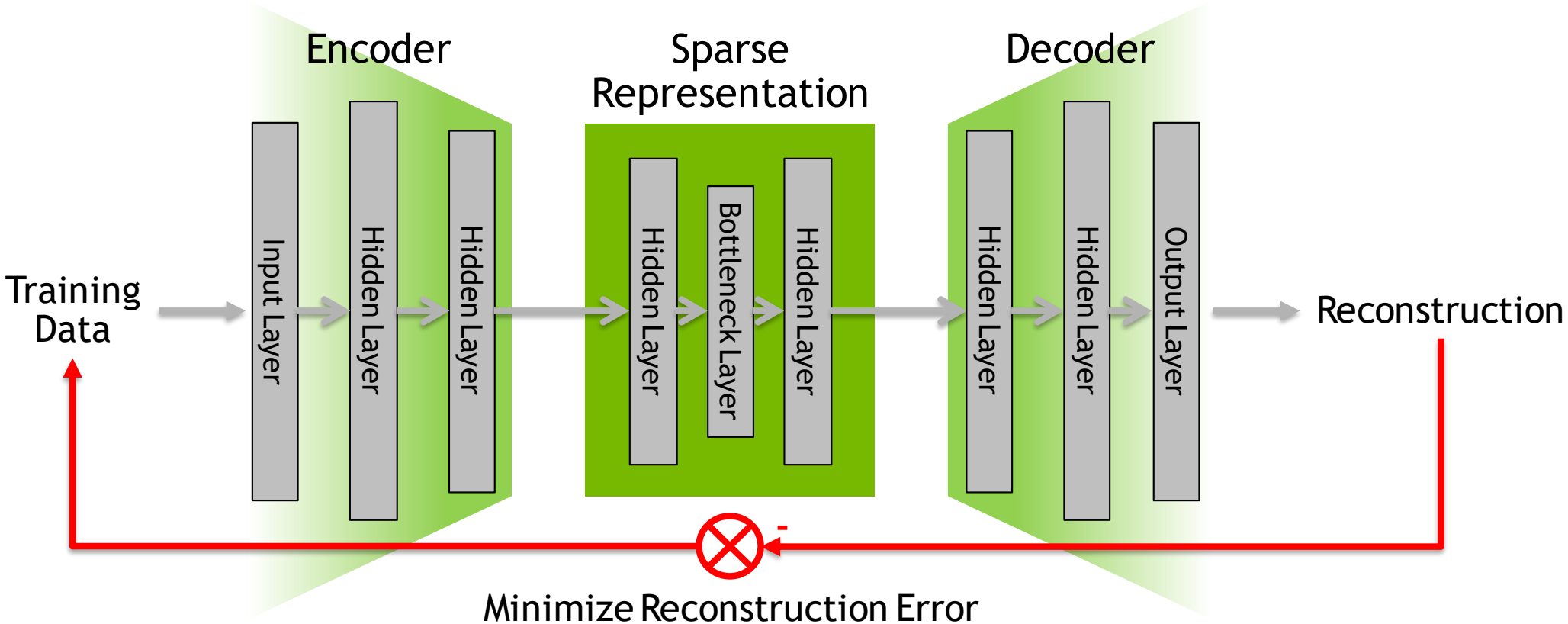
PRUNING

- Remove network connections
  - Fewer FLOPs (may not be faster)
  - Fewer connection weights



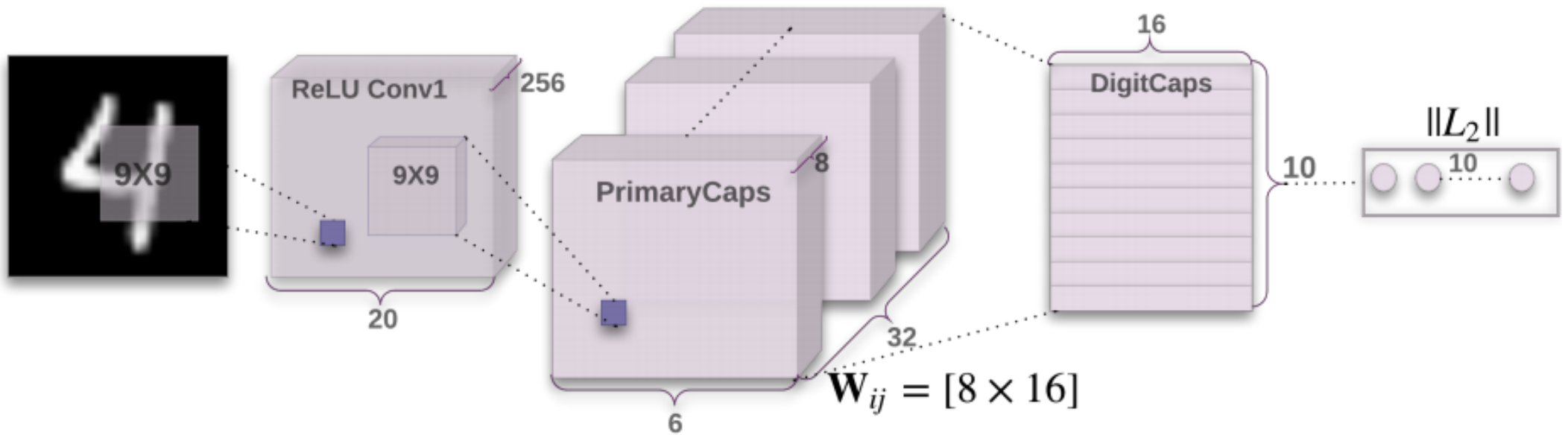
EXAMPLE: AUTOENCODERS

UNSUPERVISED feature learning



Capsules & routing with EM, Hinton et al

https://arxiv.org/pdf/1710.09829.pdf [NIPS2017]



SegCaps: capsules for object segmentation https://arxiv.org/pdf/1804.04241.pdf **University of Central Florida**

Long short-term memory (LSTM)

Hochreiter (1991) analysed vanishing gradient "*LSTM falls out of this almost naturally*"

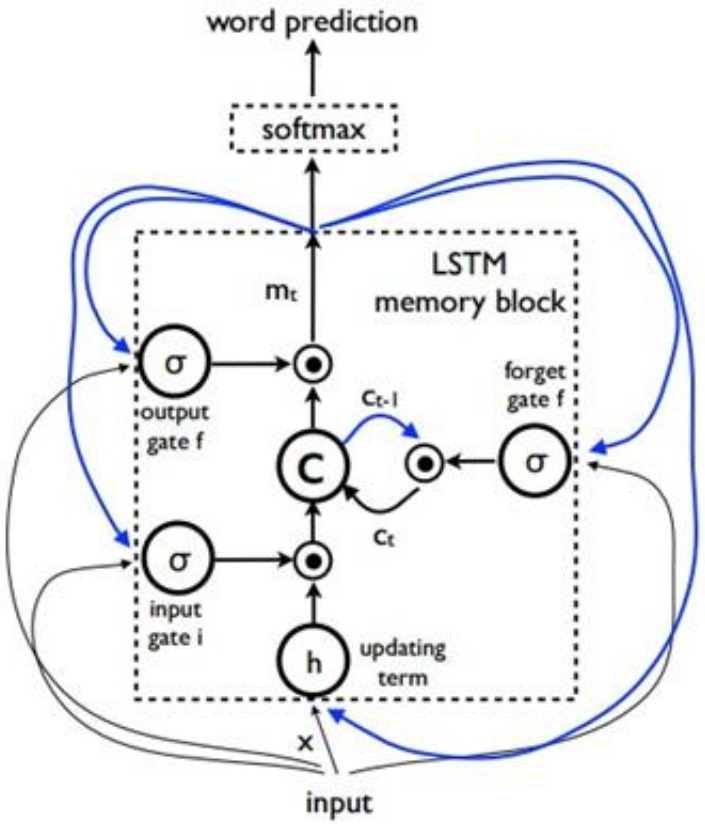
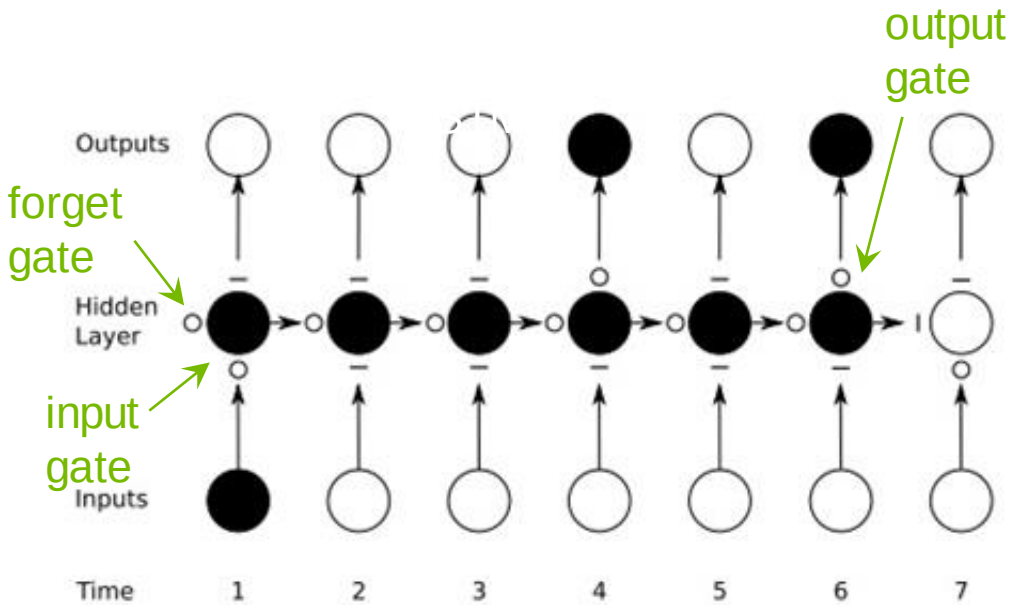


Fig from Vinyals et al, Google April 2015 NIC Generator

Training via backprop unfolded in time

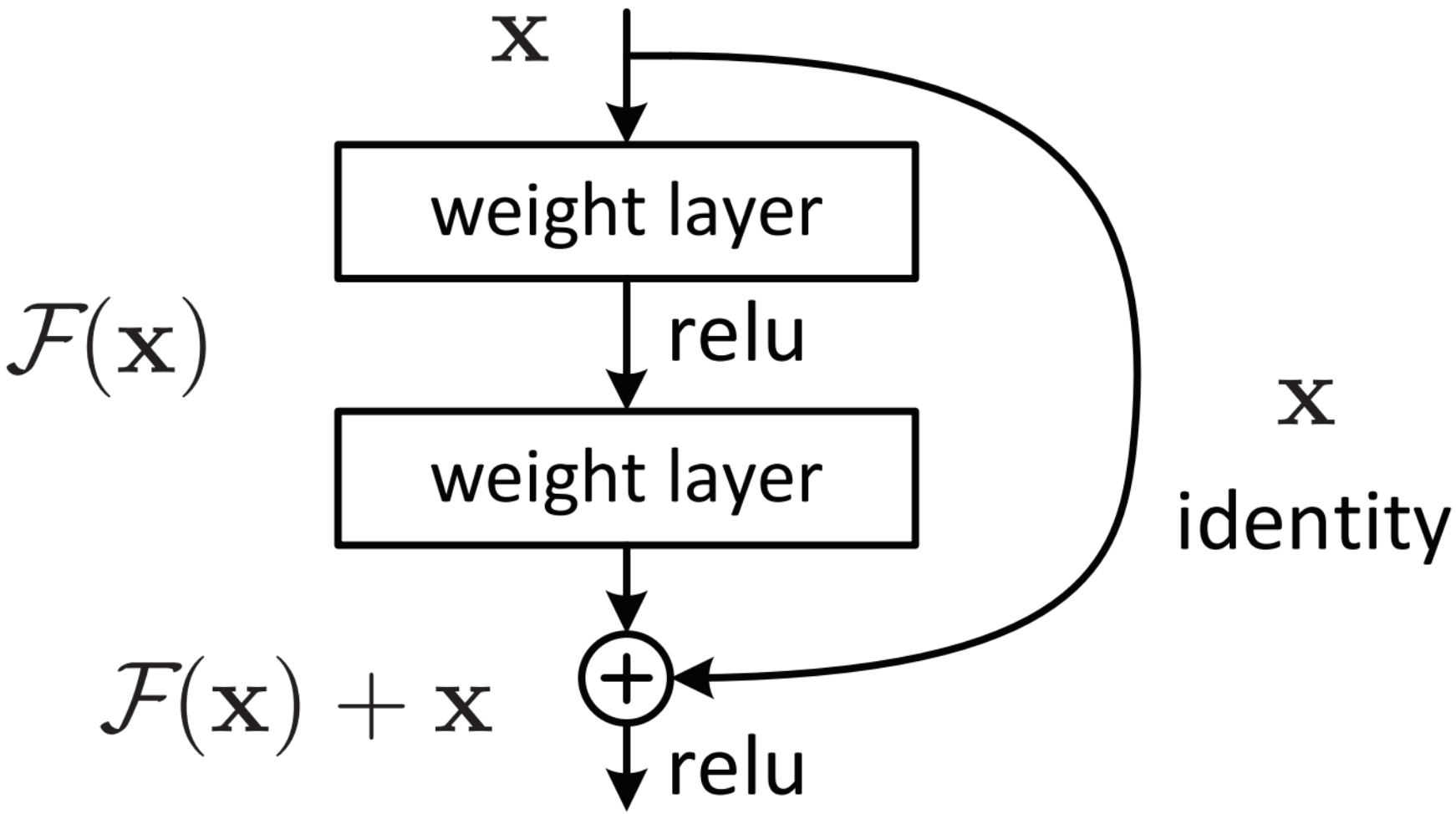


Long time dependencies are preserved until input gate is closed (-) and forget gate is open (O)

Gates control importance of the corresponding activations

Fig from Graves, Schmidhuber et al, Supervised Sequence Labelling with RNNs

Residual Networks



$\mathcal{F}(\mathbf{x})$

$\mathcal{F}(\mathbf{x}) + \mathbf{x}$

ResNets vs Highway Nets (IDSIA)

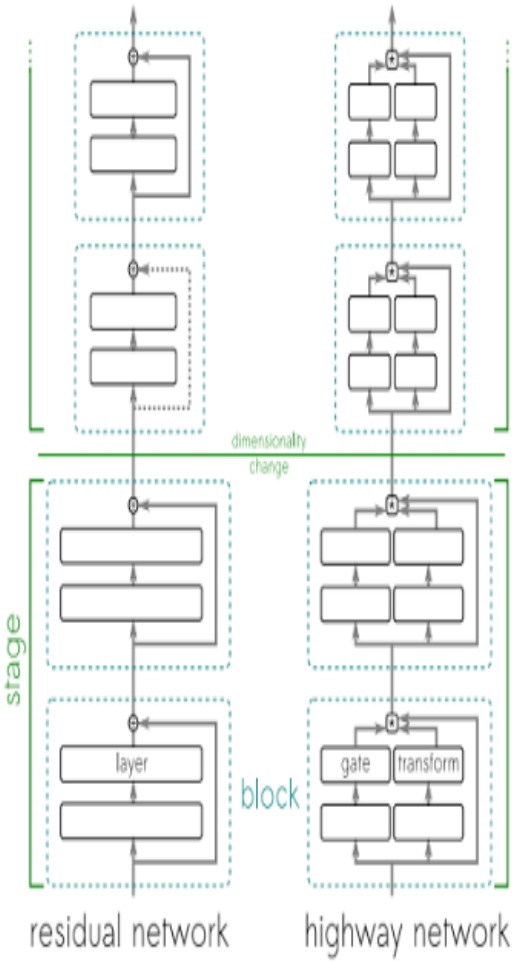
https://arxiv.org/pdf/1612.07771.pdf

Klaus Greff, Rupesh K. Srivastava

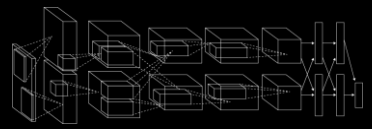
Really great explanation of "representation"

Compares the two.. shows for language modelling, translation HN >> RN.

Not quite as simple as each layer building a new level of representation from the previous - since removing any layer doesn't critically disrupt.

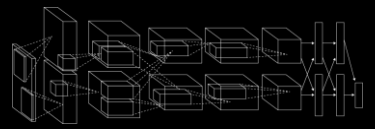


AlexNet



CAMBRIAN EXPLOSION

Convolutional Networks




Encoder/Decoder


ReLu

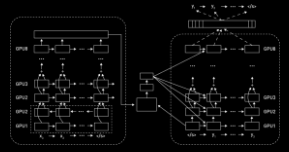

BatchNorm


Concat


Dropout

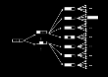

Pooling

Recurrent Networks




LSTM


GRU


Beam Search


WaveNet


CTC


Attention

Generative Adversarial Networks


3D-GAN

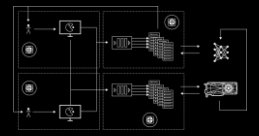

MedGAN


Conditional GAN

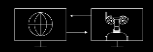

Coupled GAN


Speech Enhancement GAN

Reinforcement Learning

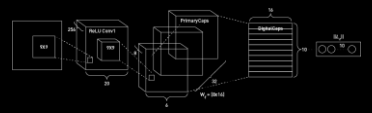


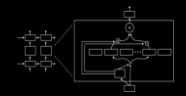

DQN


Simulation


DDPG

New Species


Capsule Nets

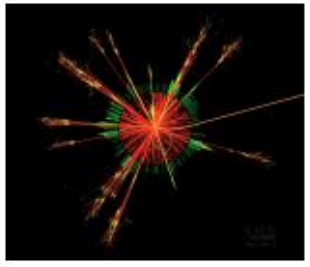

Mixture of Experts


Neural Collaborative Filtering

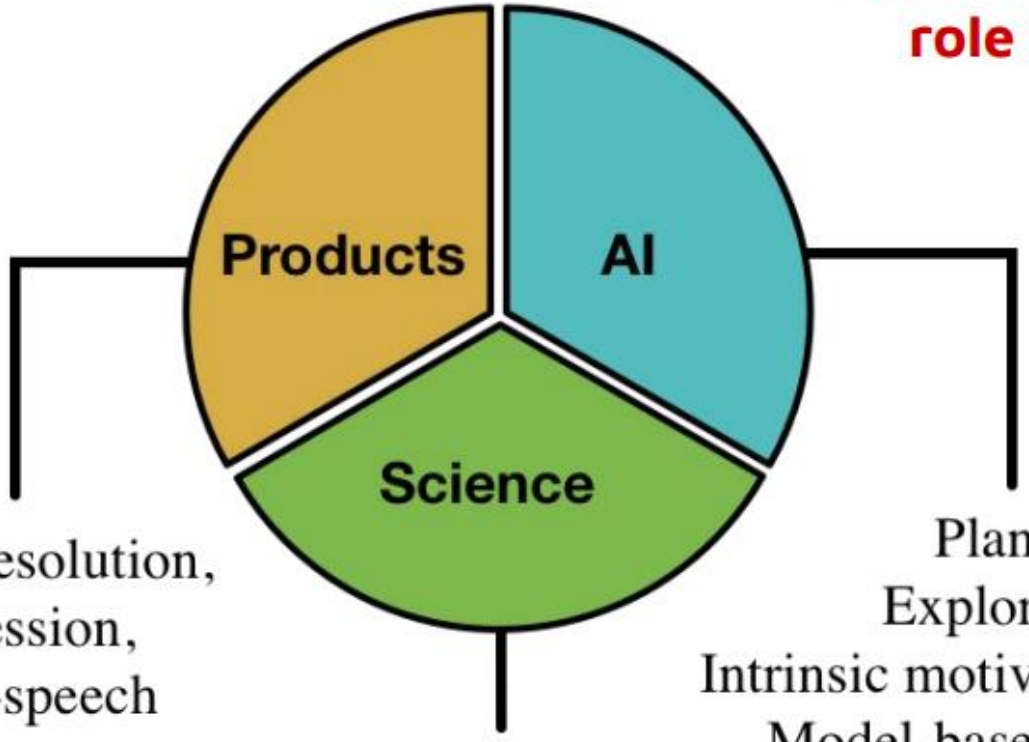

Block Sparse LSTM

Why Generative Models

Generative models have a role in many problems.



Super-resolution,
Compression,
Text-to-speech



Proteomics,
Drug Discovery,
Astronomy,
High-energy physics

Planning,
Exploration
Intrinsic motivation
Model-based RL

Shakir Mohamed and Danilo Rezende, DeepMind, UAI 2017

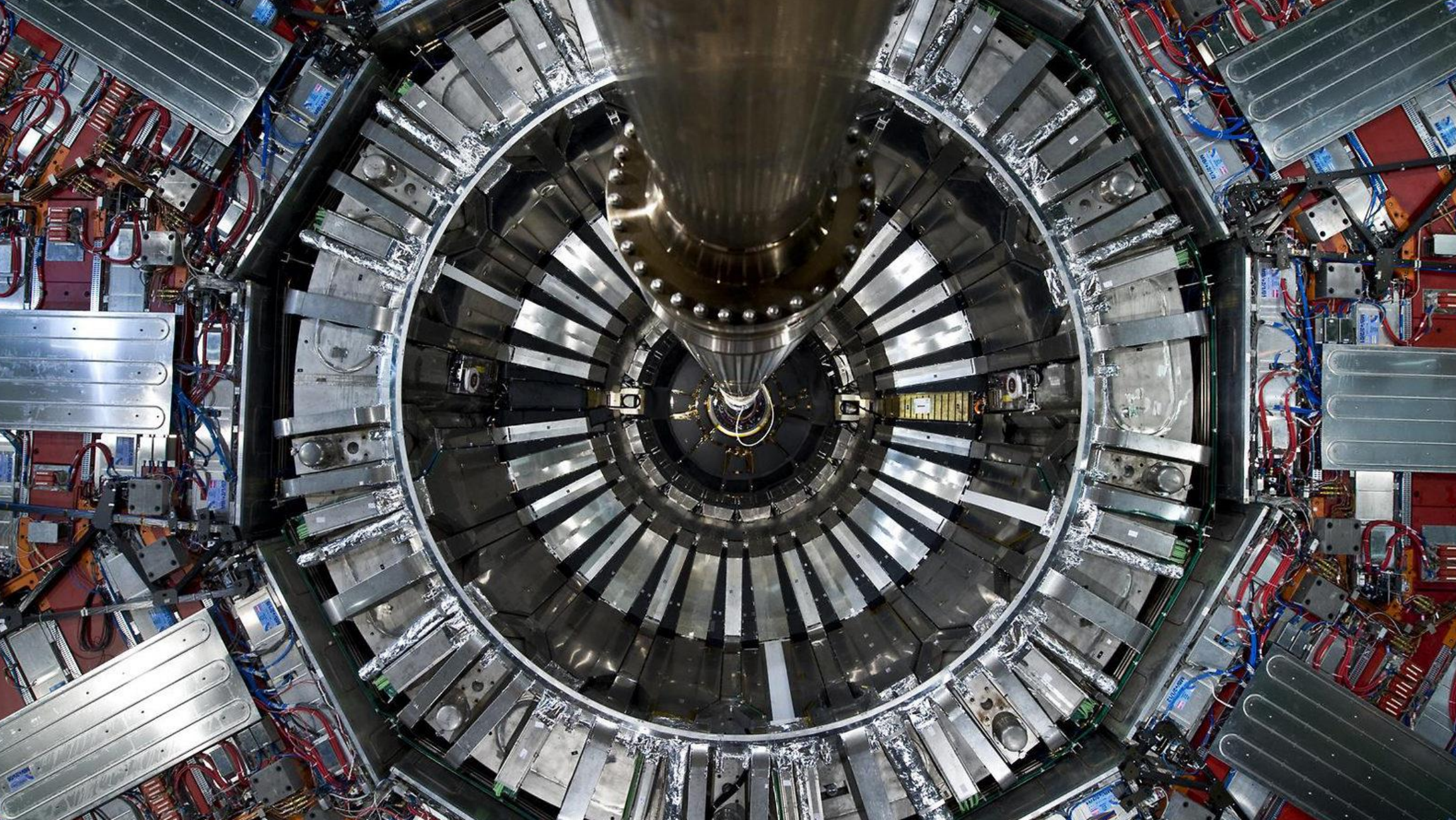
The High Altitude Water Cherenkov (HAWC) Observatory

A cosmogenic gamma ray observatory, examining some of the most energetic light in the universe

Located on Pico de Orizaba, Mexico

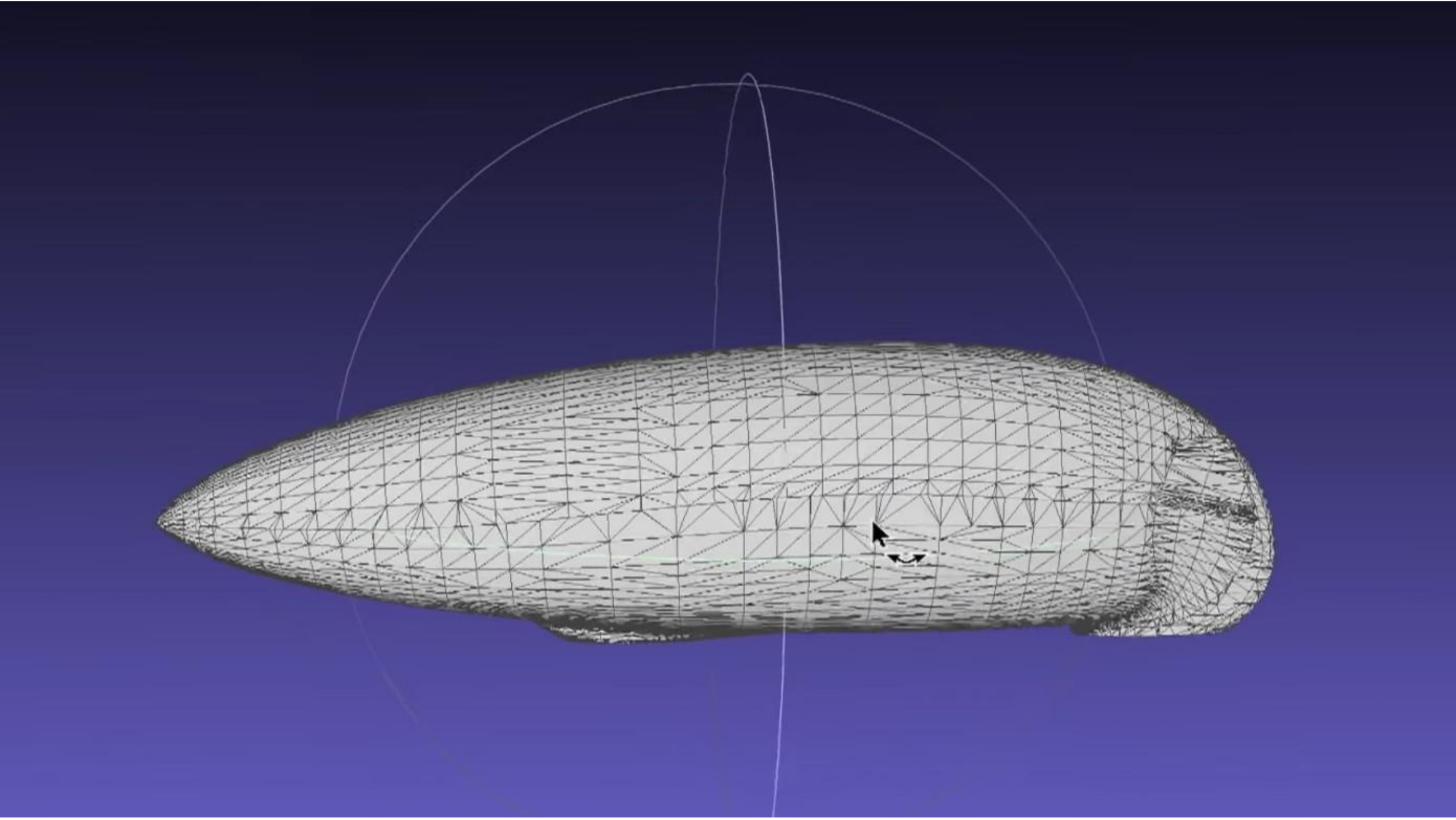
High duty cycle, high statics, high energy physics experiment

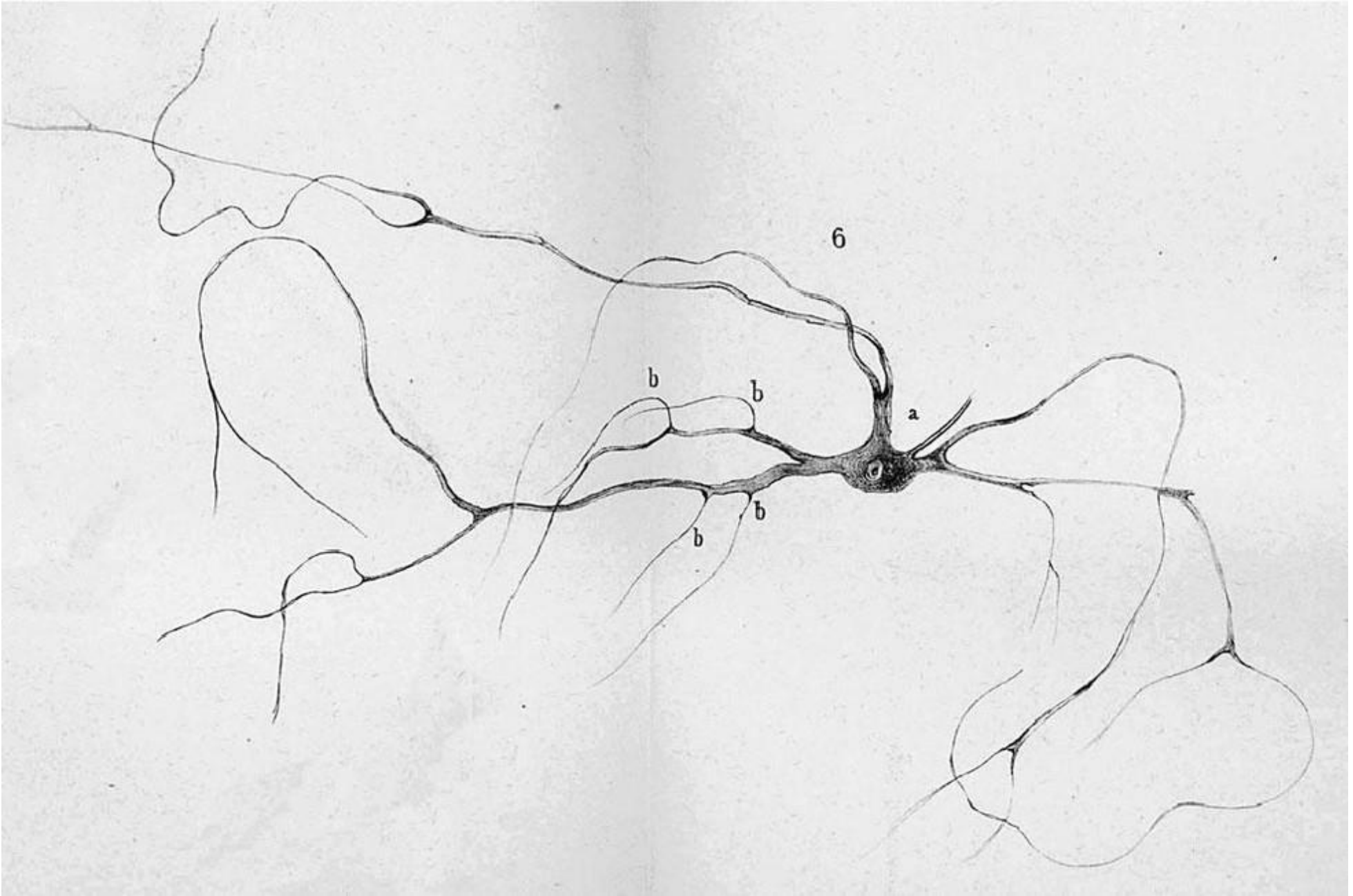
Daniel Ho, Gefen Kohavi, Michael Gussert

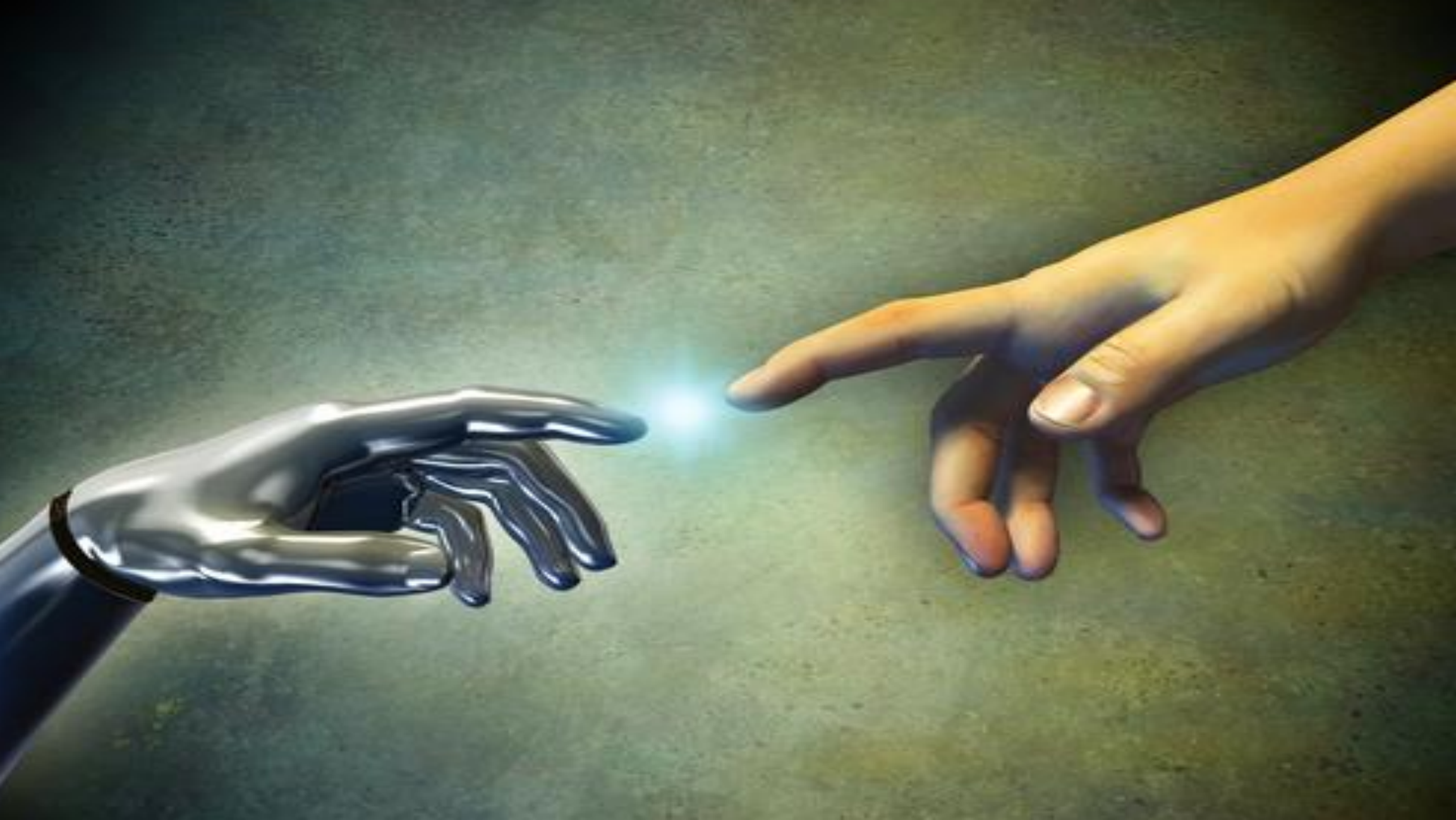


Dreamcatcher

https://autodeskresearch.com/projects/dreamcatcher

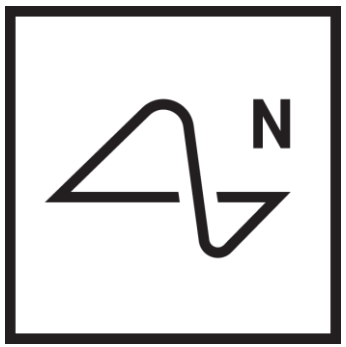



6
b
b
a
b
b



kernel

Brain Computer Interfaces
Focused on treatment for disease and dysfunction
eg epilepsy, depression, Parkinsons but ultimately to advance human intelligence by restoring and extending cognitive vibrancy.



"We're either going to have to merge with AI or be left behind"; Elon Musk

REINFORCEMENT LEARNING
“from our own mistakes”

RL TIMELINE

1984: Rich Suttons thesis

1989: Q-learning

1991: REINFORCE

2006: MCTS (Szepesvari)

2009: David Silver's thesis

2013: DQN

2015: DQN in Nature

2016: A3C & DRL

2017: Soft Q-learning/ AlphaGo Zero

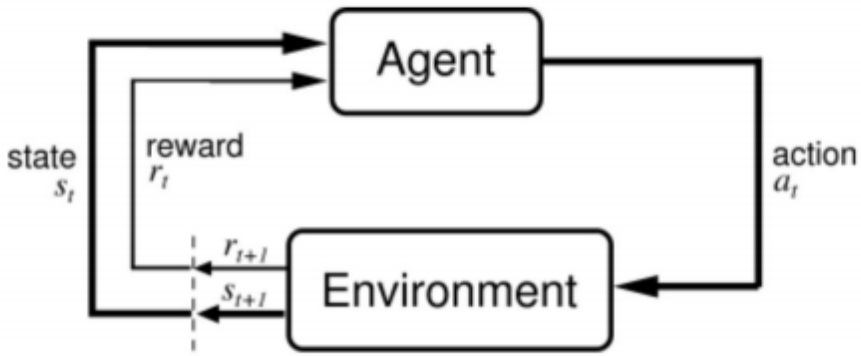
2018: GA3C

Markov Decision Processes

Provides a mathematical framework to model such tasks.

Defined as a 5-tuple : $\{\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma\}$

- S : set of states
- A : set of actions
- R : reward function
- T : transition function
- γ : discount factor



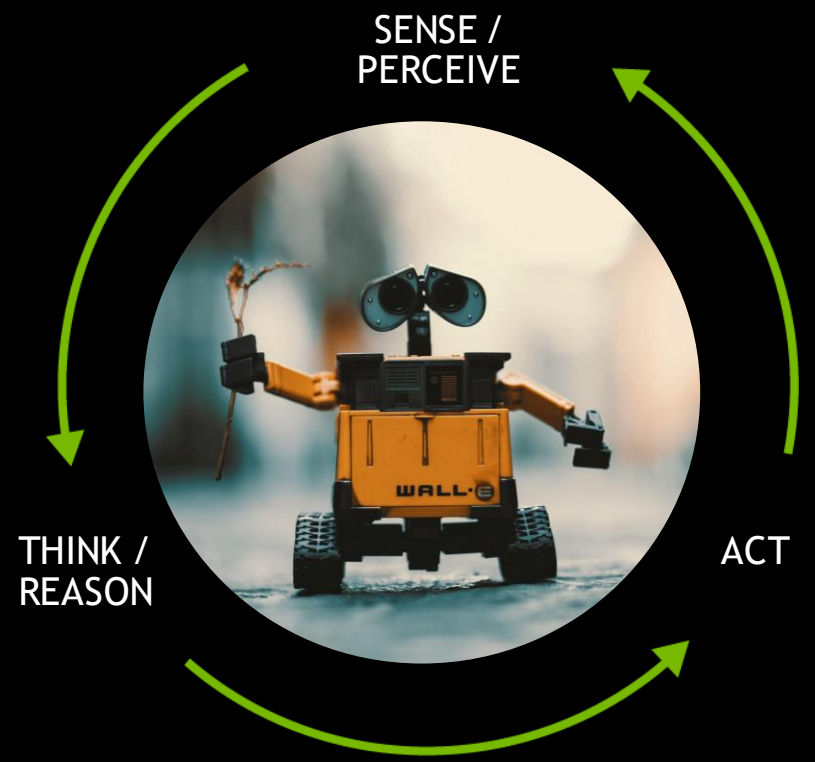
Objective : Learn/plan to act optimally so as to *MAXIMIZE* the expected long term discounted reward (value).

Tabish Rashid et al, Oxford/DeepMind QMIX: https://arxiv.org/pdf/1803.11485.pdf

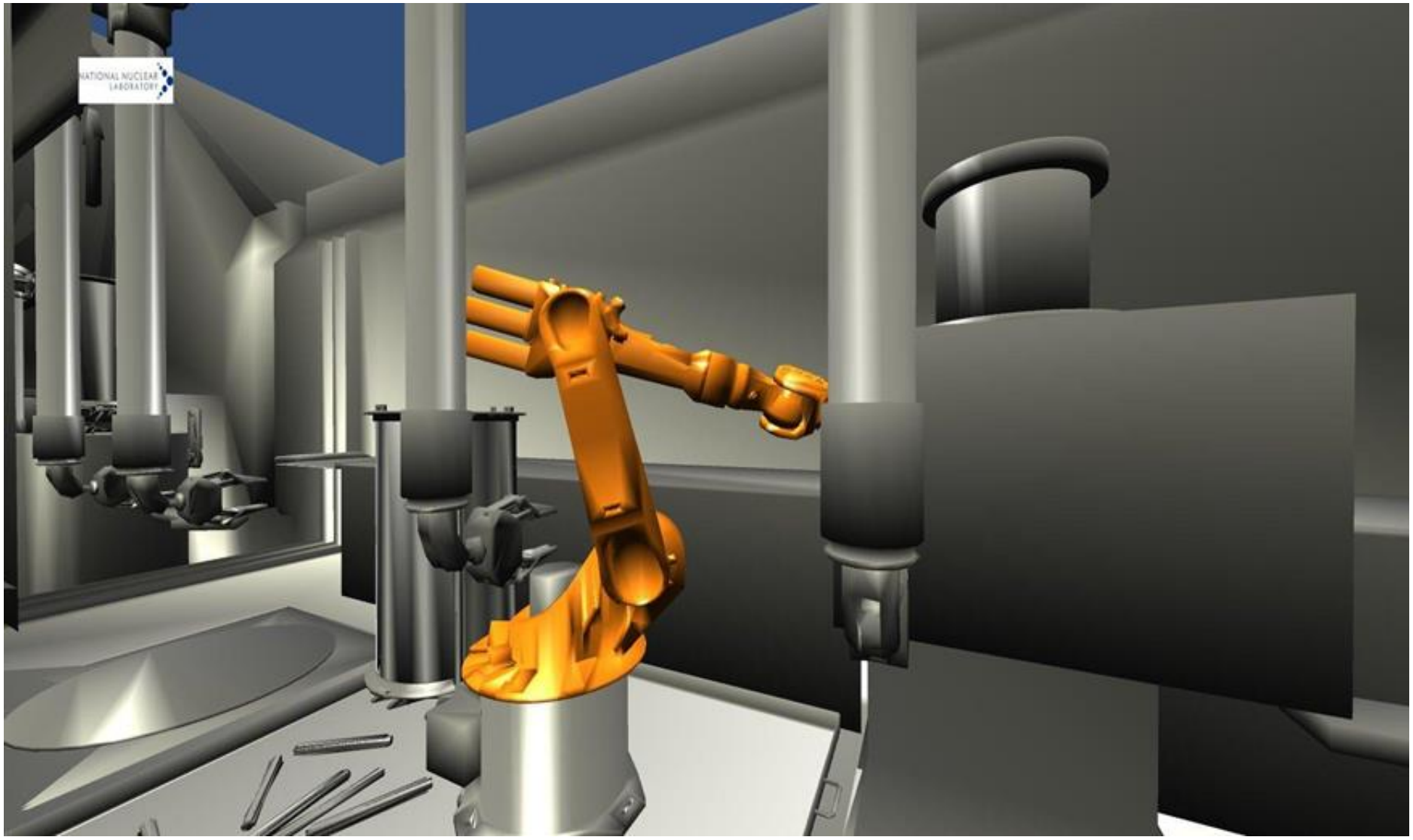
DEEPMIND ALPHA* 1.0 => 2.0 => 0.0

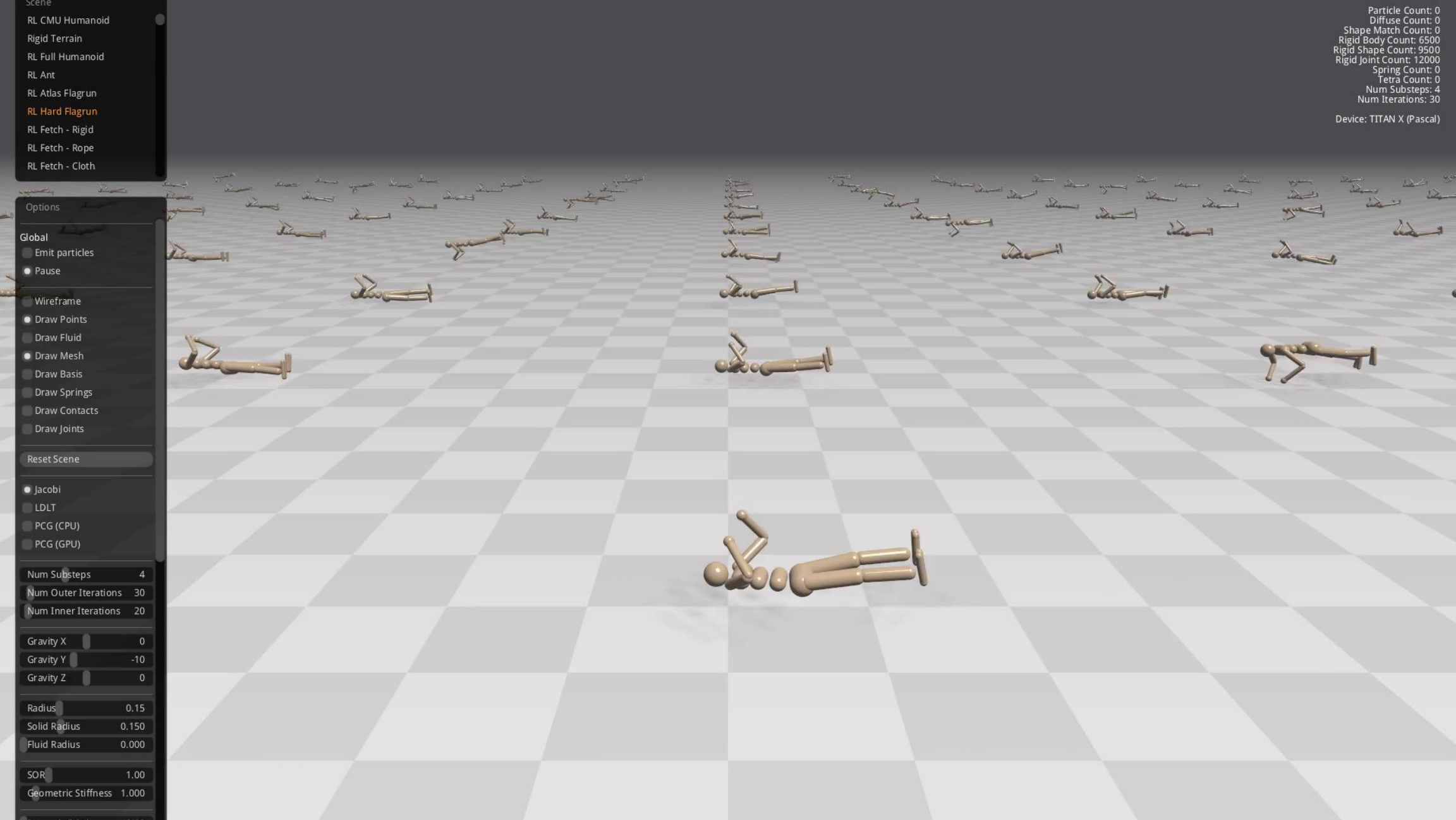
* ..a deeply structured hybrid (Gary Marcus, Jan 2018)

ROBOTS



ATLAS, BOSTON DYNAMICS


NATIONAL NUCLEAR
LABORATORY


Scene
RL CMU Humanoid
Rigid Terrain
RL Full Humanoid
RL Ant
RL Atlas Flagrun
RL Hard Flagrun
RL Fetch - Rigid
RL Fetch - Rope
RL Fetch - Cloth
Particle Count: 0
Diffuse Count: 0
Shape Match Count: 0
Rigid Body Count: 6500
Rigid Shape Count: 9500
Rigid Joint Count: 12000
Spring Count: 0
Tetra Count: 0
Num Substeps: 4
Num Iterations: 30
Device: TITAN X (Pascal)
Options
Global
Emit particles
Pause
Wireframe
Draw Points
Draw Fluid
Draw Mesh
Draw Basis
Draw Springs
Draw Contacts
Draw Joints
Reset Scene
Jacobi
LDLT
PCG (CPU)
PCG (GPU)
Num Substeps 4
Num Outer Iterations 30
Num Inner Iterations 20
Gravity X 0
Gravity Y -10
Gravity Z 0
Radius 0.15
Solid Radius 0.150
Fluid Radius 0.000
SOR 1.00
Geometric Stiffness 1.000

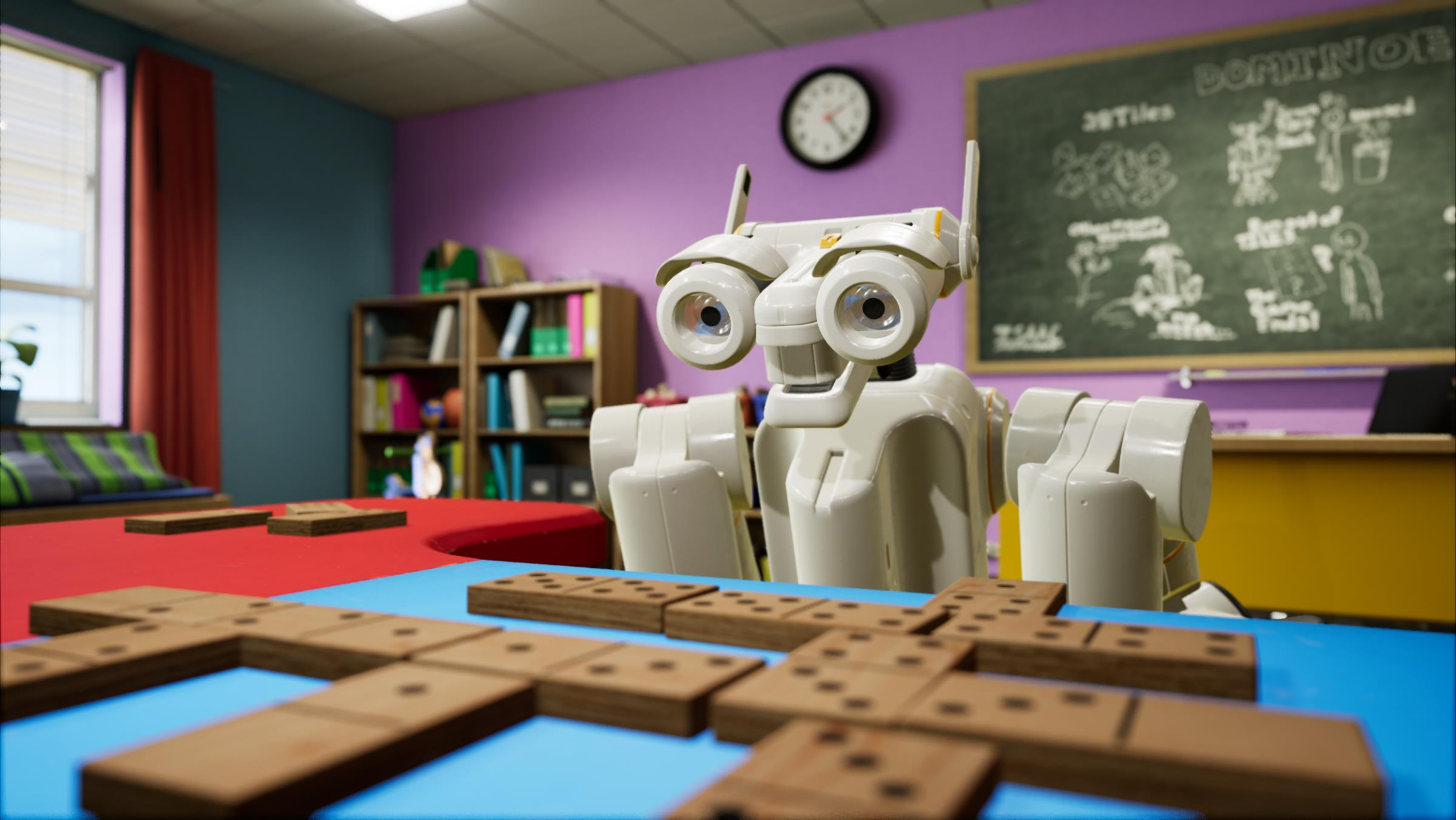
IMMERSIVE AI

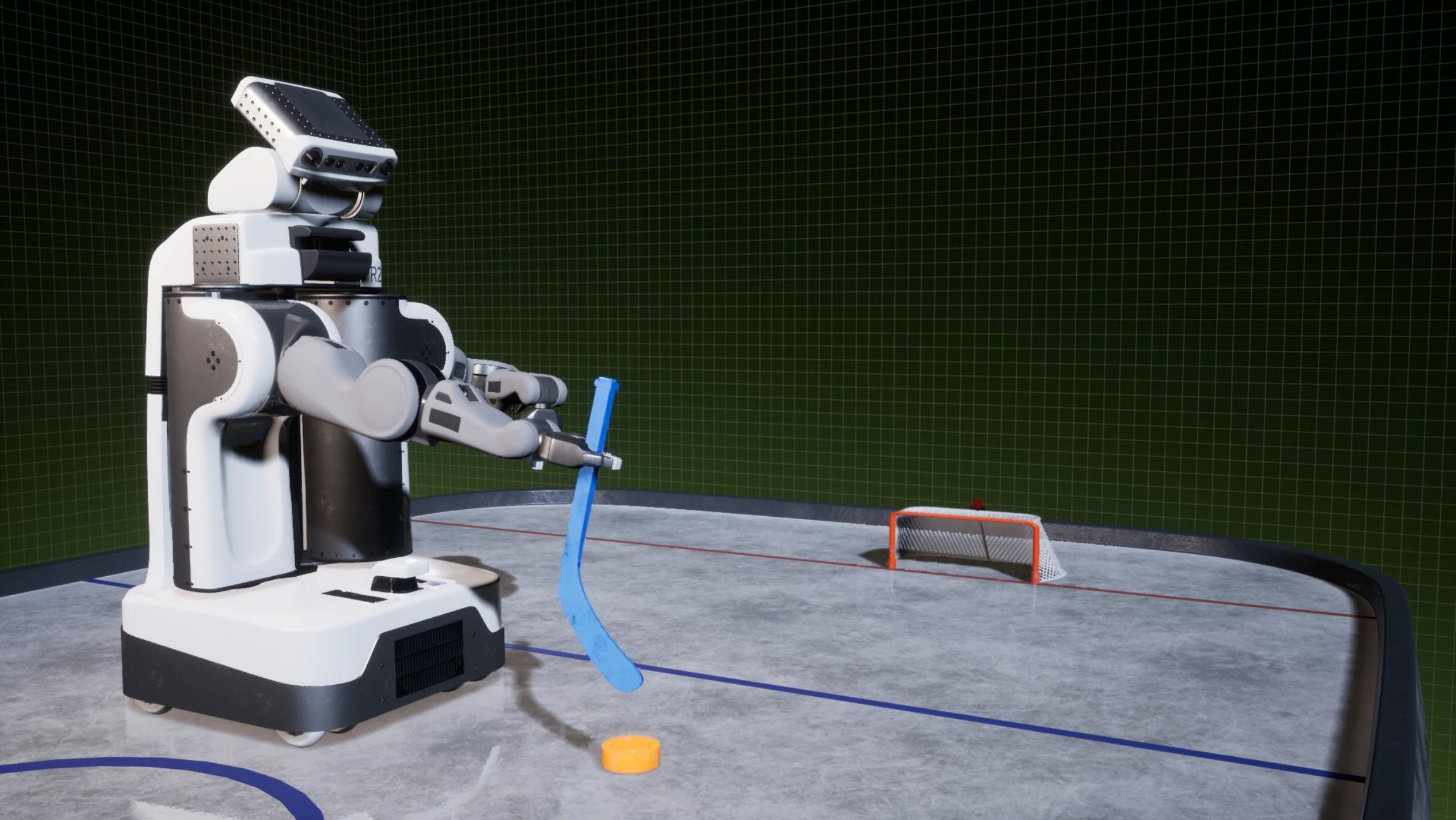
THE DESIGN LAB OF THE FUTURE

Holodeck: photorealistic, collaborative VR environment

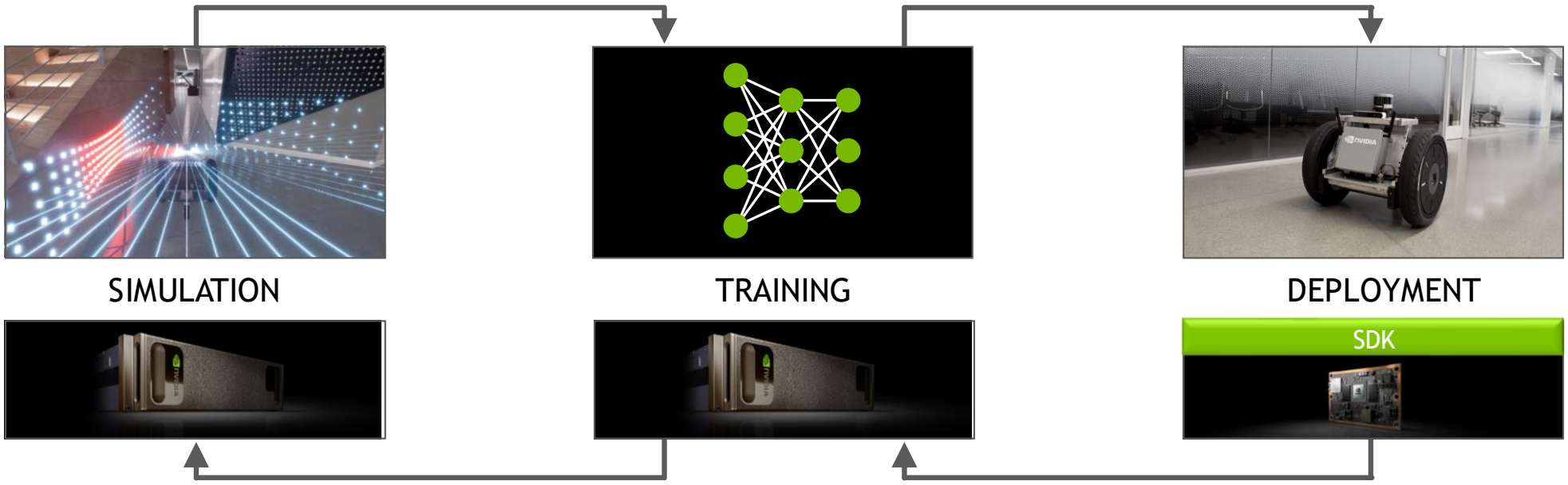
Real-world experience through sight, sound, and haptics

Explore high-fidelity, full-resolution models

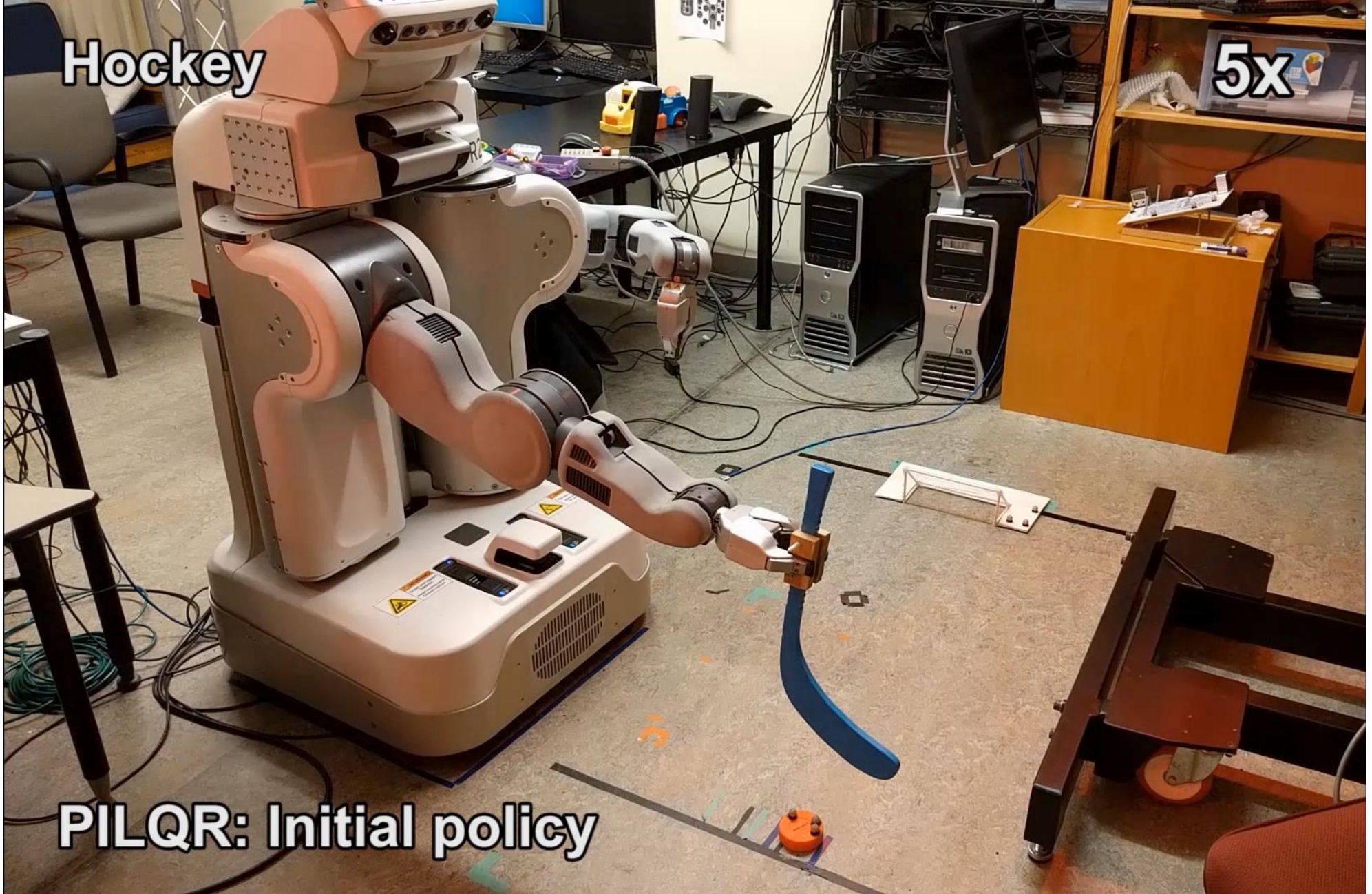


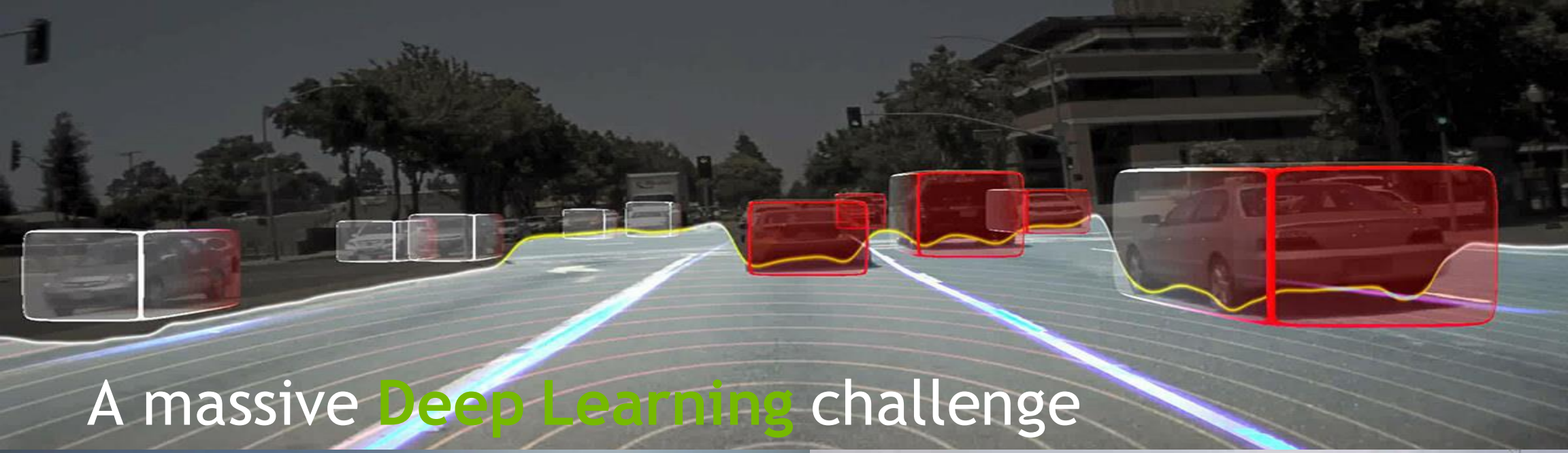


NVIDIA ISAAC ROBOTICS PLATFORM



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


Hockey
5x
PILQR: Initial policy

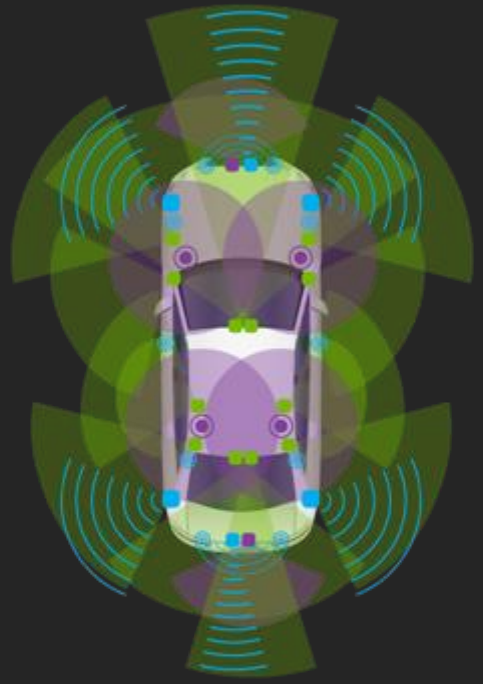


A massive Deep Learning challenge

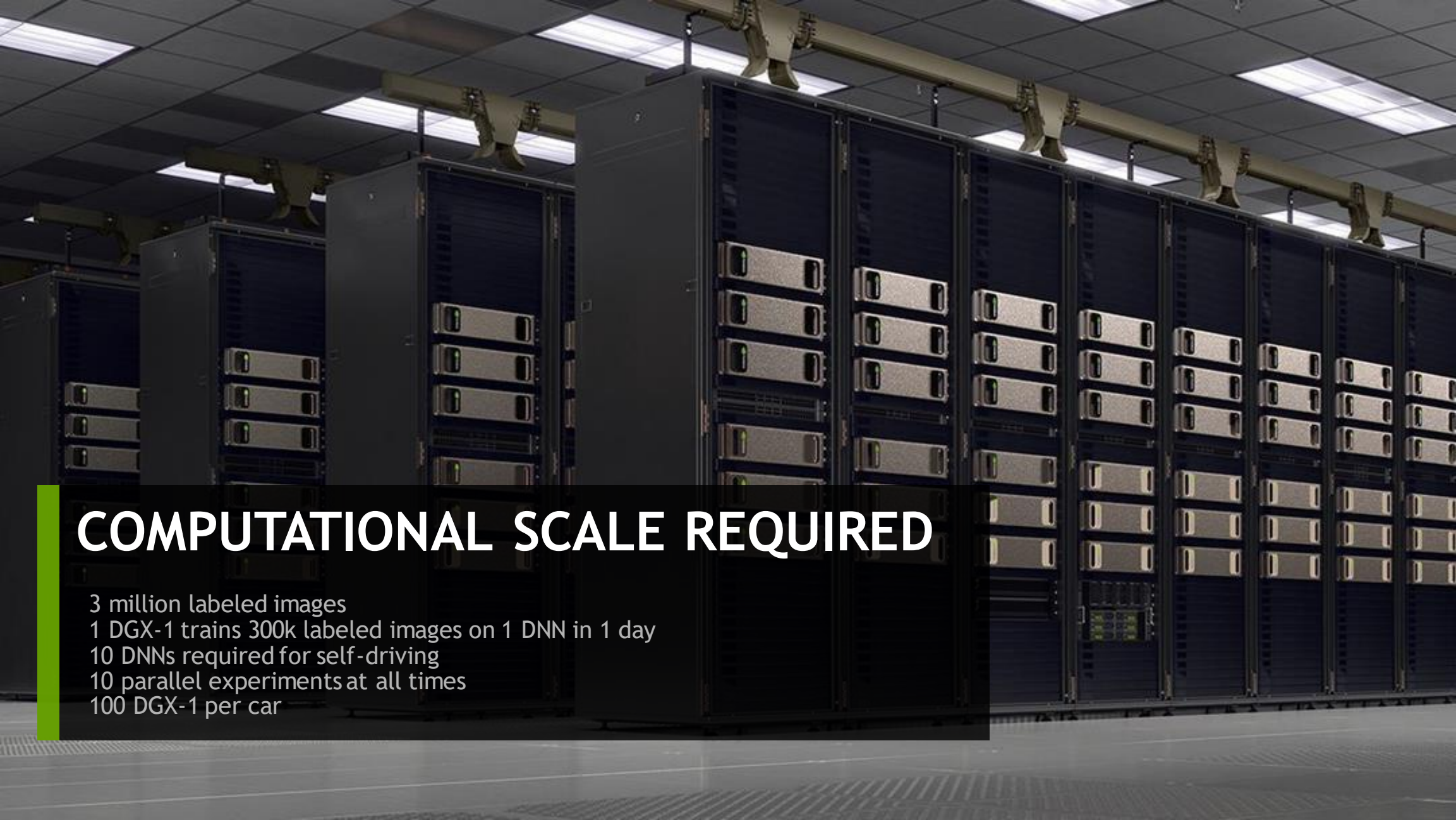
LEVEL 5 DEMANDS EXTREME COMPUTING

+ 10X camera resolution
+ Surround LIDAR point-cloud processing
+ Camera & LIDAR localization to HD map
+ Tracking all surrounding objects
+ New map generation
+ Sophisticated path planning & control
+ Algorithm diversity
+ Sensor & computing fail-operate
+ ASIL-D Functional Safety
+ Excess computing capacity

Level 2



Level 5



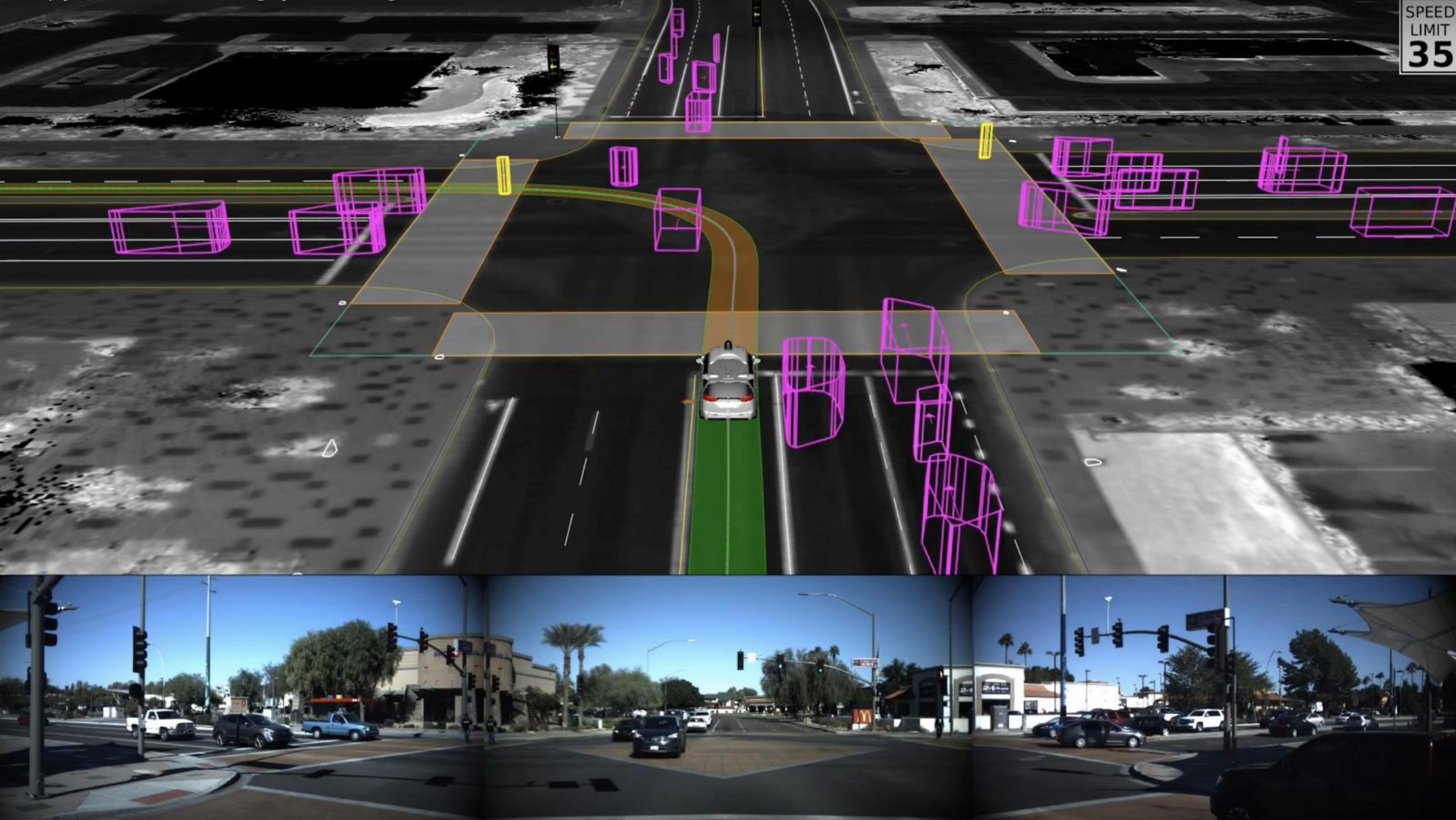
COMPUTATIONAL SCALE REQUIRED

3 million labeled images
1 DGX-1 trains 300k labeled images on 1 DNN in 1 day
10 DNNs required for self-driving
10 parallel experiments at all times
100 DGX-1 per car

BERKELEY DEEP DRIVE

https://arxiv.org/pdf/1805.04687.pdf

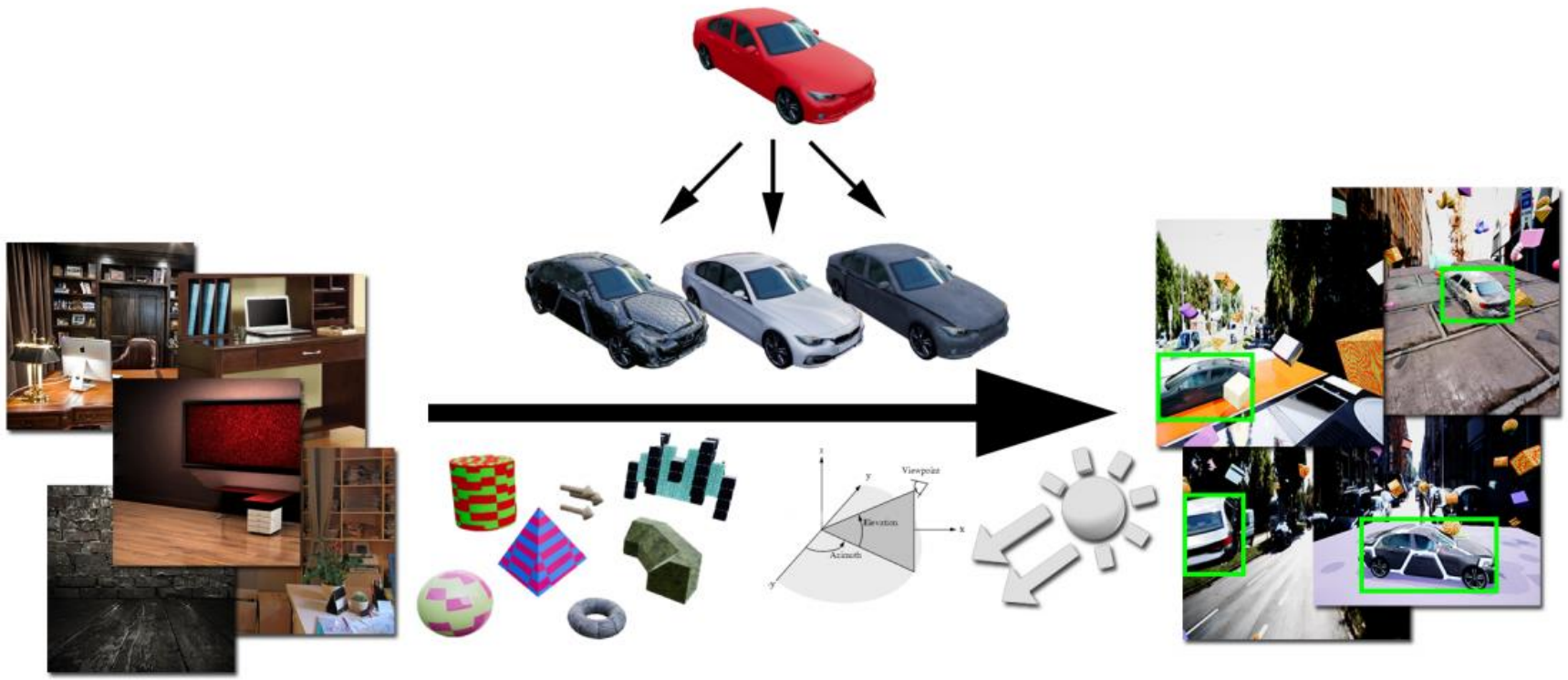
http://bdd-data.berkeley.edu


SPEED
LIMIT
35

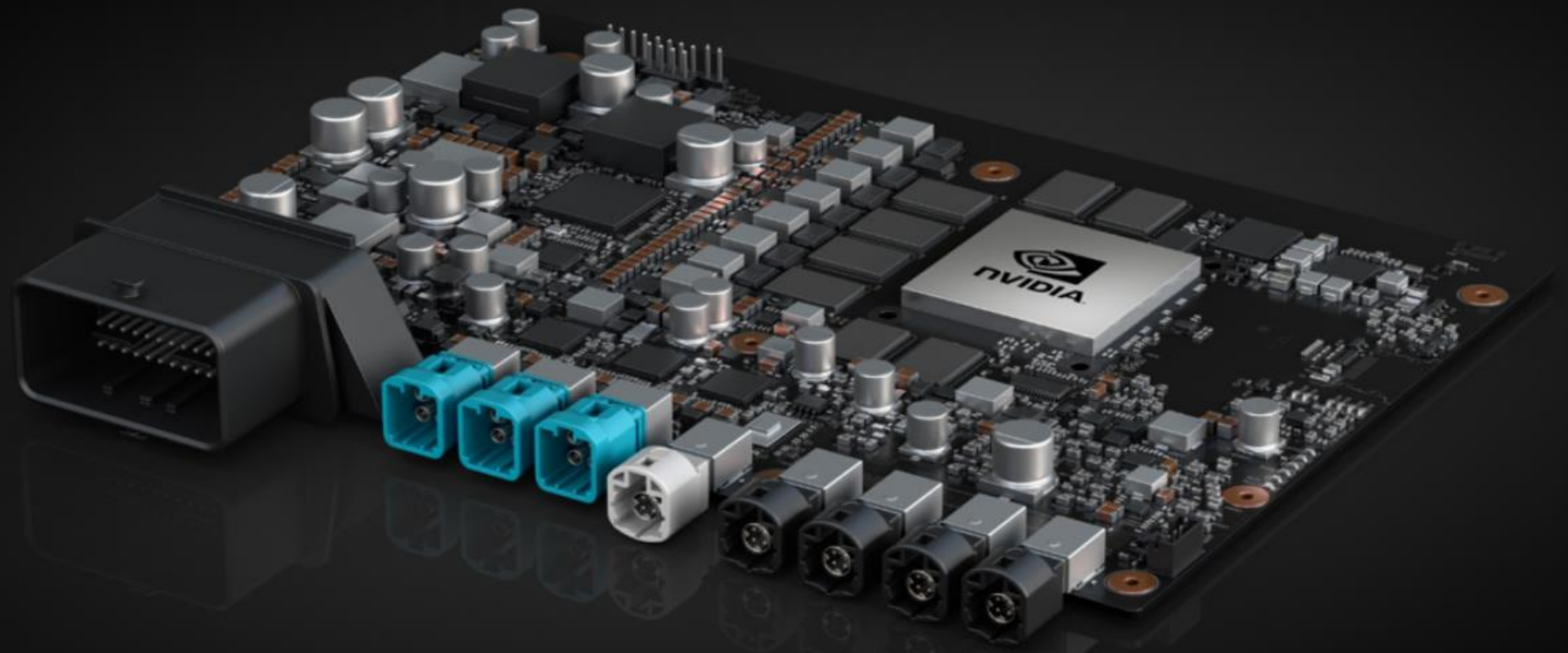
BRIDGING THE REALITY GAP

https://arxiv.org/pdf/1804.06516.pdf NVR+University of Toronto

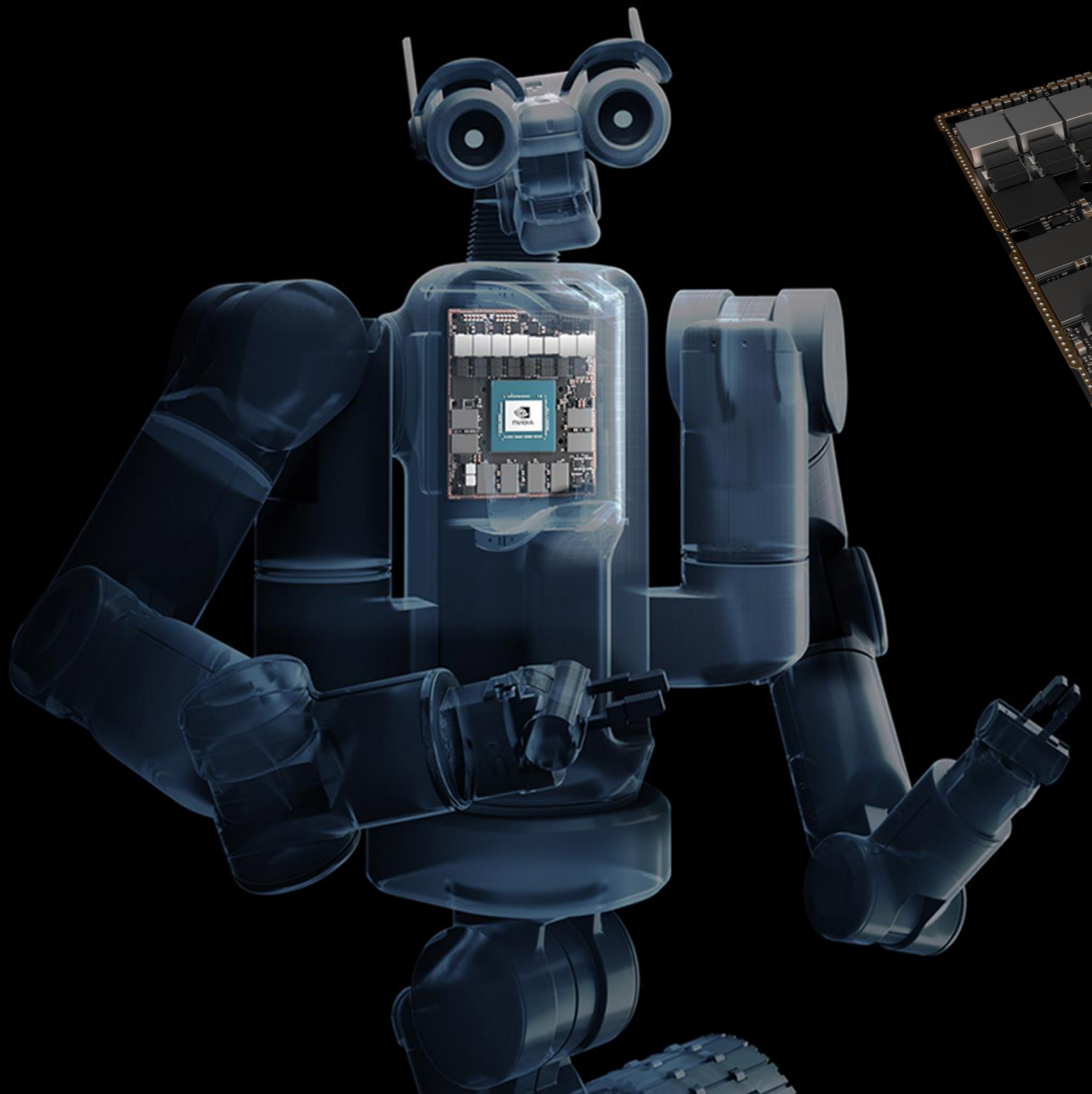
"leveraging the power of synthetic data improves upon results obtained using real data alone"

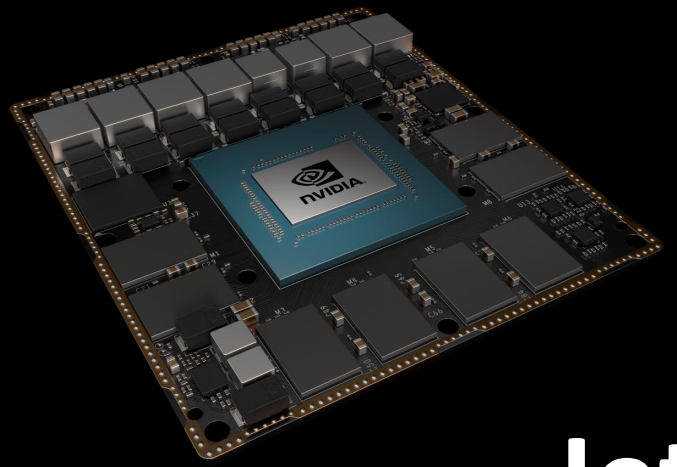


"Our in-depth technical assessment confirms the Xavier SoC architecture is suitable for use in autonomous driving applications and highlights NVIDIA's commitment to enable safe autonomous driving";
TÜV SÜD, Germany



Xavier (ASIL-D rated)



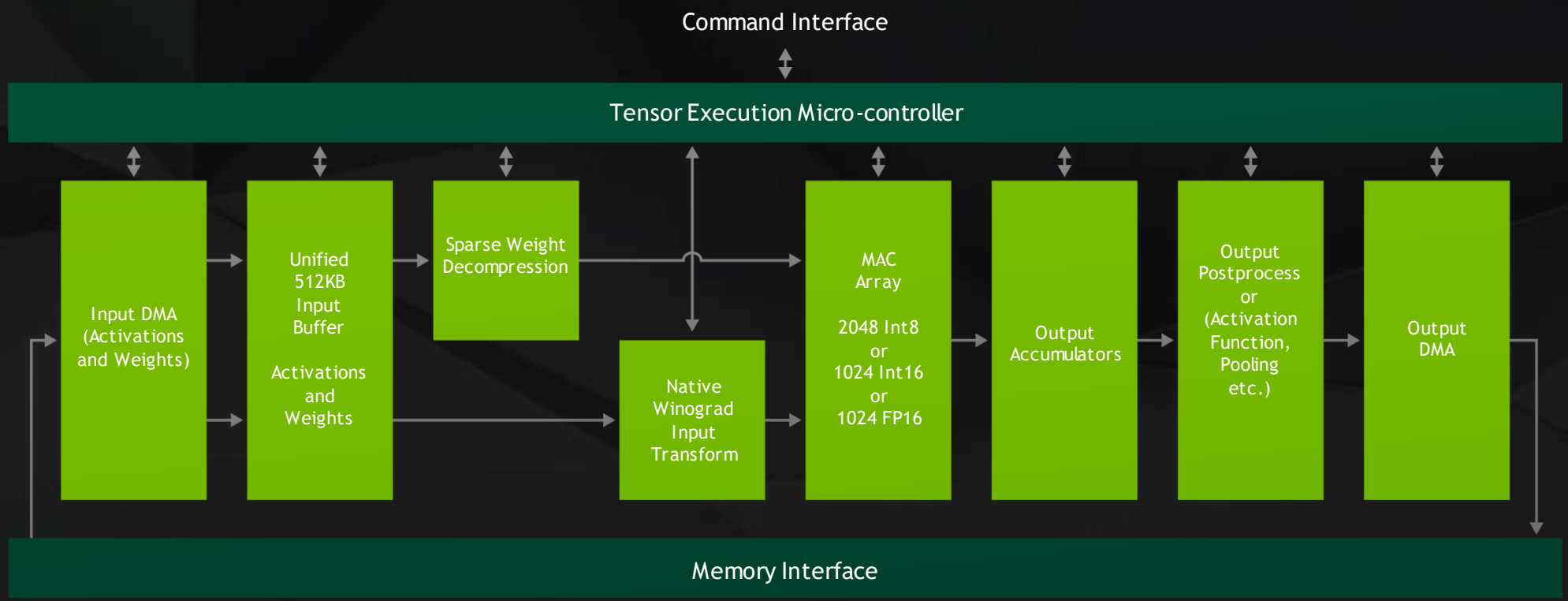


Jetson Xavier

developer.nvidia.com/
jetson-xavier

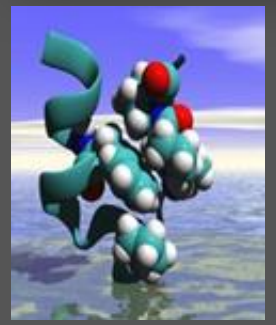
NVIDIA

XAVIER DLA
NOW OPEN SOURCE

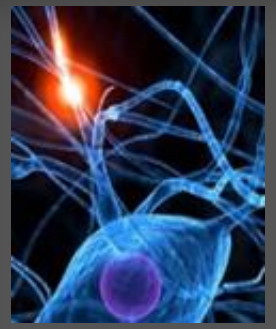


WWW.NVDLA.ORG

A PLETHORA OF HEALTHCARE STORIES

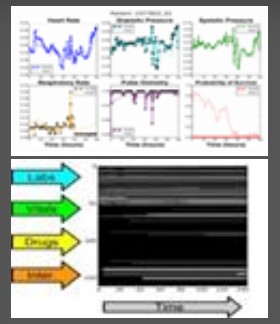


Molecular Energetics For Drug Discovery

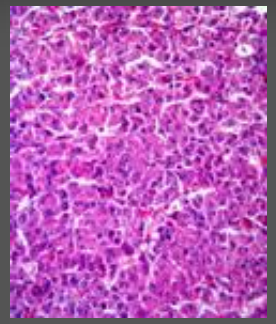


AI for Drug Discovery

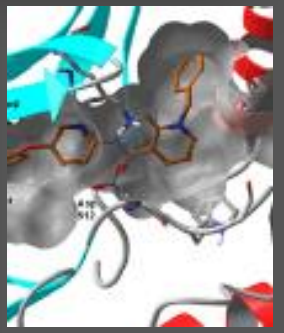
Medical Decision Making



Treatment Outcomes



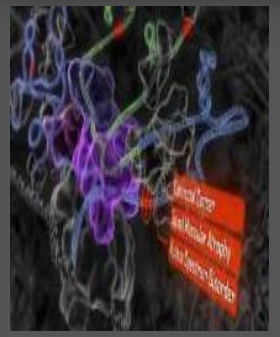
Reducing Cancer Diagnosis Errors by 85%



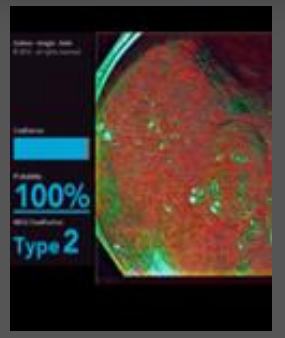
Predicting Toxicology

Predicting Growth Problems

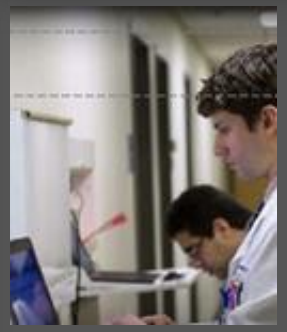
Image Processing



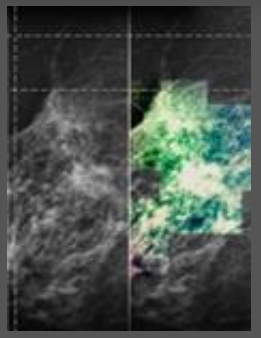
Gene Mutations



Detect Colon Polyps

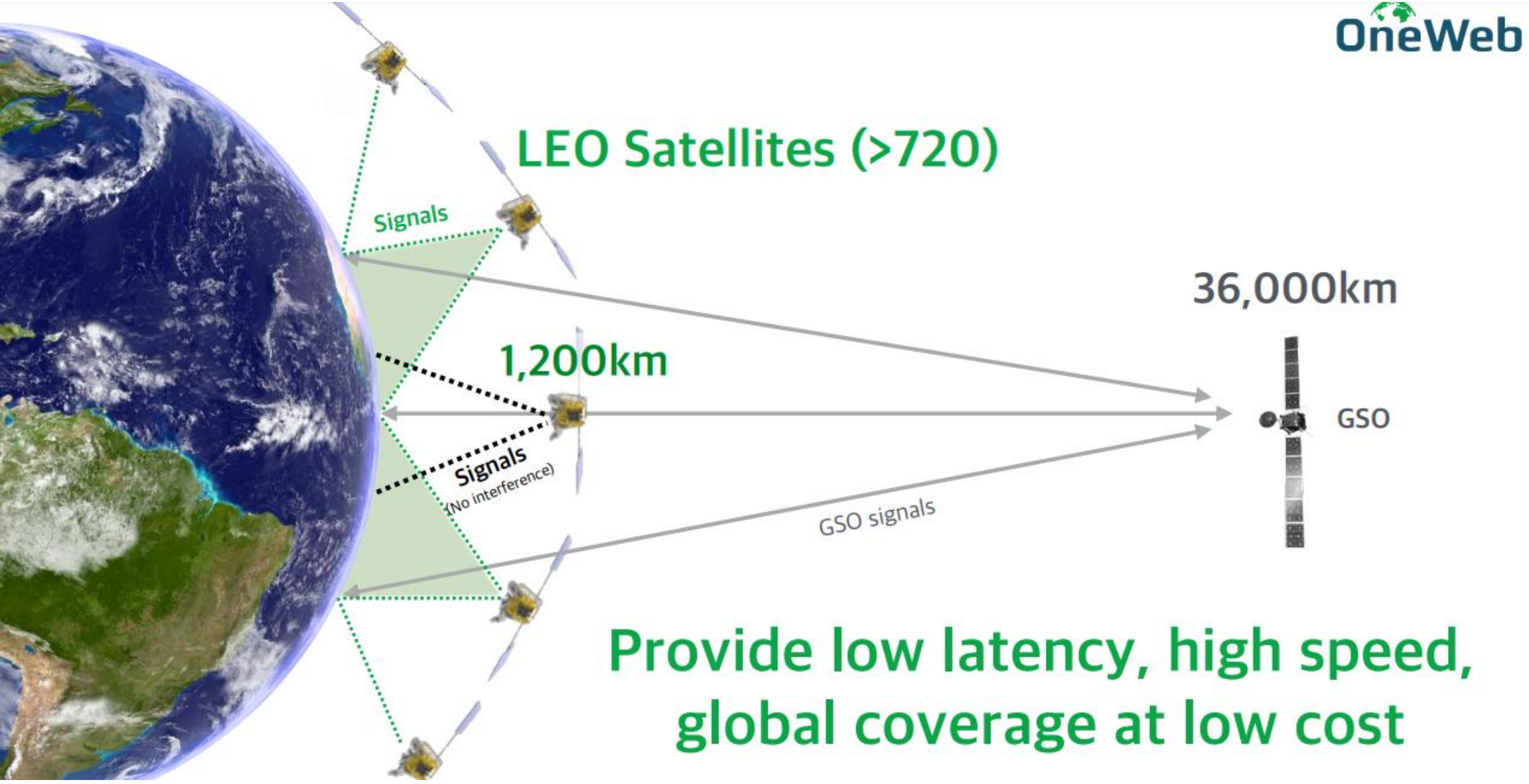


Predicting Disease from Medical Records



Enabling Detection of Fatty Acid Liver Disease

OneWeb



Provide low latency, high speed, global coverage at low cost

HUNTING "GHOST PARTICLES" WITH DEEP LEARNING

Tiny particles called neutrinos are the most abundant form of matter in the universe and understanding their properties is the focus of a world-wide campaign of experiments. Observing these 'ghost particles' in action requires instruments of incredible size and scale. Fermilab's NOvA experiment applies two enormous detectors with a total weight of 30 Million pounds spaced 500 miles apart. It is effectively one of the world's largest cameras, snapping 2 million images per second and analyzing them for neutrino activity. NOvA's scientists developed deep neural networks trained on NVIDIA GPUs with CUDA to improve the machine's detection rate by 33% - increasing the discovery potential of NOvA and other large scale experiments probing fundamental questions of the universe.

“SEEING” GRAVITY FOR THE FIRST TIME

In September 2015, 100 years after Einstein predicted them, gravitational waves were observed for the first time. Astronomers at the Laser Interferometer Gravitational-wave Observatory have since used GPU-powered deep learning to process gravitational wave data 100 times faster than previous methods, making real-time analysis possible and putting us one step closer to understanding the universe’s oldest secrets.

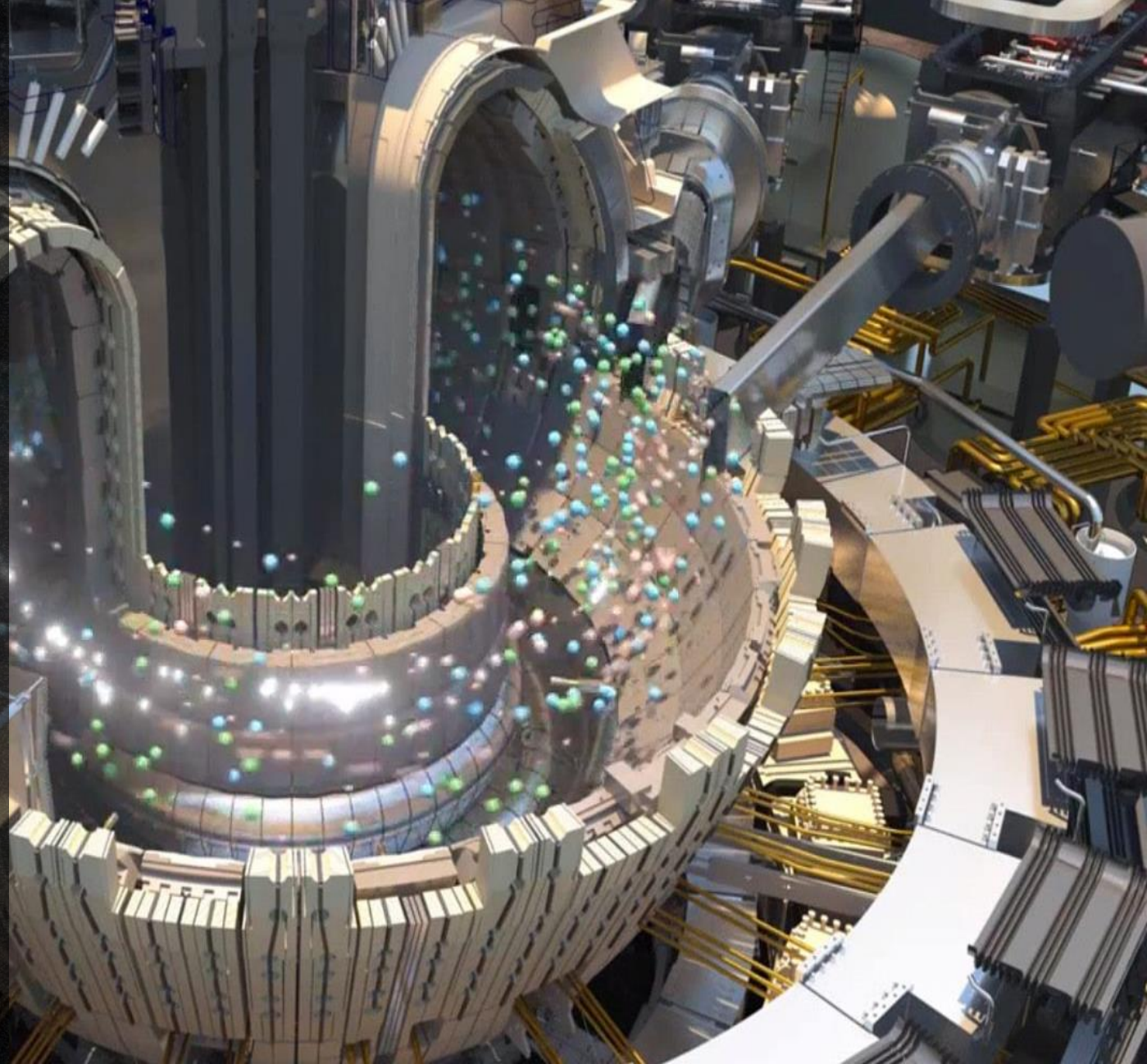
NVIDIA NCSA

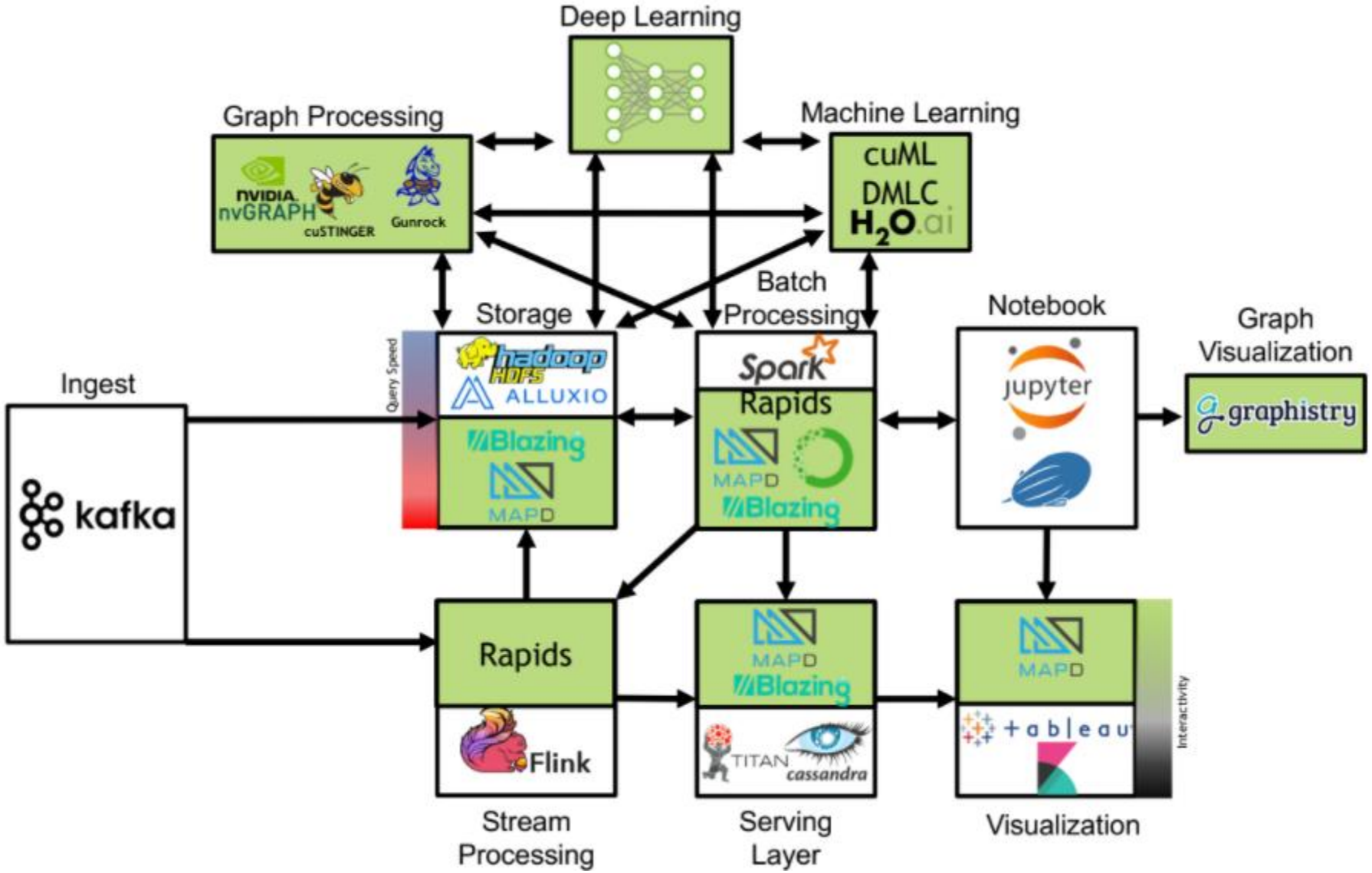
AI IS SPEEDING THE PATH TO FUSION ENERGY

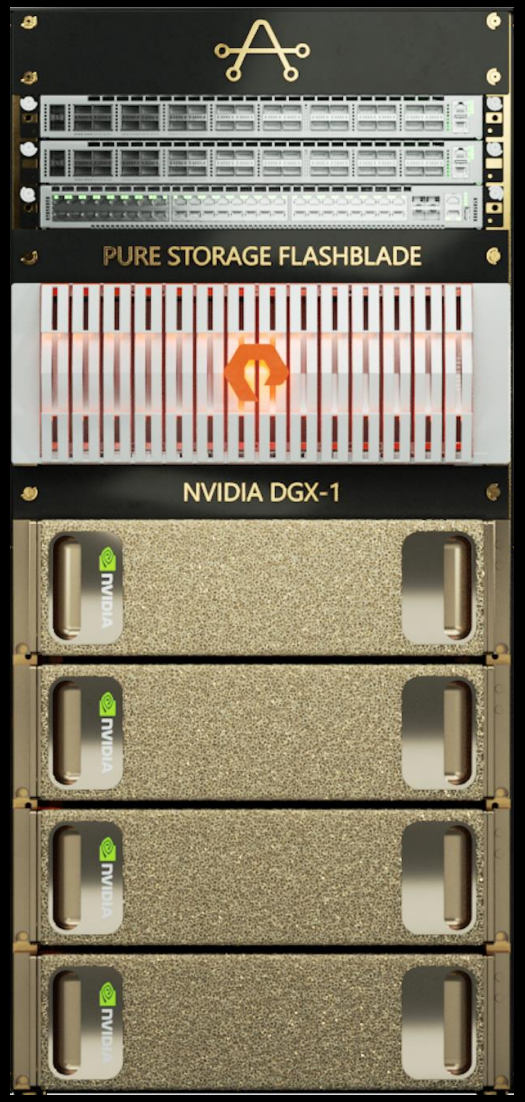
Fusion is the future of energy on Earth. But it's a highly sensitive process where even small environmental disruptions can stall reactions and even damage multi-billion $$ machines. Current models can predict the disruptions with 85% accuracy, but ITER will need something more precise. Researchers at Princeton have developed the Fusion Recurrent Neural Network (FRNN) using deep learning and NVIDIA GPUs to predict disruptions and make adjustments to minimize damage and downtime. Even a 1% improvement in the prediction accuracy can be transformative considering the immense scale and cost of fusion science. Today, FRNN is on the path to achieve 95% accuracy for ITER's tests.

PRINCETON UNIVERSITY

iter




Deep Learning
Graph Processing
Machine Learning
NVIDIA
nvGRAPH
cuSTINGER
Gunrock
cuML
DMLC
H2O.ai
Storage
Batch Processing
Notebook
Graph Visualization
Ingest
Query Speed
hadoop HDFS
ALLUXIO
Spark
Rapids
Blazing
MAPD
jupyter
graphistry
kafka
Rapids
Flink
MAPD
Blazing
TITAN
cassandra
MAPD
+ableau
Interactivity
Stream Processing
Serving Layer
Visualization



PURE STORAGE ANNOUNCES **AIRI:** AI READY INFRASTRUCTURE

- Joint Reference Architecture from Pure Storage and NVIDIA
- Simplified, converged infrastructure solution built on DGX-1 and FlashBlade
- Available through select NPN partners as a turnkey solution

---

HARDWARE

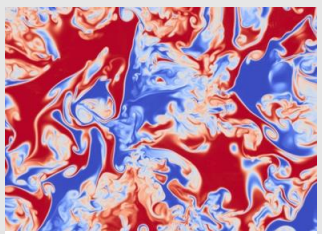
NVIDIA DGX-1 | 4x DGX-1 Systems | 4 Tensor PFLOPS
PURE FLASHBLADE™ | 15x 17TB Blades | 1.5M IOPS
ARISTA | 2x 100Gb Ethernet Switches with RDMA

SOFTWARE

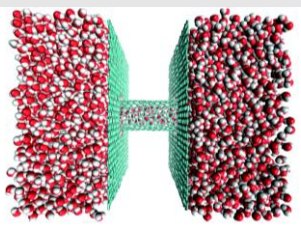
NVIDIA GPU CLOUD DEEP LEARNING STACK | NVIDIA Optimized Frameworks
AIRI SCALING TOOLKIT | Multi-node Training Made Simple

AI SUPERCOMPUTING WILL TRANSFORM HPC

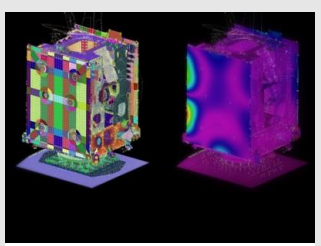
Extending Reach of HPC By Combining Computational & Data Science



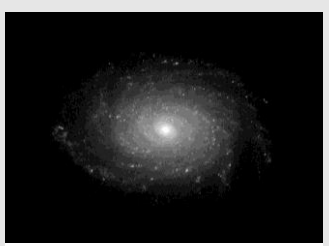
Turbulent Flow



Molecular Dynamics



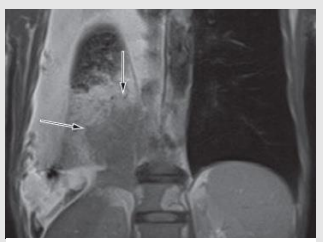
Structural Analysis



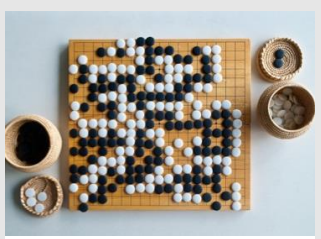
N-body Simulation

COMPUTATIONAL SCIENCE

"What's happening?"



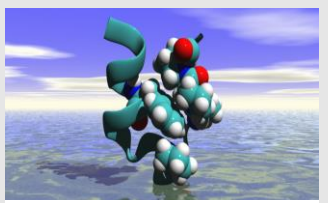
"Is there cancer?"



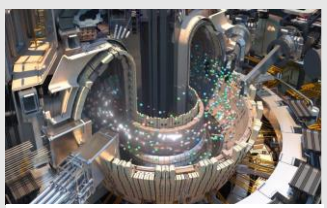
"Next move?"

"What does she mean?"

DATA SCIENCE



Drug Discovery



Clean Energy

Understanding Universe

Monitoring Climate Change

COMPUTATIONAL & DATA SCIENCE


NVIDIA
GPU-ACCELERATED
APPLICATIONS

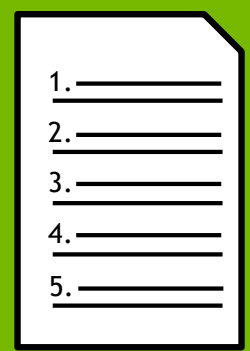
GPU READY APPS

AWARENESS : ADOPTION : UTILIZATION

Target Customers

- App Users — Life Science
- IT Managers — HPC/Data Centers
- Developers — Deep Learning

Quick Start Guide



Simple, ISV-approved, step-by-step instructions

User Benefits

- ✓ Easy setup
- ✓ Best app practices
- ✓ Optimal results
- ✓ Higher productivity
- ✓ Faster discoveries

www.nvidia.com/gpu-ready-apps

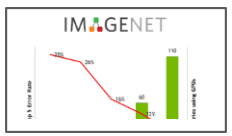
POWERING THE DEEP LEARNING ECOSYSTEM

NVIDIA SDK accelerates every major framework

COMPUTER VISION

OBJECT DETECTION

IMAGE CLASSIFICATION



SPEECH & AUDIO

VOICE RECOGNITION

LANGUAGE TRANSLATION

NATURAL LANGUAGE PROCESSING

RECOMMENDATION ENGINES

SENTIMENT ANALYSIS

DEEP LEARNING FRAMEWORKS

Caffe

Caffe2

Chainer

Microsoft Cognitive Toolkit

mxnet

PYTORCH

theano

torch

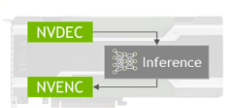
TensorFlow

NVIDIA DEEP LEARNING SDK

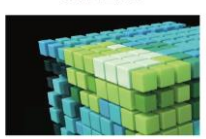
cuDNN

TensorRT

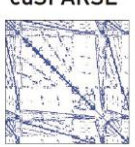
DeepStream SDK



cuBLAS



cuSPARSE

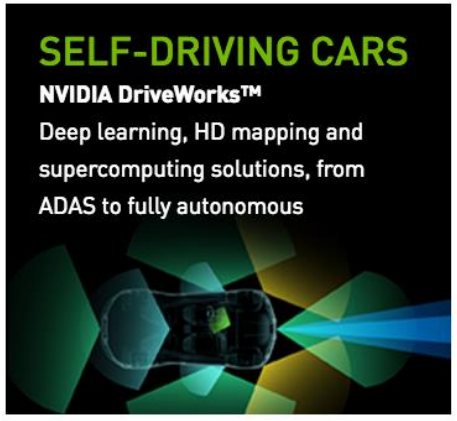


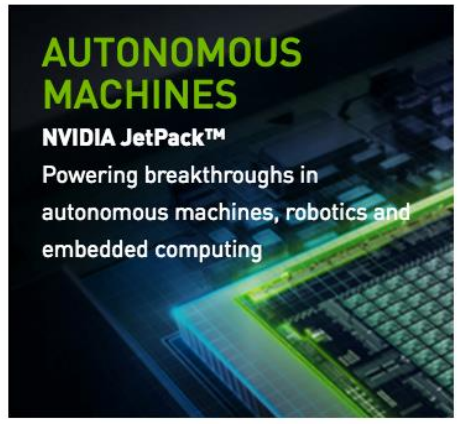
NCCL

developer.nvidia.com/deep-learning-software

NVIDIA SDK

developer.nvidia.com



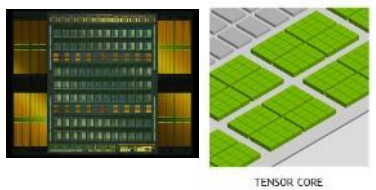


CUDA TOOLKIT 9

UNLEASHES POWER OF VOLTA

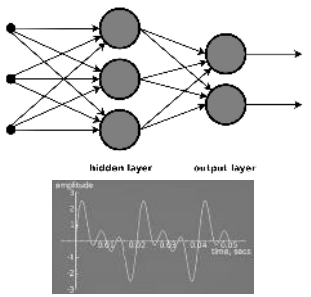
Optimized for Volta:

- Tensor Cores
- Second-Generation NVLink
- HBM2 Stacked Memory



FASTER LIBRARIES

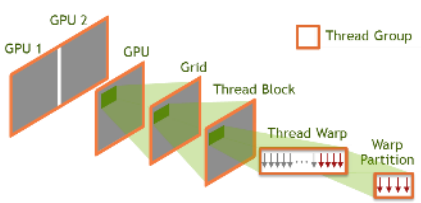
- GEMM Optimizations for RNNs (cuBLAS)
- >20x Faster Image Processing (NPP)
- FFT Optimizations Across Various Sizes (cuFFT)



COOPERATIVE THREAD GROUPS

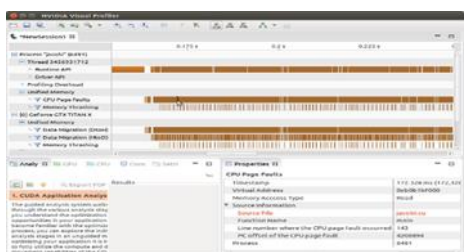
Flexible Thread Groups
Efficient Parallel Algorithms

- Synchronize Across Thread Blocks in a Single GPU or Multi-GPUs



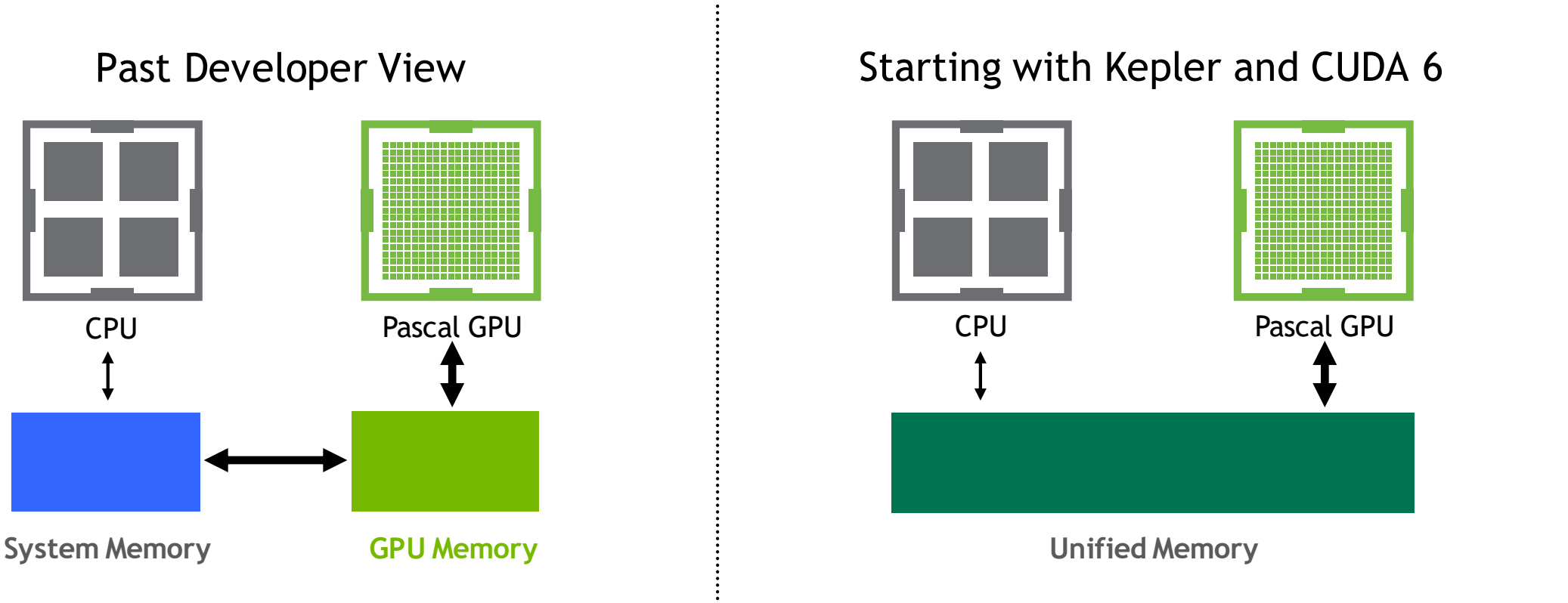
DEVELOPER TOOLS & PLATFORM UPDATES

- 1.3x Faster Compiling
- New OS and Compiler Support
- Unified Memory Profiling
- NVLink Visualization



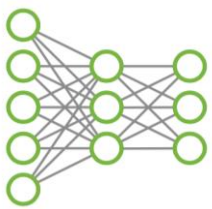
UNIFIED MEMORY

Implicit Memory Management



NVIDIA DEEP LEARNING SDK UPDATE

GPU-accelerated DL Primitives



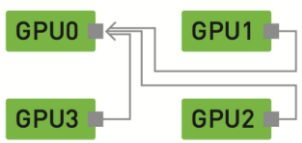
cuDNN 7

Faster training

Optimizations for RNNs

Leading frameworks support

Multi-GPU & Multi-node



NCCL 2

Multi-node distributed training (multiple machines)

Leading frameworks support

High-performance Inference Engine

TensorRT 4

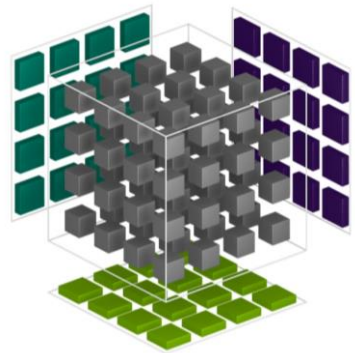
TensorFlow model reader

Object detection

INT8 RNNs support

TENSOR CORE

Mixed Precision Matrix Math
4x4 matrices



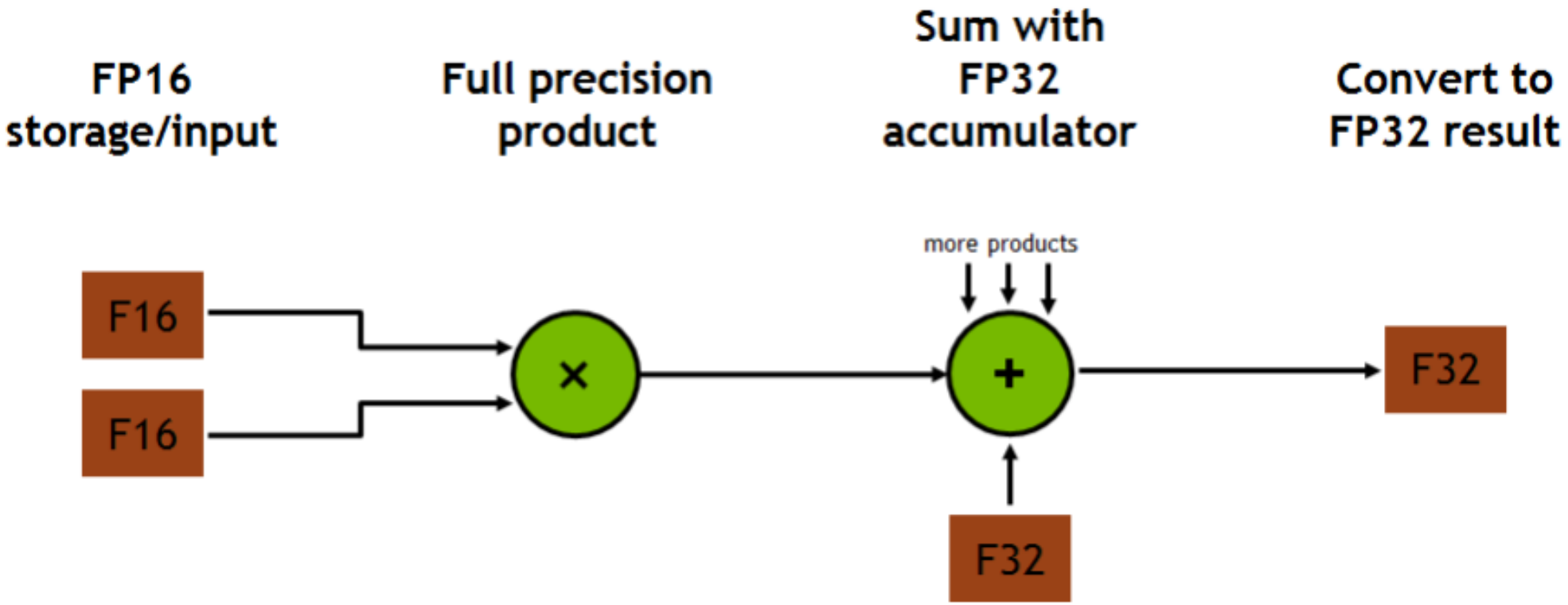
$$D = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}$$

FP16 or FP32 — FP16 — FP16 — FP16 or FP32

$$D = AB + C$$

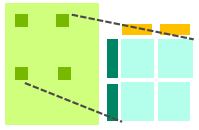
TENSORCORES

D = A * B + C (4x4)



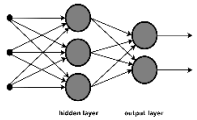
PROGRAMMING TENSOR CORES

Accelerate Training and Inference



CUTLASS

- Template library for custom matrix & linear algebra
- Achieves >90% of tuned CuBLAS



cuBLAS

- Highly optimized linear algebra
- Tensor Cores speed up GEMMs

cuDNN

- Tensor Cores speed up convolutions and RNNs

Caffe2 Chainer

mxnet PYTORCH

TensorFlow theano

Mixed Precision Best Practices
GTC Talk

CUDA Libraries
GPU Accelerated Libraries

All Major Frameworks
NGC

All Model Types
Tensor Core Optimized Examples

Tensor Cores: More resources

Examples
New mixed-precision model examples: https://developer.nvidia.com/deep-learning-examples
GitHub: https://github.com/NVIDIA/DeepLearningExamples

Tools
TensorFlow OpenSeq2seq: https://nvidia.github.io/OpenSeq2Seq/html/mixed-precision.html & arVix paper
PyTorch Apex: https://nvidia.github.io/apex/fp16_utils.html & NVIDIA developer news article

Further information
Mixed-precision blog: https://devblogs.nvidia.com/mixed-precision-training-deep-neural-networks/
Mixed-precision best practices: https://docs.nvidia.com/deeplearning/sdk/mixed-precision-training/index.html
Mixed-precision arVix paper: https://arxiv.org/abs/1710.03740
GTC 2018 Sessions: Training with Mixed Precision: Theory and Practice and Training with Mixed Precision: Real Examples

DGX/NGC deep learning framework containers: Latest versions of software stack and Tensor Core optimized examples - **DGX/NGC Registry**

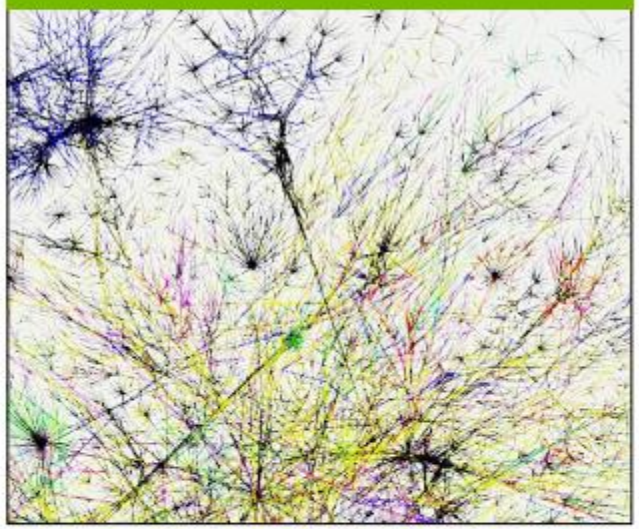
Call to action:

1) **Share these resources with your customers and partners**
2) **Provide feedback to help us prioritize new examples and improve examples**

GRAPH ANALYTICS

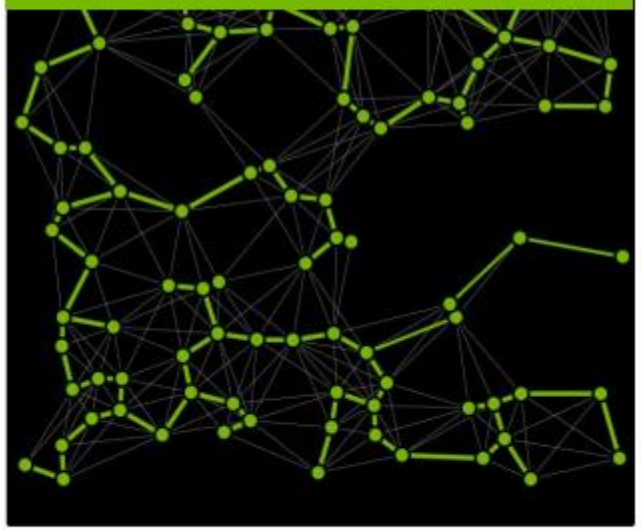
Insight from connections in big data

SOCIAL NETWORK ANALYSIS

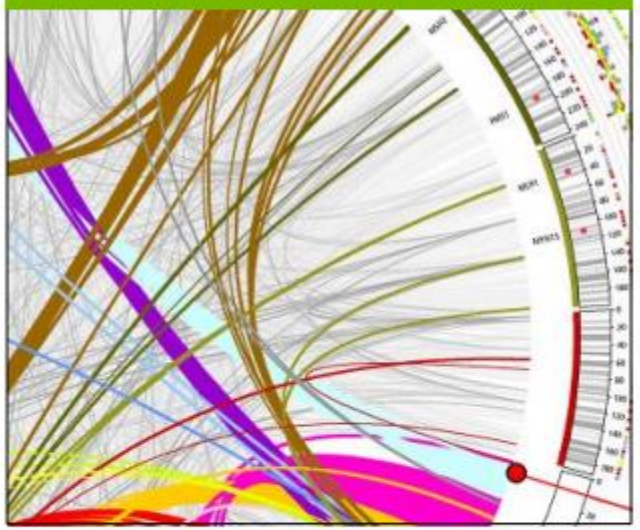


Wikimedia Commons

CYBER SECURITY / NETWORK ANALYTICS



GENOMICS



Circos.ca

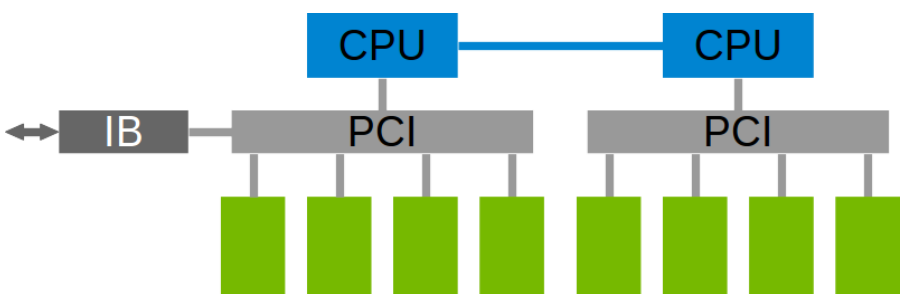
… and much more: Parallel Computing, Recommender Systems, Fraud Detection, Voice Recognition, Text Understanding, Search

NCCL 2.0

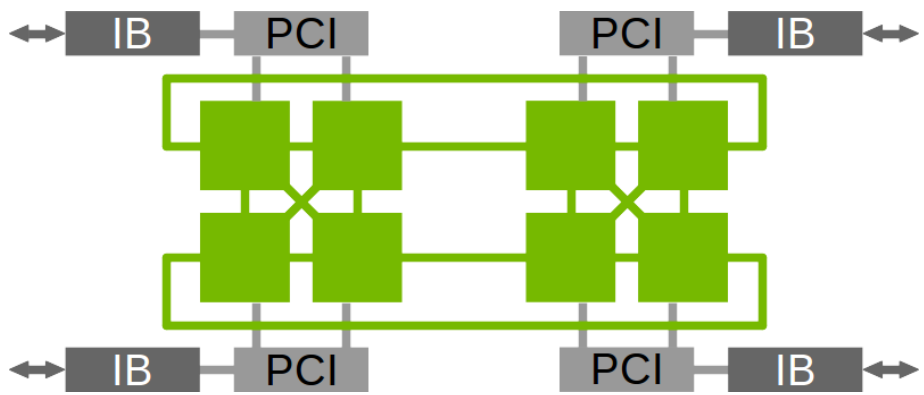
Inter-node communication

Inter-node communication using Sockets or Infiniband verbs, with multi-rail support, topology detection and automatic use of GPU Direct RDMA.

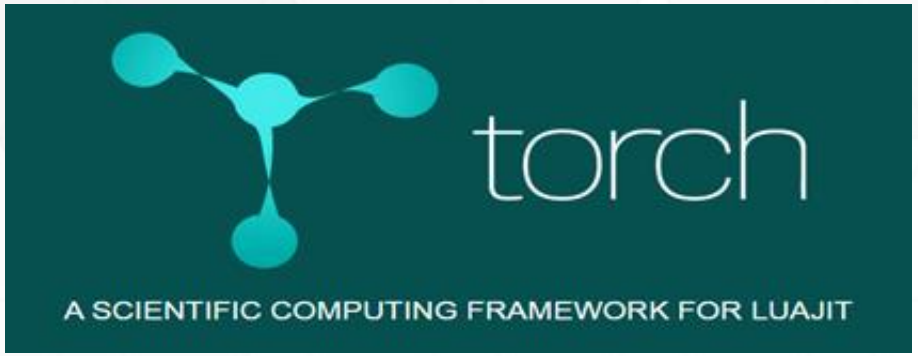
Optimal combination of NVLink, PCI and network interfaces to maximize bandwidth and create rings across nodes.



PCIe, Infiniband



DGX-1 : NVLink, 4x Infiniband



Circa 2000 - Torch7 - 4th (using odd numbers only 1,3,5,7)
Web-scale learning in speech, image and video applications

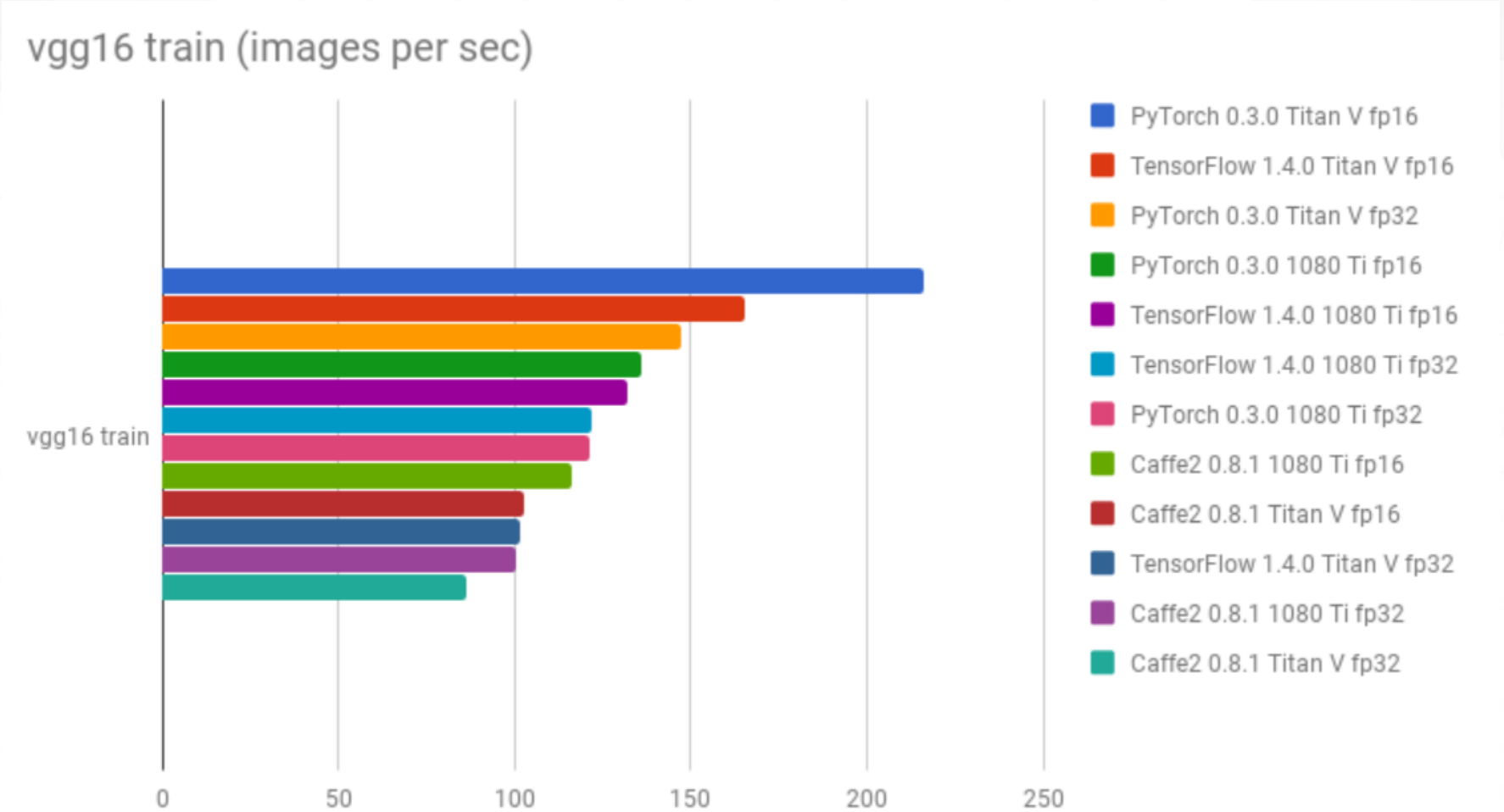
Maintained by top researchers including
Soumith Chintala - Research Engineer @ Facebook

All the goodness of **Torch7** with an intuitive Python frontend that focuses on rapid prototyping, readable code & support for a wide variety of deep learning models.

https://pytorch.org/2018/05/02/road-to-1.0.html

Benchmarks Jan 2018

https://github.com/u39kun/deep-learning-benchmark



Apex - A PyTorch Extension

Overview

- Goal: Raise PyTorch customer awareness and increase adoption of NVIDIA Tensor Cores
- Content: Provide an easy to use set of utility functions in PyTorch for mixed-precision optimizations
- Benefit: Few lines of code to achieve improved training speed while maintaining accuracy and stability of single precision (Tensor Cores)
- Target audience: Deep learning researchers and developers of PyTorch with NVIDIA Volta
- Key Features: AMP (Auditor for mixed-precision) and Optimizer Wrapper (Dynamic loss scaling and master parameters)
- Teams Involved: Leading NVIDIA PyTorch team and collaboration with external FB PyTorch team

Chainer

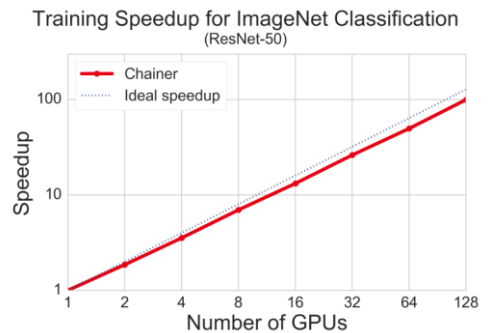
Qualities

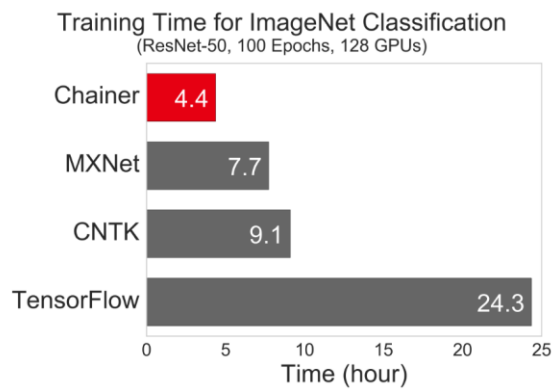
- Dynamic computation graphs with a Python API
- Models that are fast to prototype and easier to debug.
- CuPy: NumPy-equivalent multi-dimensional array-library powered by CUDA
- Extensions & Tools: ChainerRL, ChainerMN, for computer vision

End Users/ Notes

- Distributed Learning
- Expect to see Caffe2 overtake Caffe

Benchmark





MX NET + Apache

https://github.com/dmlc/
https://github.com/NVIDIA/keras

Efficiency

Portability

Flexibility

MULTI CORE – MULTI GPU – MULTI NODE

Tofu: Parallelizing Deep Learning Systems with Auto-tiling

Memonger: Training Deep Nets with Sublinear Memory Cost

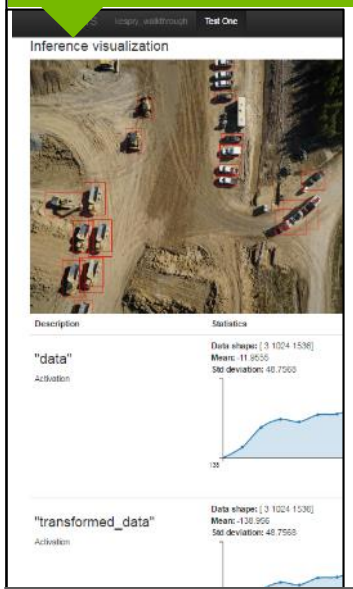
MinPy: High Performance System with NumPy Interface

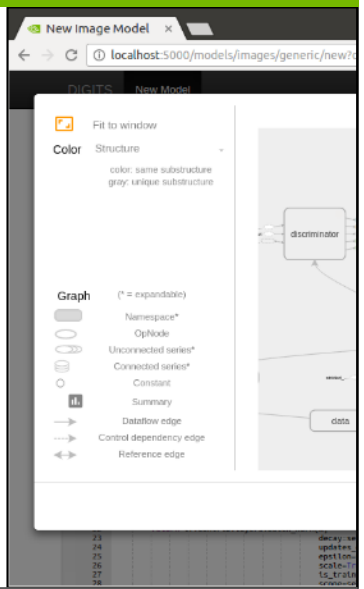
NYU

https://devblogs.nvidia.com/parallelforall/scaling-keras-training-multiple-gpus/

WHAT'S NEW IN DIGITS 6?

TENSORFLOW SUPPORT





Train TensorFlow Models Interactively with DIGITS

NEW PRE-TRAINED MODELS

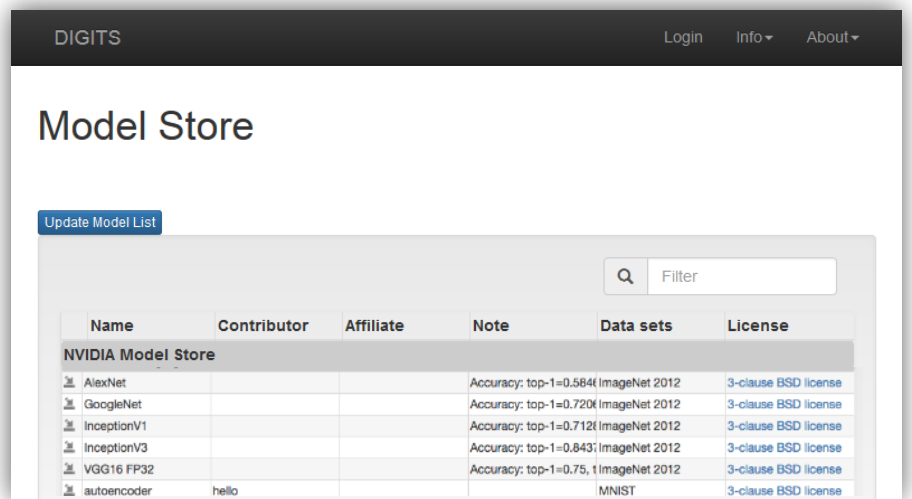
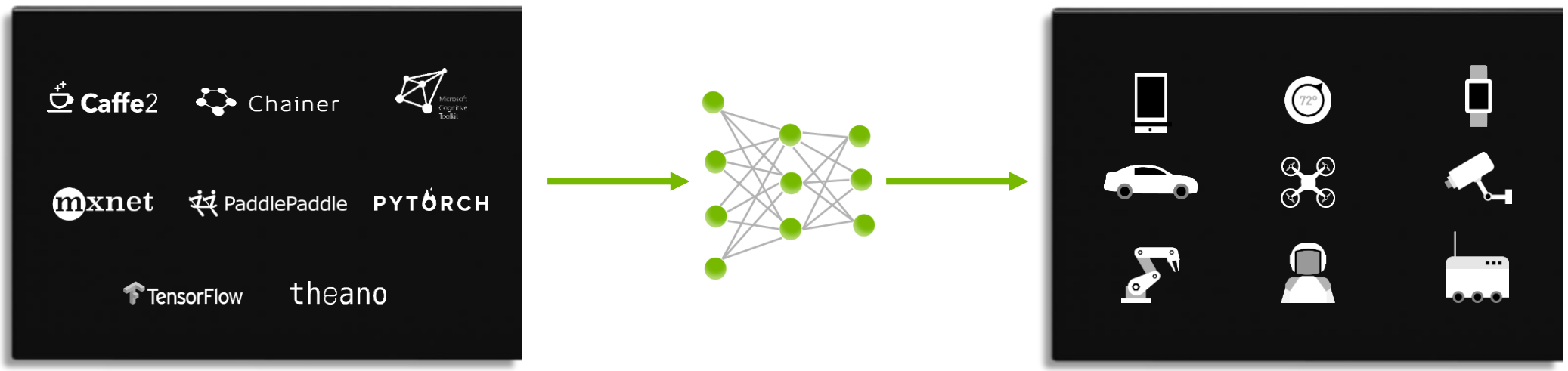


Image Classification: VGG-16, ResNet50
Object Detection: DetectNet

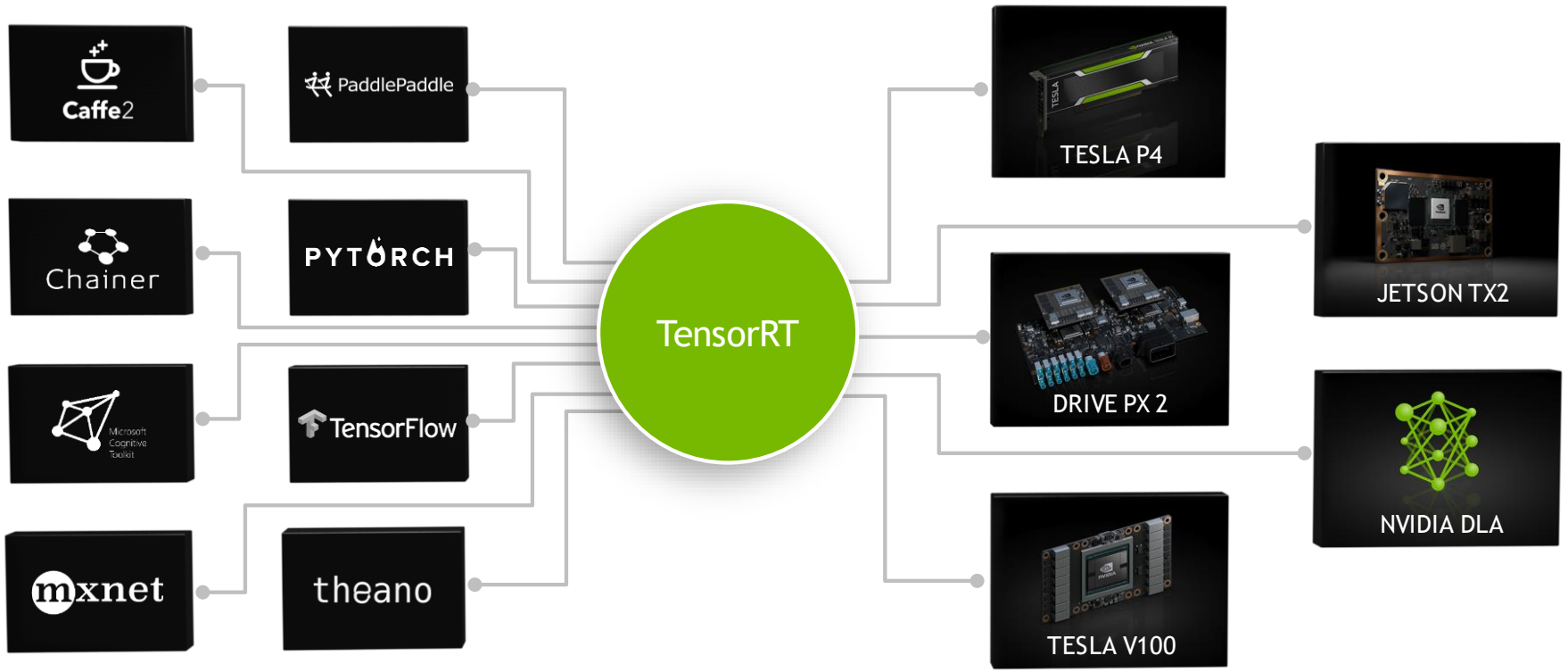
AI INFERENCE NEEDS TO RUN EVERYWHERE



Training DNN Model Inferencing

NEW NVIDIA TENSORRT 4

Programmable Inference Accelerator



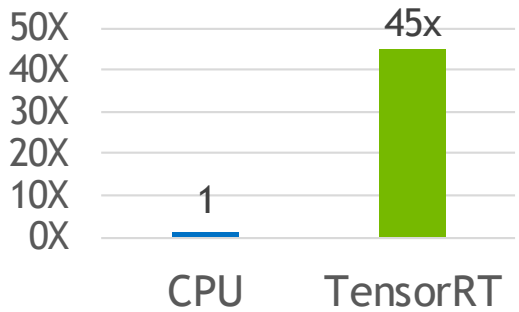
Compile and Optimize Neural Networks | Support for Every Framework
Optimize for Each Target Platform


Skylake
TensorFlow on CPU
Images Per Sec: 5.0

NVIDIA TensorRT 4

RNN and MLP Layers • ONNX Import • NVIDIA DRIVE Support

Recommender, Speech & Machine Translation



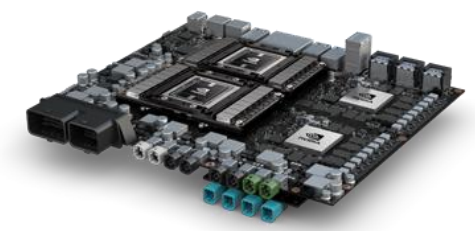
Accelerate inference of recommenders, speech and machine translation apps with new layers and optimizations

50x Faster Inference for ONNX Models

ONNX
Caffe2
Chainer
Microsoft CNTK
mxnet
PYTORCH

Easily import and accelerate inference for ONNX frameworks with native ONNX parser in TensorRT

Support for NVIDIA DRIVE Xavier

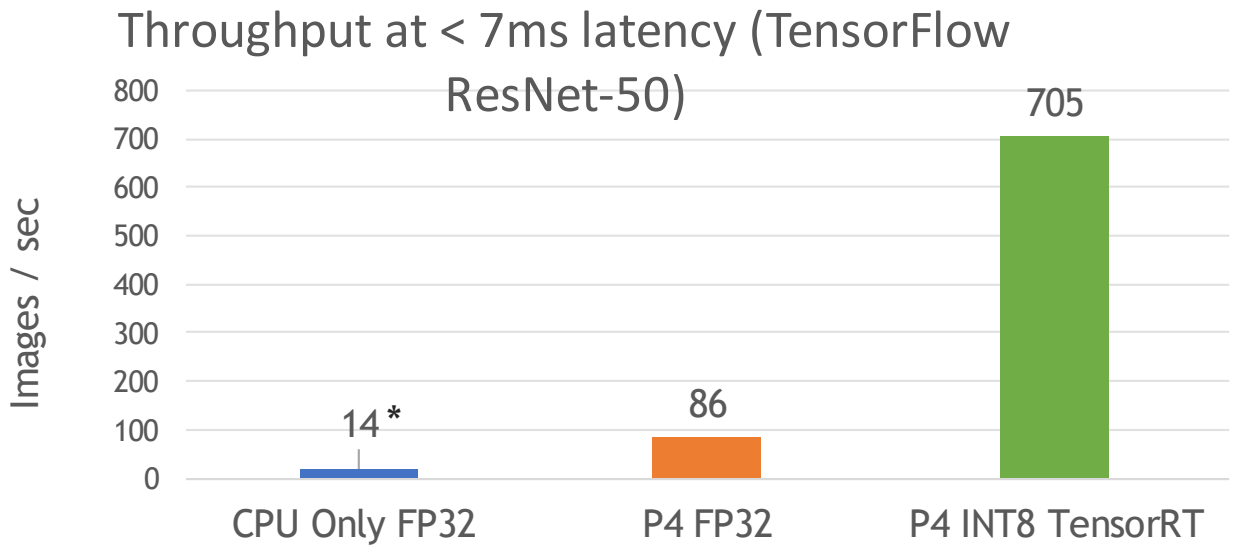


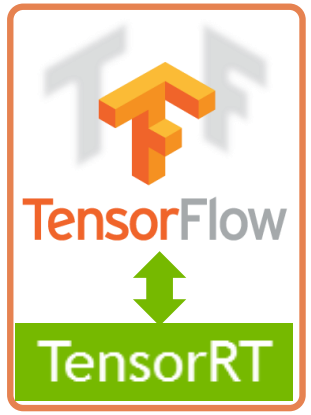
Deploy optimized deep learning inference models NVIDIA DRIVE Xavier

Free download to members of NVIDIA Developer Program soon at
developer.nvidia.com/tensorrt

TensorRT INTEGRATED WITH TensorFlow

8x faster Inference Than TensorFlow Only





Available in TensorFlow 1.7

https://github.com/tensorflow/tensorflow

* Min CPU latency measured was 70 ms. It is not < 7 ms.
CPU: Skylake Gold 6140, 2.5GHz, Ubuntu 16.04; 18 CPU threads.
Pascal P4; CUDA (384.111; v9.0.176);
Batch size: CPU=1, TF_GPU=1 (latency 12 ms) , TF-TRT=4 w/ latency=6ms


arm
NVIDIA

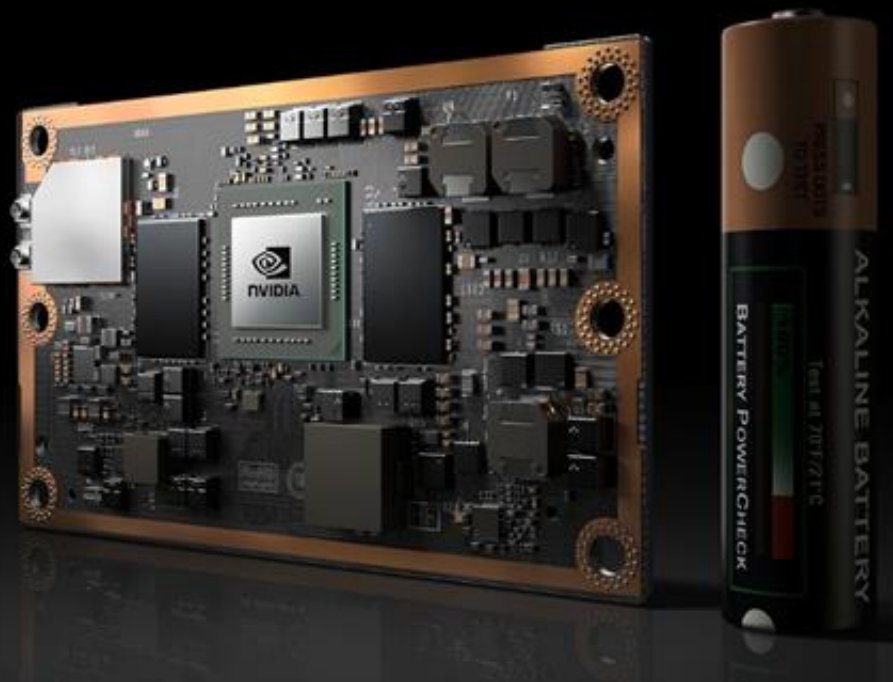
JETSON TX2

SUPERCOMPUTER FOR AI AT THE EDGE

2 Core i7 PCs in <10W

256 CUDA cores

>1 TFLOPS



cuDNN, TensorRT

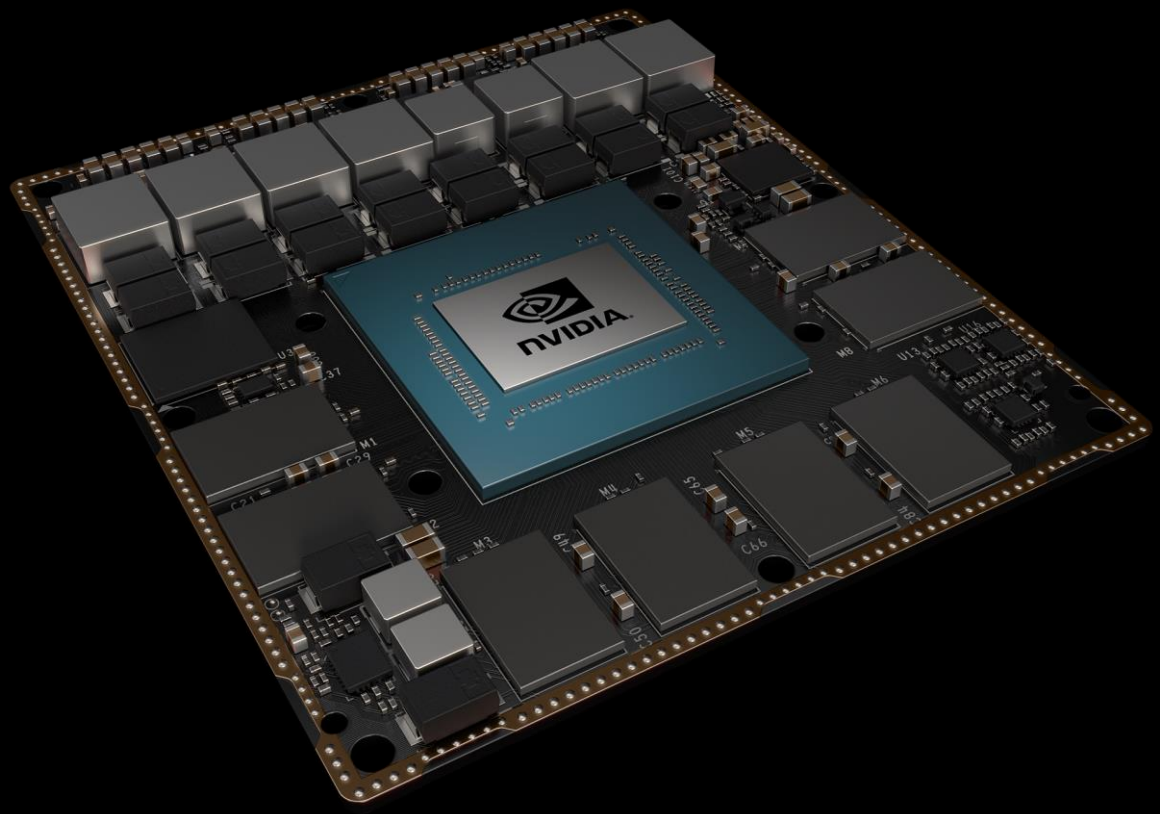
CUDA

Linux, ROS

ANNOUNCING:
JETSON XAVIER

Computer for Autonomous Machines

AI Server Performance in 30W • 15W • 10W
512 Volta CUDA Cores • 2x NVDLA
8 core CPU
30 DL TOPS



JETSON XAVIER DEVELOPER KIT

$1299 (US)

Available from distributors WW

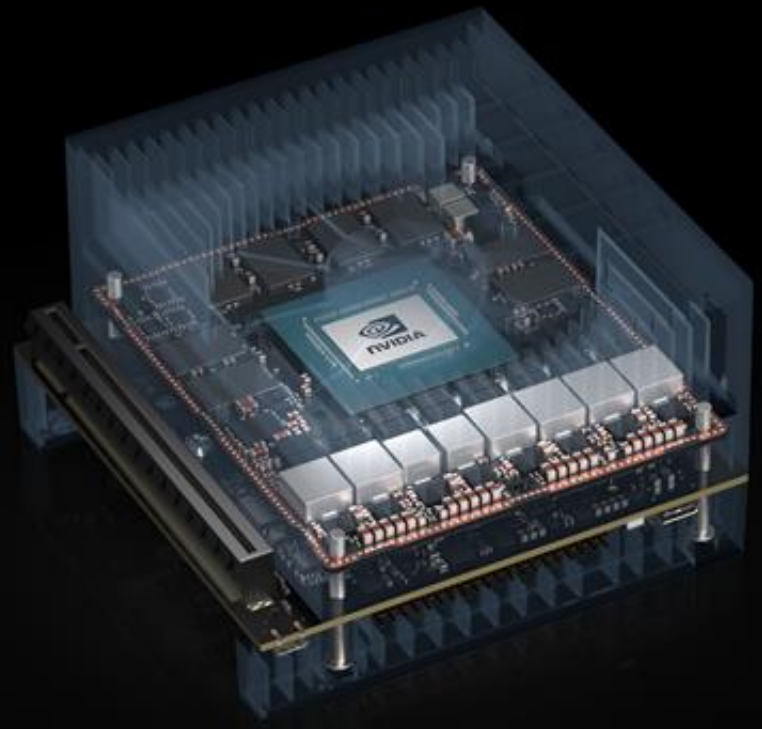
Early access August 2018

JETSON XAVIER DEVELOPER KIT

$1299 (US)

Available from distributors WW

Early access August 2018



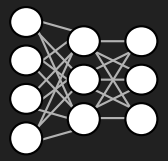
JETPACK SDK FOR AI @ THE EDGE

DEVELOPER.NVIDIA.COM/EMBEDDED-COMPUTING

Sample Code

Nsight Developer Tools

Multimedia API



TensorRT
cuDNN

Deep Learning

VisionWorks
OpenCV

Computer Vision

Vulkan
OpenGL

Graphics

libargus
Video API

Media

CUDA, Linux4Tegra, ROS

Jetson Embedded Supercomputer: Advanced GPU, 64-bit CPU, Video CODEC, VIC, ISP

NVIDIA

NEW YORK UNIVERSITY DEEP LEARNING

Robotics Teaching Kit with ‘Jet’ - ServoCity

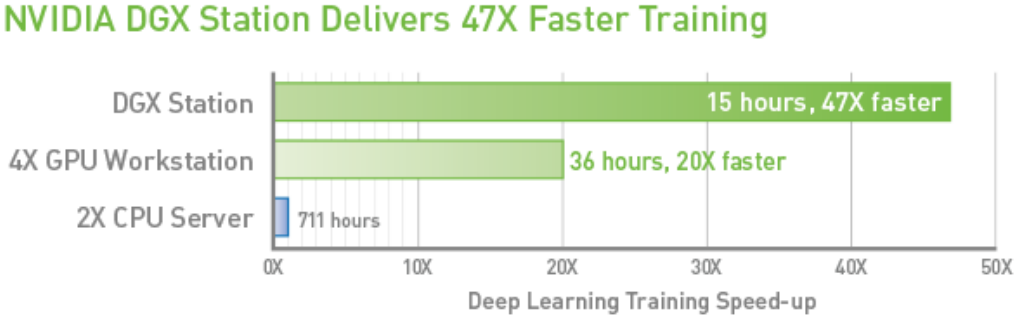
Available to Instructors Now!

developer.nvidia.com/teaching-kits

Now with 32GB per GPU, same cost

GROUNDBREAKING AI AT YOUR DESK

Designed for your office DGX Station is the world's first personal supercomputer for leading-edge AI development. Built on the same Deep Learning Stack powering all NVIDIA DGX™ Systems, you can now experiment at your desk and extend your work across DGX Systems and the cloud.



DGX-1 POWERS FASTER, MORE EFFICIENT DRUG DISCOVERY

The high cost of drug discovery is driving researchers and pharmaceutical companies to turn to AI as a faster, more efficient way to develop new drugs.

Professor Okuno, Kyoto University and RIKEN, have formed the Life INtelligence Consortium (LINC) to build an AI drug discovery ecosystem in Japan. LINC uses the NVIDIA DGX-1 AI supercomputer—the DGX-1 delivers the extreme performance LINC needs to solve complex problems and speed drug discovery.

Now with 32GB per GPU, same cost

DGX POD ARCHITECTURE

a single data center rack containing up to 9x NVIDIA DGX-1 servers, storage, networking & NVIDIA AI software

Nine DGX-1 servers

12 storage servers

10 GbE (min) storage & management switch

Mellanox 100 Gpps intra-rack high speed network switches.

ANNOUNCING NVIDIA DGX-2

THE LARGEST GPU EVER CREATED

2 PFLOPS | 512GB HBM2 | 10 kW | 350 lbs

NVIDIA GPU CLOUD

Optimized Stacks for Every Cloud

GPU CONTAINERS

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

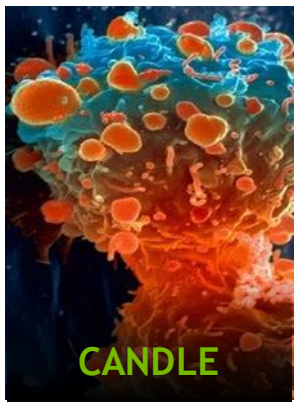
HPC
GAMESS
Gromacs
LAMMPS
NAMD
RELION
CHROMA
MILC
CANDLE
Lattice Microbes

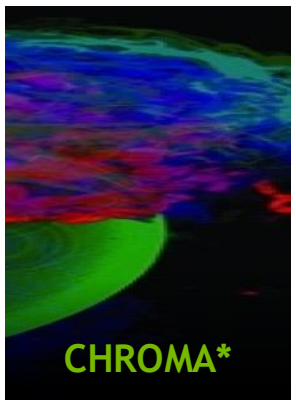
HPC Viz
ParaView Holodeck
ParaView IndeX
ParaView OptiX
IndeX
VMD

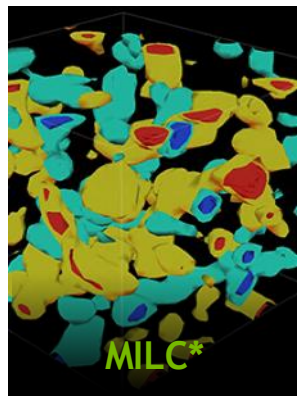
ANALYTICS
MapD
Kinetica

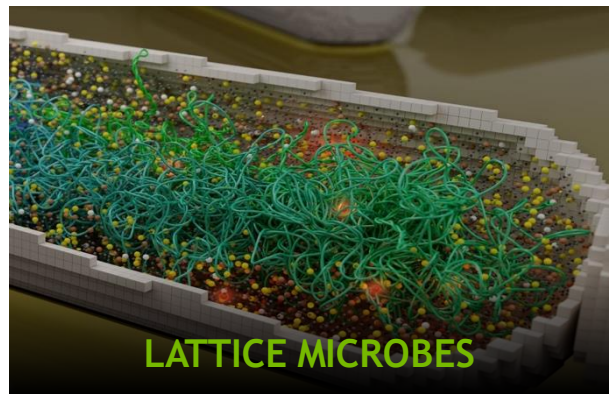
20,000+ Registered Organizations | 30 Containers
NOW on AWS, GCP, AliCloud, Oracle Cloud, DGX

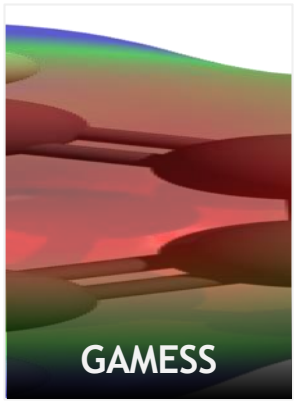
HPC APPS CONTAINERS ON NVIDIA GPU CLOUD

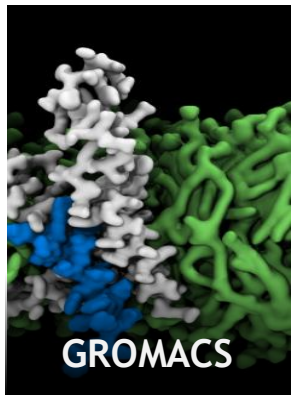


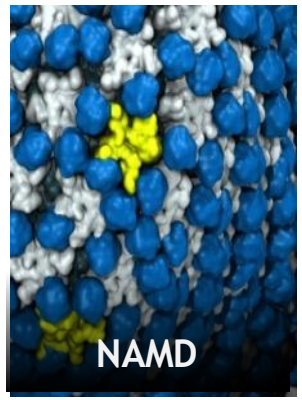


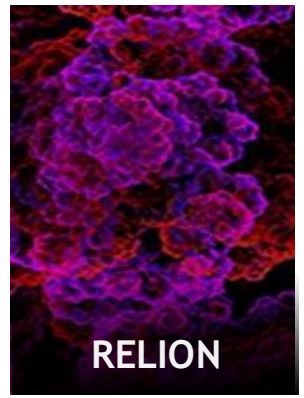












DKRZ DEUTSCHES KLIMARECHENZENTRUM

JOHNS HOPKINS UNIVERSITY

KAUST

京都大学 KYOTO UNIVERSITY

MONASH University

Pfizer

PennState

Schlumberger

Stanford University

UNIVERSITY OF CAMBRIDGE

東京大学 THE UNIVERSITY OF TOKYO

YONSEI UNIVERSITY

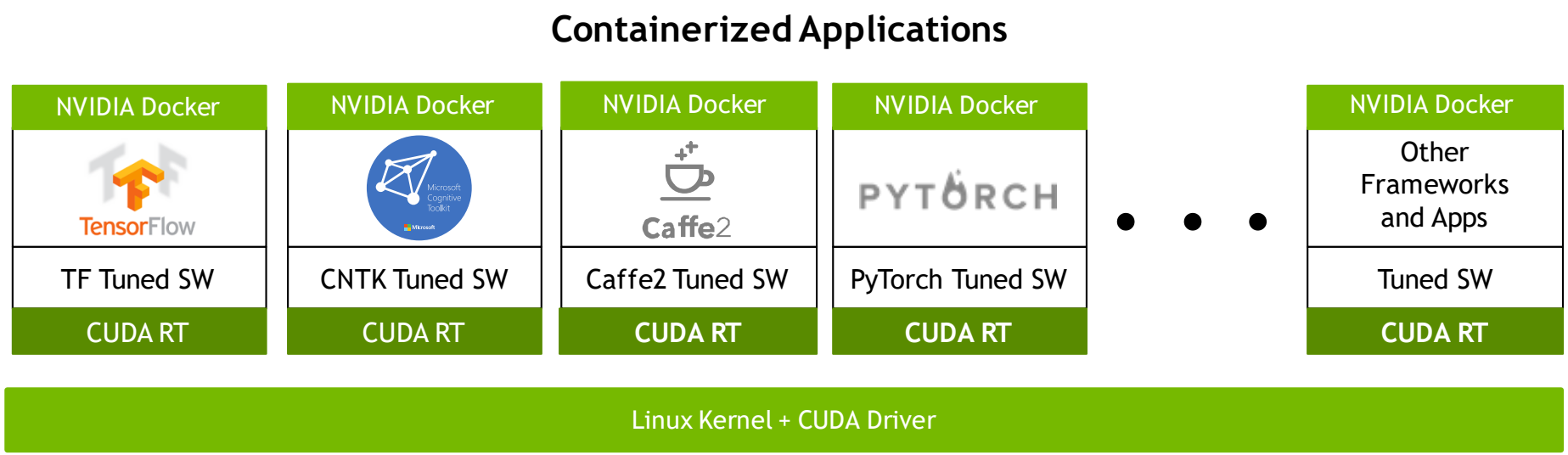
RAPID CONTAINER ADDITION

RAPID USER ADOPTION

*Coming soon

ALWAYS UP-TO-DATE

Monthly Updates from NVIDIA to Frameworks and Containers



NVIDIA DEEP LEARNING INSTITUTE

Online self-paced labs and instructor-led workshops on deep learning and accelerated computing

Take self-paced labs at **www.nvidia.co.uk/dlilabs**

View upcoming workshops and request a workshop onsite at **www.nvidia.co.uk/dli**

Educators can join the University Ambassador Program to teach DLI courses on campus and access resources. Learn more at **www.nvidia.com/dli**

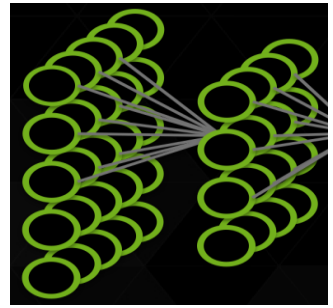
Caffe2

Microsoft Cognitive Toolkit

mxnet

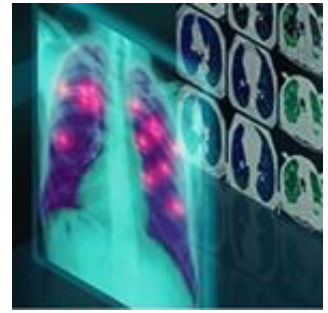
TensorFlow

PYTORCH



Fundamentals

Autonomous Vehicles

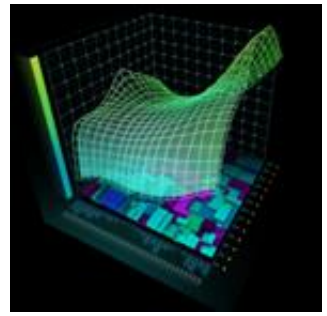


Healthcare

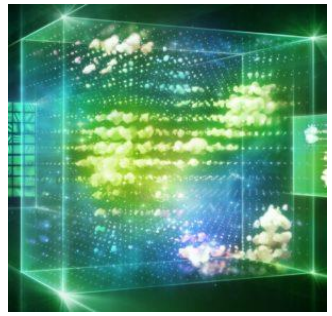
Intelligent Video Analytics

Robotics

Game Development & Digital Content



Finance



Accelerated Computing

Virtual Reality

NVIDIA DEEP LEARNING INSTITUTE

DEEP LEARNING INSTITUTE

University Ambassador Program

Preparing today's students and researchers for tomorrow's AI computing challenges

Want to bring DLI to your campus?

DLI recognizes qualified academics as applied deep learning experts, enabling them to bring free DLI exclusively to university students and staff

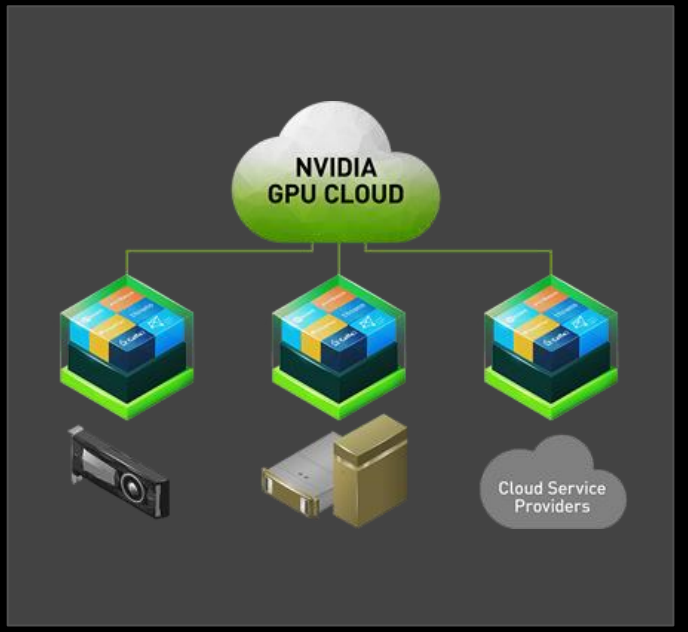
DLI University Ambassador is an additional status on top of DLI instructor certification with additional benefits

Interested candidates can apply via www.nvidia.com/dli

NEXT STEPS

GTC Europe | October 10-12 2018
www.nvidia.com/en-us/gtc

NVIDIA Deep Learning Institute
www.nvidia.co.uk/dli



NGC
www.nvidia.com/en-us/gpu-cloud


NVIDIA®