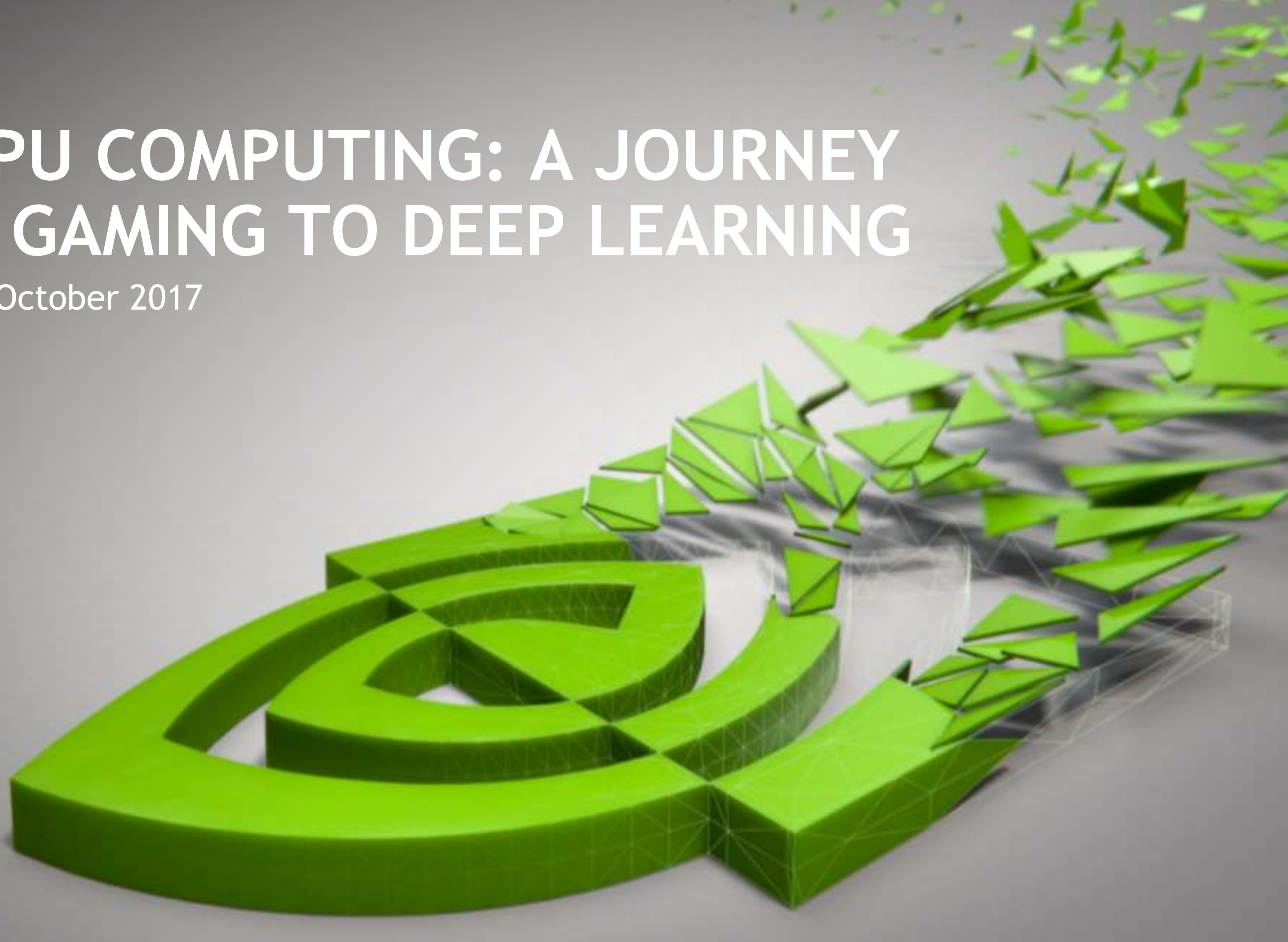
NVIDIA GPU COMPUTING: A JOURNEY FROM PC GAMING TO DEEP LEARNING

Stuart Oberman | October 2017



GAMING

PRO VISUALIZATION

DATA CENTER

AUTO

NVIDIA ACCELERATED COMPUTING

GEFORCE: PC Gaming

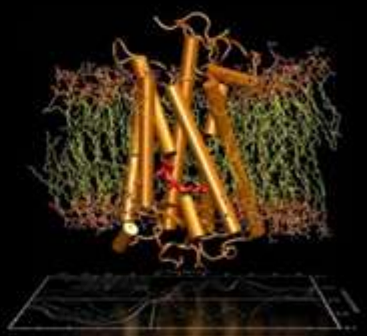
200M GeForce gamers worldwide

Most advanced technology

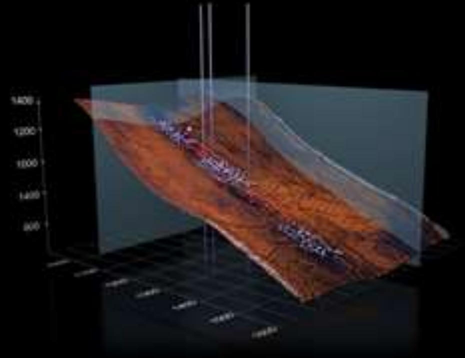
Gaming ecosystem: More than just chips

Amazing experiences & imagery

GPU COMPUTING

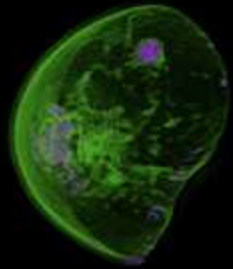


Drug Design
Molecular Dynamics
15x speed up



Seismic Imaging
Reverse Time Migration
14x speed up

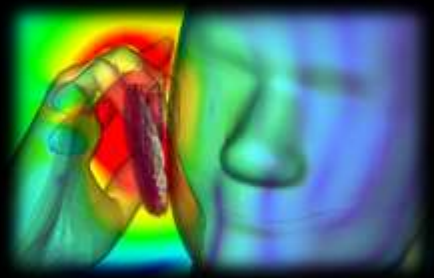
Automotive Design
Computational Fluid Dynamics



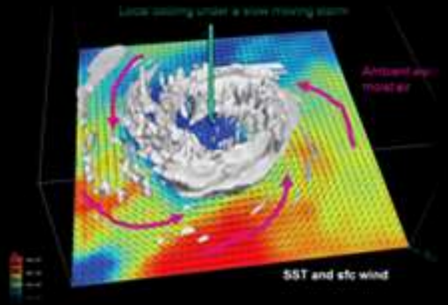
Medical Imaging
Computed Tomography
30-100x speed up

Astrophysics
n-body

Options Pricing
Monte Carlo
20x speed up



Product Development
Finite Difference Time Domain



Weather Forecasting
Atmospheric Physics

NVIDIA GPUS: 1999 TO NOW

https://youtu.be/I25dLTIPREA

SOUL OF THE GRAPHICS PROCESSING UNIT

GPU: Changes Everything

- **Accelerate computationally-intensive applications**
- NVIDIA introduced GPU in 1999
  - A single chip processor to accelerate PC gaming and 3D graphics
- Goal: approach the image quality of movie studio offline rendering farms, but in real-time
  - Instead of hours per frame, > 60 frames per second
- Millions of pixels per frame can all be operated on in parallel
  - 3D graphics is often termed *embarrassingly parallel*
- Use large arrays of floating point units to exploit wide and deep parallelism

CLASSIC GEFORCE GPUS

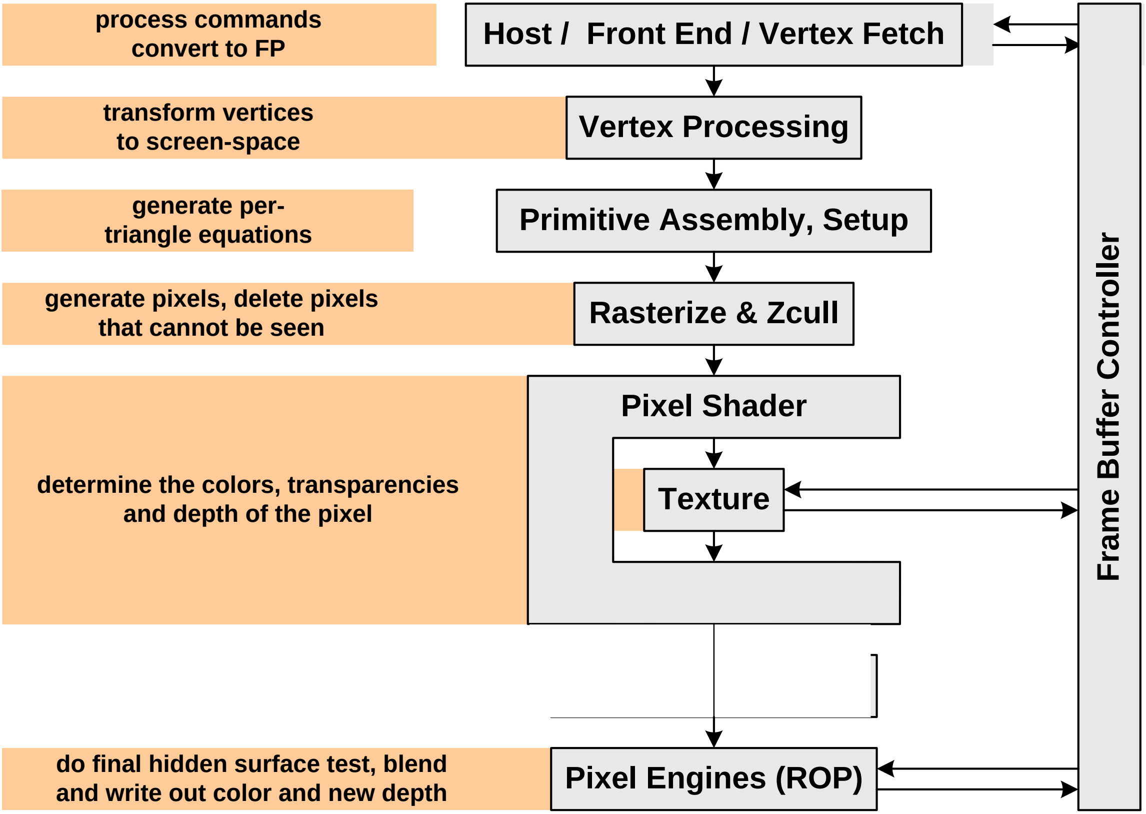
GEFORCE 6 AND 7 SERIES

2004-2006

- Example: GeForce 7900 GTX
- 278M transistors
- 650MHz pipeline clock
- 196mm^2 in 90nm
- >300 GFLOPS peak, single-precision

THE LIFE OF A TRIANGLE IN A GPU

Classic Edition



process commands convert to FP

transform vertices to screen-space

generate per-triangle equations

generate pixels, delete pixels that cannot be seen

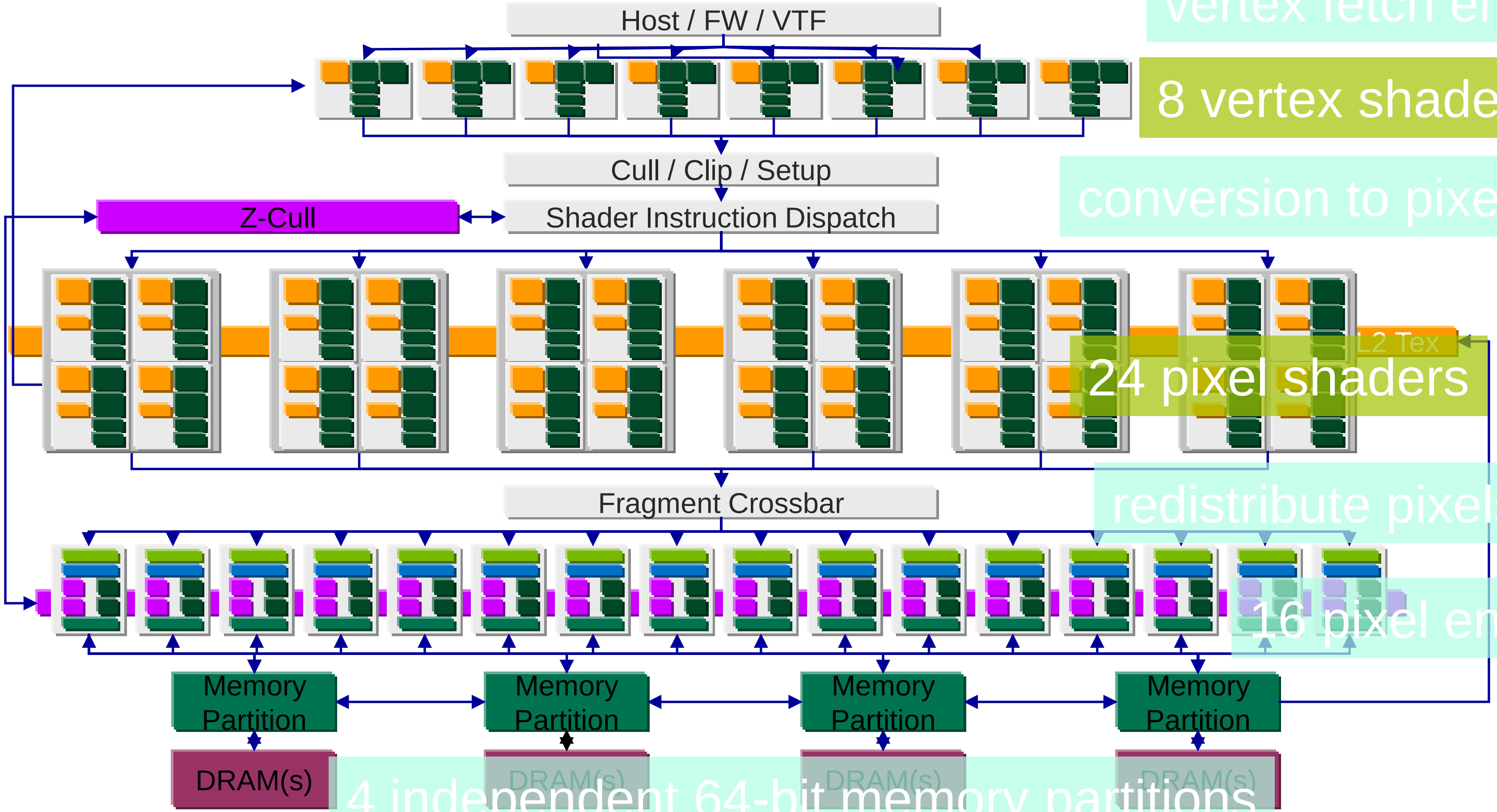
determine the colors, transparencies and depth of the pixel

do final hidden surface test, blend and write out color and new depth

NUMERIC REPRESENTATIONS IN A GPU

- Fixed point formats
  - u8, s8, u16, s16, s3.8, s5.10, ...
- Floating point formats
  - fp16, fp24, fp32, ...
  - Tradeoff of dynamic range vs. precision
- Block floating point formats
  - Treat multiple operands as having a common exponent
  - Allows a tradeoff in dynamic range vs storage and computation

INSIDE THE 7900GTX GPU



vertex fetch engine

8 vertex shaders

conversion to pixels

24 pixel shaders

redistribute pixels

16 pixel engines

4 independent 64-bit memory partitions

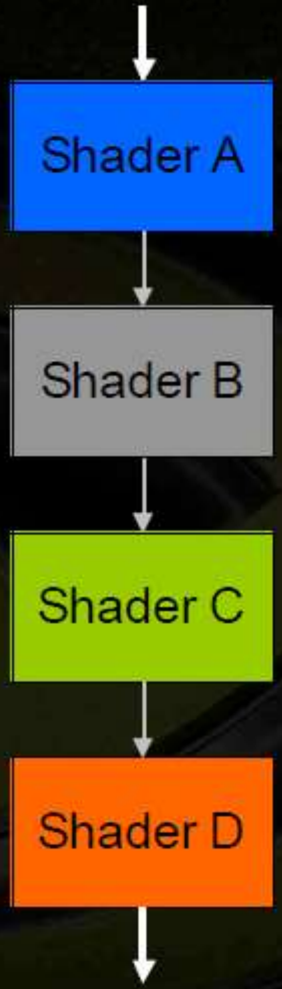
G80: REDEFINED THE GPU

G80

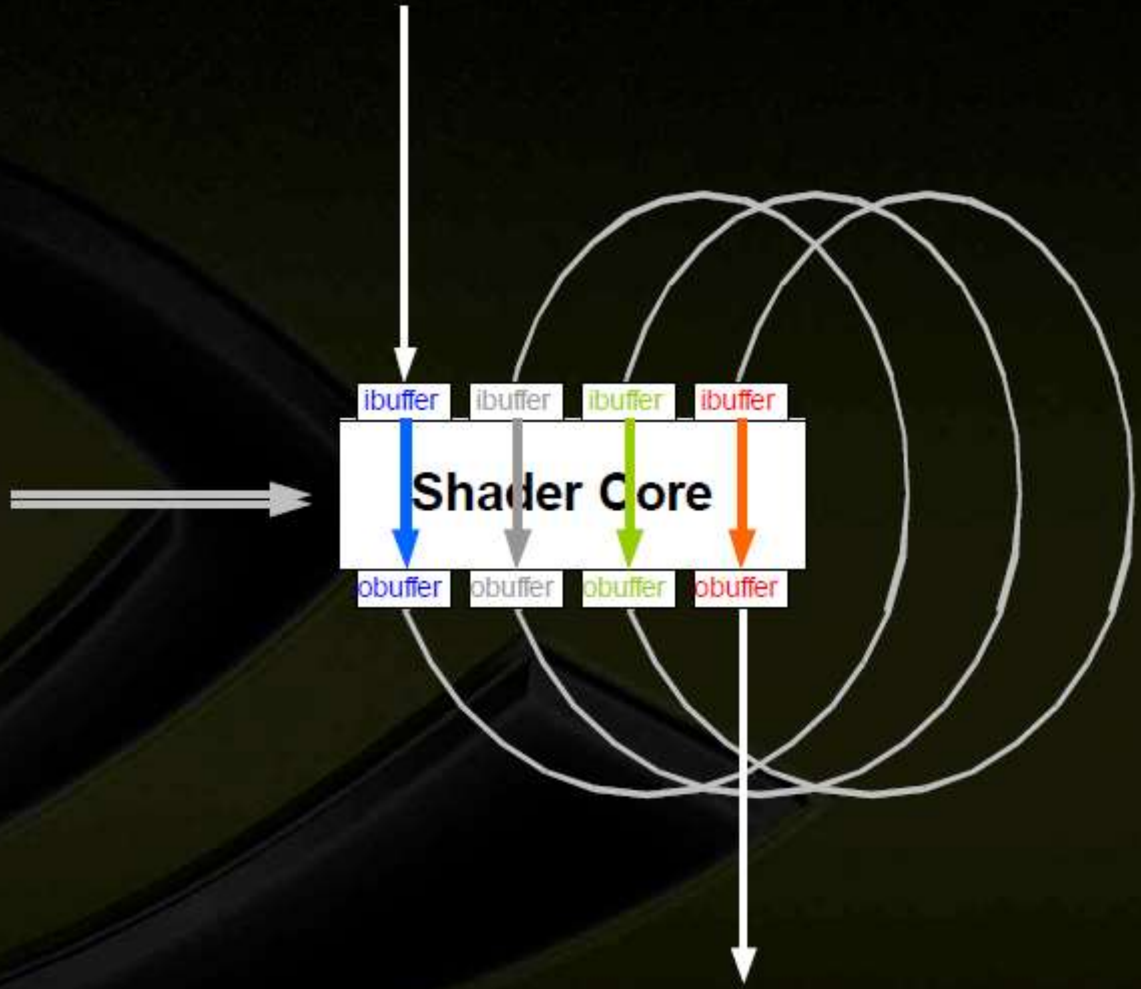
GeForce 8800 released 2006

- G80 first GPU with a unified shader processor architecture
  - Introduced the SM: Streaming Multiprocessor
    - Array of simple streaming processor cores: SPs or CUDA cores
  - All shader stages use the same instruction set
  - All shader stages execute on the same units
- Permits better sharing of SM hardware resources
- Recognized that building dedicated units often results in under-utilization due to the application workload

Dedicated Design



Unified Design



G80 FEATURES

- 681M transistors
- 470mm2 in 90nm
- First to support Microsoft DirectX10 API
- Invested a little extra (epsilon) HW in SM to also support general purpose throughput computing
  - Beginning of CUDA everywhere
- SM functional units designed to run at 2x frequency, half the number of units
  - 576 GFLOPs @ 1.5GHz , IEEE 754 fp32 FADD and FMUL
- 155W

BEGINNING OF GPU COMPUTING

Throughput Computing

- Latency Oriented
  - Fewer, bigger cores with out-of-order, speculative execution
  - Big caches optimized for latency
  - Math units are small part of the die
- Throughput Oriented
  - Lots of simple compute cores and hardware scheduling
  - Big register files. Caches optimized for bandwidth.
  - Math units are most of the die

CUDA

Most successful environment for throughput computing

C++ for throughput computers

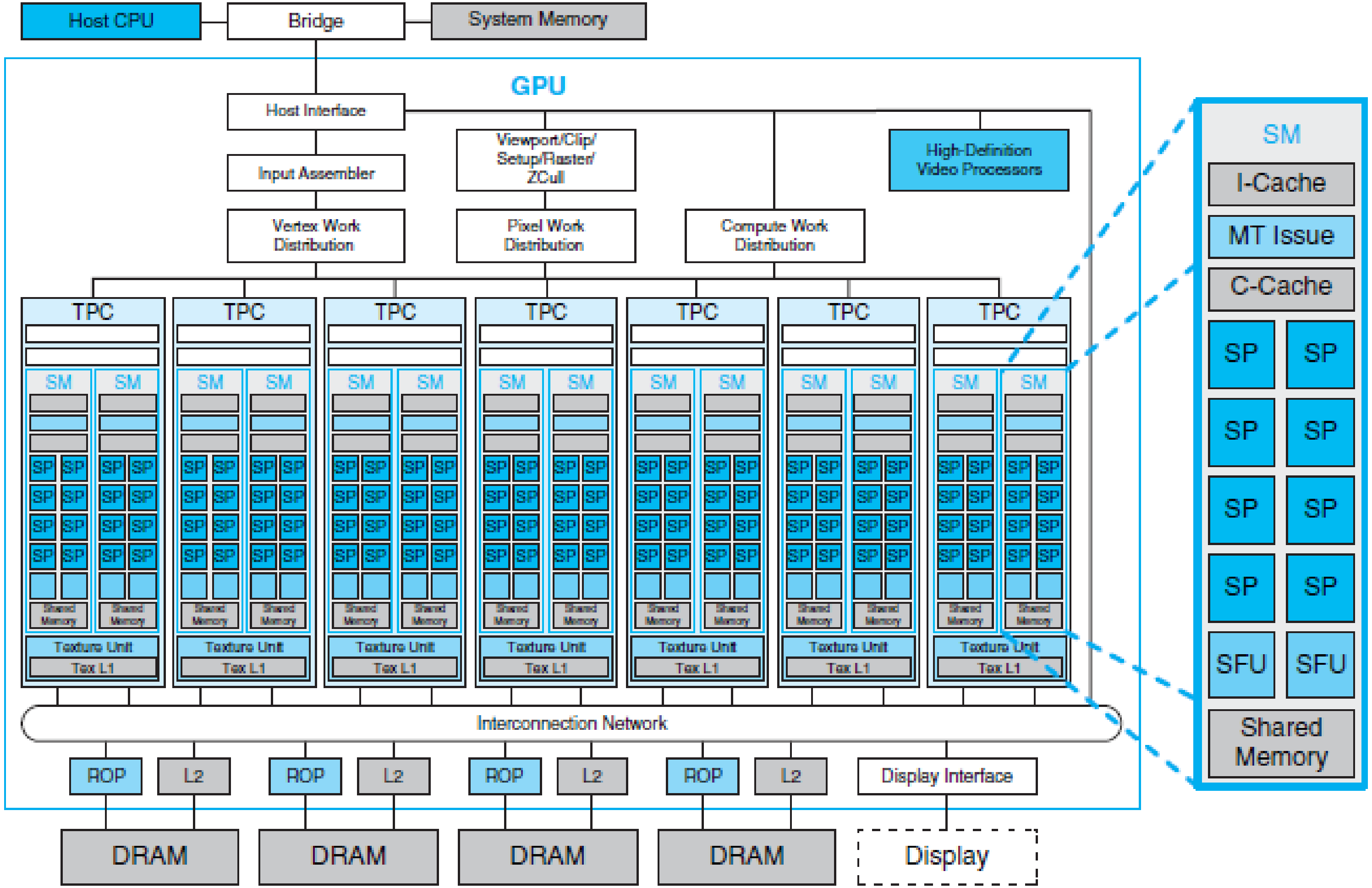
On-chip memory management

Asynchronous, parallel API

Programmability makes it possible to innovate

New layer type? No problem.

G80 ARCHITECTURE

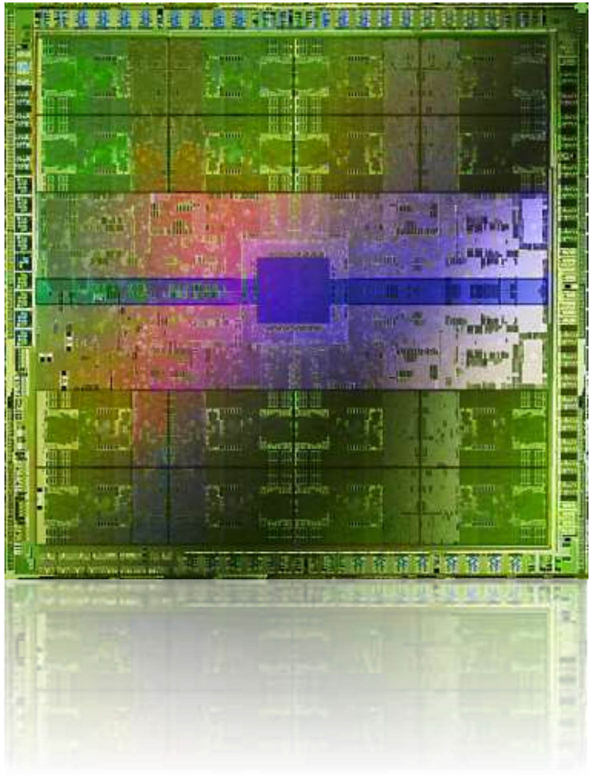


FROM FERMI TO PASCAL

FERMI GF100

Tesla C2070 released 2011

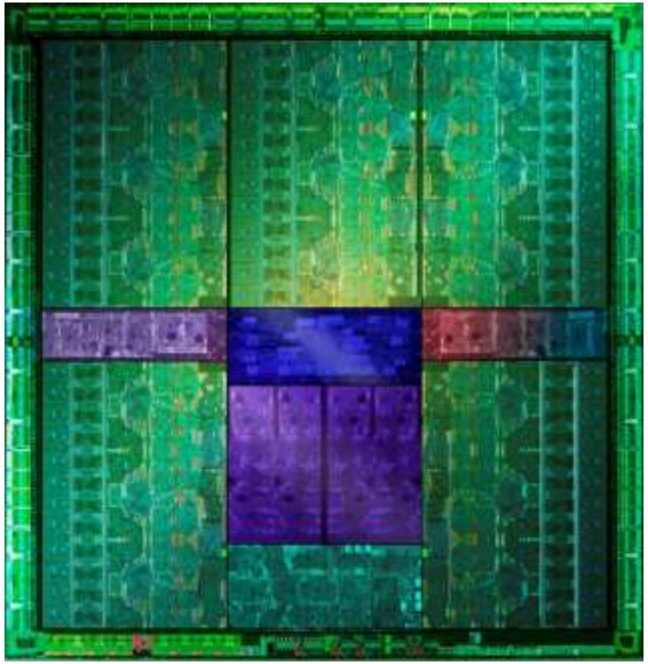
- 3B transistors
- 529 mm2 in 40nm
- 1150 MHz SM clock
- 3rd generation SM, each with configurable L1/shared memory
- IEEE 754-2008 FMA
- 1030 GFLOPS fp32, 515 GFLOPS fp64
- 247W



KEPLER GK110

Tesla K40 released 2013

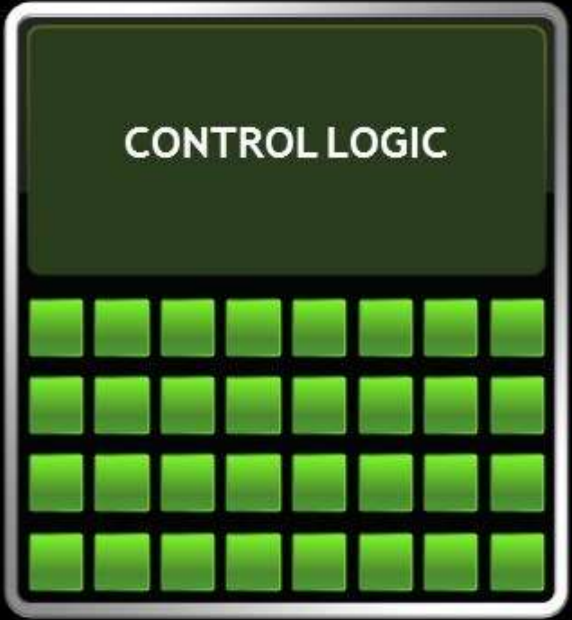
- 7.1B transistors
- 550 mm2 in 28nm
- Intense focus on power efficiency, operating at lower frequency
  - 2880 CUDA cores at 810 MHz
- Tradeoff of area efficiency vs. power efficiency
- 4.3 TFLOPS fp32, 1.4 TFLOPS fp64
- 235W



Kepler: Fast & Efficient

SM

Fermi



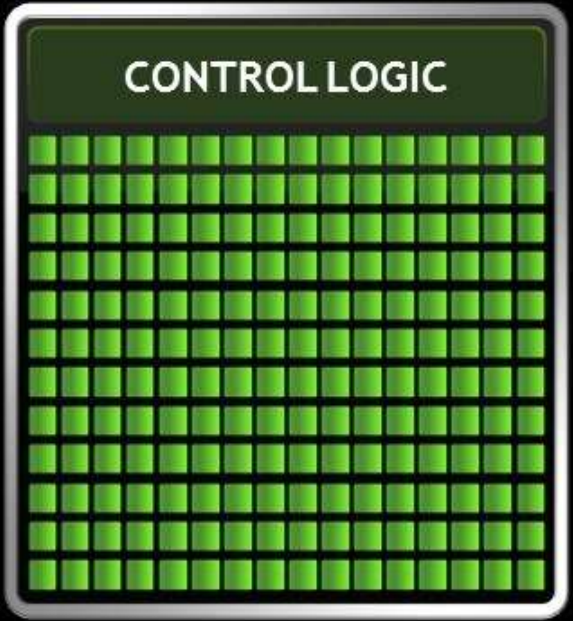
32 cores

3x

Perf / Watt

SMX

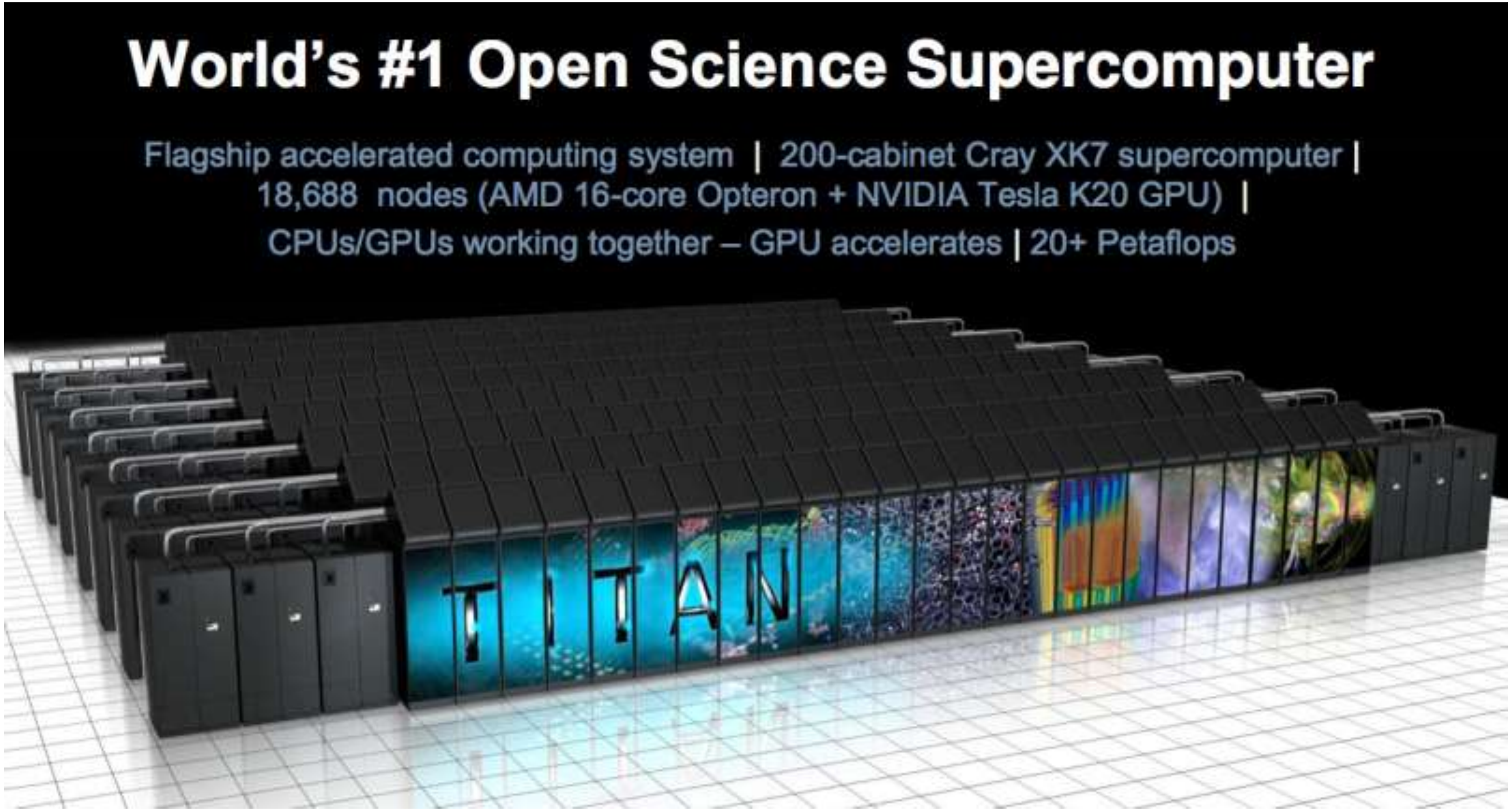
Kepler



192 cores

TITAN SUPERCOMPUTER

Oak Ridge National Laboratory



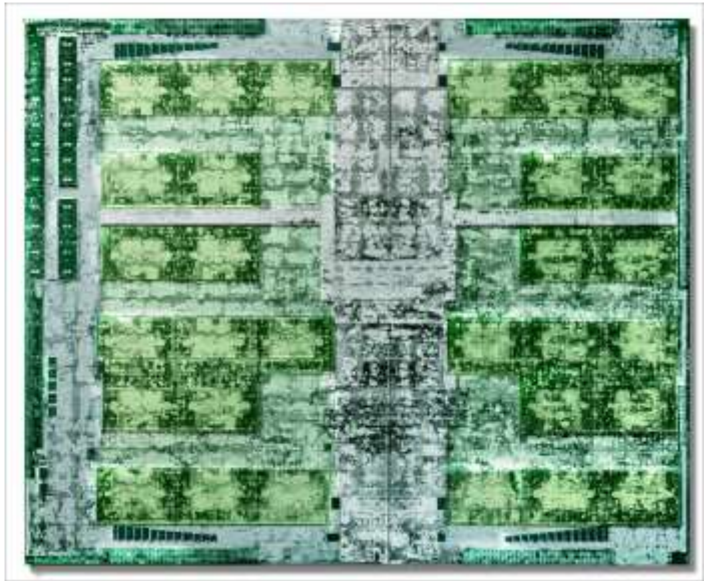
World's #1 Open Science Supercomputer

Flagship accelerated computing system | 200-cabinet Cray XK7 supercomputer | 18,688 nodes (AMD 16-core Opteron + NVIDIA Tesla K20 GPU) | CPUs/GPUs working together – GPU accelerates | 20+ Petaflops

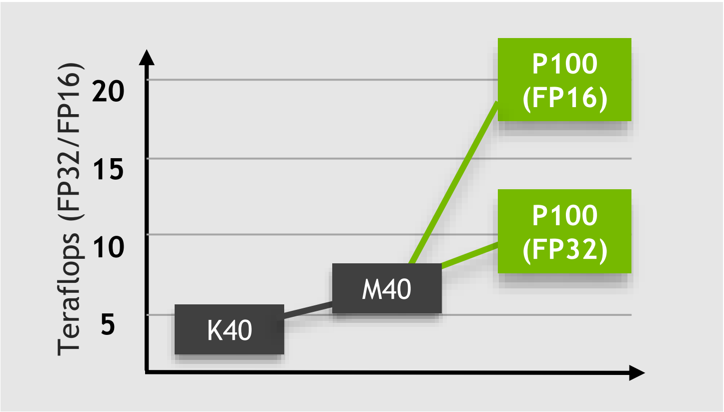
PASCAL GP100

released 2016

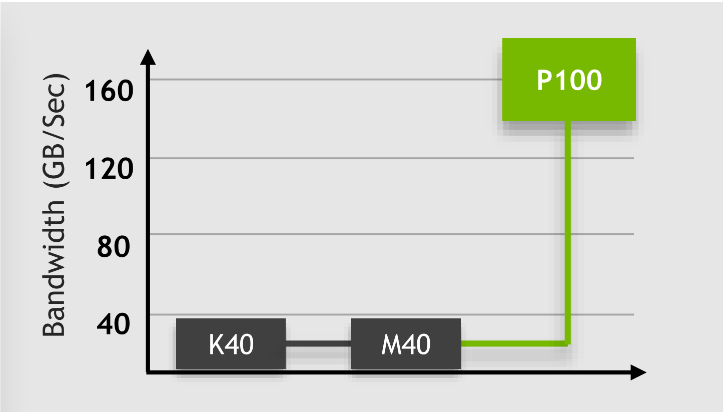
- 15.3B transistors
- 610 mm2 in 16ff
- 10.6 TFLOPS fp32, 5.3 TFLOPS fp64
- 21 TFLOPS fp16 for Deep Learning training and inference acceleration
- New high-bandwidth NVLink GPU interconnect
- HBM2 stacked memory
- 300W



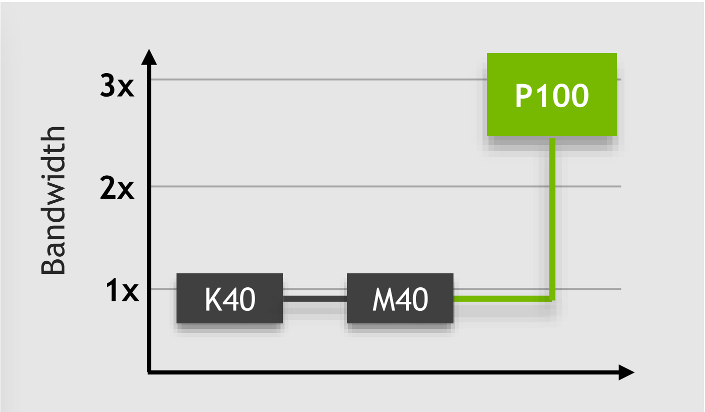
MAJOR ADVANCES IN PASCAL



3x Compute

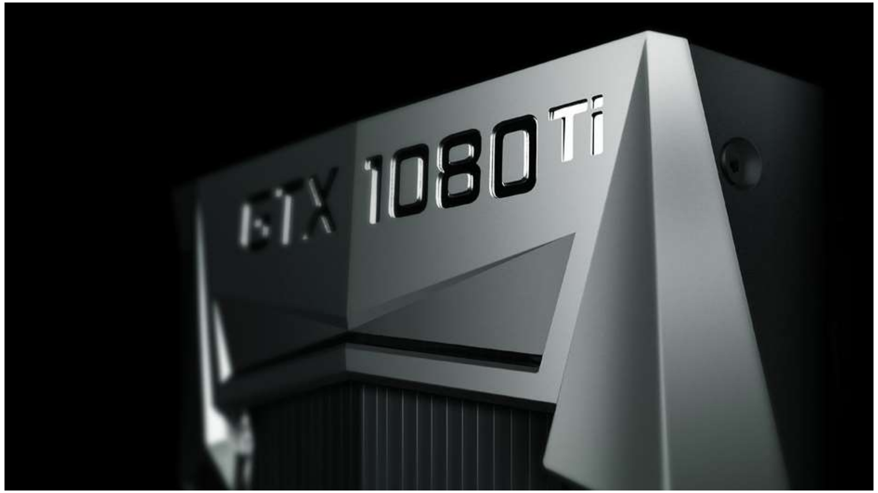


5x GPU-GPU BW



3x GPU Mem BW

GEFORCE GTX 1080TI



https://www.nvidia.com/en-us/geforce/products/10series/geforce-gtx-1080-ti/

https://youtu.be/2c2vN736V60

FINAL FANTASY XV PREVIEW DEMO WITH GEFORCE GTX 1080TI

https://www.geforce.com/whats-new/articles/final-fantasy-xv-windows-edition-4k-trailer-nvidia-gameworks-enhancements

https://youtu.be/h0o3fctwXw0

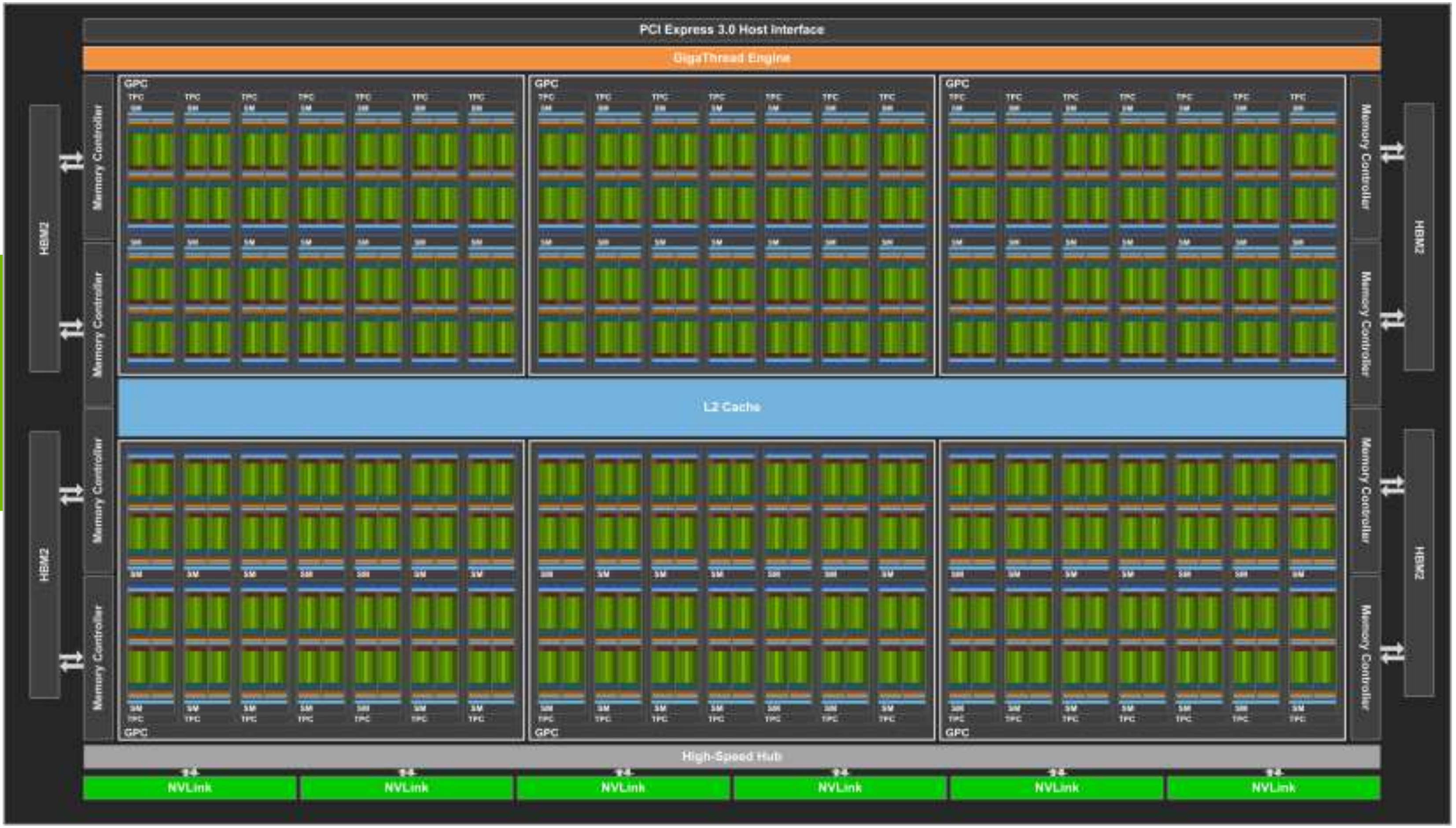
2017: VOLTA

TESLA V100: 2017

21B transistors
815 mm^2 in 16ff

80 SM
5120 CUDA Cores
640 Tensor Cores

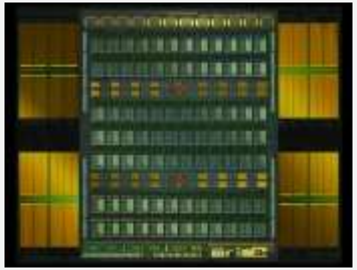
16 GB HBM2
900 GB/s HBM2
300 GB/s NVLink



*full GV100 chip contains 84 SMs

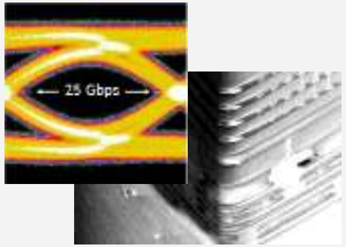
TESLA V100

Volta Architecture



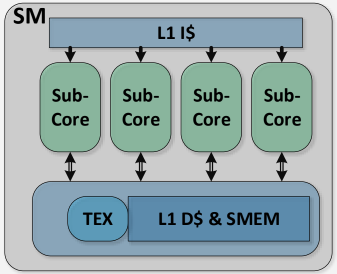
Most Productive GPU

Improved NVLink & HBM2



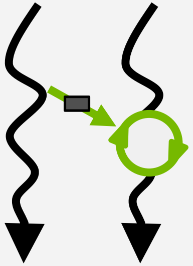
Efficient Bandwidth

New SM Core



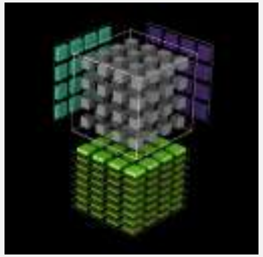
Performance & Programmability

Independent Thread Scheduling



New Algorithms

Tensor Core



120 Programmable TFLOPS Deep Learning

More V100 Features: 2x L2 atomics, int8, new memory model, copy engine page migration, MPS acceleration, and more …

The Fastest and Most Productive GPU for Deep Learning and HPC

GPU PERFORMANCE COMPARISON

	P100	V100	Ratio
DL Training	10 TFLOPS	120 TFLOPS	12x
DL Inferencing	21 TFLOPS	120 TFLOPS	6x
FP64/FP32	5/10 TFLOPS	7.5/15 TFLOPS	1.5x
HBM2 Bandwidth	720 GB/s	900 GB/s	1.2x
STREAM Triad Perf	557 GB/s	855 GB/s	1.5x
NVLink Bandwidth	160 GB/s	300 GB/s	1.9x
L2 Cache	4 MB	6 MB	1.5x
L1 Caches	1.3 MB	10 MB	7.7x

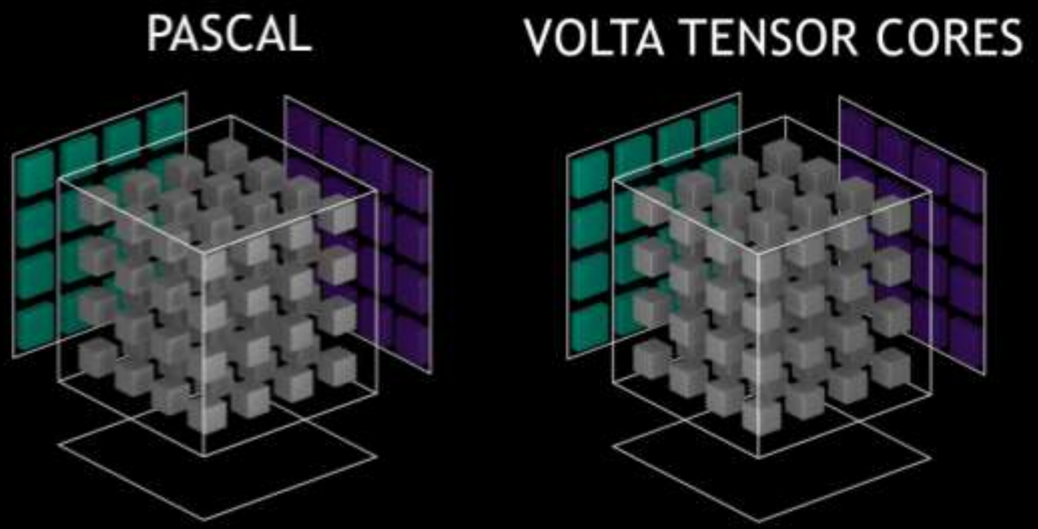
TENSOR CORE

CUDA TensorOp instructions & data formats

4x4 matrix processing array

D[FP32] = A[FP16] * B[FP16] + C[FP32]

Optimized for deep learning



🟩 Activation Inputs 🟪 Weights Inputs 🟩 Output Results

TENSOR CORE

Mixed Precision Matrix Math
4x4 matrices

$$\underset{\text{FP16 or FP32}}{D} = \underset{\text{FP16}}{\begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix}} \underset{\text{FP16}}{\begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix}} + \underset{\text{FP16 or FP32}}{\begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}}$$

$$D = AB + C$$

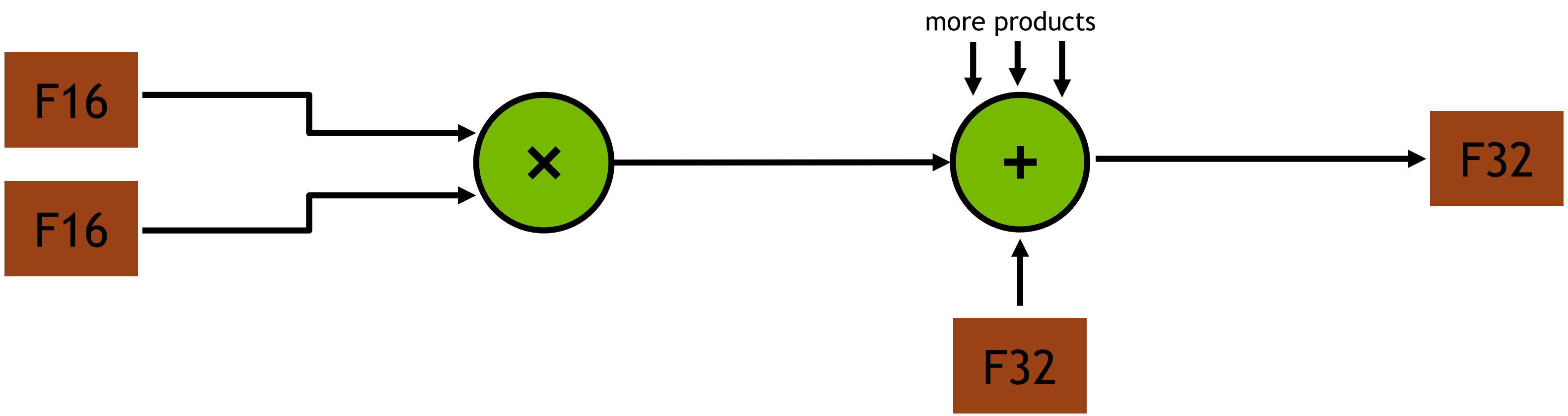
VOLTA TENSOR OPERATION

FP16 storage/input

Full precision product

Sum with FP32 accumulator

Convert to FP32 result



more products

F16

F16

×

+

F32

F32

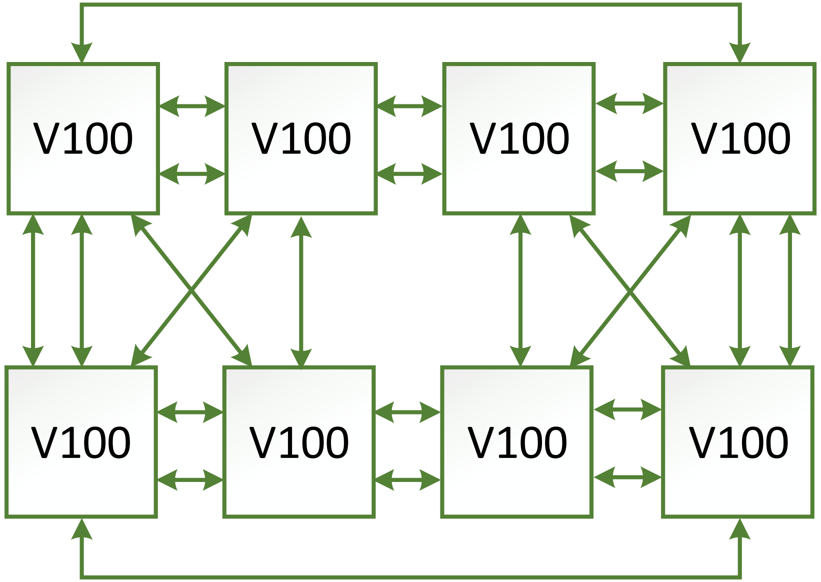
Also supports FP16 accumulator mode for inferencing

NVLINK - PERFORMANCE AND POWER

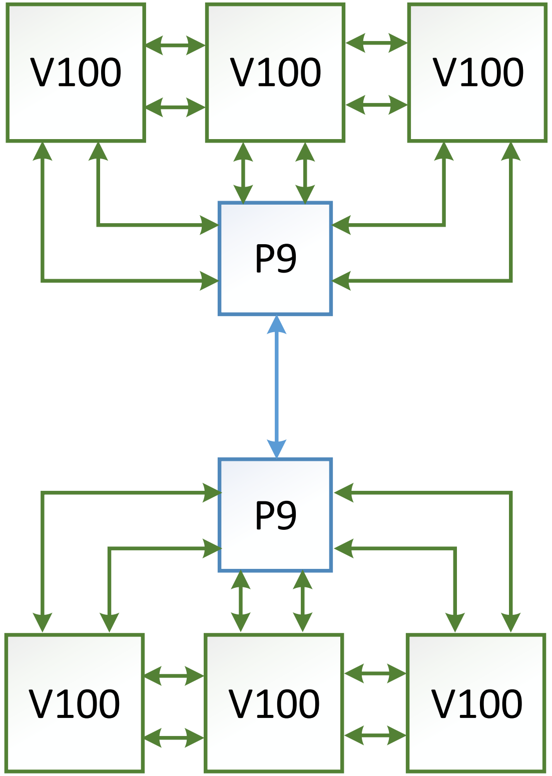
Bandwidth	25Gbps signaling 6 NVLinks for GV100 1.9 x Bandwidth improvement over GP100
Coherence	Latency sensitive CPU caches GMEM Fast access in local cache hierarchy Probe filter in GPU
Power Savings	Reduce number of active lanes for lightly loaded link

NVLINK NODES

DL – HYBRID CUBE MESH – DGX-1 w/ Volta



HPC – P9 CORAL NODE – SUMMIT



NARROWING THE SHARED MEMORY GAP

with the GV100 L1 cache

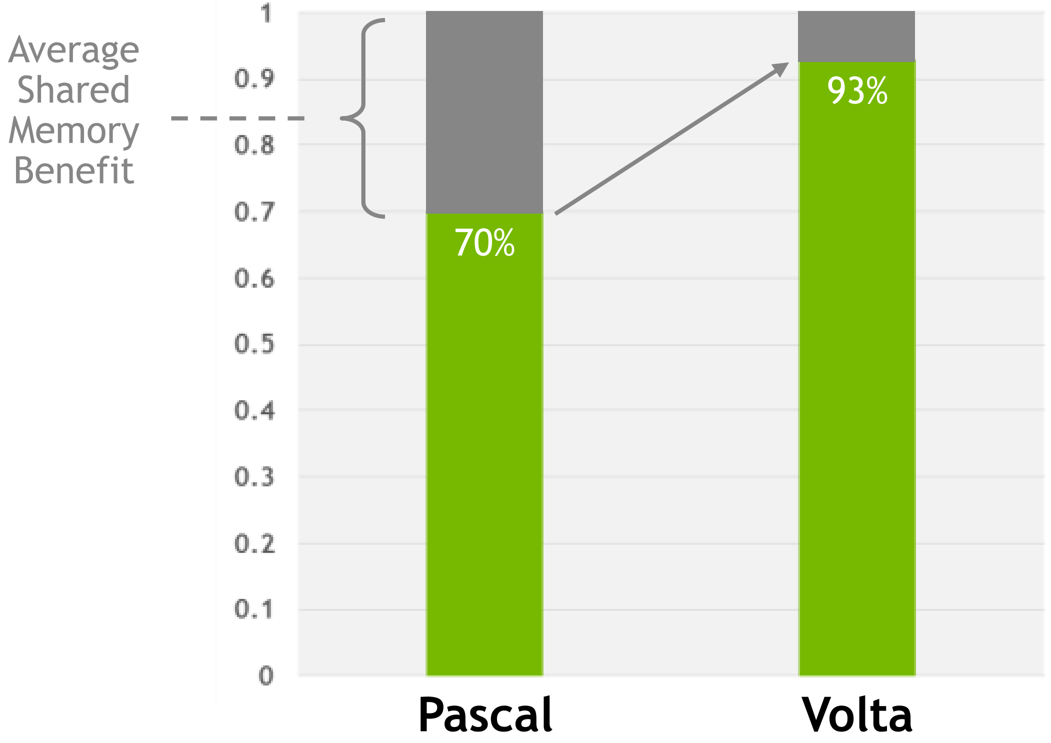
Cache: vs shared

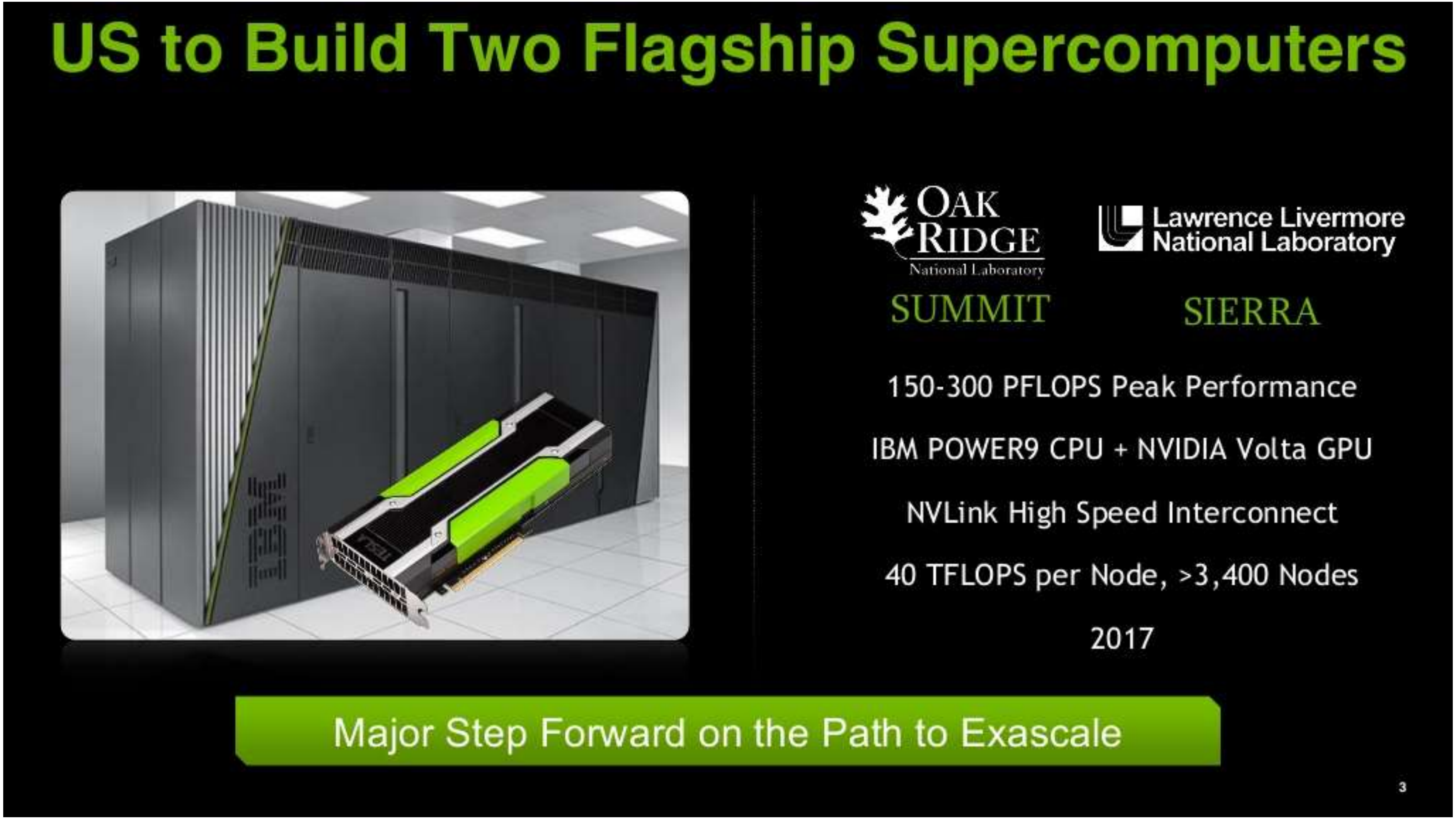
- Easier to use
- 90%+ as good

Shared: vs cache

- Faster atomics
- More banks
- More predictable

Directed testing: shared in global





US to Build Two Flagship Supercomputers

Oak Ridge National Laboratory

SUMMIT

Lawrence Livermore National Laboratory

SIERRA

150-300 PFLOPS Peak Performance

IBM POWER9 CPU + NVIDIA Volta GPU

NVLink High Speed Interconnect

40 TFLOPS per Node, >3,400 Nodes

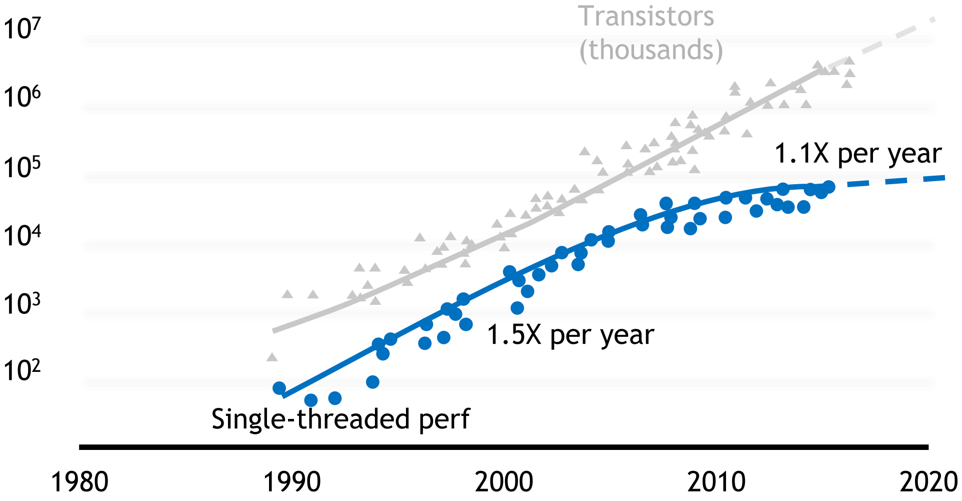
2017

Major Step Forward on the Path to Exascale

3

GPU COMPUTING AND DEEP LEARNING

TWO FORCES DRIVING THE FUTURE OF COMPUTING

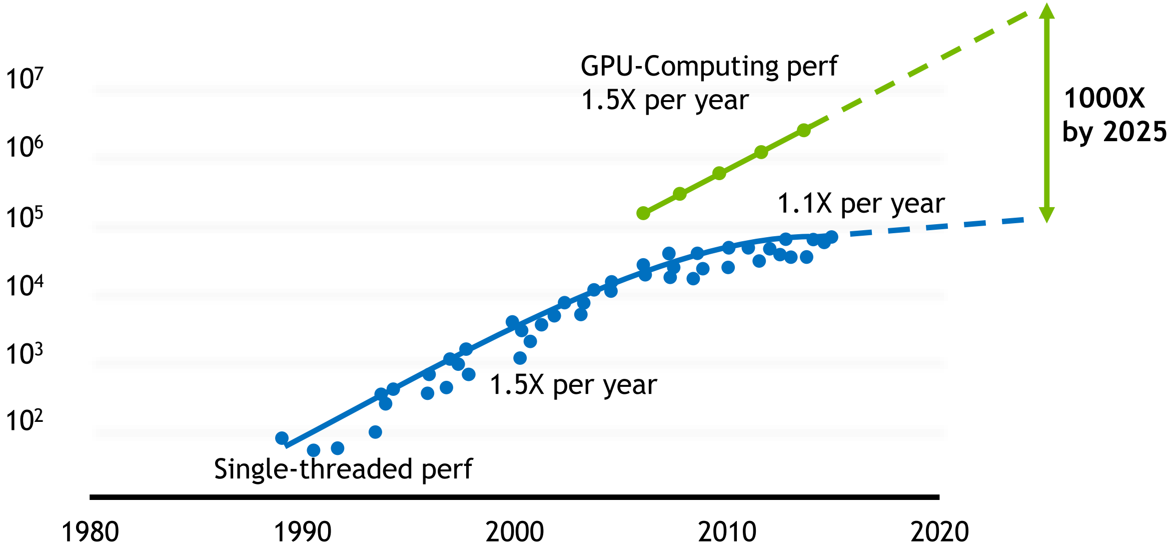


40 Years of Microprocessor Trend Data

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

The Big Bang of Deep Learning

RISE OF NVIDIA GPU COMPUTING



40 Years of Microprocessor Trend Data

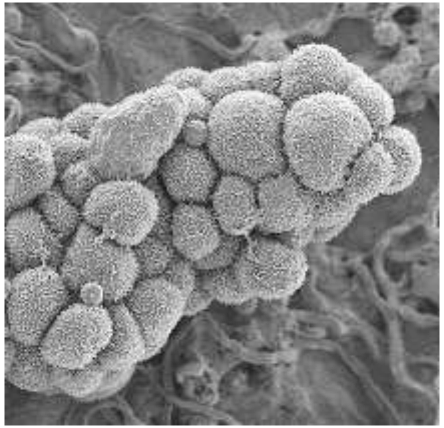
Original data up to the year 2010 collected and plotted by M. Horowitz,
F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

The Big Bang of Deep Learning

DEEP LEARNING EVERYWHERE

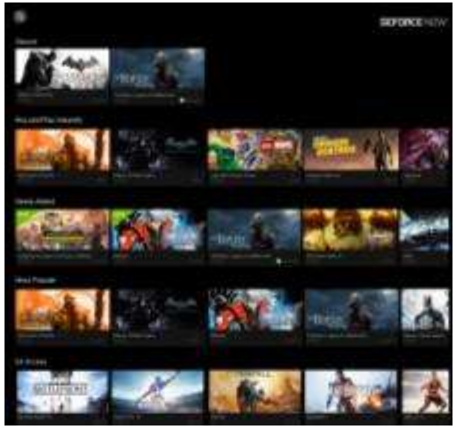
INTERNET & CLOUD

Image Classification
Speech Recognition
Language Translation
Language Processing
Sentiment Analysis
Recommendation



MEDICINE & BIOLOGY

Cancer Cell Detection
Diabetic Grading
Drug Discovery

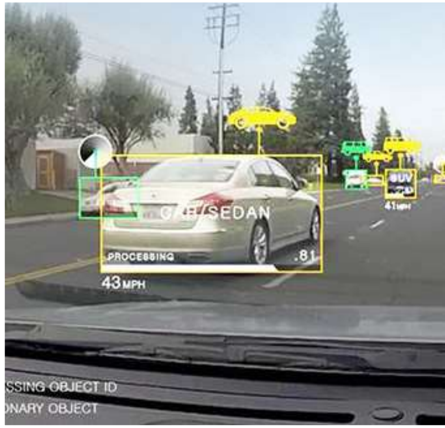


MEDIA & ENTERTAINMENT

Video Captioning
Video Search
Real Time Translation

SECURITY & DEFENSE

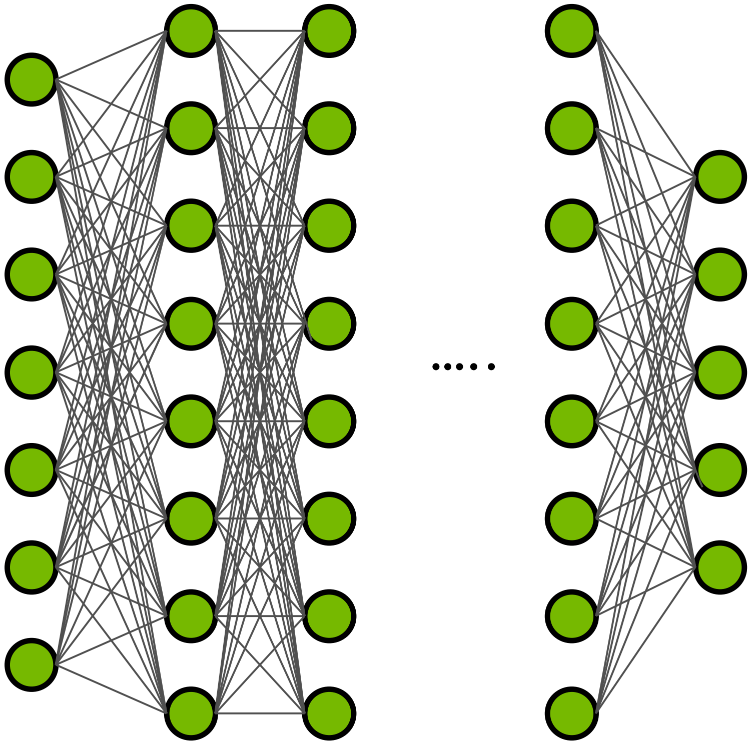
Face Detection
Video Surveillance
Satellite Imagery

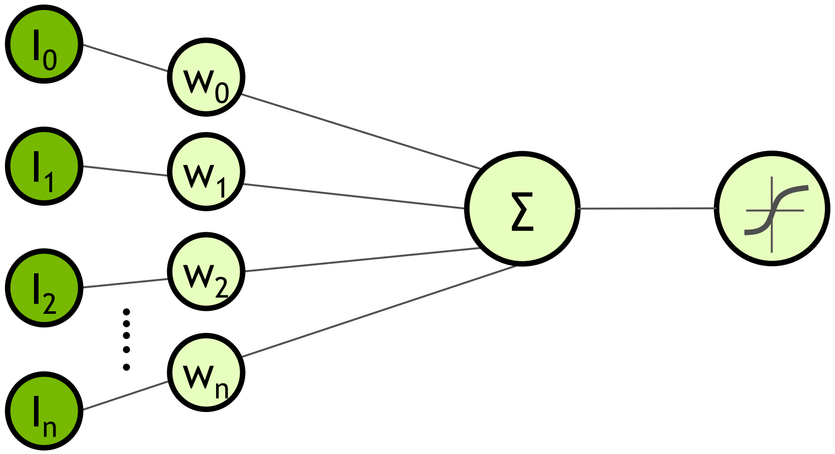


AUTONOMOUS MACHINES

Pedestrian Detection
Lane Tracking
Recognize Traffic Sign

DEEP NEURAL NETWORK



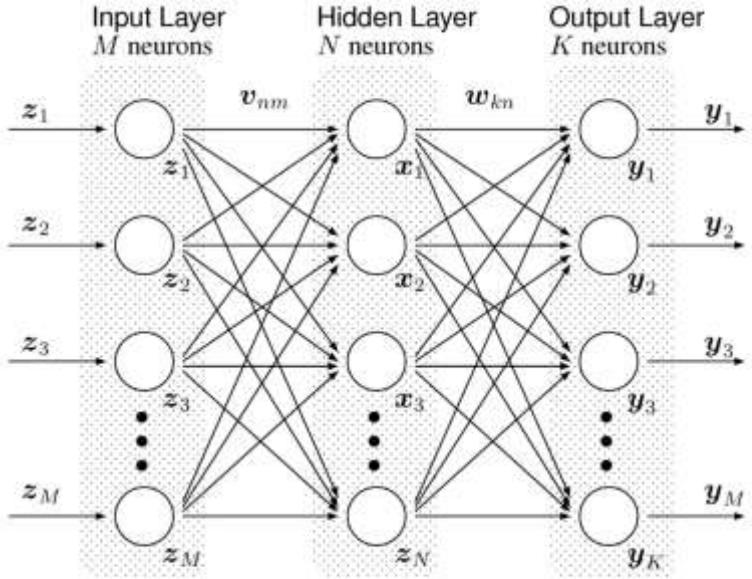


ANATOMY OF A FULLY CONNECTED LAYER

Lots of dot products

Each neuron calculates a dot product, M in a layer

$$x_1 = g(\boldsymbol{v}_{x_1} * \boldsymbol{z})$$



COMBINE THE DOT PRODUCTS

What if we assemble the weights into a matrix?

Each neuron calculates a dot product, M in a layer

$$x_1 = g(\boldsymbol{v}_{x_1} * \boldsymbol{z})$$

What if we assemble the weights as [M, K] matrix?

Matrix-vector multiplication (GEMV)

Unfortunately …

M*K+2*K elements load/store

M*K FMA math operations

This is memory bandwidth limited!

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

BATCH TO GET MATRIX MULTIPLICATION

Making the problem math limited

Can we turn this into a GEMM?

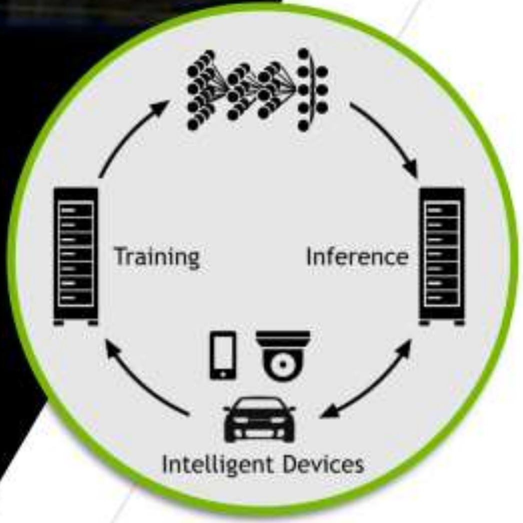
“Batching”: process several inputs at once

Input is now a matrix, not a vector

Weight matrix remains the same

1 <= N <= 128 is common

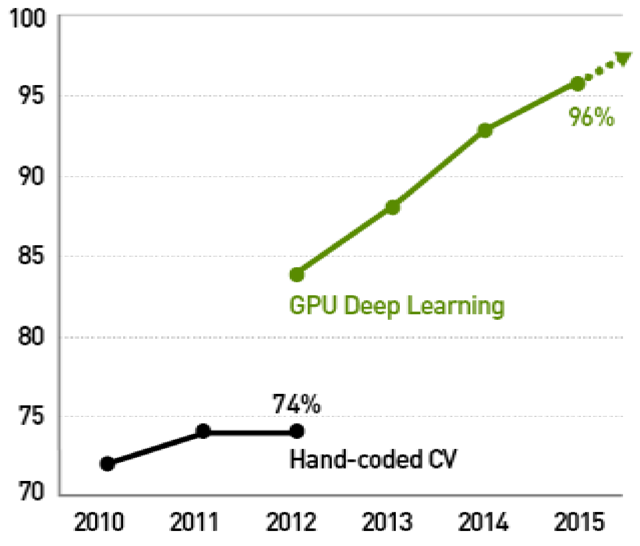
$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$



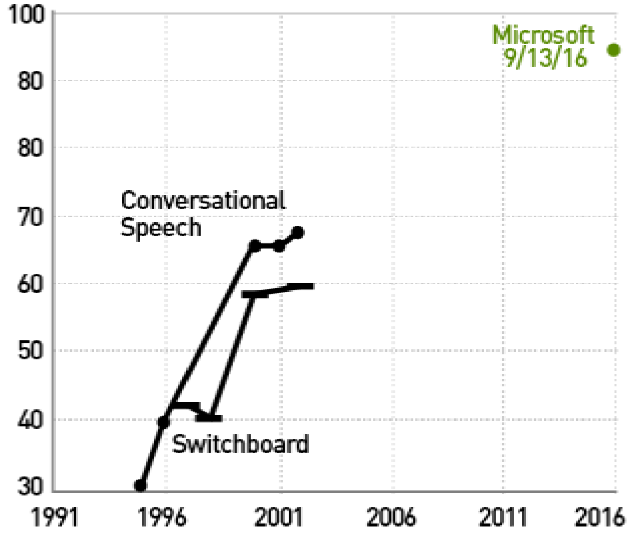
GPU DEEP LEARNING —
A NEW COMPUTING MODEL

AI IMPROVING AT AMAZING RATES

IMAGENET ACCURACY

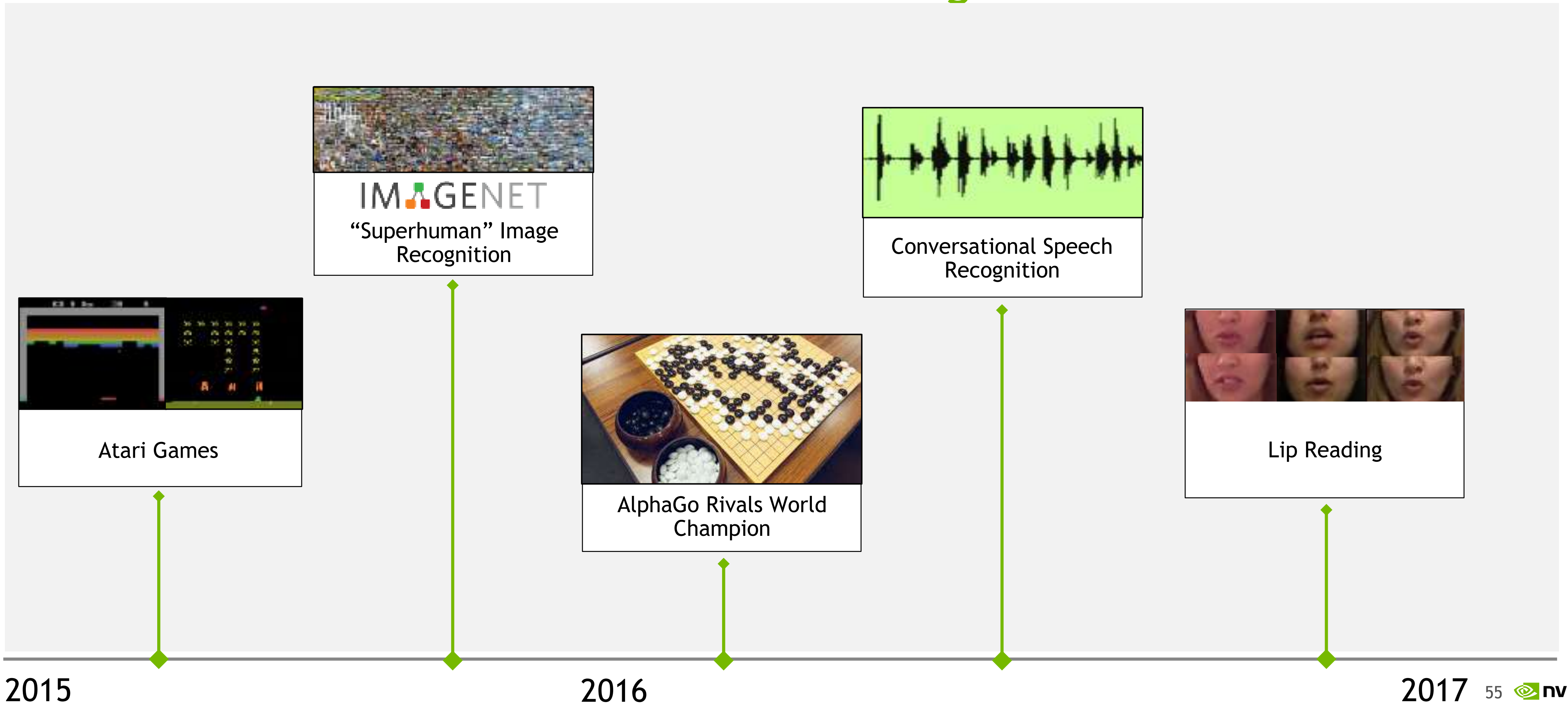


SPEECH RECOGNITION ACCURACY



AI BREAKTHROUGHS

Recent Breakthroughs



Atari Games

IMAGENET

“Superhuman” Image Recognition

AlphaGo Rivals World Champion

Conversational Speech Recognition

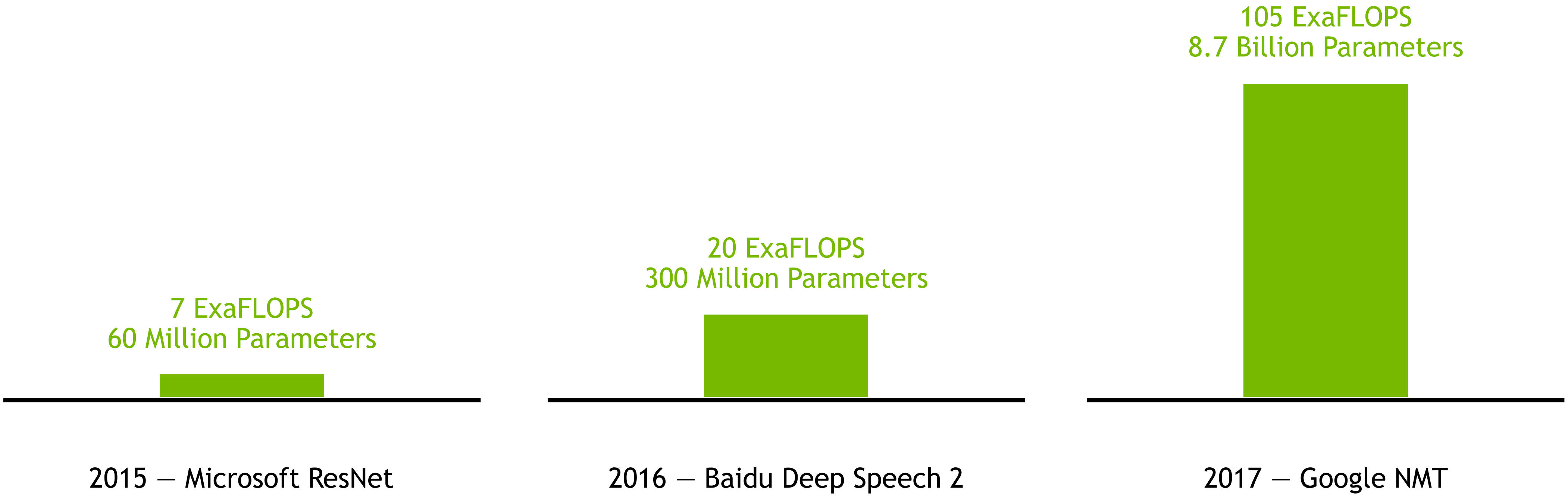
Lip Reading

2015

2016

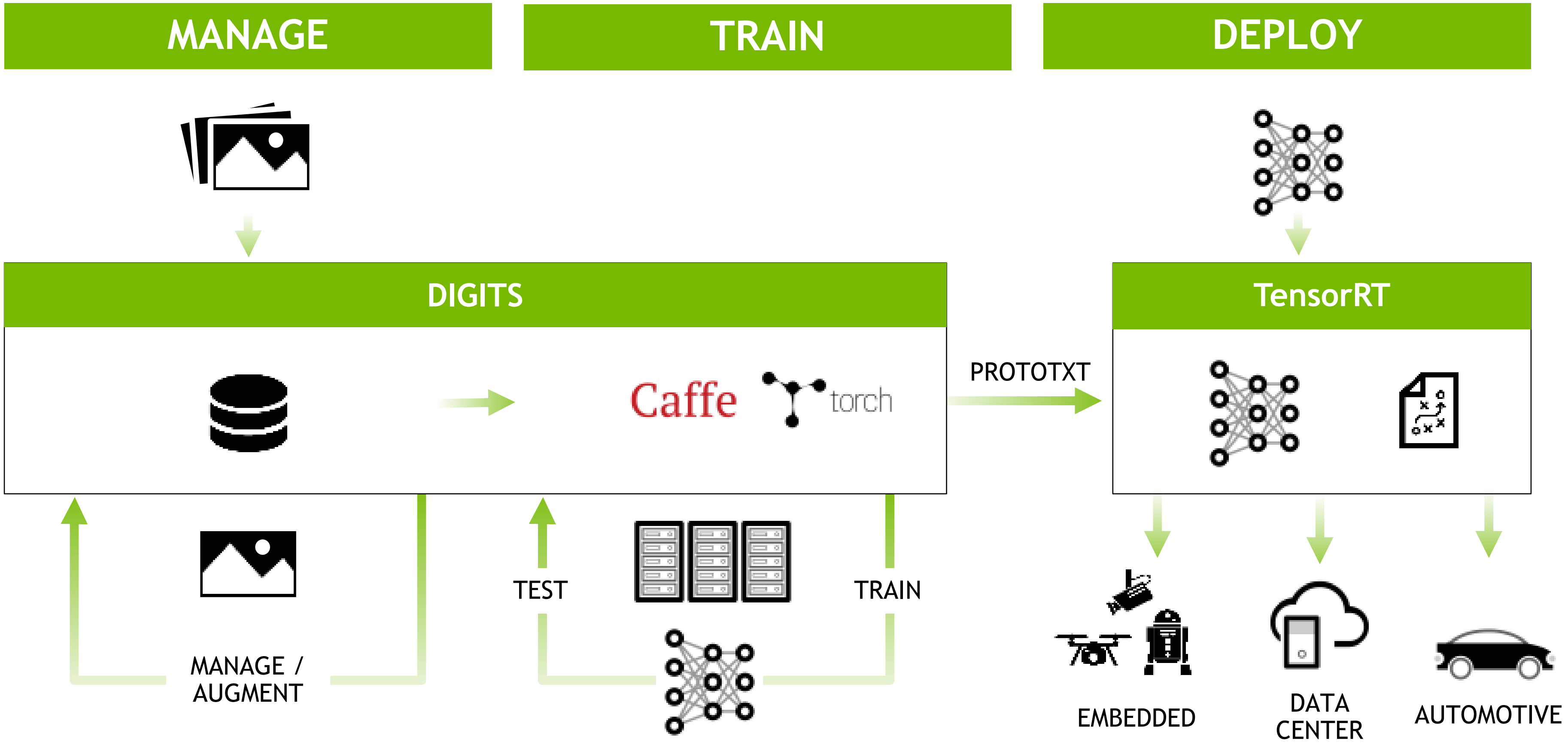
2017

MODEL COMPLEXITY IS EXPLODING



NVIDIA DNN ACCELERATION

A COMPLETE DEEP LEARNING PLATFORM



DNN TRAINING

NVIDIA DGX STATION

PERSONAL DGX

480 Tensor TFLOPS | 4x Tesla V100 16GB

NVLink Fully Connected | 3x DisplayPort

1500W | Water Cooled

NVIDIA DGX STATION

PERSONAL DGX

480 Tensor TFLOPS | 4x Tesla V100 16GB

NVLink Fully Connected | 3x DisplayPort

1500W | Water Cooled

$69,000

NVIDIA DGX-1 WITH TESLA V100

ESSENTIAL INSTRUMENT OF AI RESEARCH

960 Tensor TFLOPS | 8x Tesla V100 | NVLink Hybrid Cube

From 8 days on TITAN X to 8 hours

400 servers in a box

NVIDIA DGX-1 WITH TESLA V100

ESSENTIAL INSTRUMENT OF AI RESEARCH

960 Tensor TFLOPS | 8x Tesla V100 | NVLink Hybrid Cube

From 8 days on TITAN X to 8 hours

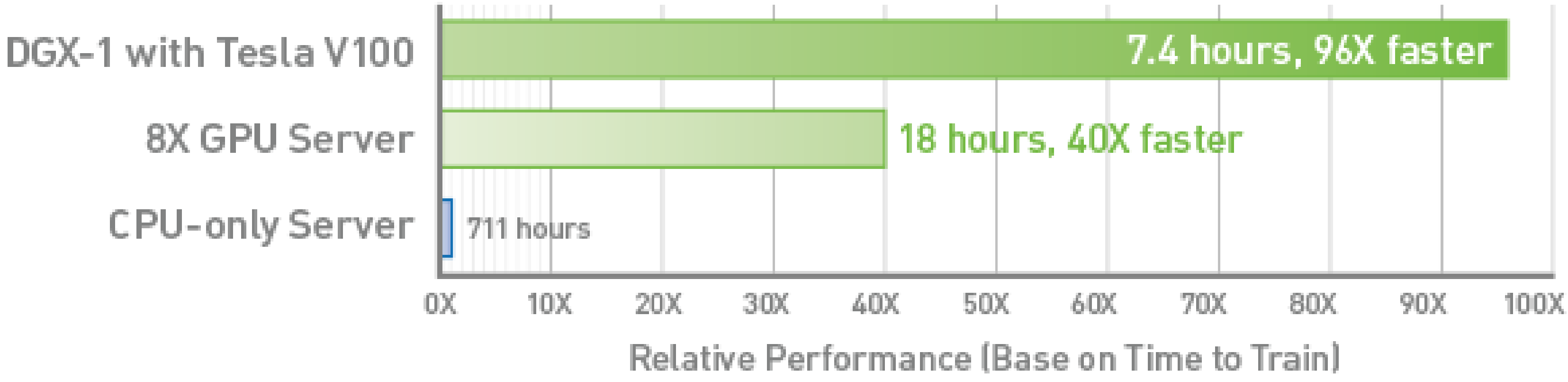
400 servers in a box

$149,000

DNN TRAINING WITH DGX-1

Iterate and Innovate Faster

NVIDIA DGX-1 Delivers 96X Faster Training



Workload: ResNet50, 90 epochs to solution | CPU Server: Dual Xeon E5-2699 v4, 2.6GHz

DNN INFERENCE

TensorRT

High-performance framework makes it easy to develop GPU-accelerated inference

Production deployment solution for deep learning inference

Optimized inference for a given trained neural network and target GPU

Solutions for Hyperscale, ADAS, Embedded

Supports deployment of fp32,fp16,int8* inference

* int8 support will be available from v2

TensorRT for Data Center

Image Classification	Object Detection	Image Segmentation

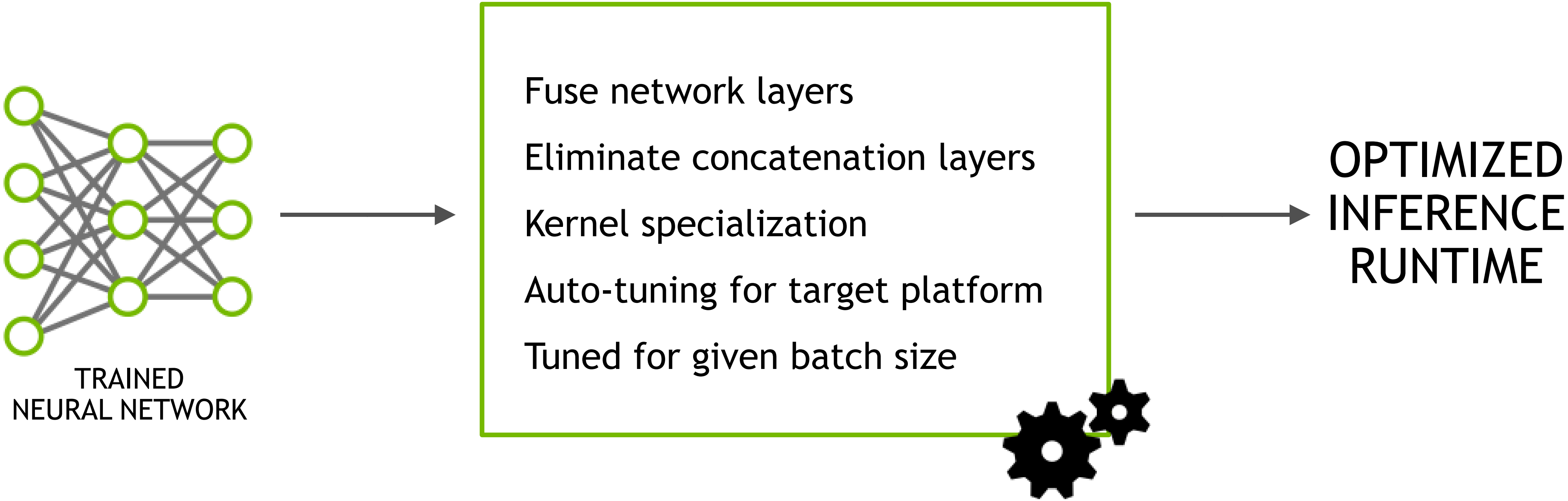
TensorRT for Automotive

Pedestrian Detection	Lane Tracking	Traffic Sign Recognition

NVIDIA DRIVE PX 2

TensorRT

Optimizations



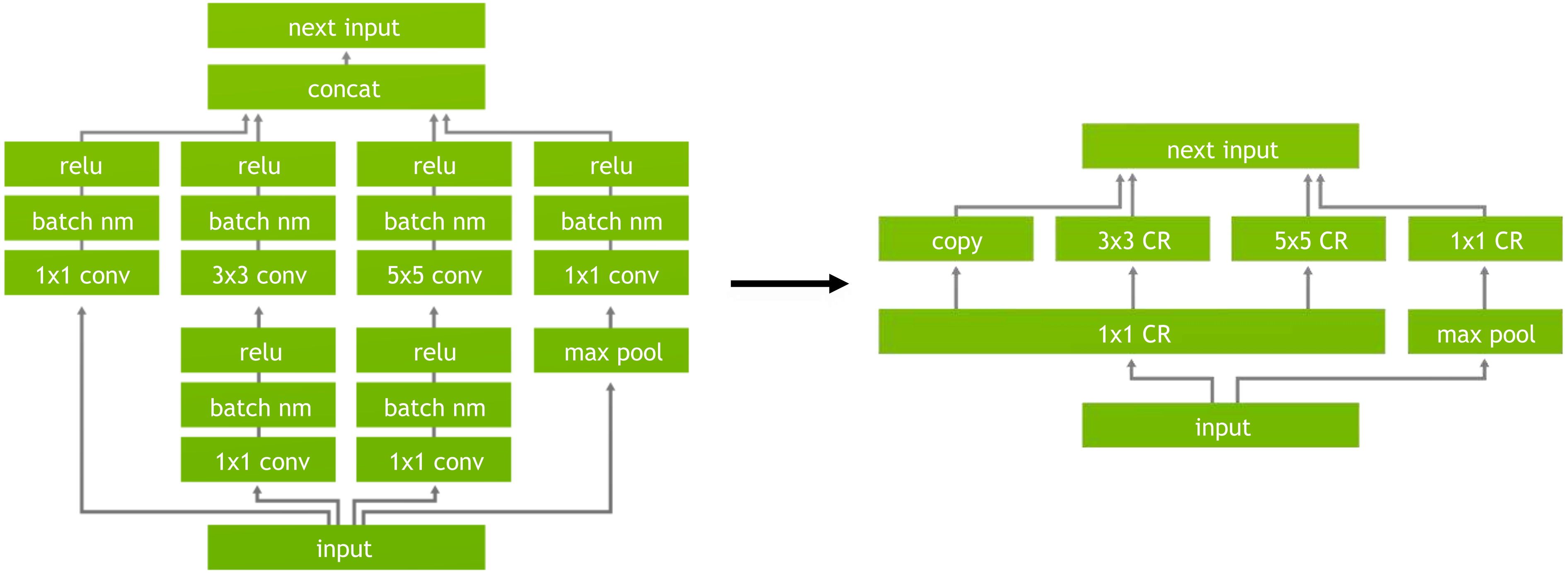
TRAINED NEURAL NETWORK

- Fuse network layers
- Eliminate concatenation layers
- Kernel specialization
- Auto-tuning for target platform
- Tuned for given batch size

OPTIMIZED INFERENCE RUNTIME

NVIDIA TENSORRT

Programmable Inference Accelerator

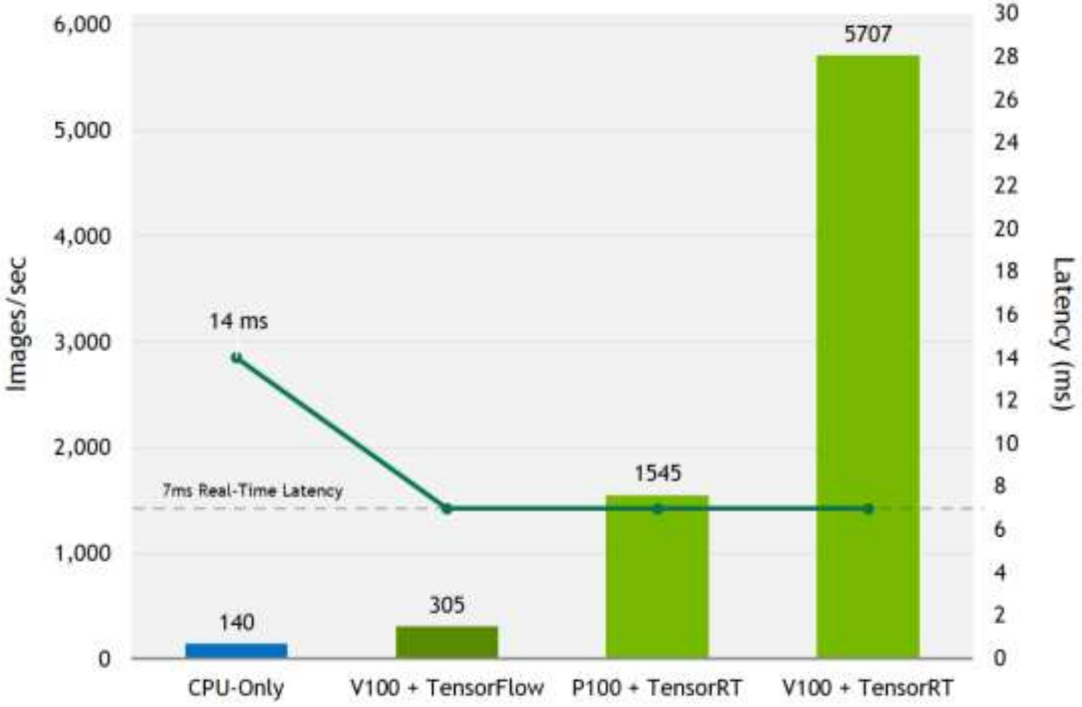


Weight & Activation Precision Calibration | Layer & Tensor Fusion
Kernel Auto-Tuning | Multi-Stream Execution

V100 INFERENCE

Datacenter Inference Acceleration

- 3.7x faster inference on V100 vs. P100
- 18x faster inference on TensorFlow models on V100
- 40x faster than CPU-only



Inference throughput (images/sec) on ResNet50. **V100 + TensorRT**: NVIDIA TensorRT (FP16) @ 6.97 ms latency, batch size 39, Tesla V100-SXM2-16GB, E5-2690 v4@2.60GHz 3.5GHz Turbo (Broadwell) HT On. **P100 + TensorRT**: NVIDIA TensorRT (FP16) @ 6.47 ms latency, batch size 10, Tesla P100-PCIE-16GB, E5-2690 v4@2.60GHz 3.5GHz Turbo (Broadwell) HT On **V100 + TensorFlow**: Preview of volta optimized TensorFlow (FP16) @ 6.67 ms latency, batch size 2, Tesla V100-PCIE-16GB, E5-2690 v4@2.60GHz 3.5GHz Turbo (Broadwell) HT On. **CPU-Only:** Intel Xeon-D 1587 Broadwell-E CPU and Intel DL SDK. Score doubled to comprehend Intel's stated claim of 2x performance improvement on Skylake with AVX512.

AUTONOMOUS VEHICLE TECHNOLOGY

AI IS THE SOLUTION TO SELF DRIVING CARS

PERCEPTION

REASONING

DRIVING

HD MAP

MAPPING

AI COMPUTING

PARKER

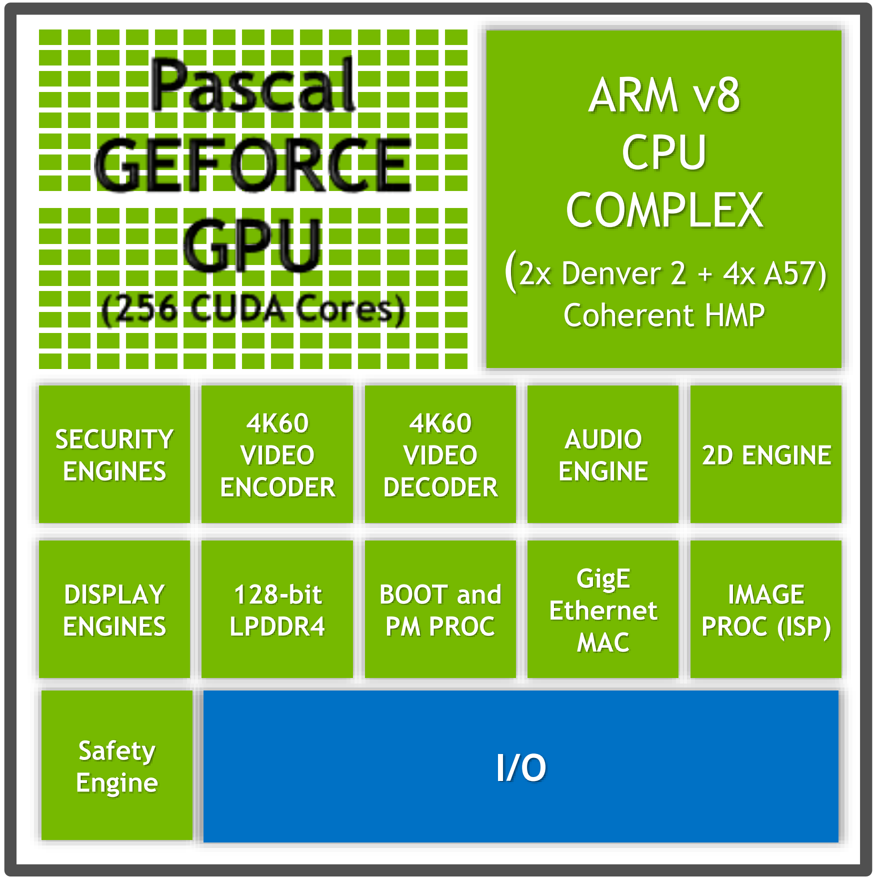
Next-Generation System-on-Chip

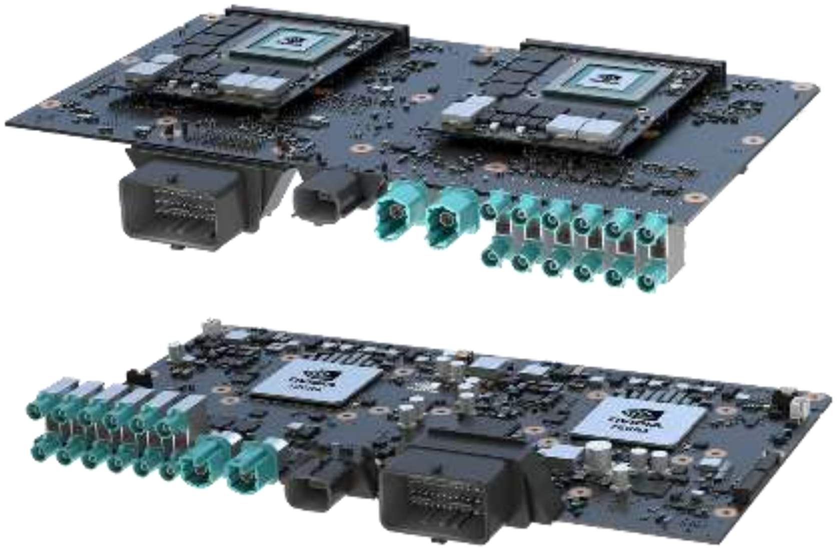
NVIDIA’s next-generation Pascal graphics architecture

1.5 teraflops

NVIDIA’s next-generation ARM 64b Denver 2 CPU

Functional safety for automotive applications





DRIVE PX 2 COMPUTE COMPLEXES

2 Complete AI Systems

Pascal Discrete GPU
- 1,280 CUDA Cores
- 4 GB GDDR5 RAM

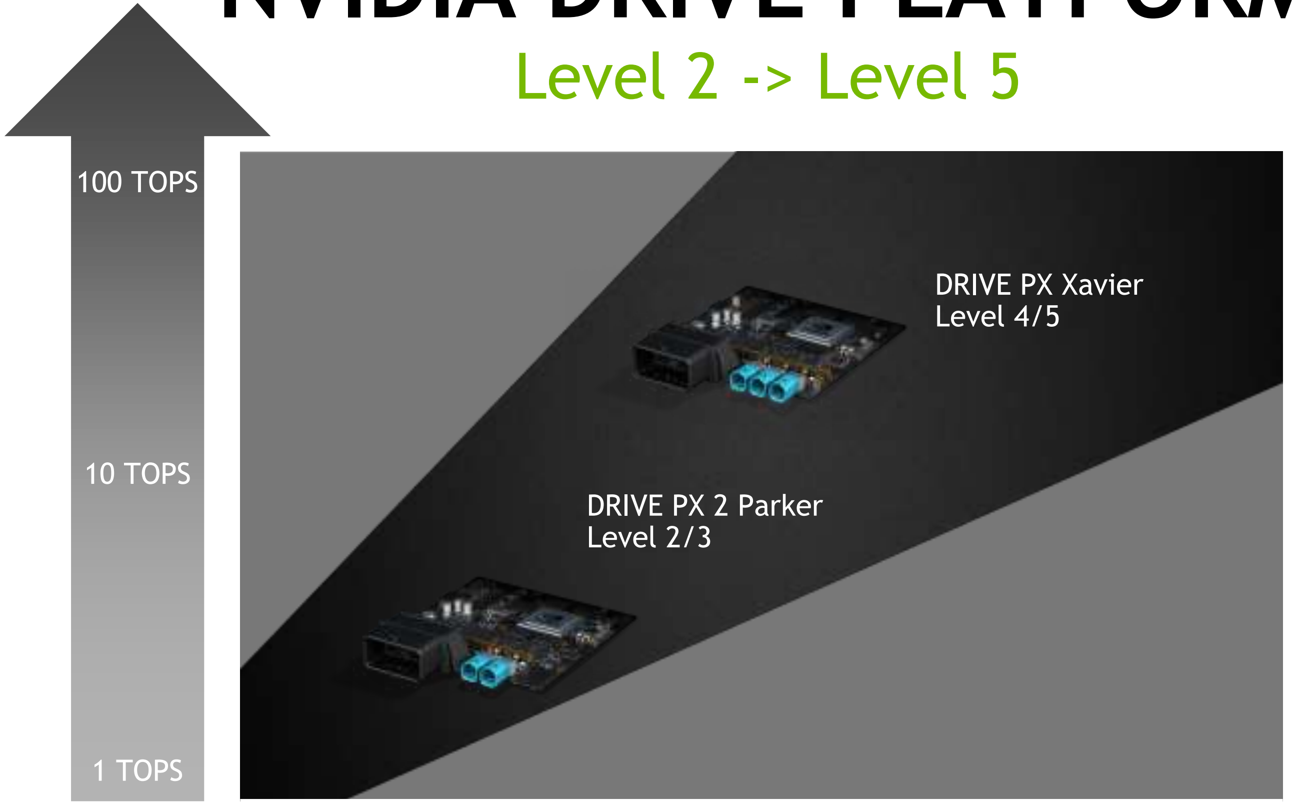
Parker SOC Complex
- 256 CUDA Cores
- 4 Cortex A57 Cores
- 2 NVIDIA Denver2 Cores
- 8 GB LPDDR4 RAM
- 64 GB Flash

Safety Microprocessor

Infineon AURIX Safety Microprocessor
- ASIL D

NVIDIA DRIVE PLATFORM

Level 2 -> Level 5



ONE ARCHITECTURE

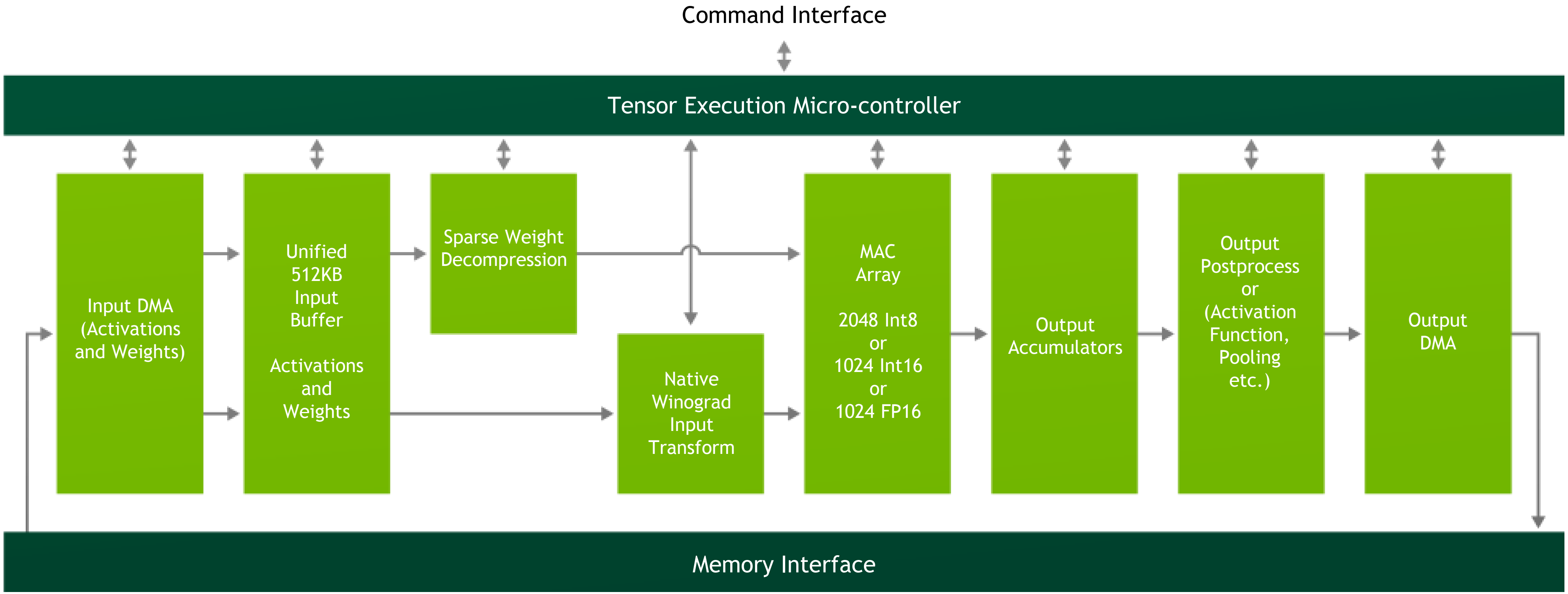
DRIVE PX 2

2 PARKER + 2 PASCAL GPU | 20 TOPS DL | 120 SPECINT | 80W

DRIVE PX (Xavier)

30 TOPS DL | 160 SPECINT | 30W

ANNOUNCING XAVIER DLA NOW OPEN SOURCE


Command Interface
Tensor Execution Micro-controller
Input DMA (Activations and Weights)
Unified 512KB Input Buffer
Activations and Weights
Sparse Weight Decompression
Native Winograd Input Transform
MAC Array
2048 Int8 or 1024 Int16 or 1024 FP16
Output Accumulators
Output Postprocess or (Activation Function, Pooling etc.)
Output DMA
Memory Interface

http://nvdla.org/

NVIDIA DRIVE
END TO END SELF-DRIVING CAR PLATFORM

Caffe

CNTK

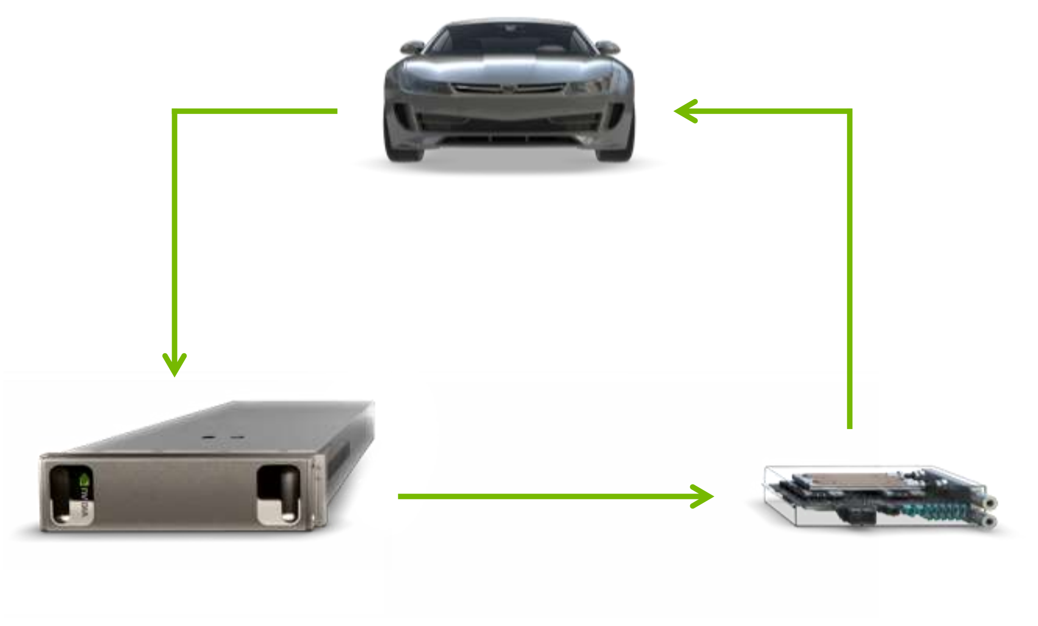
KALDI

TensorFlow

theano

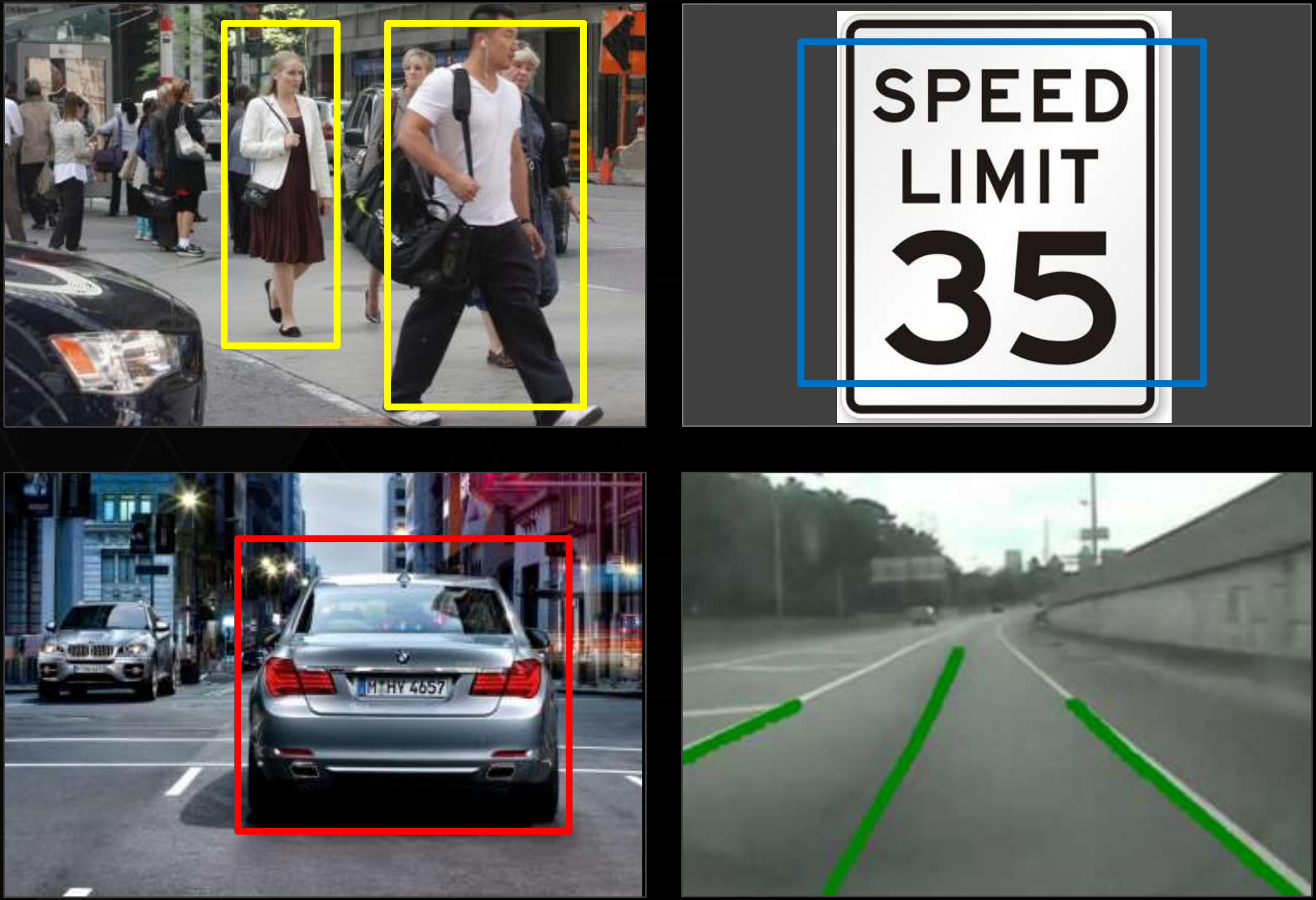
torch

Training on DGX-1

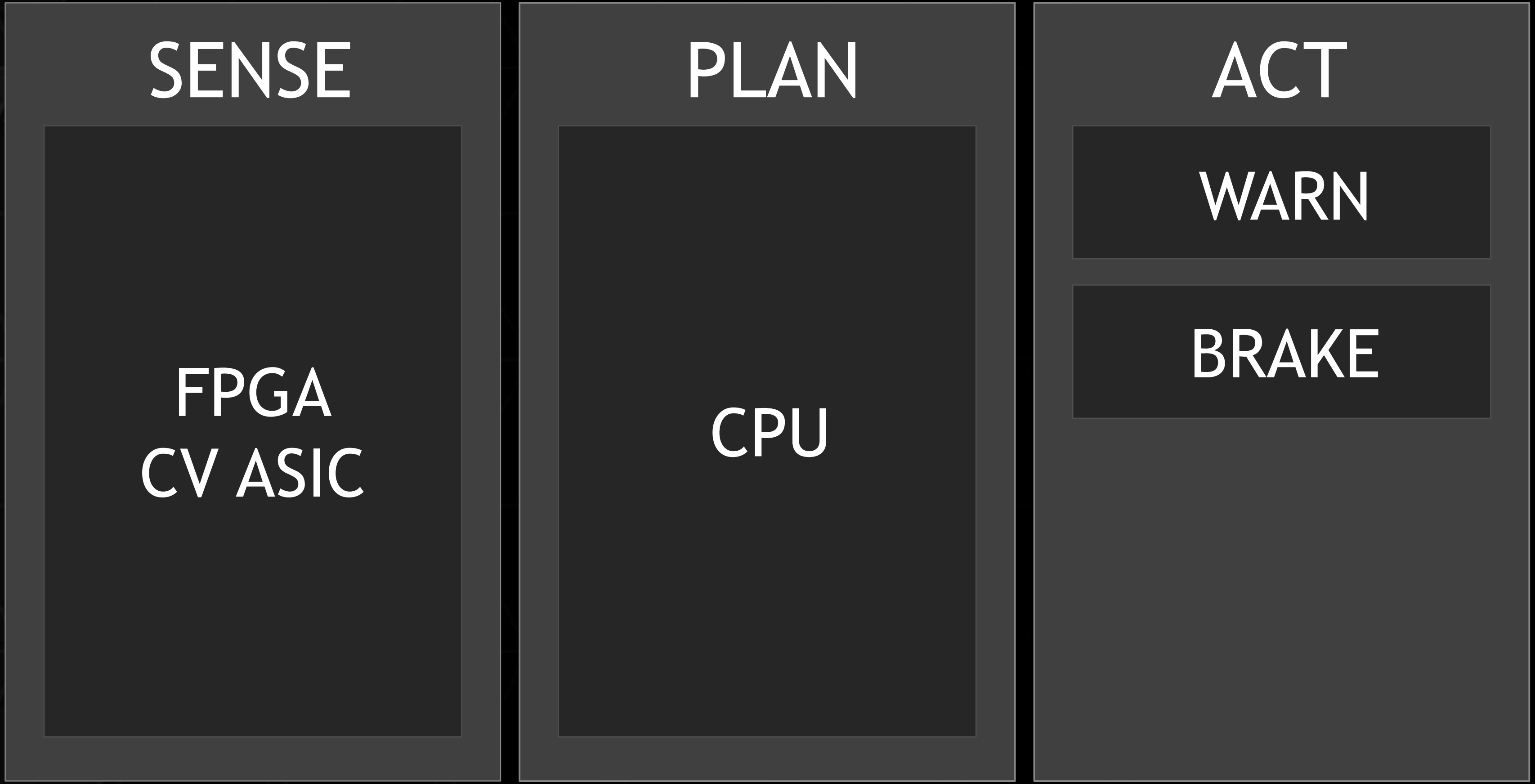


Driving with DriveWorks

DRIVING AND IMAGING



CURRENT DRIVER ASSIST



SENSE

FPGA
CV ASIC

PLAN

CPU

ACT

WARN

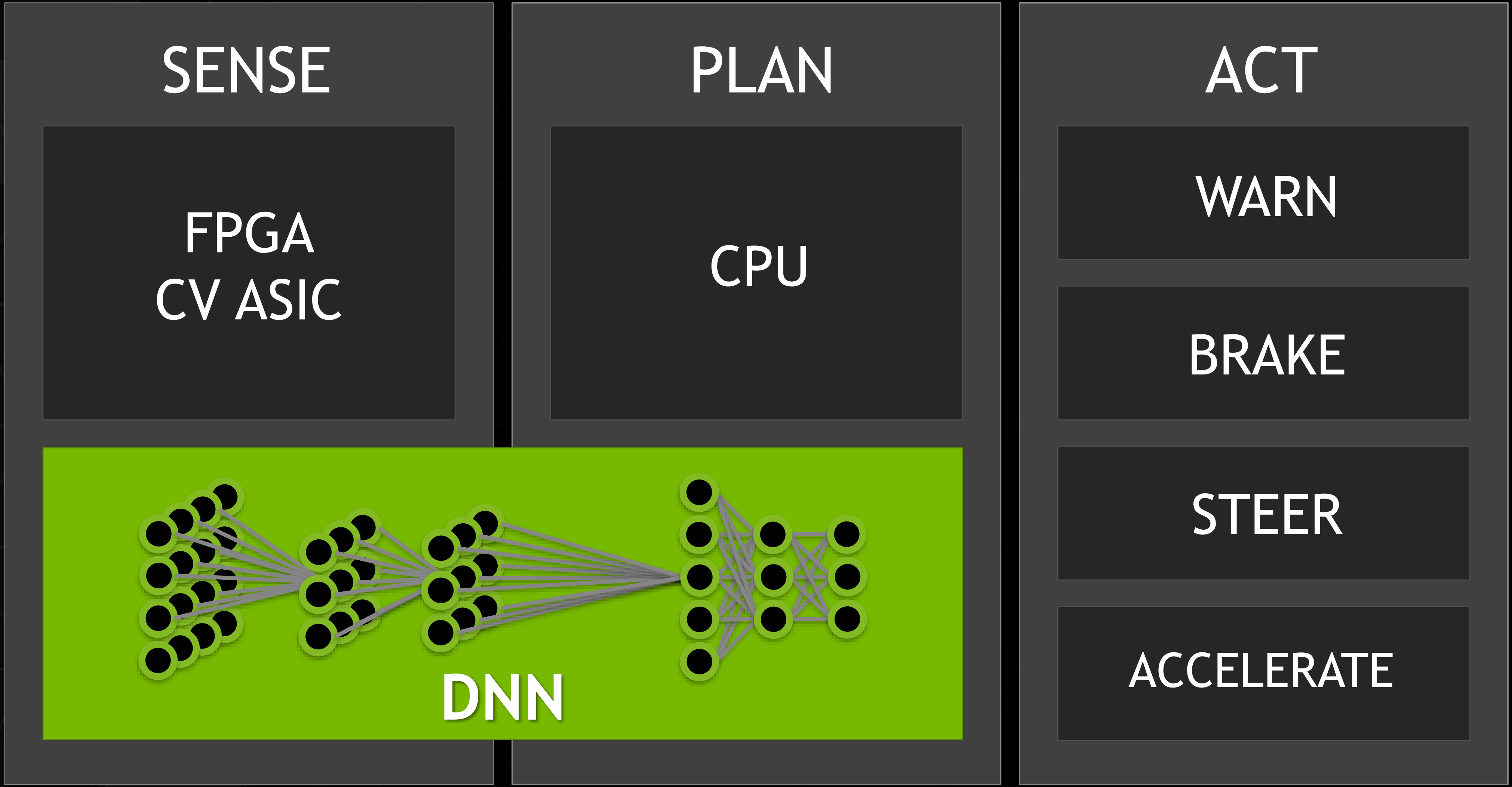
BRAKE

STOP
25

SCHOOL BUS
STOP
STOP
25

STOP
25

FUTURE AUTONOMOUS DRIVING SYSTEM



SENSE

FPGA
CV ASIC

PLAN

CPU

ACT

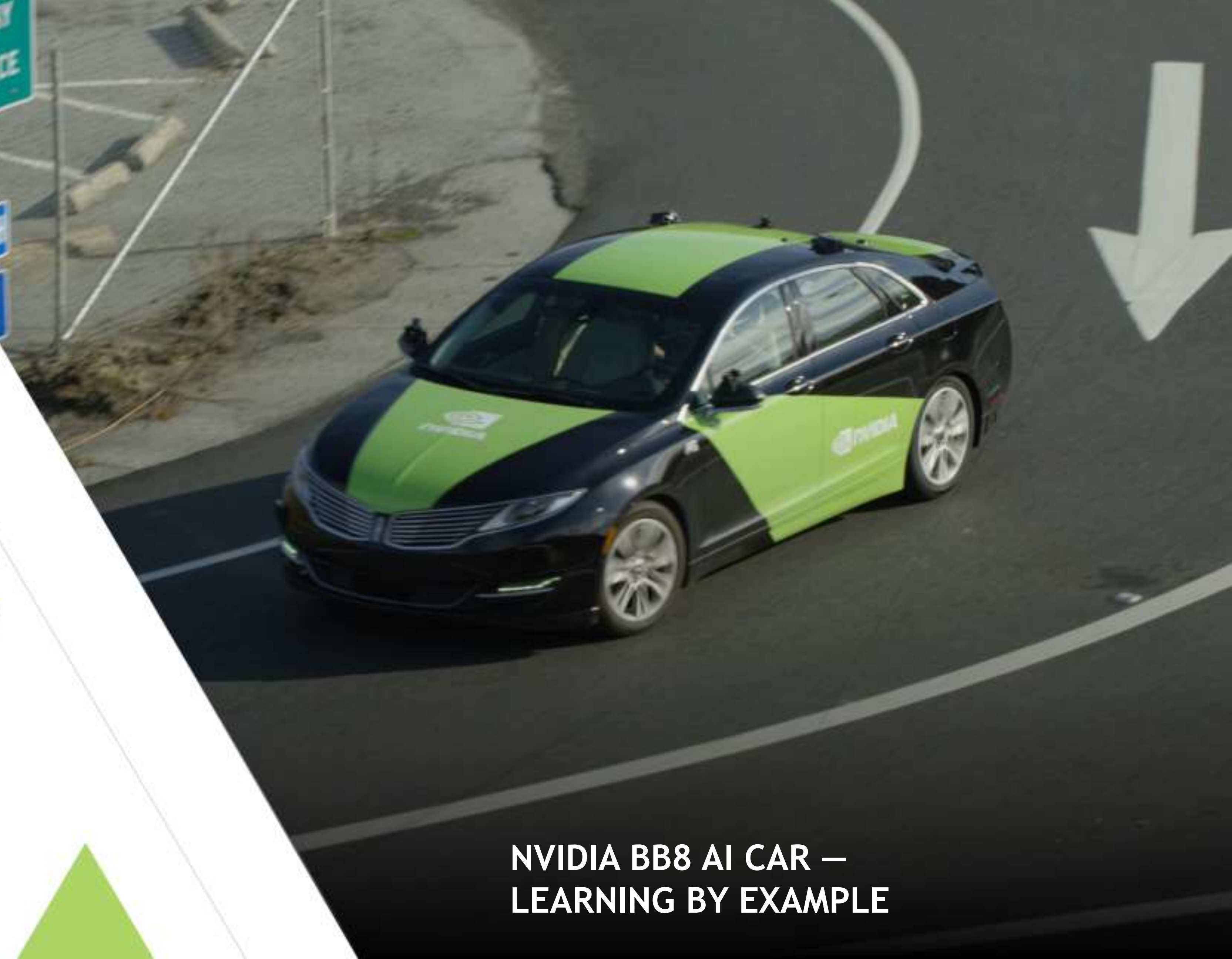
WARN

BRAKE

STEER

ACCELERATE

DNN



NVIDIA BB8 AI CAR — LEARNING BY EXAMPLE

Jetson™ Xavier™
NVIDIA® Isaac™

2018.06.03

OUR CULTURE

A LEARNING MACHINE

INNOVATION

“willingness to take risks”

ONE TEAM

“what’s best for the company”

INTELLECTUAL HONESTY

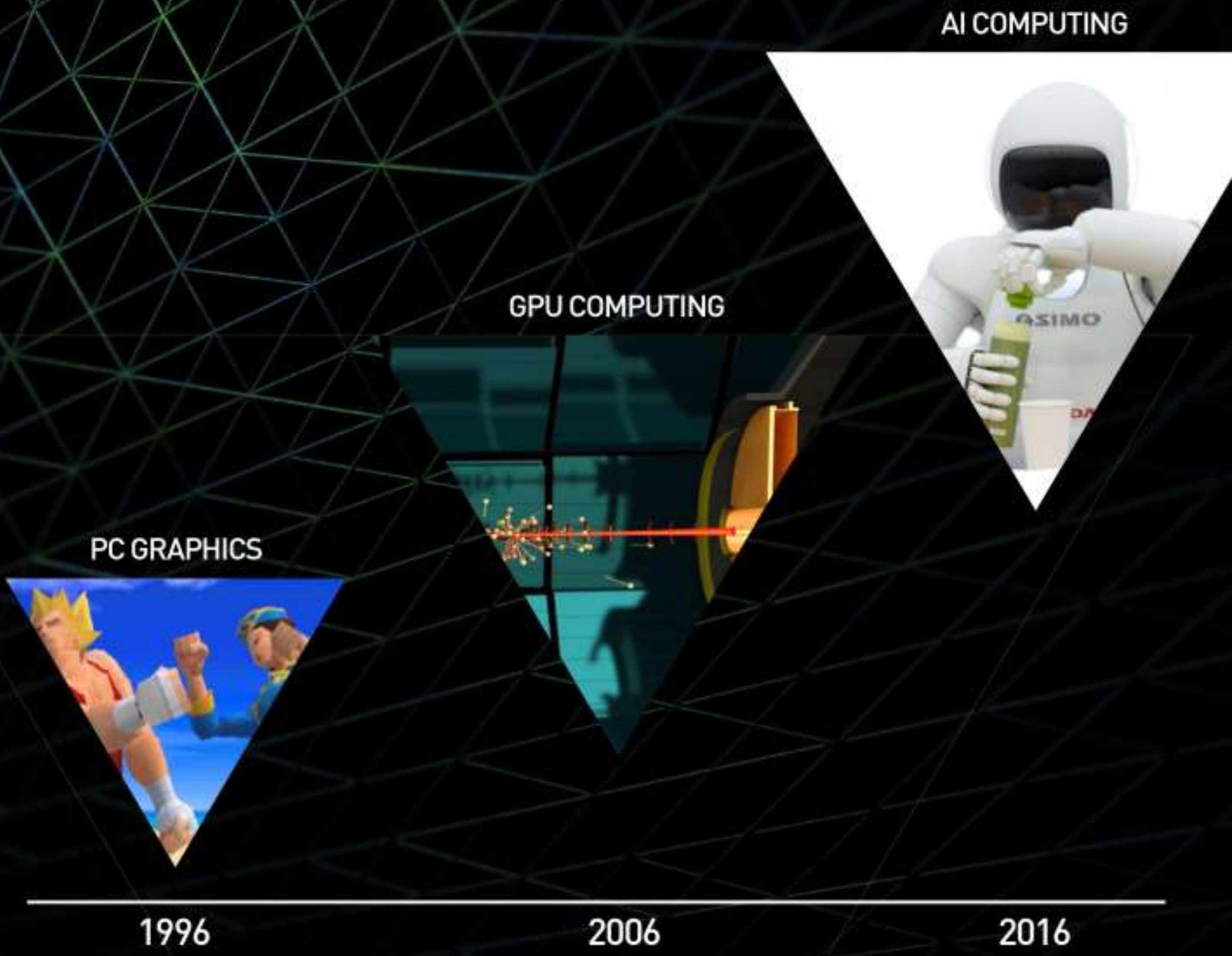
“admit mistakes, no ego”

SPEED & AGILITY

“the world is changing fast”

EXCELLENCE

“hold ourselves to the highest standards”


AI COMPUTING
GPU COMPUTING
PC GRAPHICS
1996
2006
2016