ULTRA-FAST DATA ACQUISITION: GPUS

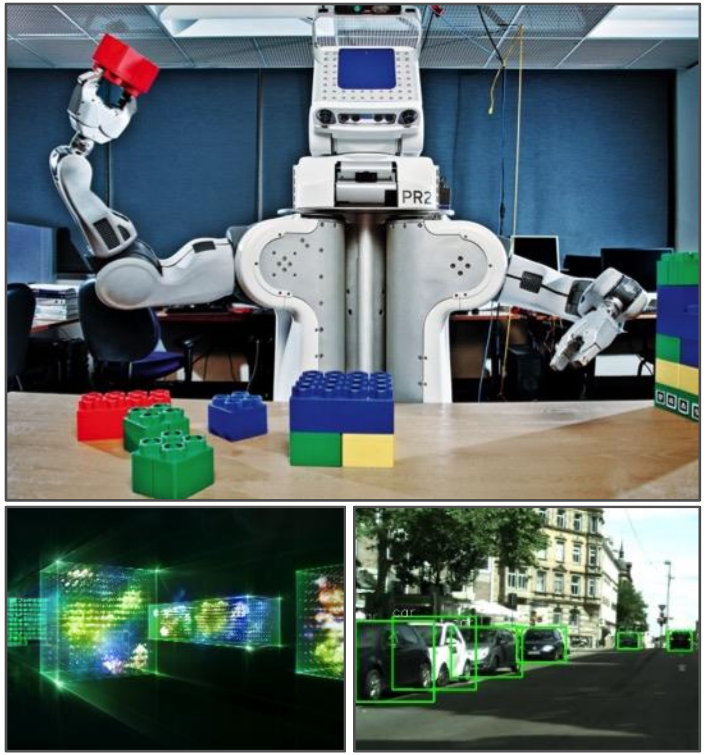
Peter Messmer, Sr. Manager HPC Vis/DevTech

pmessmer@nvidia.com

NVIDIA - AI COMPUTING COMPANY

Computer Graphics

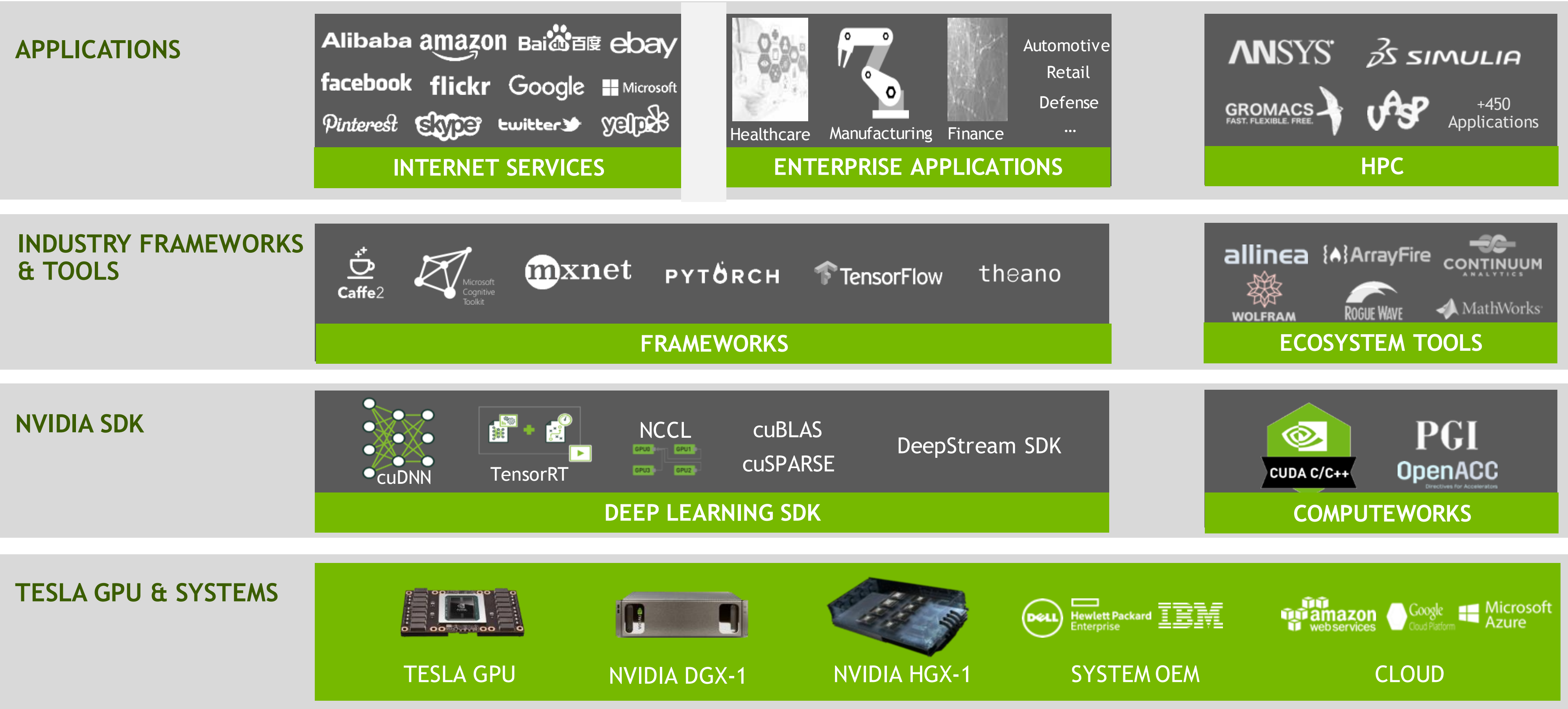
GPU Computing



Artificial Intelligence

TESLA PLATFORM

Leading Data Center Platform for Accelerating HPC and AI



APPLICATIONS

Alibaba amazon Baidu百度 ebay facebook flickr Google Microsoft Pinterest skype twitter yelp

INTERNET SERVICES

Healthcare Manufacturing Finance

Automotive
Retail
Defense
…

ENTERPRISE APPLICATIONS

ANSYS SIMULIA GROMACS FAST. FLEXIBLE. FREE. VASP

+450 Applications

HPC

INDUSTRY FRAMEWORKS & TOOLS

Caffe2 Microsoft Cognitive Toolkit mxnet PYTORCH TensorFlow theano

FRAMEWORKS

allinea ArrayFire CONTINUUM ANALYTICS WOLFRAM ROGUE WAVE MathWorks

ECOSYSTEM TOOLS

NVIDIA SDK

cuDNN TensorRT NCCL cuBLAS cuSPARSE DeepStream SDK

DEEP LEARNING SDK

CUDA C/C++ PGI OpenACC

COMPUTEWORKS

TESLA GPU & SYSTEMS

TESLA GPU NVIDIA DGX-1 NVIDIA HGX-1

DELL Hewlett Packard Enterprise IBM

SYSTEM OEM

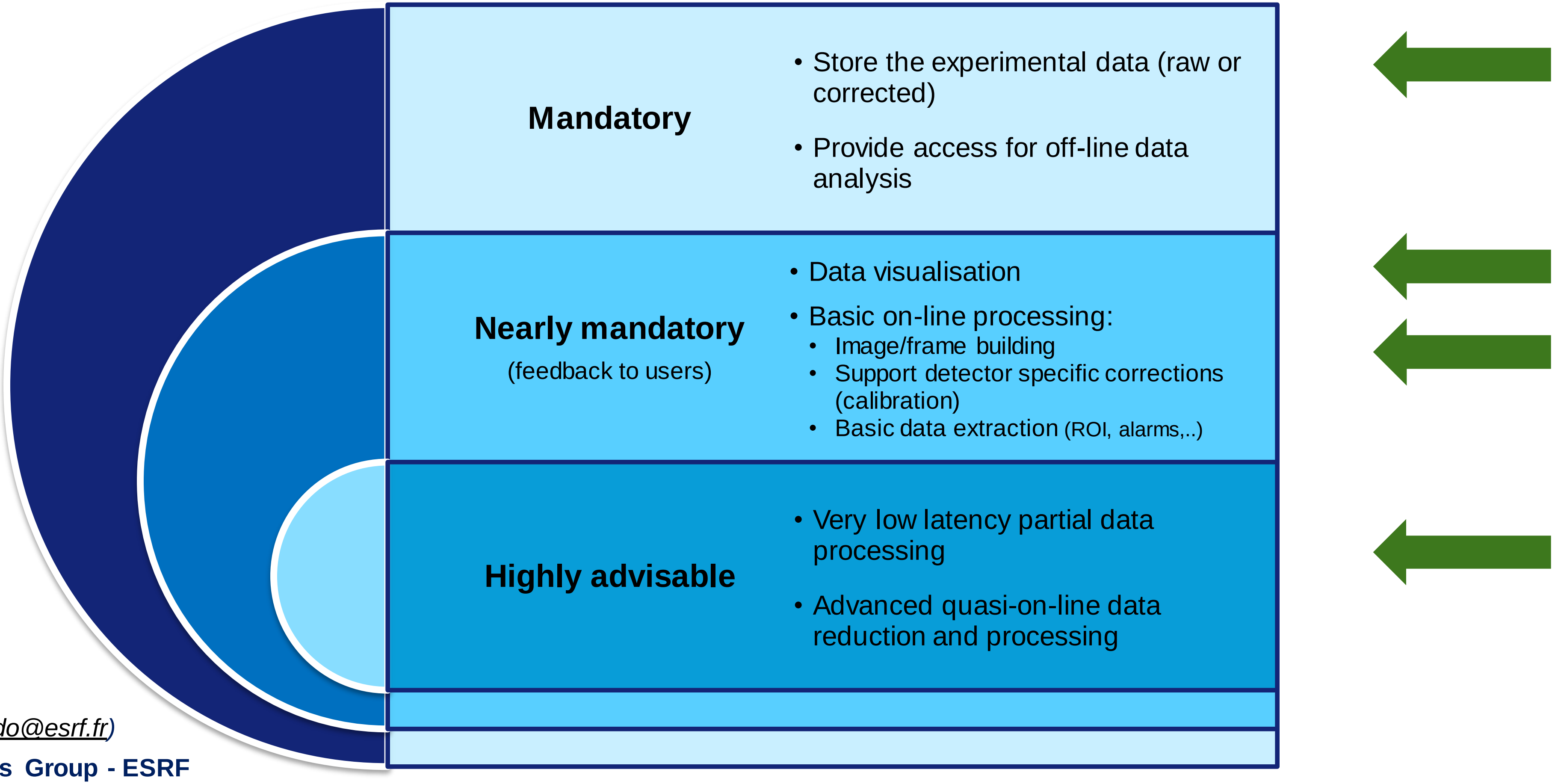
amazon web services Google Cloud Platform Microsoft Azure

CLOUD

Long-term support. E.g. for CT Scanners; ORNL Titan installed in 2012, still running

WHAT THE ACQUISITION SYSTEM SHOULD DO?

Main functional features of the DAQ system



Mandatory

- Store the experimental data (raw or corrected)
- Provide access for off-line data analysis

Nearly mandatory

(feedback to users)

- Data visualisation
- Basic on-line processing:
  - Image/frame building
  - Support detector specific corrections (calibration)
  - Basic data extraction (ROI, alarms,..)

Highly advisable

- Very low latency partial data processing
- Advanced quasi-on-line data reduction and processing

Slide courtesy of

Pablo Fajardo *(fajardo@esrf.fr)*

Detector & Electronics Group - ESRF

AGENDA

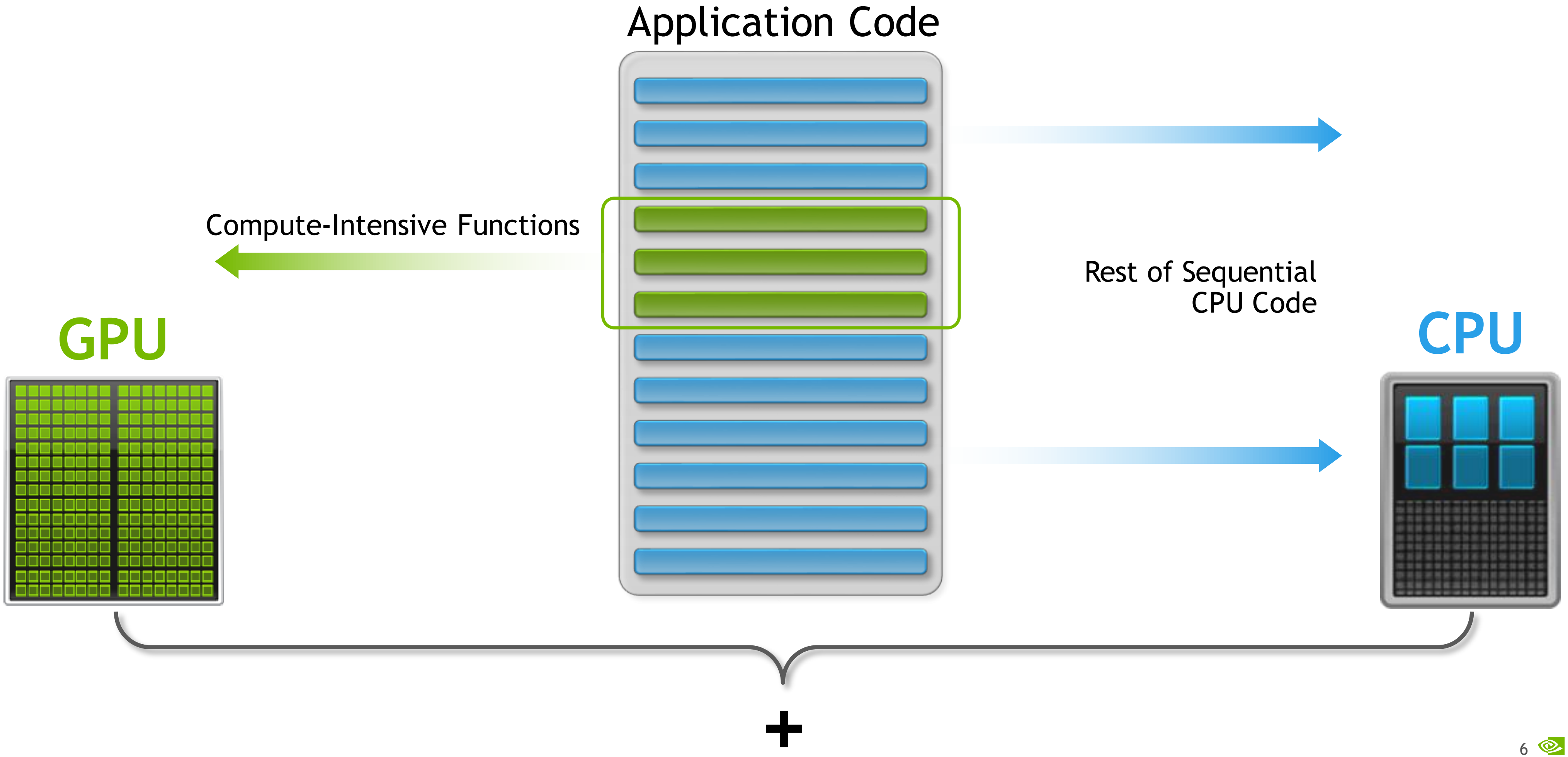
GPU overview

Getting data onto the GPU

Basic data processing

Advanced data processing

HOW GPU ACCELERATION WORKS



Application Code

Compute-Intensive Functions

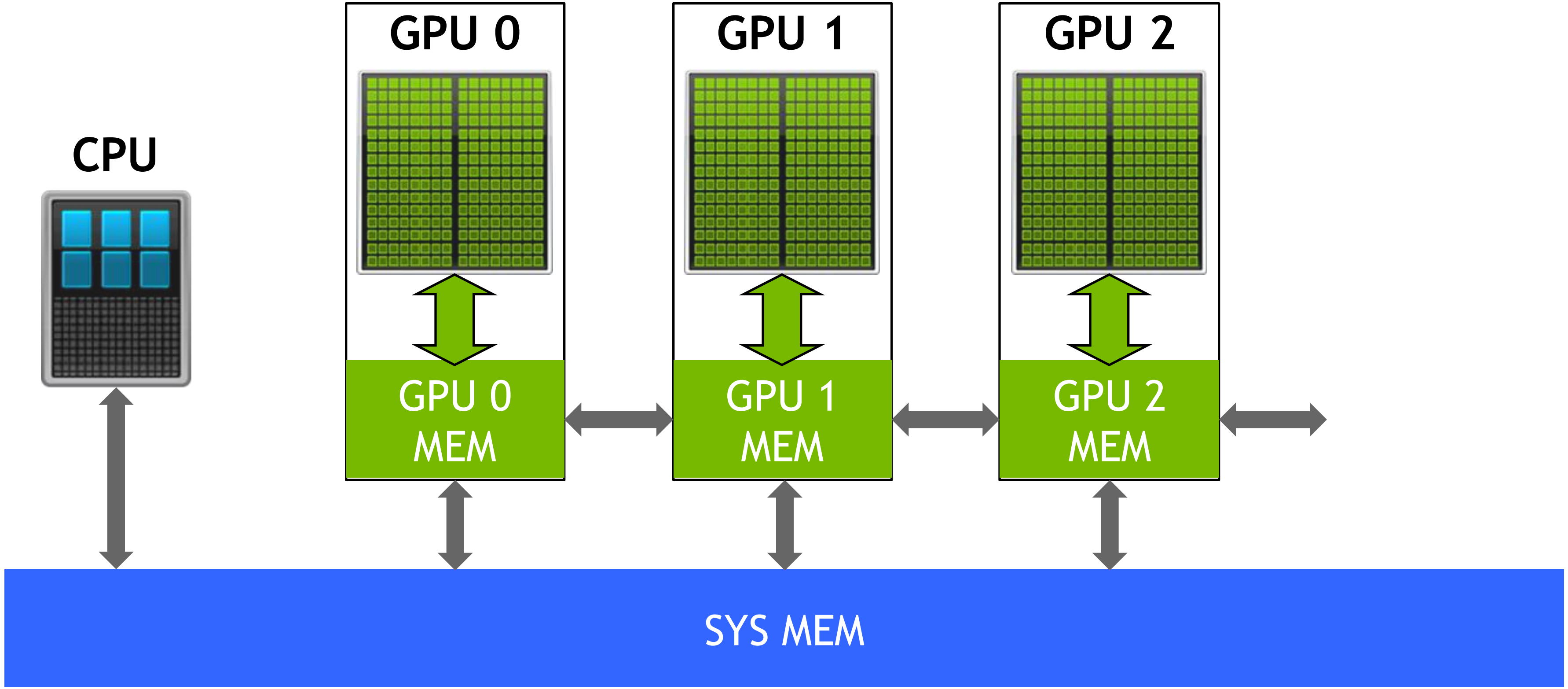
Rest of Sequential
CPU Code

GPU

CPU

+

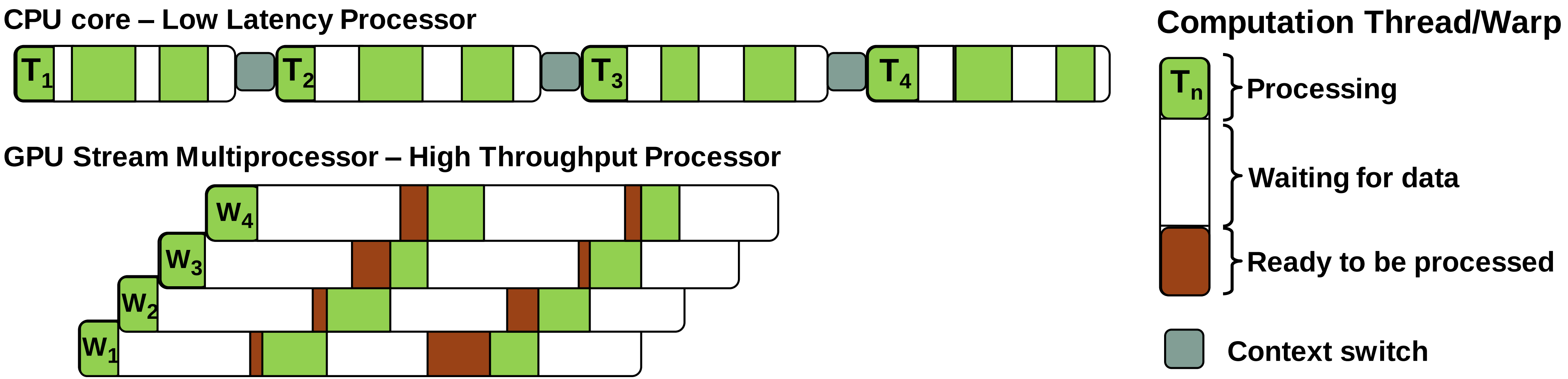
HETEROGENEOUS ARCHITECTURES



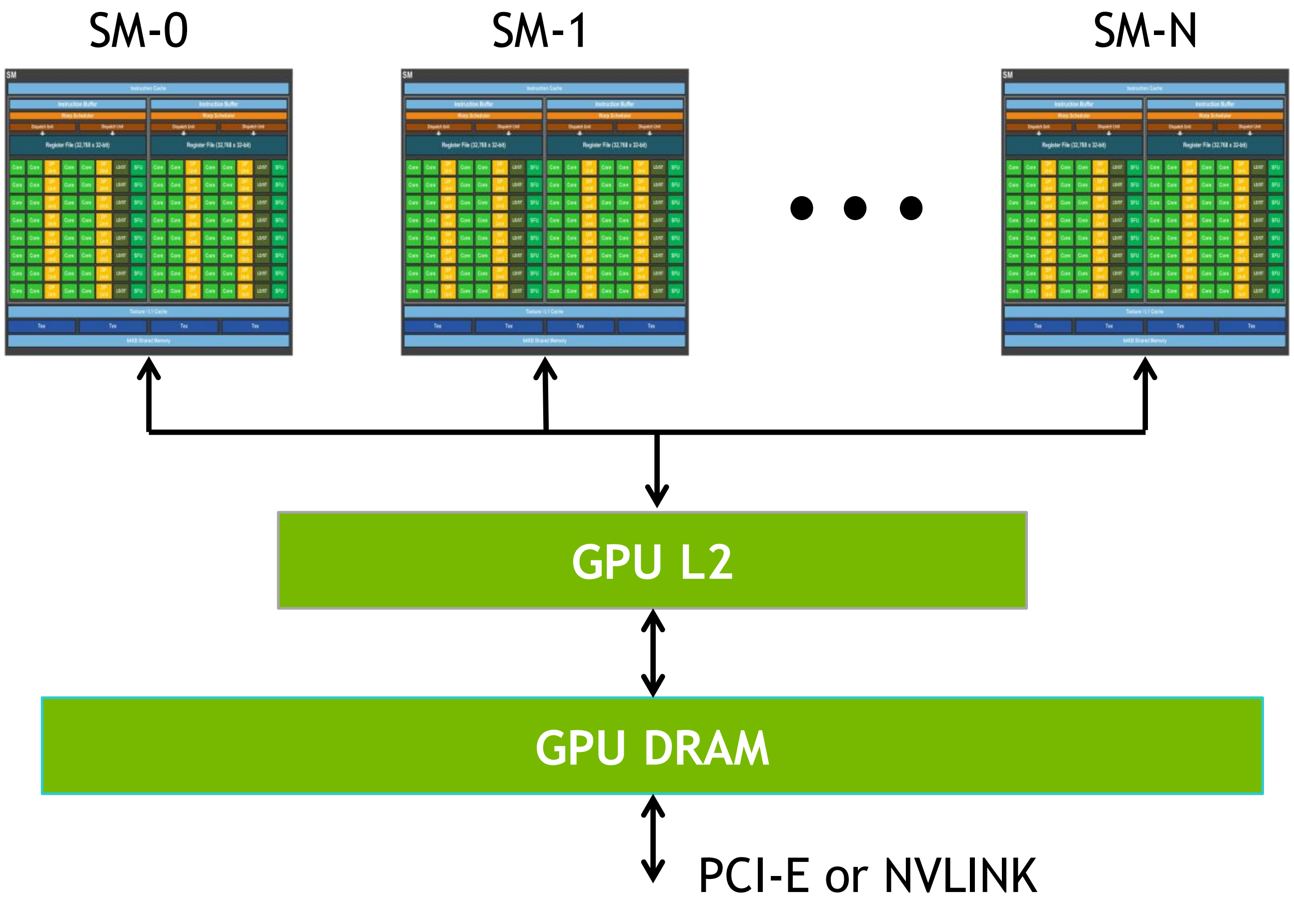
LOW LATENCY OF HIGH THROUGHPUT?

CPU architecture must **minimize latency** within each thread

GPU architecture **hides latency** with computation from other threads (warps)



GPU ARCHITECTURE



SM-0

SM-1

SM-N

GPU L2

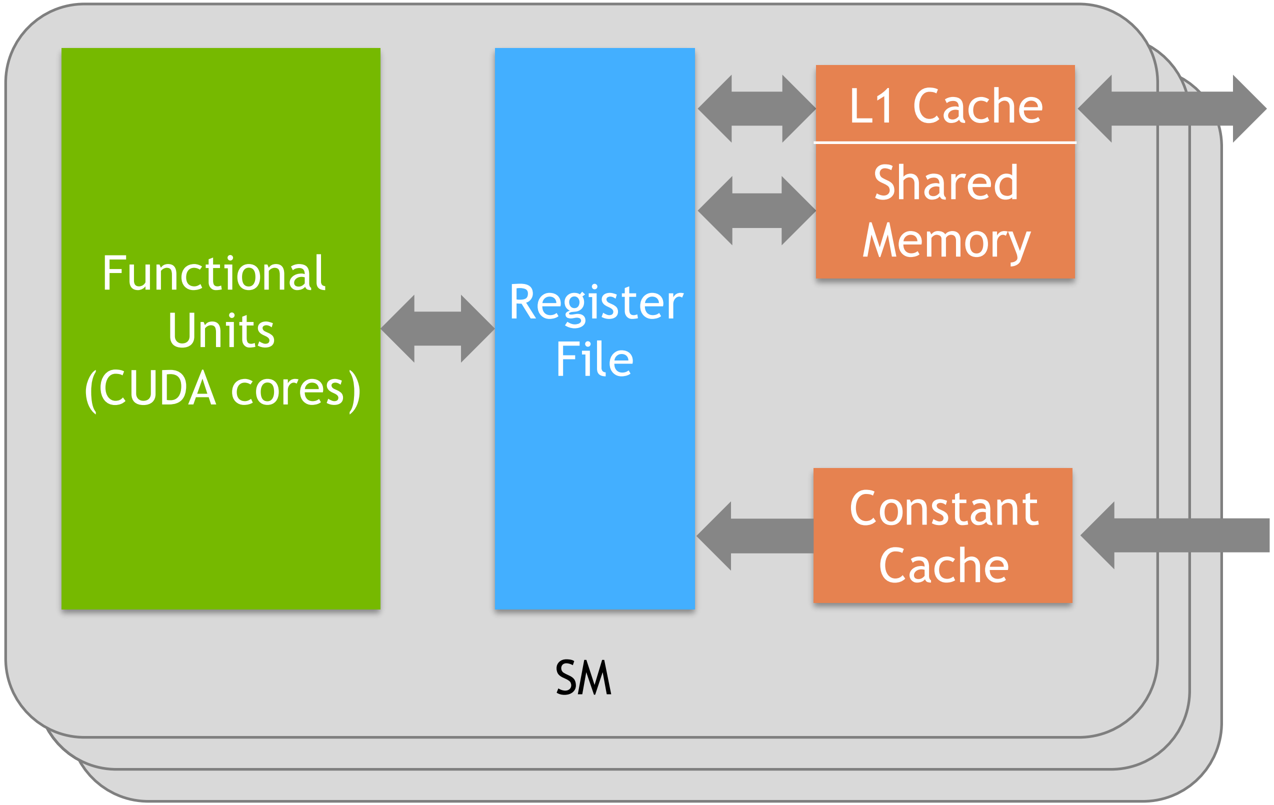
GPU DRAM

PCI-E or NVLINK

GPU SM ARCHITECTURE

Volta SM

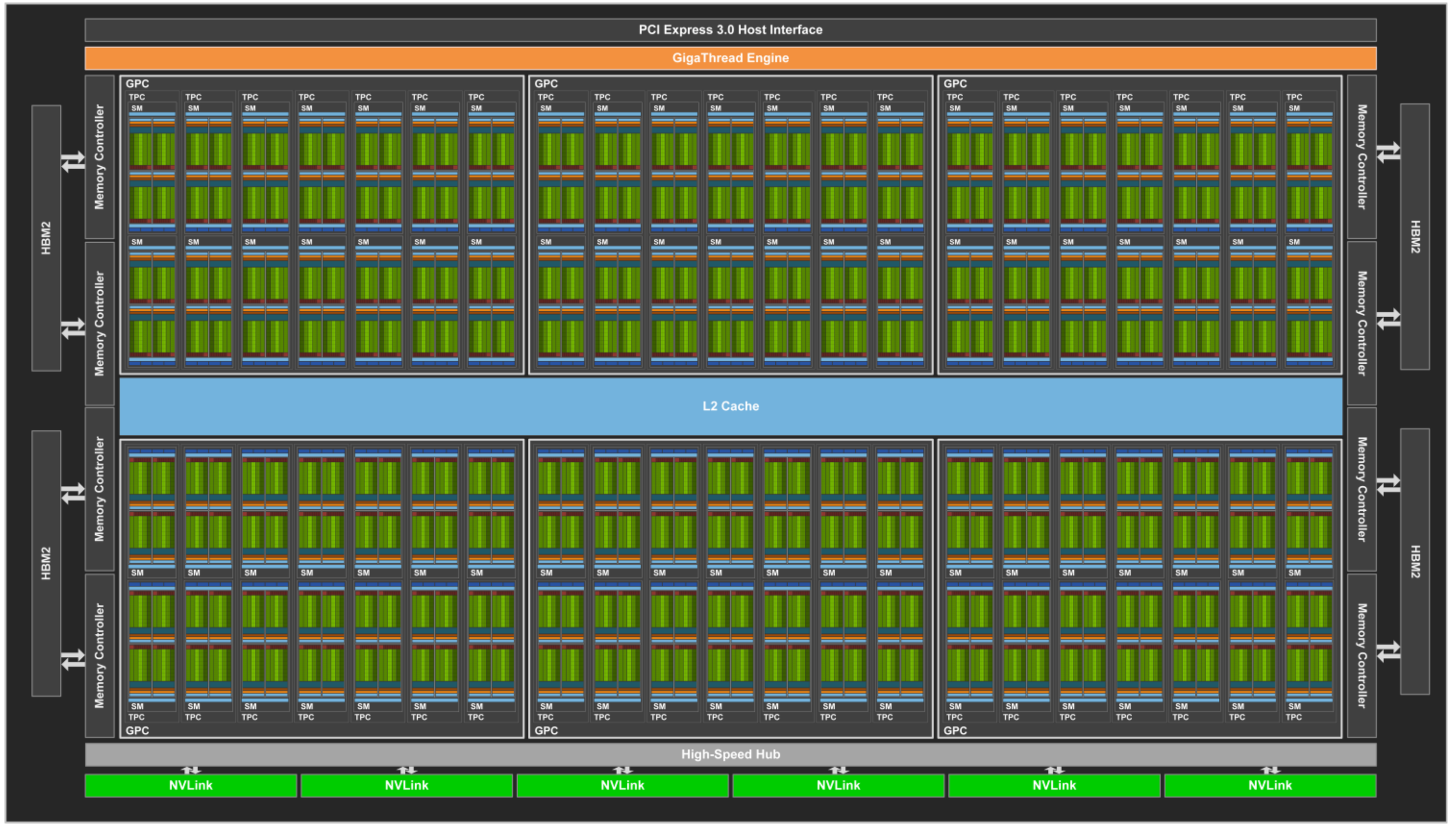
	GV100
FP32 Cores	64
FP64 Cores	32
Tensor Cores	8
Register File	256 KB
Shared Memory	up to 96 KB



80 SMs on Tesla V100

NVIDIA TESLA V100

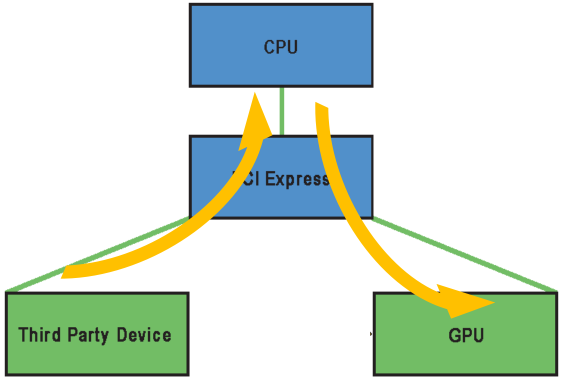
- 21B transistors
  815 mm^{2}, 12nm FFN
- 80 SM
  5120 CUDA Cores
  640 Tensor Cores
- 7.8 FP64 TFLOPS
- 15.6 FP32 TFLOPS
- 125 Tensor TFLOPS
- 16 GB HBM2
  900 GB/s memory bandwidth
- 300 GB/s NVLink bandwidth



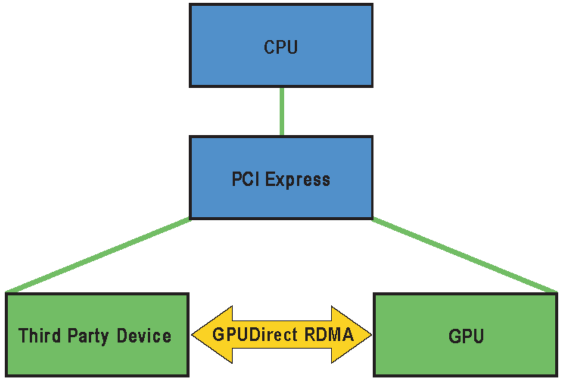
*full GV100 chip contains 84 SMs

GETTING DATA ONTO THE GPU

- Transfer via host-memory
  - Simple if PCIe not saturated
  - 3rd Party Device/GPU not necessarily on same PCI root
  - Use asynchronous mem copies



- GPUDirect RDMA
  - Pin 3rd party device physical BAR addresses against GPU userspace addresses (3rd party device driver mod)
  - Devices on same PCI complex
  - Tesla/Quadro only



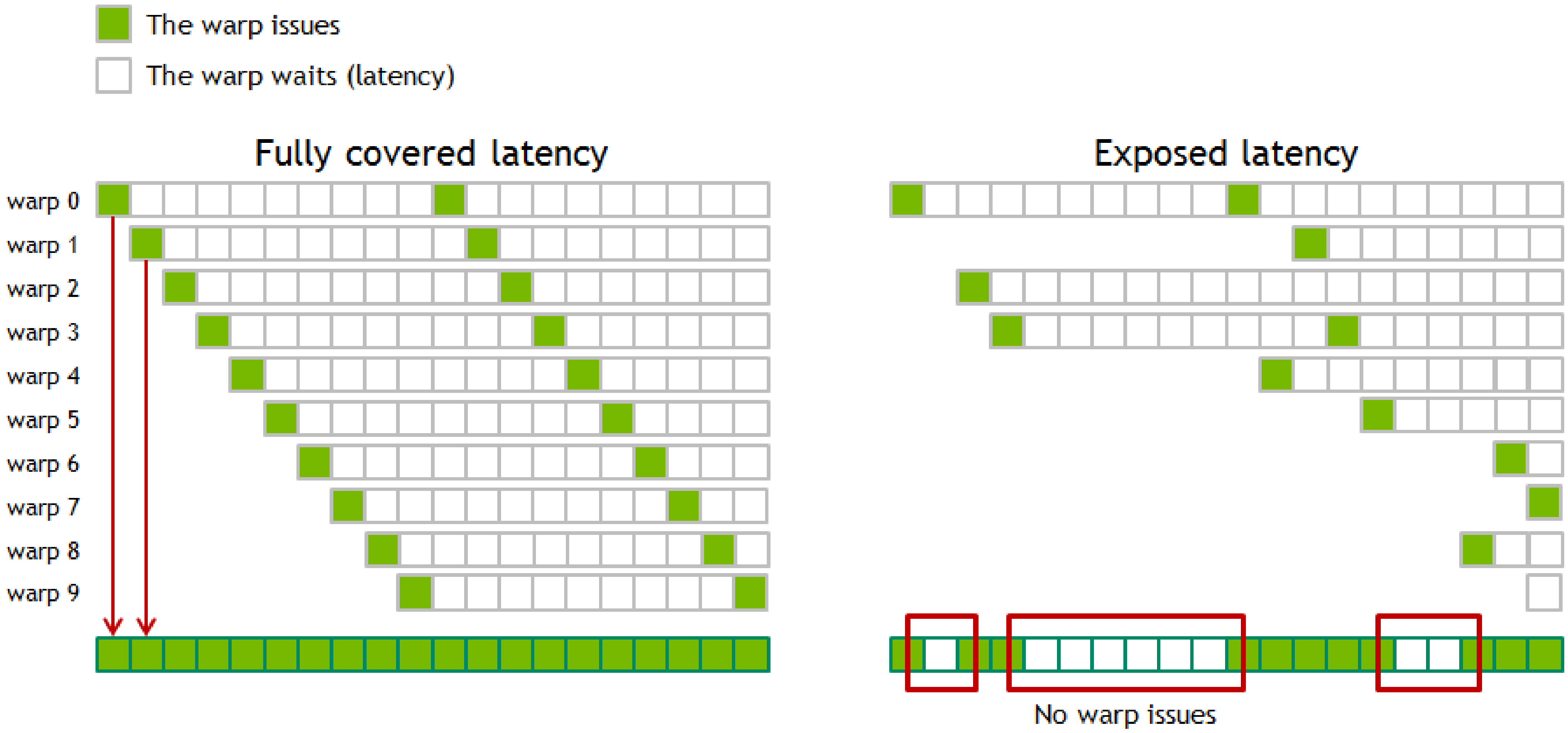
Docu: http://docs.nvidia.com/cuda/gpudirect-rdma
Sample: https://github.com/NVIDIA/gdrcopy

TESLA FAMILY

GPU comparison (boost clocks)

	Tesla K40	Tesla P100	Tesla V100
Peak FP32 (TFLOP/s)	5.04	10.6	15
Peak FP64 (TFLOP/s)	1.68	5.3	7.5
Peak Tensor Core (TFLOP/s)	N/A	N/A	120
Memory Size (GB)	12	16	16
Memory Bandwidth (GB/s)	288	732	900

LATENCY HIDING



🟩 The warp issues

□ The warp waits (latency)

Fully covered latency

Exposed latency

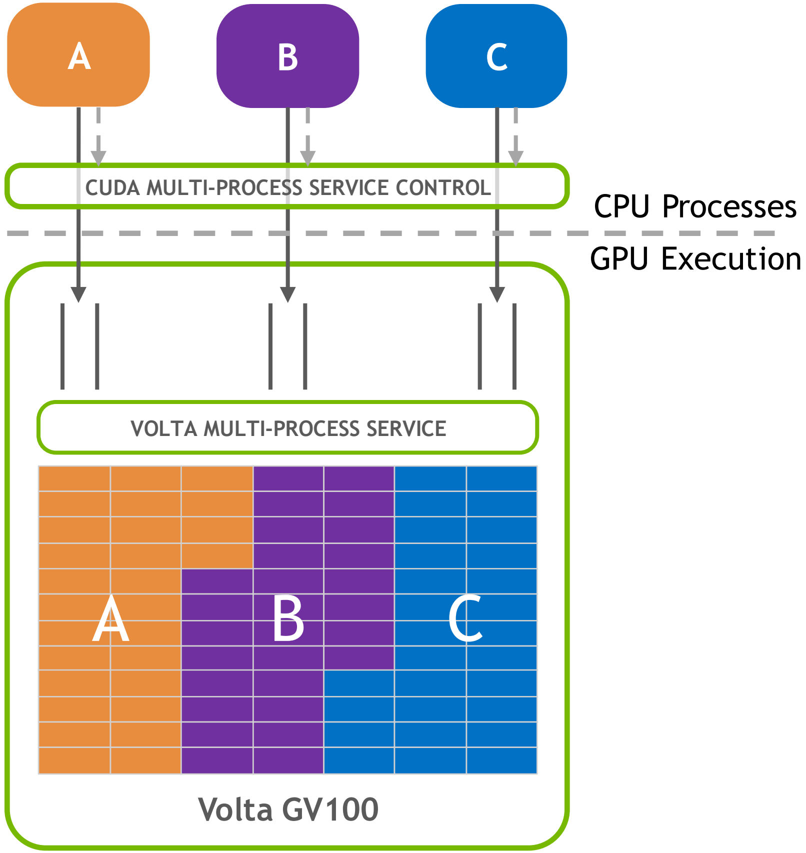
Need sufficient work! > 120000 threads in flight at any time

VOLTA MULTI-PROCESS SERVICE

- GPUs getting wide
- Some problems too small to fill entire GPU
- Share GPU amongst multiple CPU processes

Hardware Accelerated Work Submission

Hardware Isolation


A
B
C
CUDA MULTI-PROCESS SERVICE CONTROL
CPU Processes
GPU Execution
VOLTA MULTI-PROCESS SERVICE
A
B
C
Volta GV100

GPU FOR SIGNAL PROCESSING

Bit Twiddling

- Integer INT32 instruction throughput: same as floating point (FP32) throughput *
- Half precision FP16 throughput 2xFP32 (sign, 5 exp, 10 mant). 10 bit integers.
- FP16 matrix-product cores (Tensor Cores), implicit FP16 -> FP32

http://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html#arithmetic-instructions

Rearranging data

- Relaxed coalescence constraints (32B transaction size)
- Shared memory on SM for fast, irregular access

Visualizing

HARDWARE ACCELERATED COMPRESSION ENGINES

Video compression and more

Multiple encoder/decoder per chip

(3 enc, 1 dec on V100)

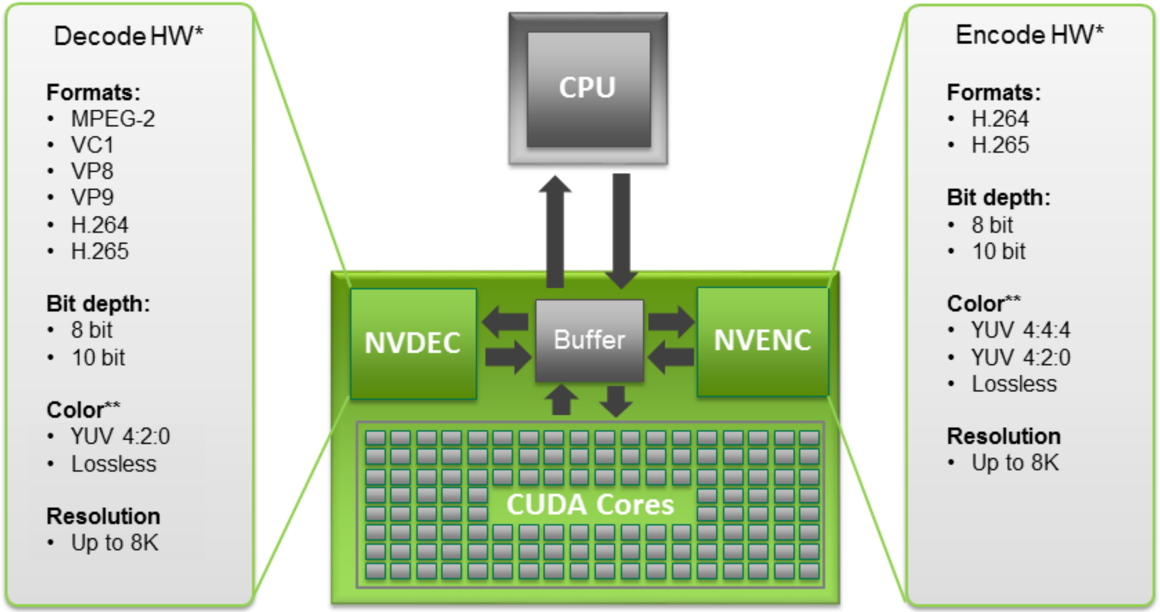
Transparently managed

Lots of formats supported, incl. lossless

I-frame, P-frame support

No CUDA/Rendering resources used

API: NVCODEC



Decode HW*

Formats:
- MPEG-2
- VC1
- VP8
- VP9
- H.264
- H.265

Bit depth:
- 8 bit
- 10 bit

Color**
- YUV 4:2:0
- Lossless

Resolution
- Up to 8K

Encode HW*

Formats:
- H.264
- H.265

Bit depth:
- 8 bit
- 10 bit

Color**
- YUV 4:4:4
- YUV 4:2:0
- Lossless

Resolution
- Up to 8K

https://developer.nvidia.com/nvidia-video-codec-sdk

VIDEO FOR DATA COMPRESSION

Not only for streaming

Measured data cubes exhibit lots of coherency

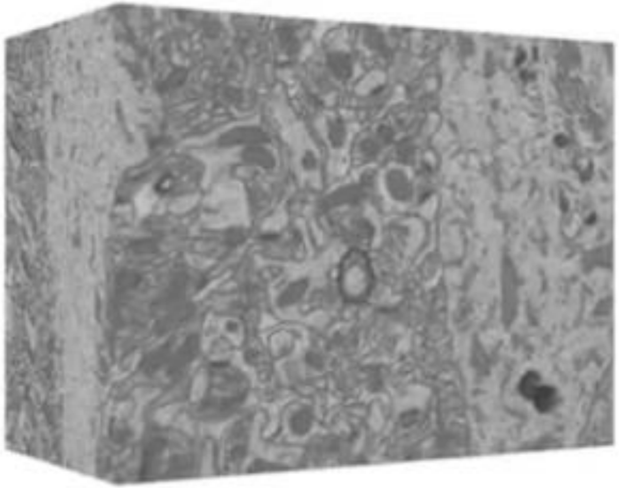
Huge compression possible

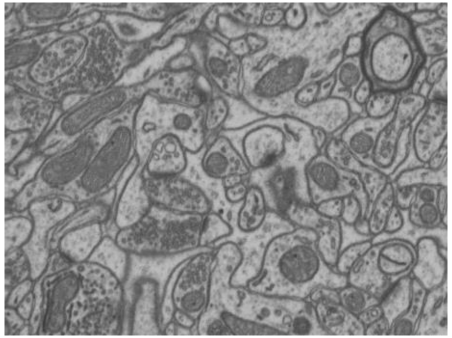
3.2GB -> 51 MB (-> 21MB)

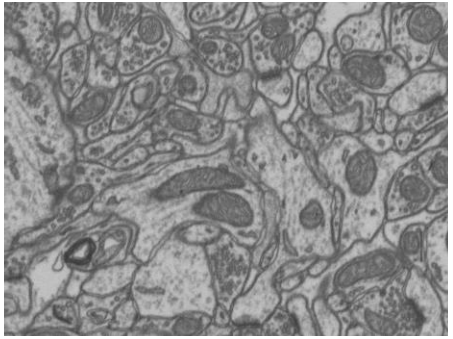
Opportunity for large volume data

GPU-GPU use case not yet supported

Get in touch for more details



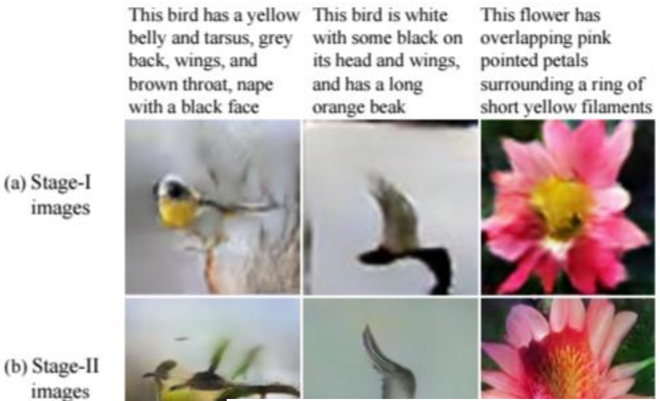




Data and images: http://cvlab.epfl.ch/data/em

WHY THE AI EXCITEMENT?

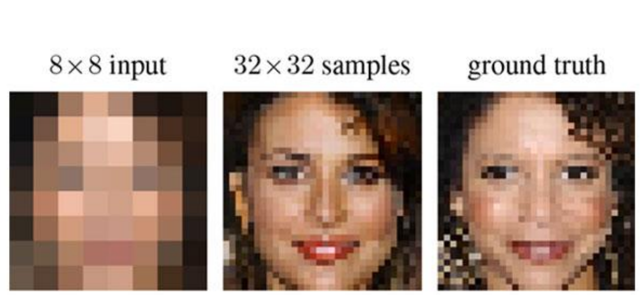
GPUs as Enablers of Breakthrough Results



This bird has a yellow belly and tarsus, grey back, wings, and brown throat, nape with a black face

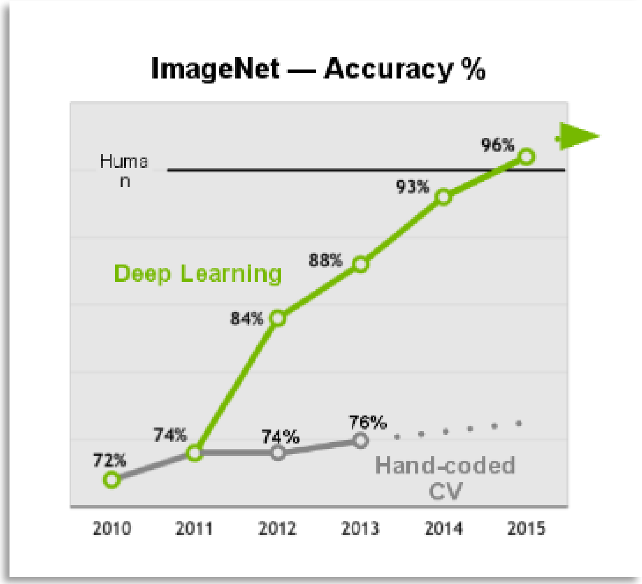
This bird is white with some black on its head and wings, and has a long orange beak

This flower has overlapping pink pointed petals surrounding a ring of short yellow filaments

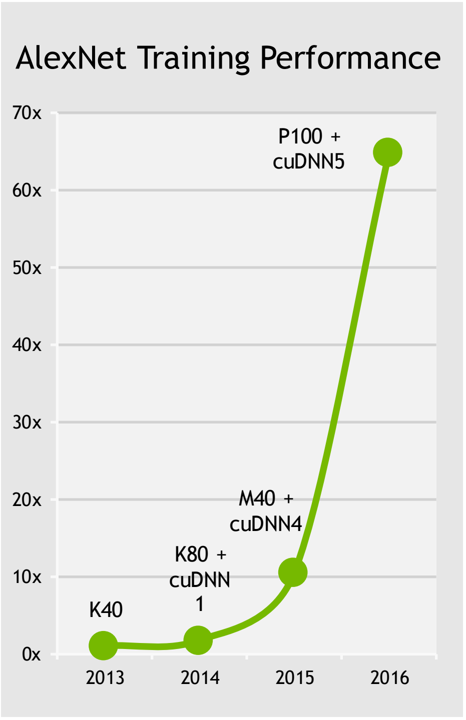


Dahl et al. 2017

We can generate photorealistic images from textual descriptions and super-enhance blurry photos!



Achieve super-human accuracy in classification



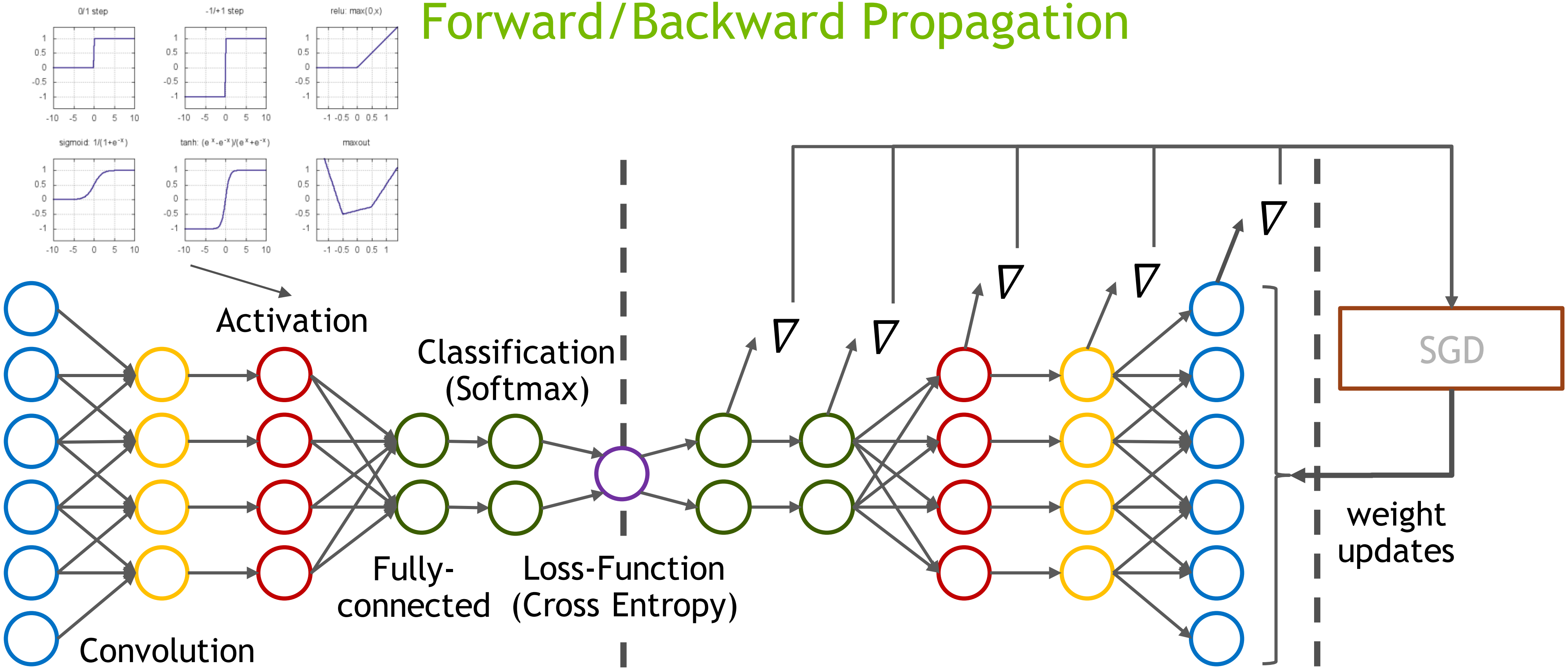
65x in 3 Years

And we are getting faster fast

Paper: H.Zhang et al. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks, arXiv:1612.03242

1-SLIDE INTRO TO CONVOLUTIONAL NEURAL NETS

Forward/Backward Propagation



Input

Forward-propagation

Backward-propagation (gradient computation)

All layers are differentiable

Optimization Algorithm

AI INFERENCING IS EXPLODING

2 Trillion

Messages Per Day On LinkedIn

PERSONALIZATION

500M

Daily active users of iFlyTek

SPEECH

140 Billion

Words Per Day Translated by Google

TRANSLATION

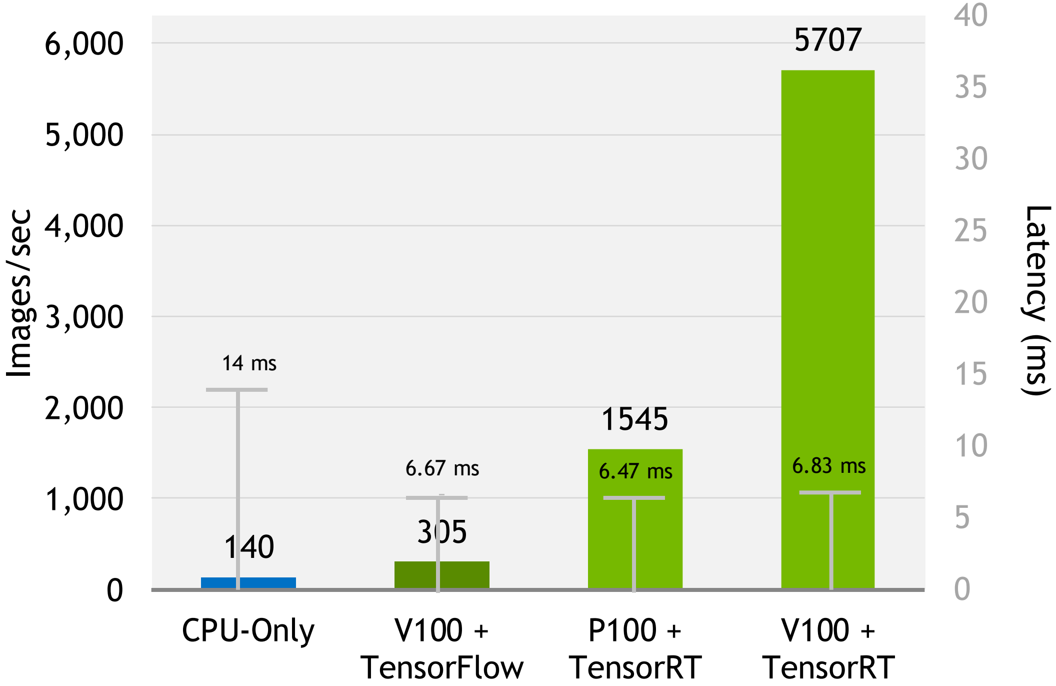
60 Billion

Video frames/day uploaded on Youtube

VIDEO

TENSORRT 3 PERFORMANCE

40x Faster CNNs on V100 vs. CPU-Only Under 7ms Latency (ResNet50)



Inference throughput (images/sec) on ResNet50. **V100 + TensorRT**: NVIDIA TensorRT (FP16), batch size 39, Tesla V100-SXM2-16GB, E5-2690 v4@2.60GHz 3.5GHz Turbo (Broadwell) HT On. **P100 + TensorRT**: NVIDIA TensorRT (FP16), 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), 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-ECPU and Intel DL SDK. Score doubled to comprehend Intel's stated claim of 2x performance improvement on Skylake with AVX512.

Image classification network

224x224 pixels images

1000 categories

50 layers, mostly convolutions

Under 7ms latency from pixels to class

NORMALIZING AND PRE-PROCESSING

DL trained on jet images vs. physically-motivated feature driven approaches

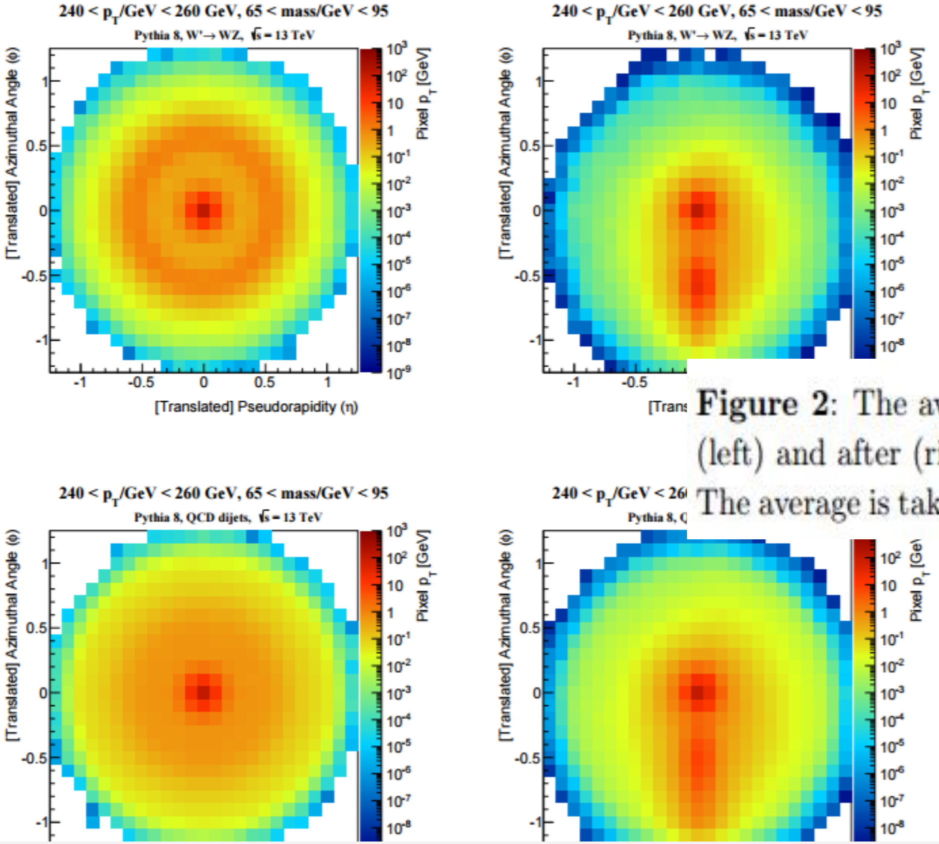


Figure 2: The average jet image for signal W jets (top) and background QCD jets (bottom) before (left) and after (right) applying the rotation, re-pixelation, and inversion steps of the pre-processing. The average is taken over images of jets with 240 GeV $< p_T <$ 260 GeV and 65 GeV $<$ mass $<$ 95 GeV.

From: L. de Oliveira et al., Jet-Images -- Deep Learning Edition, JHEP07, 2016

From: L. de Oliveira et al., Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis, 2017

Jetson TX1 Module

January 2016 availability

	JETSON TX1 MODULE
GPU	1 TFLOP/s 256-core Maxwell
CPU	64-bit ARM A57 CPUs
Memory	4 GB LPDDR4 | 25.6 GB/s
Video decode	4K 60Hz
Video encode	4K 30Hz
CSI	Up to 6 cameras | 1400 Mpix/s
Display	2x DSI, 1x eDP 1.4, 1x DP 1.2/HDMI
Wifi	802.11 2x2 ac
Networking	1 Gigabit Ethernet
PCIE	Gen 2 1x1 + 1x4
Storage	16 GB eMMC, SDIO, SATA
Other	3x UART, 3x SPI, 4x I2C, 4x I2S, GPIOs

JETSON EMBEDDED PLATFORM

A comprehensive platform for development and deployment of advanced embedded products

Jetson SDK

L4T kernel
Bootloader
Reference FS

Platform drivers

Library support

cuDNN
CUDA libs
VisionWorks

OpenVX
OpenCV, ROS
OpenGL

Developer tools

Nsight
NVTX
Jetpack installer

Debuggers
Profilers

Developer portal

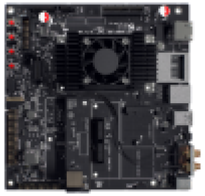
Downloads
Training
Forum

White papers
Success stories
Partner pages

Jetson TX1 module

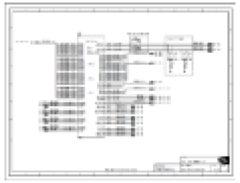
Unmatched performance
Production-ready
Long lifetime

Developer kit



Module
Breakout board
5MP camera

Design collateral



Ref. design files
Design guides
TRM

Reference diags
Factory users guide

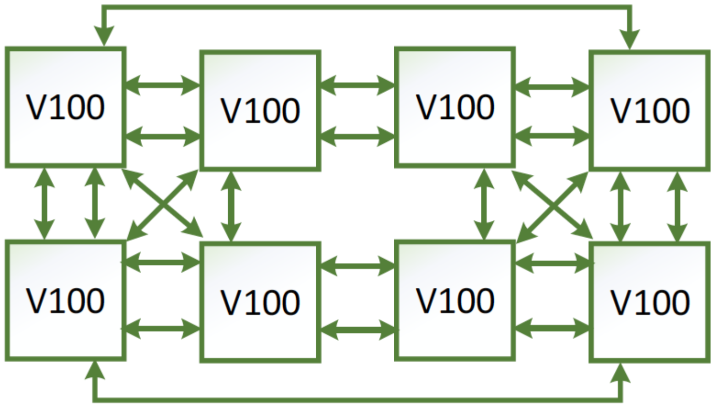
Ecosystem

Design partners
IHVs
ISVs

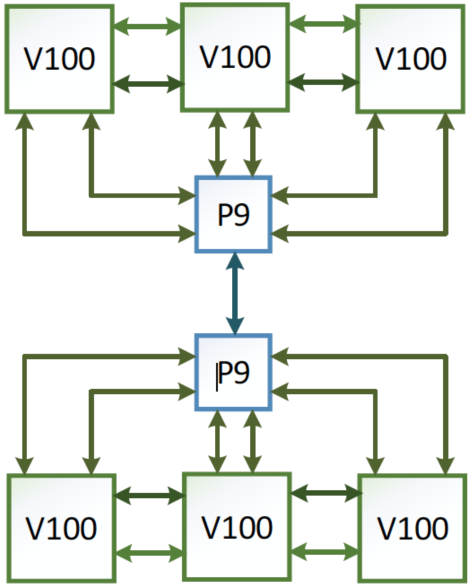
Higher Ed
Hacker/maker
OSS

VOLTA NVLINK

- 6 NVLINKS @ 50 GB/s bidirectional
- Reduce number of lanes for lightly loaded link (Power savings)
- Coherence features for NVLINK enabled CPUs



Hybrid cube mesh
(eg. DGX1V)



POWER9 based node

SUMMARY

- Getting data to GPU directly via RDMA-Direct
- Perform processing, even bit-twiddling, on the GPU. Can FP16 be used?
- GPUs are wide and have lots of memory. Run many things concurrently
- Compression via hardware video encoders.
- Advanced processing on the fly via Deep Learning methods
- Jetson for compute on the edge
- NVLInk for fast exchange of data between GPUs

NVIDIA