Nvidia GPU Architecture for AI

Georg Zitzlsberger georg.zitzlsberger@vsb.cz

12-11-2019

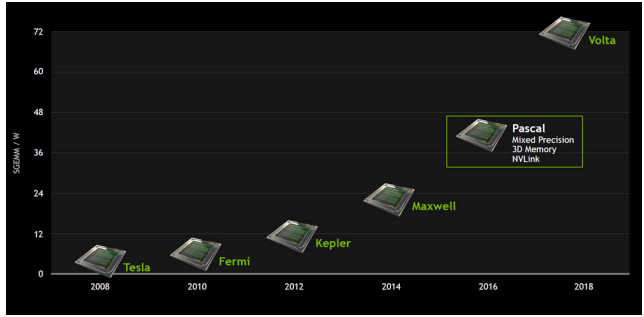
EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research, Development and Education

MINISTRY OF EDUCATION, YOUTH AND SPORTS

Agenda

- Nvidia Tesla Roadmap
- Nvidia Tesla Benchmarks
- Nvidia DGX
- NVLink
- Nvidia Collective Communications Library
- Deep Learning Containers
- CUDA for AI
- Questions

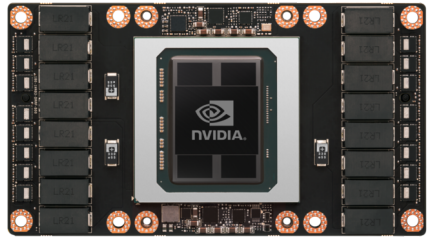
Nvidia Tesla Roadmap



(Image: Nvidia)

- **Kepler**: K10, K20(X), K40, K80
- **Maxwell**: M4, M6, M10, M40, M60
- **Pascal**: P4, P6, P40, P100
- **Volta**: V100
- **Turing**: T4

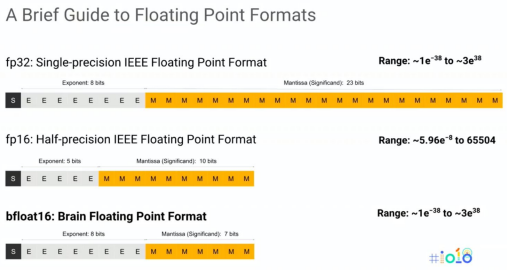
Nvidia Tesla P100 (Pascal)



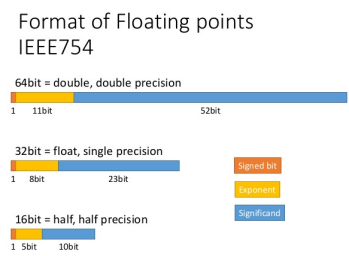
(Image: Nvidia)

- Peak 4.7 TFLOPS (double precision), 9.3 TFLOPS (single precision), 18.7 (half precision)
- 12/16 GB HBM2 (CoWoS)
- Peak memory bandwidth: 549 GB/s (12 GB), 732 GB/s (16 GB)
- 3584 CUDA cores
- NVLink v1 (SXM2) or PCIe x16 Gen3
- 250 Watts (PCIe), 300 (SXM2) Watts

Excursion: Floating Point Types



(Image: Google)

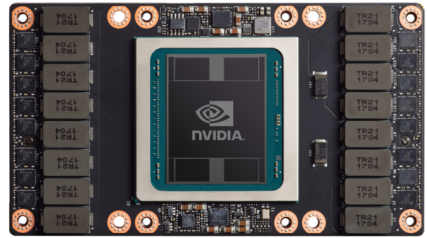


(Image: Nvidia)

- FP32, FP16 and BFloat16[1] are of interest for Deep Learning (DL)
- FP64 is not common for DL but other Machine Learning algorithms
- Also integer types can be used (e.g. INT4, INT8)

[1] Note that we use the term "half precision" for FP16, which does not include BFloat16

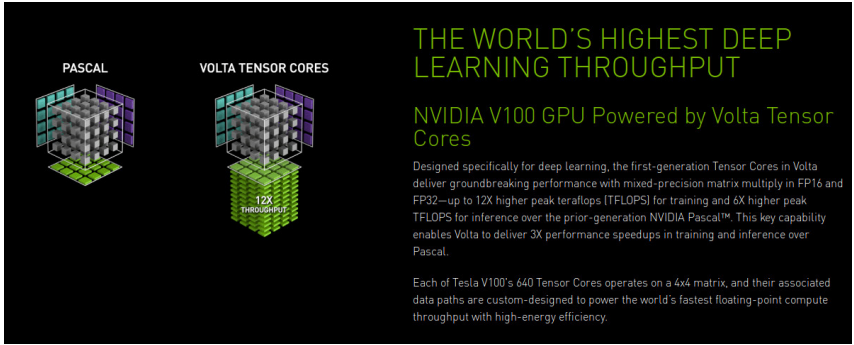
Nvidia Tesla V100 (Volta)

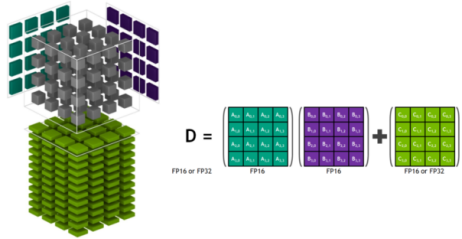


(Image: Nvidia)

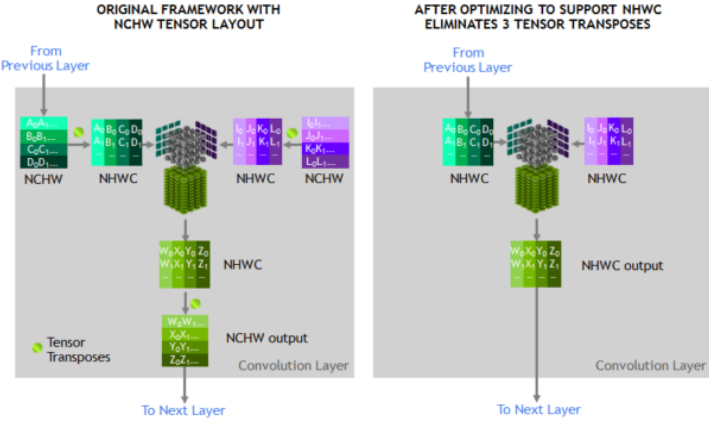
- Peak (SXM2) 7.8 TFLOPS (double precision), 15.7 TFLOPS (single precision), 125 (half precision)
- Up to 125 "TensorTFLOPS" for Deep Learning
- 16/32 GB HBM2 (CoWoS)
- Peak 900 GB/s memory bandwidth
- 5120 CUDA cores + 620 Tensor Cores
- NVLink v2 (SXM2) or PCIe x16 Gen3
- 250 (PCIe), 300 (SXM2) Watts

Nvidia Tesla V100 Tensor Cores





Nvidia Tesla V100 Tensor Cores



(Image: Nvidia)

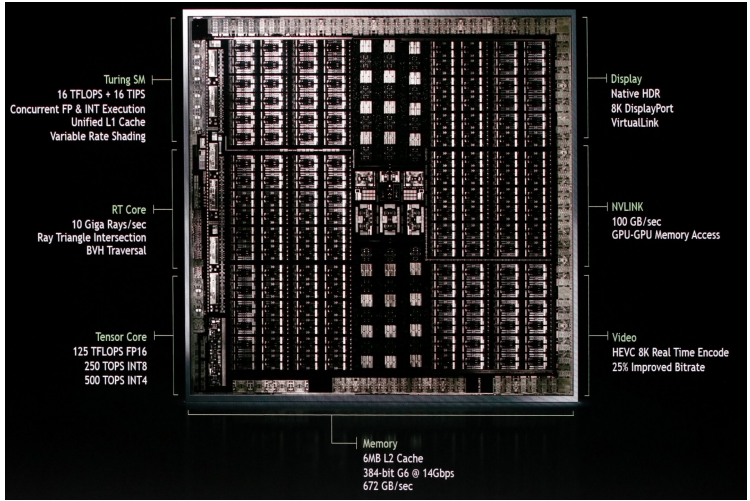
More information in the blog ▸ Volta Tensor Core GPU Achieves New AI Performance Milestones

Nvidia Tesla T4 (Turing)

(Image: Nvidia)

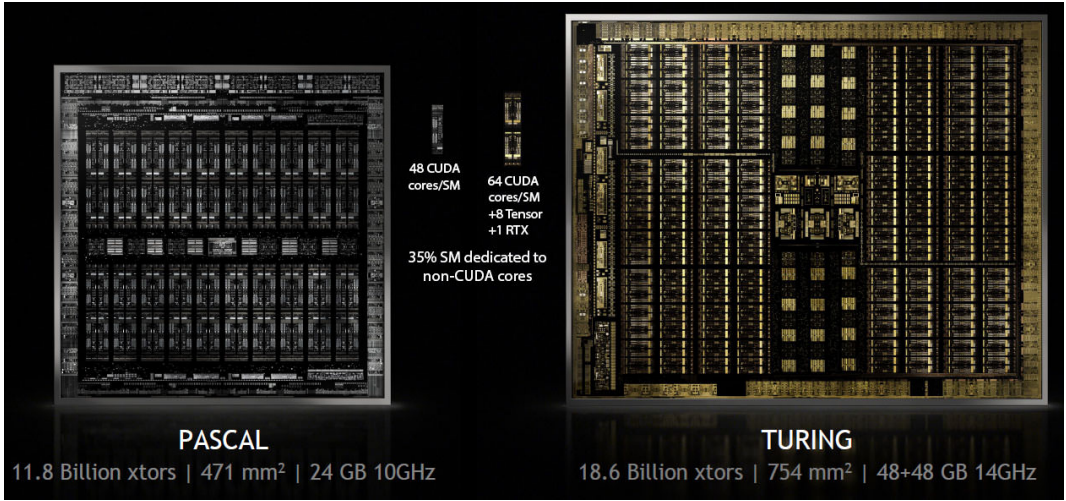
- Peak 8.1 TFLOPS (single precision)
- Up to 65 "TensorTFLOPS" for Deep Learning, 130/260 INT8/4 TOPS for inference
- 16 GB GDDR6
- Peak 300 GB/s memory bandwidth
- 2560 CUDA cores + 320 Tensor Cores
- PCIe x16 Gen3
- 70 Watts

Nvidia Turing Outlook



(Image: Nvidia)

Nvidia Turing Outlook - cont'd



(Image: Nvidia)

Nvidia Turing Outlook - cont'd

TURING TENSOR CORE

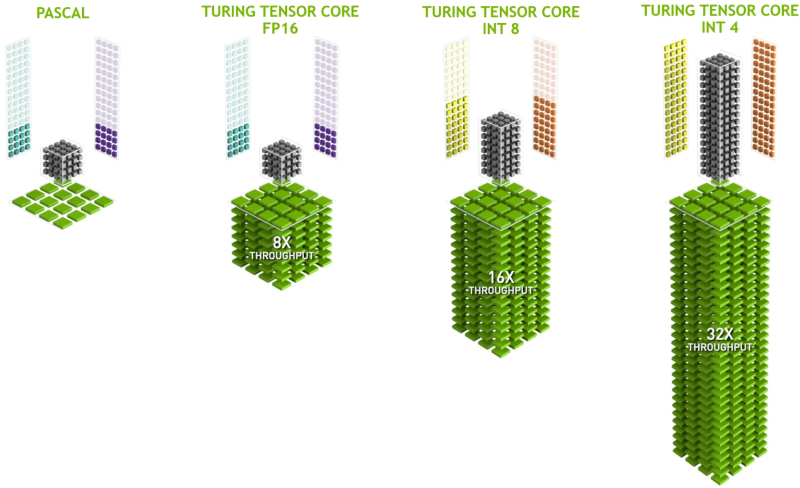
114 TFLOPS FP16

228 TOPS INT8

455 TOPS INT4

(Image: Nvidia)

Nvidia Turing Outlook - cont'd



(Image: Nvidia)

Nvidia Tesla Benchmarks

- **Single Precision (FP32):**

GPU	P100	T100	T4
Max. TFLOPS	9.32	14.02	8.07
FFT TFLOPS	1.51	2.30	0.66
GEMM TFLOPS	8.79	13.48	3.29
Theoretic Peak TFLOPS	9.3	15.7	8.1

- **Double Precision (FP64):**

GPU	P100	T100	T4
Max. TFLOPS	4.74	7.07	0.25
FFT TFLOPS	0.76	1.15	0.13
GEMM TFLOPS	4.26	5.92	0.25
Theoretic Peak TFLOPS	4.7	7.8	N/A

Results published at ▸ microway.com

Deep Learning Performance

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

(Image: Nvidia)

Heads-up:

- Real performance highly dependent on layers/operations used
- Uses half/mixed precision parameters that might or might not deliver the same/comparable model accuracy
- Assumes that the entire model and data fits into GPU memory

Nvidia DGX

(Image: Nvidia)

Nvidia DGX-1:

- 8x V100 GPUs (also exists with P100 GPUs)
- Deep Learning: 1,000 TFLOPS (peak)
- CUDA cores: 40,960
- Tensor cores: 5,120
- Host system: 2x 20 core Intel E5-2698v4, 256 GB memory
- Power consumption: 3,500 Watts

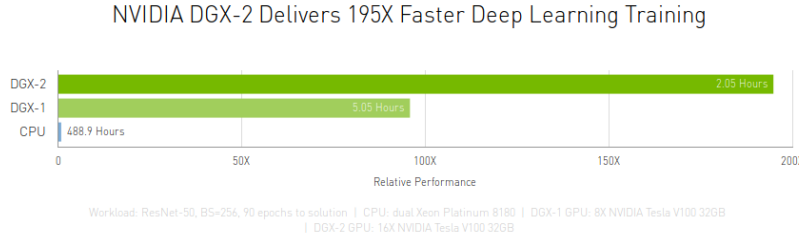
Nvidia DGX

(Image: Nvidia)

Nvidia DGX-2:

- 16x V100 GPUs
- Deep Learning: 2,000 TFLOPS (peak)
- CUDA cores: 81,920
- Tensor cores: 10,240
- Host system: 2x 24 core Intel Xeon Platinum 8168 Processor, 512 GB memory
- 12 NVSwitches (for NVLinks)
- Power consumption: 10,000 Watts

Nvidia DGX-1 vs. DGX-2

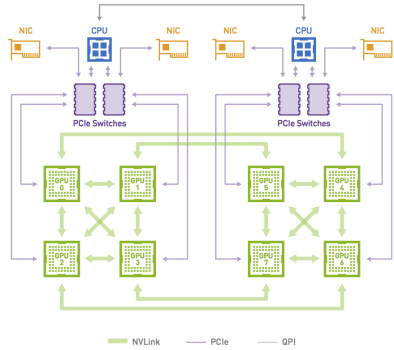


(Image: Nvidia)

More benchmark results can be found at the Nvidia Tesla Deep Learning Product Performance web page.

Also compare the official MLPerf results for training and inference here

NVLink

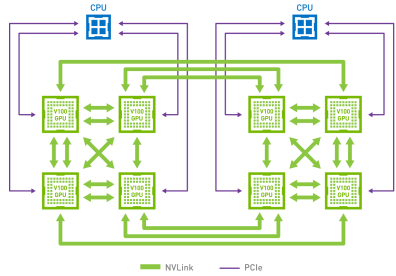


(Image: Nvidia)

NVLink (v1):

- Tesla P100
- Hybrid Cube Mesh
- 4 links per GPU, 20 GB/s per direction/per NVLink (160 GB/s aggregated)

NVLink

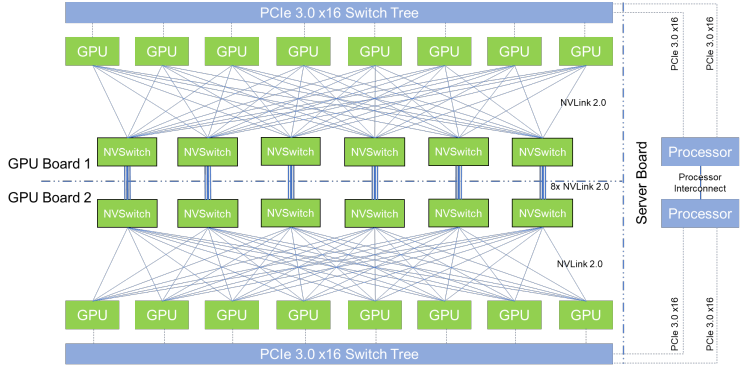


(Image: Nvidia)

NVLink 2.0:

- Tesla V100
- Hybrid Cube Mesh (also switch configuration possible)
- 6 links per GPU, 25 GB/s per direction/per NVLink2 (300 GB/s aggregated)

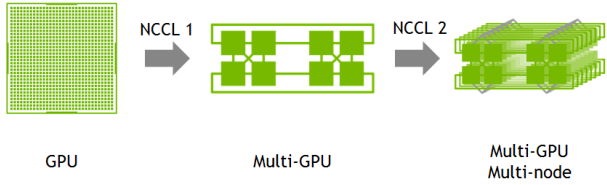
DGX-2 NVSwitch



(Image: TIRIAS Research)

- Uses (crossbar/XBAR) NVSwitch, enabling full GPU to GPU bandwidth
- One port per GPU (2 ports reserved)
- 2.4 TB/s bi-section bandwidth
- More information in ▸ Technical Overview

Nvidia Collective Communications Library

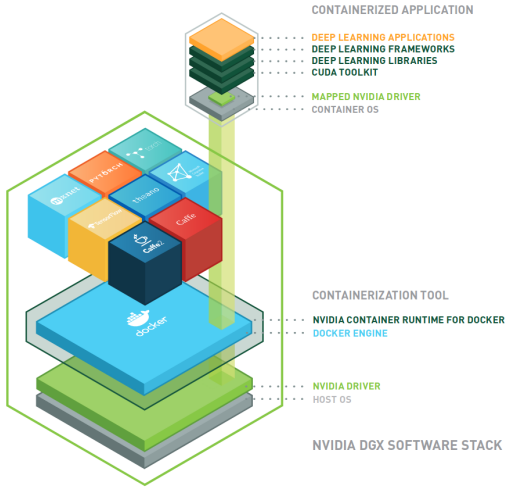


(Image: Nvidia)

Nvidia Collective Communications Library (NCCL):

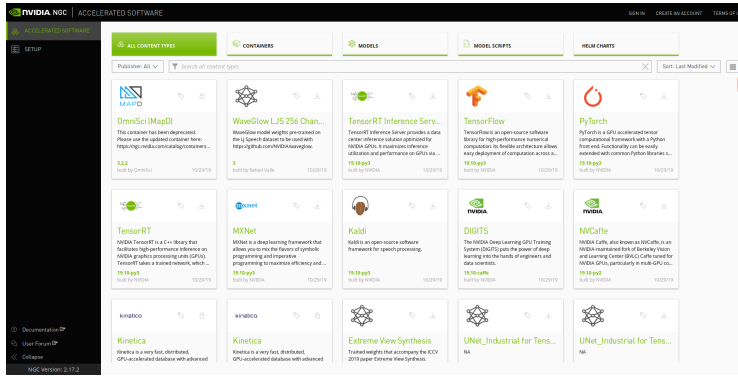
- Similarity to MPI collectives
- Pre NCCL 2.4: ring communication
- NCCL 2.4: double binary tree communication with improved latency
- Supported by major Deep Learning frameworks, e.g. Tensorflow, PyTorch
- Supported by distributed training framework Horovod

Deep Learing Containers



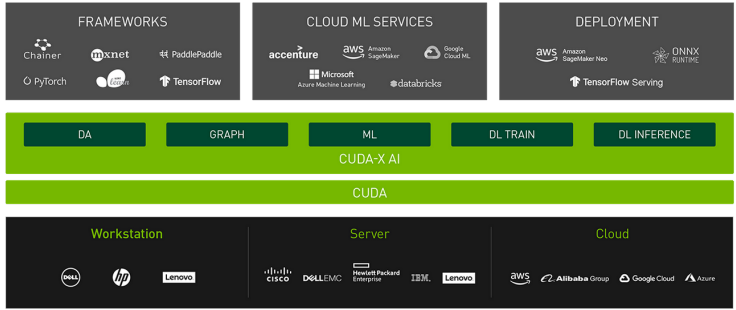
(Image: Nvidia)

Nvidia GPU Cloud (NGC)



- Ready made containers for different needs
- Optimized for Nvidia GPUs
- Available via the NGC Catalog

CUDA for AI

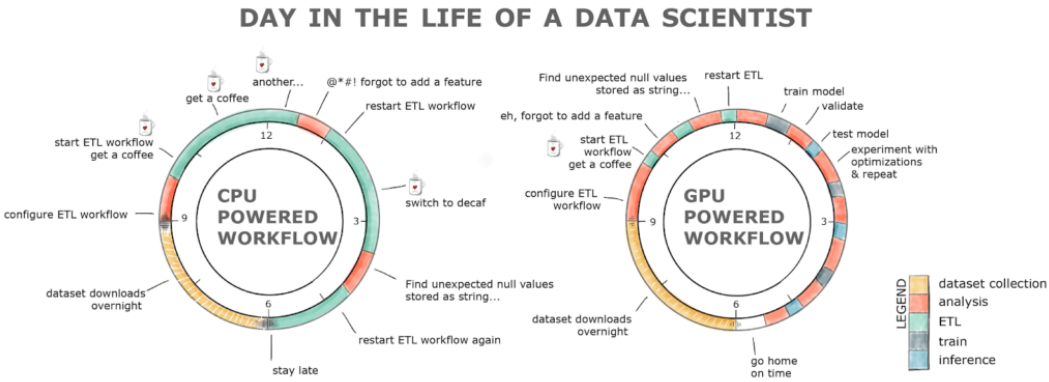


(Image: Nvidia)

CUDA-X AI:

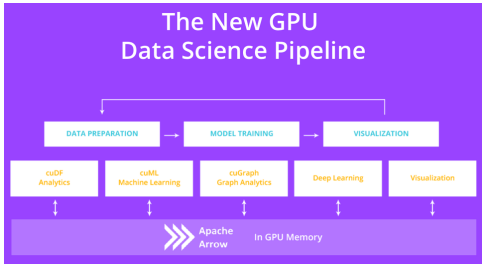
- cuDNN: Deep Learning primitives
- cuML: Machine Learning algorithms
- cuDF: Data Frame library (similar to Pandas)
- cuGraph: Graph Analytics
- Nvidia TensorRT: Optimized inference
- and others...

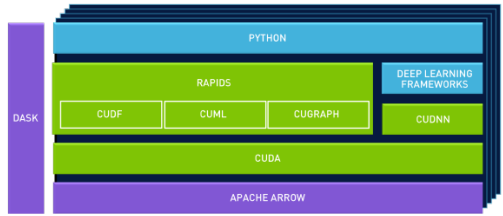
For Data Scientists. . .



(Image: Nvidia)

Nvidia RAPIDS

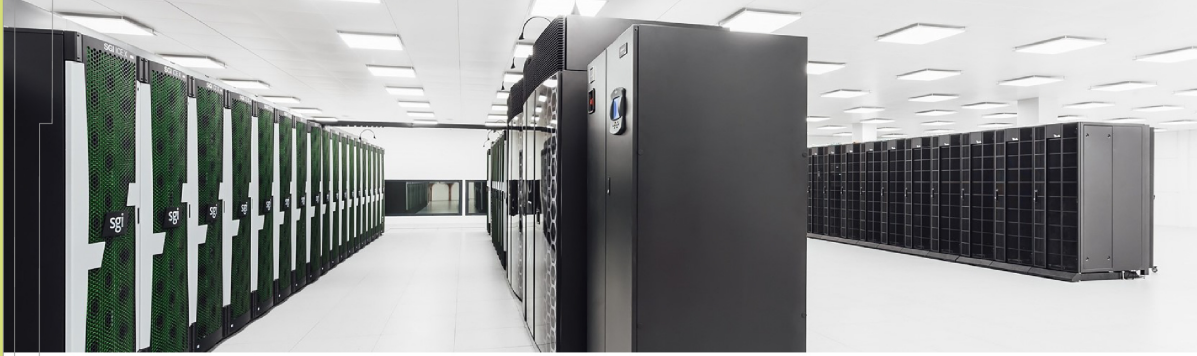




(Images: Nvidia)

- Apache Arrow based Data Processing Pipeline
- Open source on Github: ▸ RAPIDS projects

Q&A



VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

IT4Innovations National Supercomputing Center

VŠB – Technical University of Ostrava
Studentská 6231/1B
708 00 Ostrava-Poruba, Czech Republic
www.it4i.cz

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION, YOUTH AND SPORTS