

# Představení NVIDIA Ampere

Kamila Jeřábková, Petr Plodík – M Computers



# NVIDIA webináře



- 20. května 2020, 15:00: [Úvod do akcelerovaných výpočtů](#)
- 21. května 2020, 9:00: [Novinky NVIDIA, architektura Ampere](#)
- 27. května 2020, 15:00: [Novinky NVIDIA, architektura Ampere](#) (stejný obsah jako 21. května)
- 3. června 2020, 15:00: [NVIDIA GPU Cloud \(NGC\) prakticky](#)
- 9. června 2020, 15:00: [Do nitra NVIDIA A100 a DGX A100](#)

Tyto webináře jsou zdarma, stačí se jen zaregistrovat výše uvedených odkazech nebo na stránce <https://www.mcomputers.cz/webinare>,“ doplňuje.



# Nový webinář: Do nitra NVIDIA A100 a NVIDIA DGX A100

Termín: 9. června 2020 15:00

Nových technologií představených v NVIDIA Ampere architektuře (akcelerátor NVIDIA A100 a systém NVIDIA DGX A100) je tolik, že si zaslouží podrobnější technické představení:

- o hardwarová architektura NVIDIA A100

- o detailní popis nových Tensor jader, FP64, FP32, TF32, BF16, FP16, ....

- o akcelerace pomocí Fine-Grained Structured Sparsity

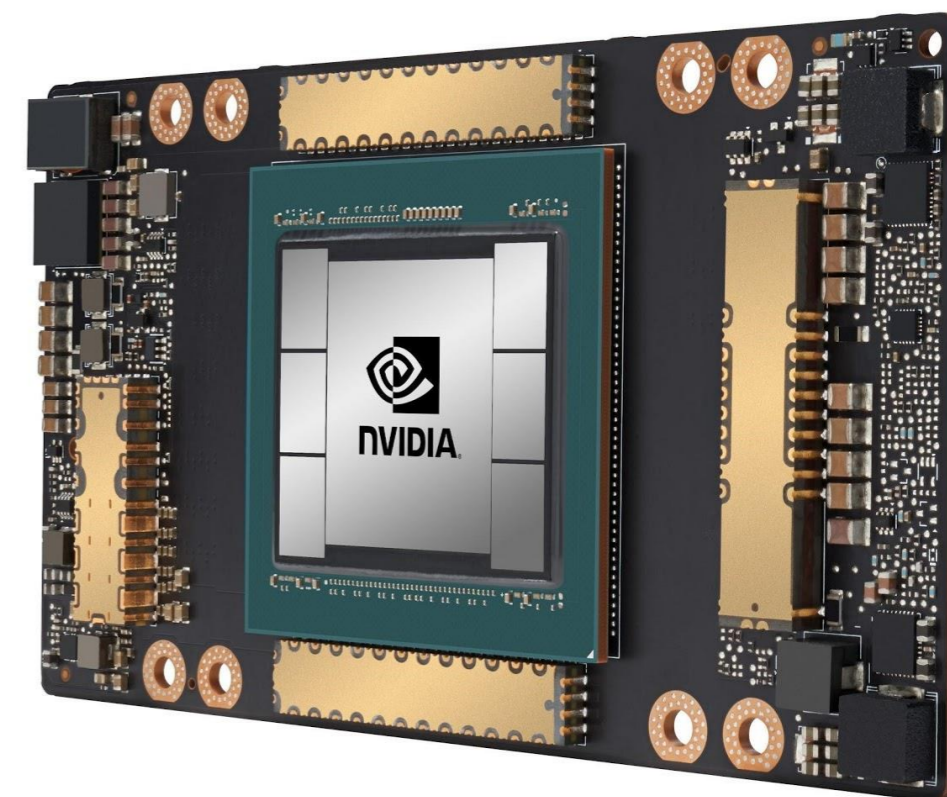
- o popis systému NVIDIA DGX A100

- o porovnání s předchozí generací NVIDIA karet

- o novinky v NVIDIA softwaru -- CUDA, knihovny

Registrace:

[https://us02web.zoom.us/webinar/register/9815899239796/WN\\_ZS34IzWDStSFwruy8loFBA](https://us02web.zoom.us/webinar/register/9815899239796/WN_ZS34IzWDStSFwruy8loFBA)



# NVIDIA ELITE PARTNER

Jsme jediný ELITE partner ve střední a východní Evropě  
Nabízíme slevy pro vysoké školy, výzkumné instituce a start-upy

Demo pool nejnovějších NVIDIA technologií:

NVIDIA DGX Station (Intel + 4x V100)

IBM AC922 (POWER9 + 4x V100)

NVIDIA Tesla akcelerátory (V100, T4)

NVIDIA Jetson produkty

Tým certifikovaných specialistů

Reference na vysokých školách, ve státní správě



NVIDIA DGX-2 installation at IT4Innovations  
288 views · Sep 26, 2019

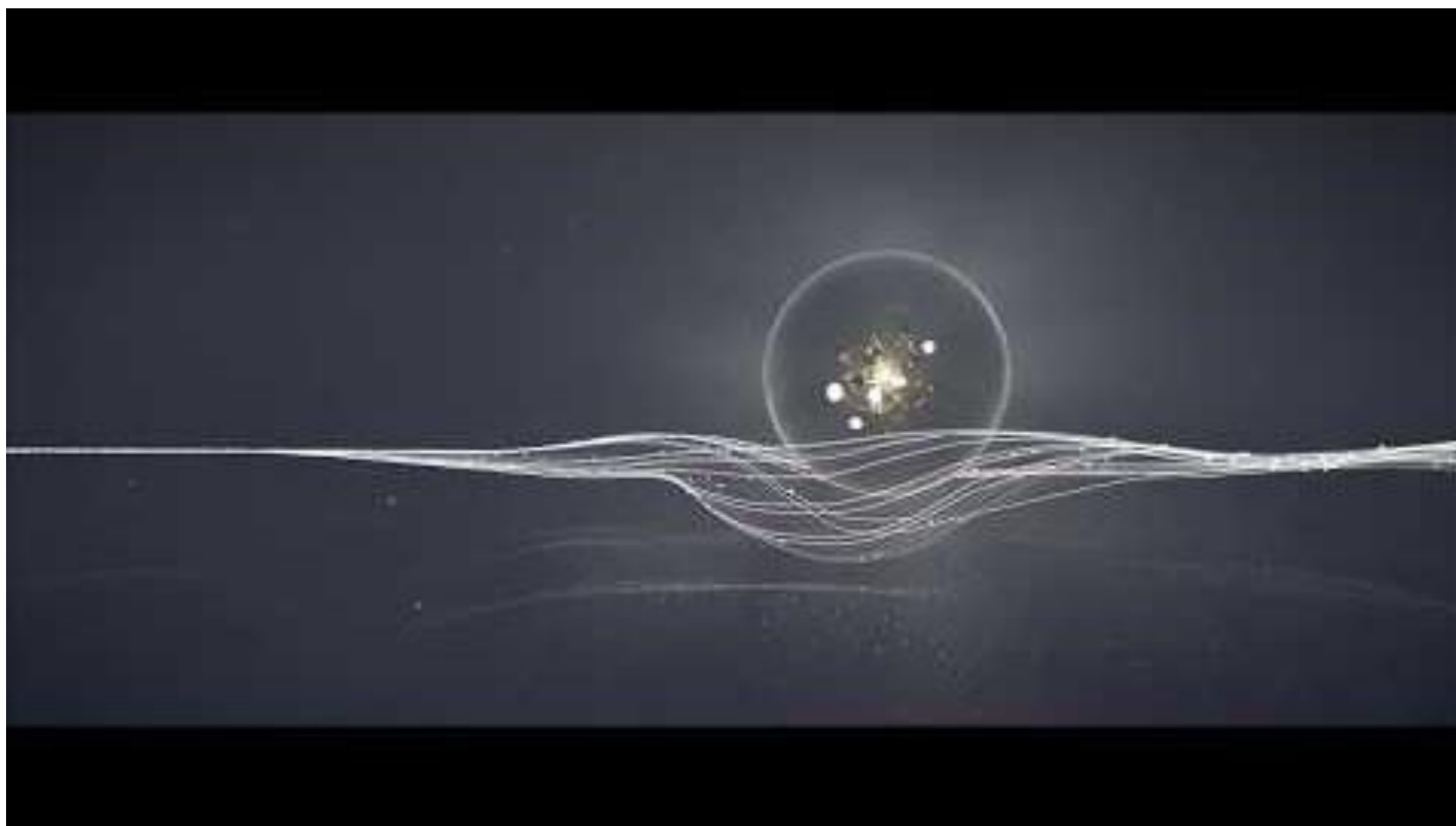
M Computers s.r.o.  
7 subscribers

**NOVINKY**  
ZE SVĚTA NVIDIA  
A SUPERPOČÍTAČŮ  
27. 11. 2019  
ČVUT CIAC  
společných partnerů 15000  
Praha 6, poslechárna 240

Ralph Hinsche, NVIDIA  
Marcin Gaczor, KINGSTON  
a další

REGISTRACE NA: AI DAYS.CZ

# I am AI, GTC 2020



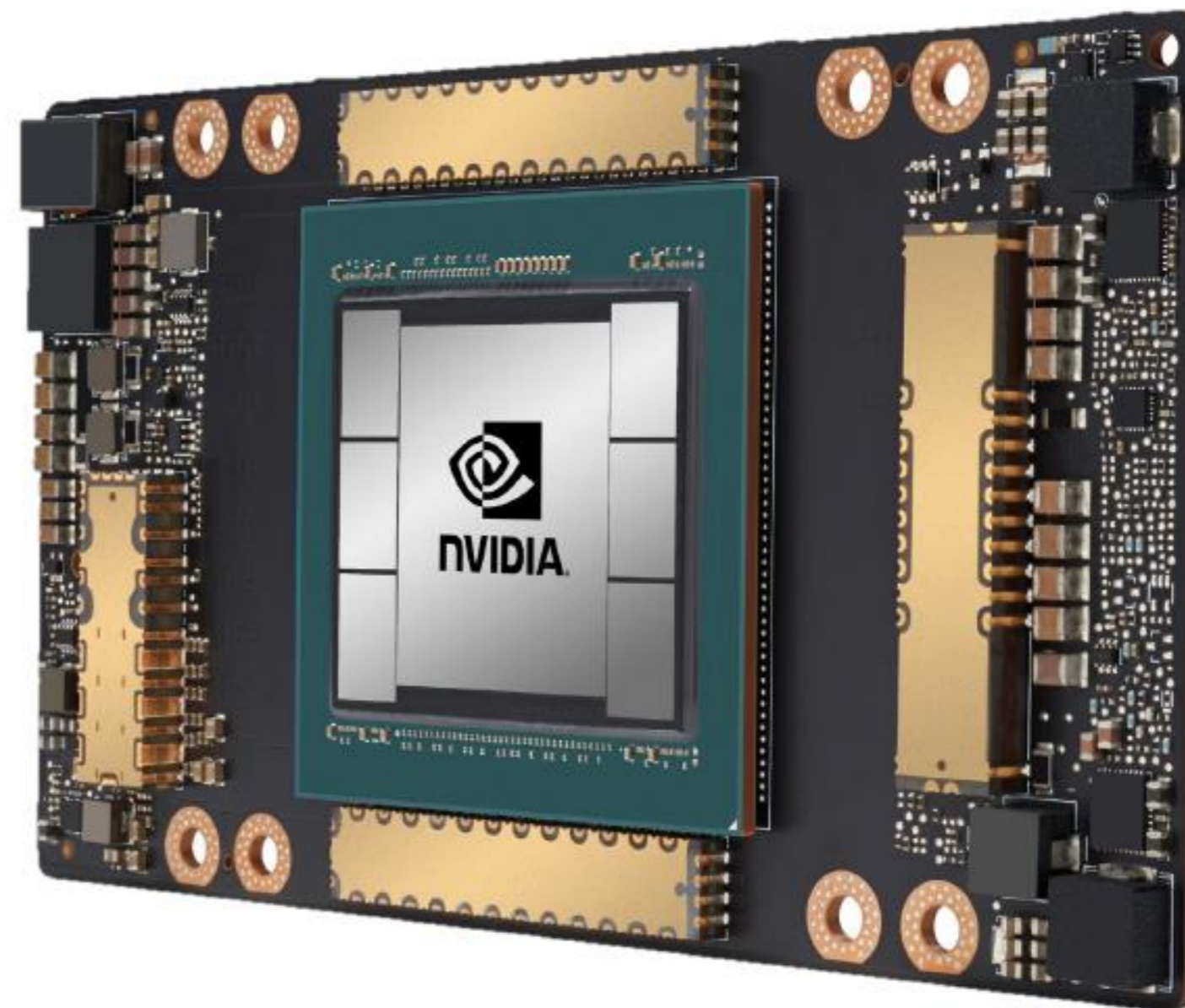
I am an Explorer a objevuji cizí světy,  
I am a helper a pomáhám potřebným,  
I am a healer a formuji budoucnost medicíny,  
I am a visioanary a hledám příběhy ve vzdálených galaxiích,  
I am a builder a pracuji v plně automatizovaných továrnách,  
I am a narrator a vyprávím vymyšlené příběhy,  
I am a composer a skládám hudbu,  
I am AI -- jsem umělá inteligence.

[https://youtu.be/e2\\_hsjpTi4w](https://youtu.be/e2_hsjpTi4w)

# NVIDIA A100

## NVIDIA A100 for NVLink

Peak FP64	9.7 TF
Peak FP64 Tensor Core	19.5 TF
Peak FP32	19.5 TF
Peak TF32 Tensor Core	156 TF   312 TF*
Peak BFLOAT16 Tensor Core	312 TF   624 TF*
Peak FP16 Tensor Core	312 TF   624 TF*
Peak INT8 Tensor Core	624 TOPS   1,248 TOPS*
Peak INT4 Tensor Core	1,248 TOPS   2,496 TOPS*
GPU Memory	40 GB
GPU Memory Bandwidth	1,555 GB/s
Interconnect	NVIDIA NVLink 600 GB/s PCIe Gen4 64 GB/s
Multi-instance GPUs	Various instance sizes with up to 7MIGs @5GB
Form Factor	4/8 SXM on NVIDIA HGX™ A100
Max TDP Power	400W



54B XTOR | 826mm<sup>2</sup> | TSMC 7N | 40GB Samsung HBM2 | 600 GB/s NVLink

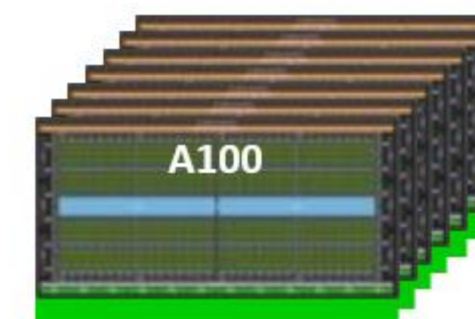
# A100 TENSOR-CORE GPU

54 billion transistors in 7nm



**Scale OUT**  
Multi-Instance GPU

7x



**Scale UP**  
2x BW

3<sup>rd</sup> gen.  
NVLINK

108 SMs  
6912 CUDA Cores

40MB L2  
6.7x capacity

1.56 TB/s HBM2  
1.7x bandwidth

# NVIDIA A100 SM (Streaming Multiprocessors)



**Third-generation Tensor Core**  
*Faster and more efficient*  
*Comprehensive data types*  
*Sparsity acceleration*

**Asynchronous data movement  
 and synchronization**

**Increased L1/SMEM capacity**



# Porovnání výkonnosti A100 vs. V100

	V100	A100	A100 Sparsity <sup>1</sup>	A100 Speedup	A100 Speedup with Sparsity
<b>A100 FP16 vs. V100 FP16</b>	31.4 TFLOPS	78 TFLOPS	N/A	2.5x	N/A
<b>A100 FP16 TC vs. V100 FP16 TC</b>	125 TFLOPS	312 TFLOPS	624 TFLOPS	2.5x	5x
<b>A100 BF16 TC vs. V100 FP16 TC</b>	125 TFLOPS	312 TFLOPS	624 TFLOPS	2.5x	5x
<b>A100 FP32 vs. V100 FP32</b>	15.7 TFLOPS	19.5 TFLOPS	N/A	1.25x	N/A
<b>A100 TF32 TC vs. V100 FP32</b>	15.7 TFLOPS	156 TFLOPS	312 TFLOPS	10x	20x
<b>A100 FP64 vs. V100 FP64</b>	7.8 TFLOPS	9.7 TFLOPS	N/A	1.25x	N/A
<b>A100 FP64 TC vs. V100 FP64</b>	7.8 TFLOPS	19.5 TFLOPS	N/A	2.5x	N/A
<b>A100 INT8 TC vs. V100 INT8</b>	62 TOPS	624 TOPS	1248 TOPS	10x	20x
<b>A100 INT4 TC</b>	N/A	1248 TOPS	2496 TOPS	N/A	N/A
<b>A100 Binary TC</b>	N/A	4992 TOPS	N/A	N/A	N/A

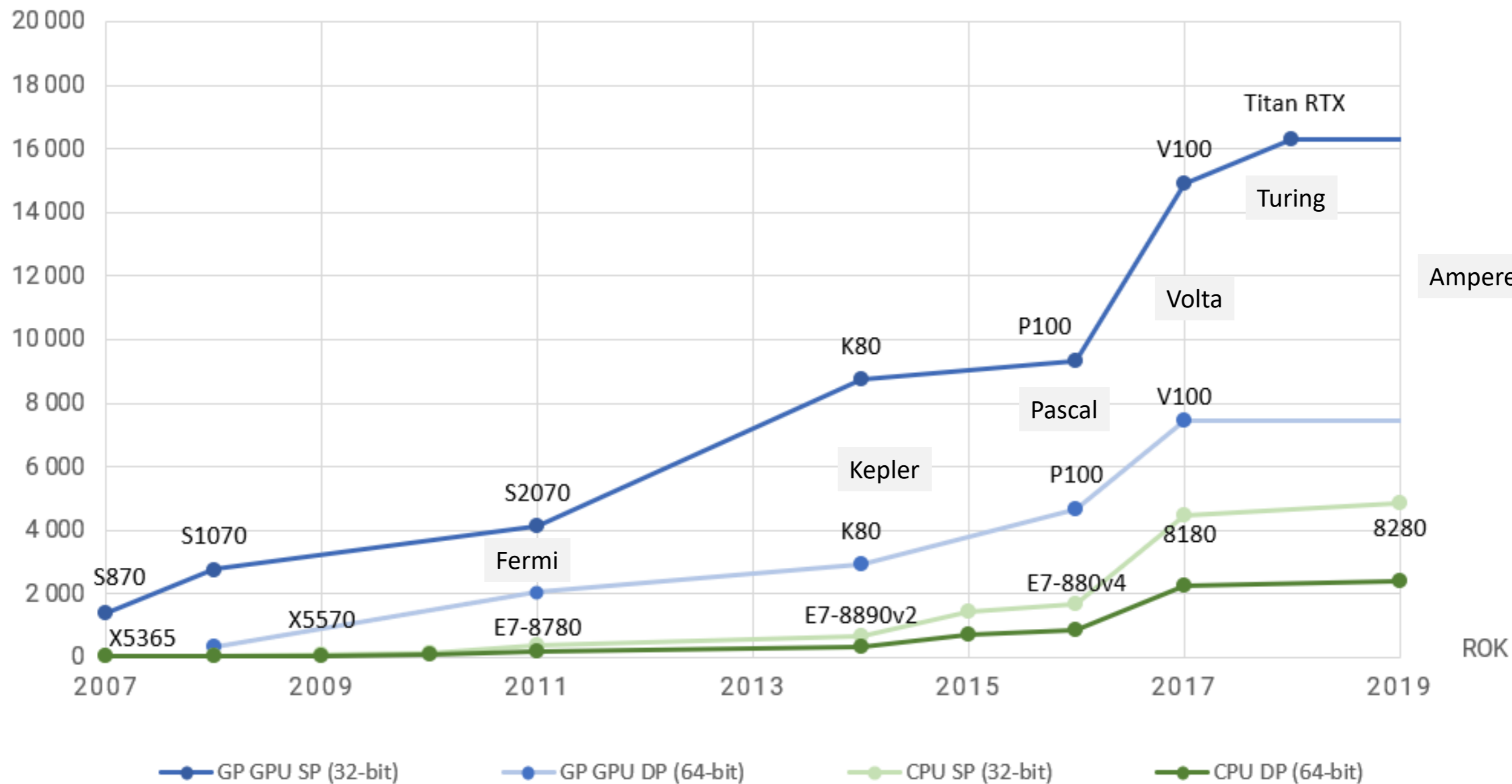
Table 2. A100 speedup over V100 (TC=Tensor Core, GPUs at respective clock speeds).

1) Effective TOPS / TFLOPS using the new Sparsity feature

# POROVNÁNÍ VÝKONNOSTI CPU (INTEL) A GP GPU (NVIDIA)

FP32 (SINGLE PRECISSION) A FP64 (DOUBLE) VÝPOČTY DEKADECKÁ OSA

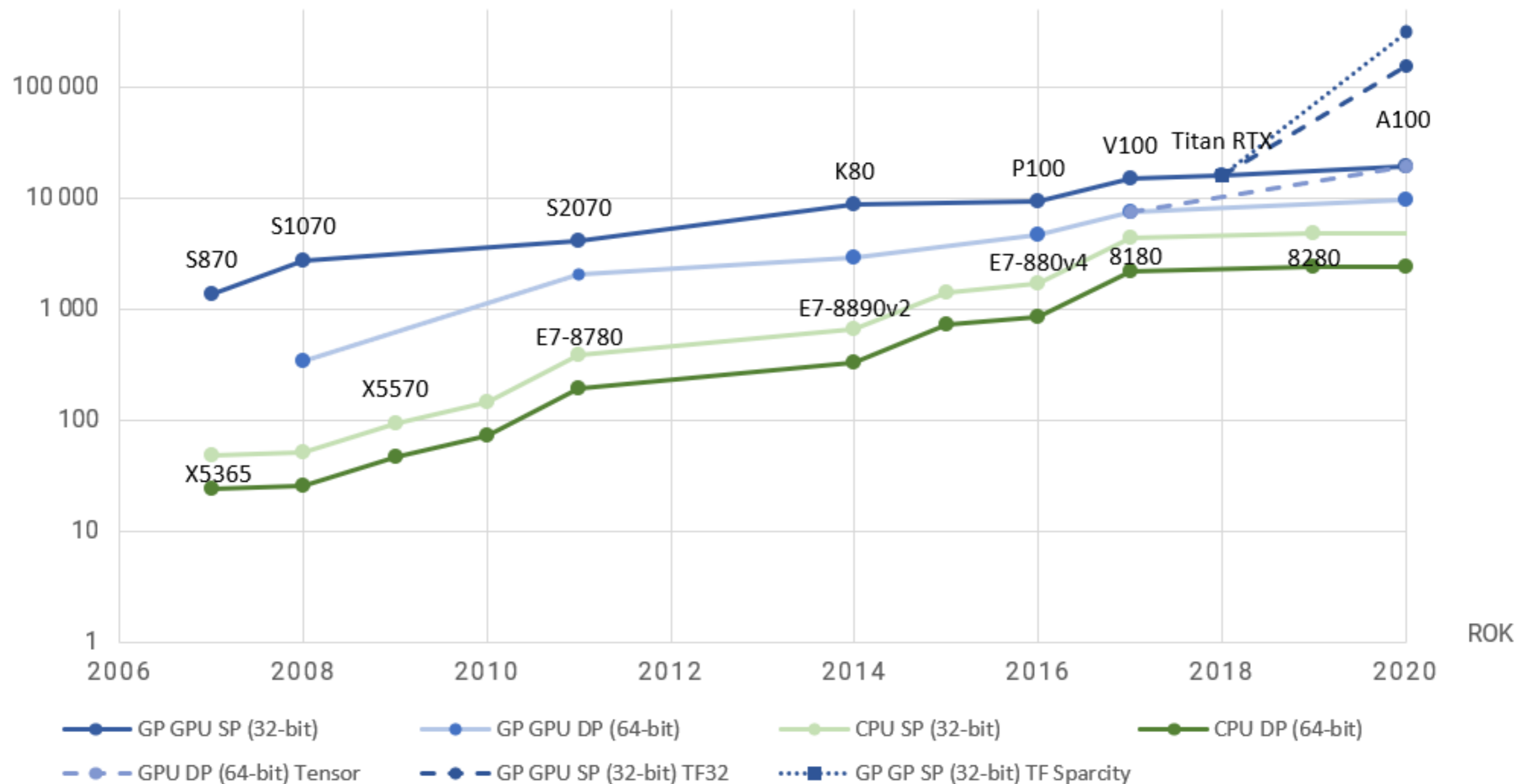
RPEAK / GFLOPS



# POROVNÁNÍ VÝKONNOSTI CPU (INTEL) A GP GPU (NVIDIA)

FP32 (SINGLE PRECISSION) A FP64 (DOUBLE) VÝPOČTY LOGARITMICKÁ OSA

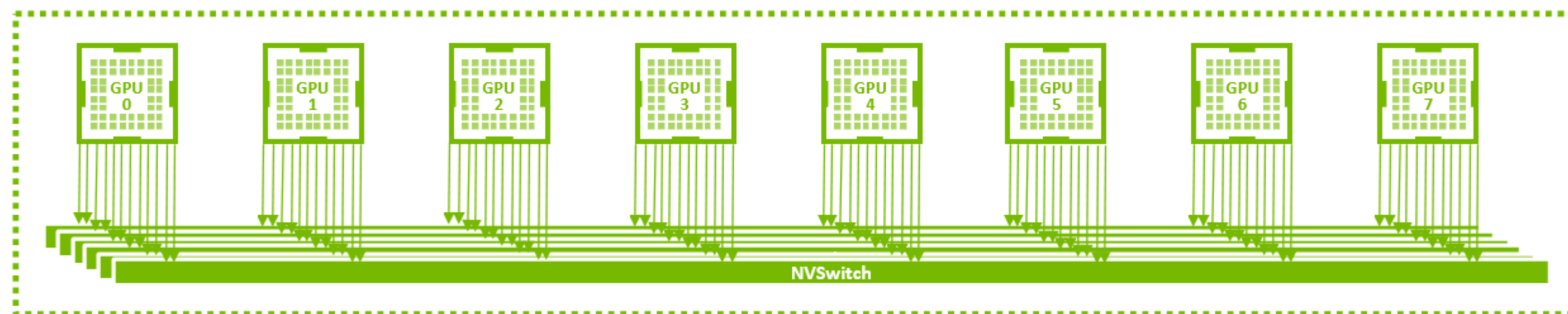
RPEAK / GFLOPS



ROK

# HGX A100: 3<sup>RD</sup> GEN NVLINK & SWITCH

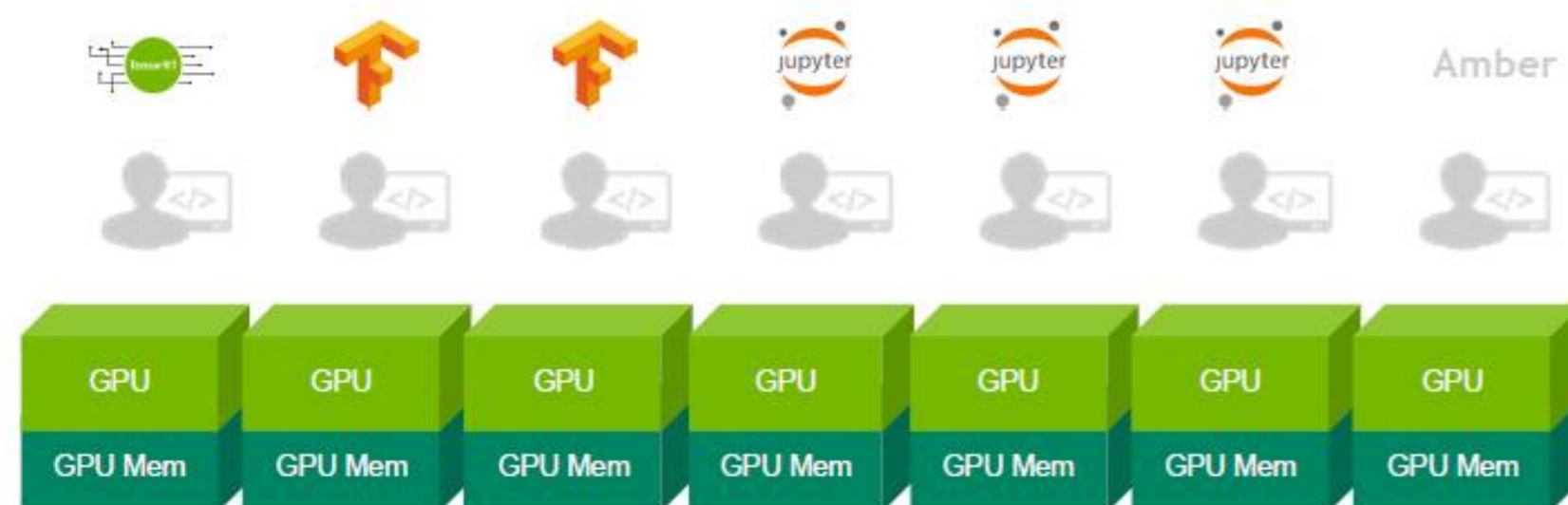
- ▶ **HGX A100 4-GPU:** fully-connected system with 100GB/s all-to-all BW
- ▶ **New NVSwitch:** 6B transistors in TSMC 7FF, 36 ports, 25GB/s each, per direction
- ▶ **HGX A100 8-GPU:** 6x NVSwitch in a fat tree topology, 2.4TB/s full-duplex bandwidth



Hardware consistency

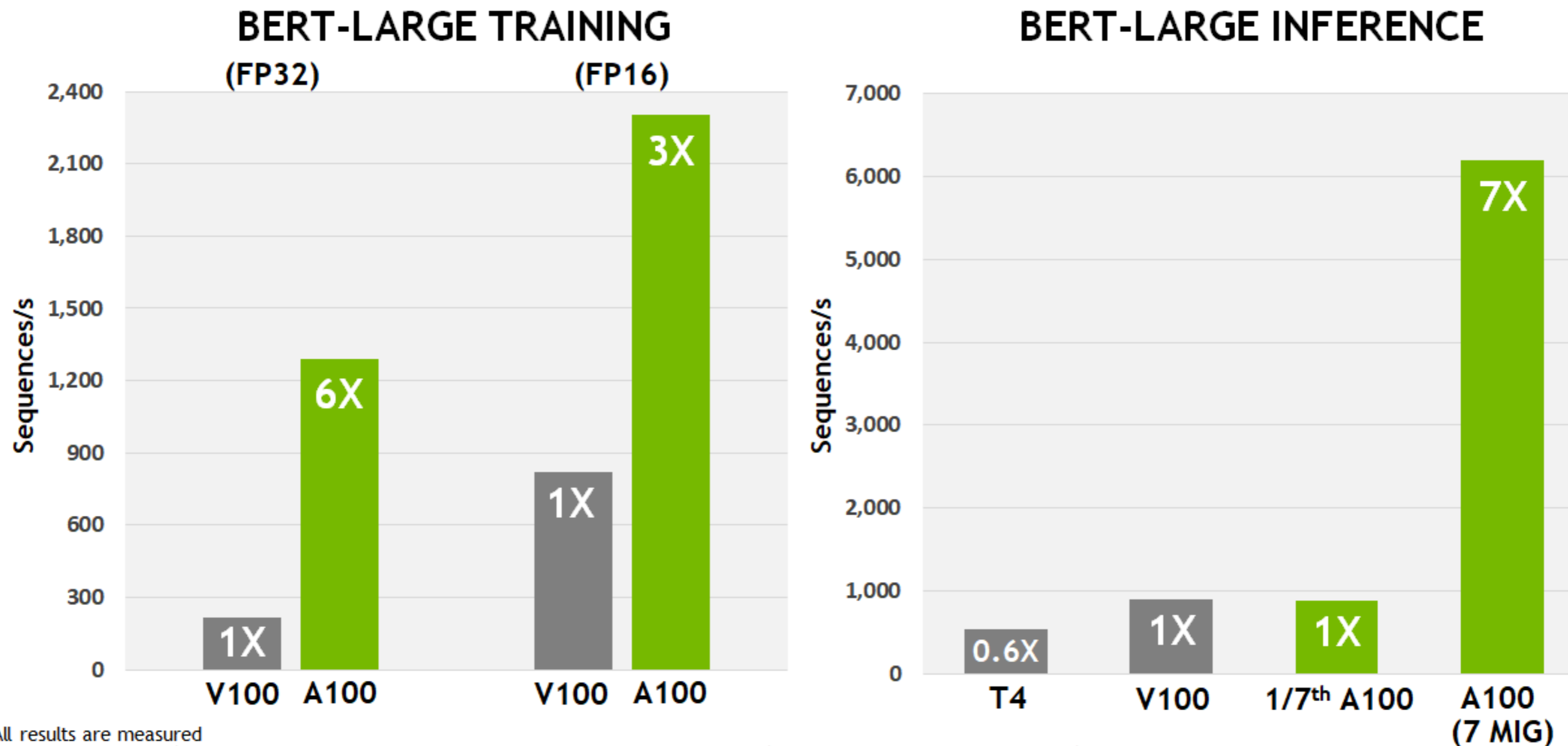
# MOST FLEXIBLE AI PLATFORM WITH MULTI-INSTANCE GPU (MIG)

Optimize GPU Utilization, Expand Access to More Users with Guaranteed Quality of Service



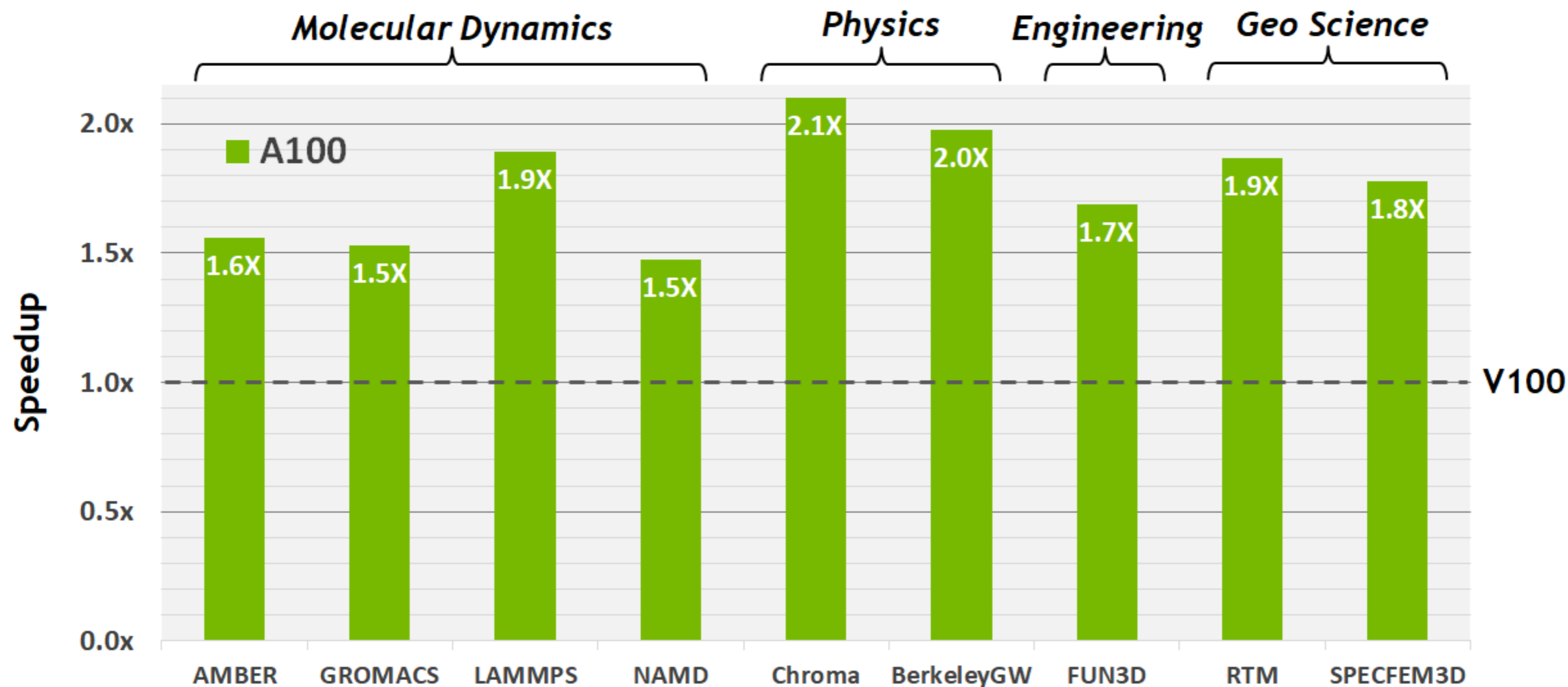
- Up To 7 GPU Instances In a Single A100
- Simultaneous Workload Execution With Guaranteed Quality Of Service
- All MIG instances run in parallel with predictable throughput & latency
- **Flexibility** to run any type of workload on a MIG instance
- **Right Sized GPU Allocation**  
Different sized MIG instances based on target workloads

# UNIFIED AI ACCELERATION



All results are measured  
 BERT Large Training (FP32 & FP16) measures Pre-Training phase, uses PyTorch including (2/3) Phase1 with Seq Len 128 and (1/3) Phase 2 with Seq Len 512,  
 V100 is DGX1 Server with 8xV100, A100 is DGX A100 Server with 8xA100, A100 uses TF32 Tensor Core for FP32 training  
 BERT Large Inference uses TRT 7.1 for T4/V100, with INT8/FP16 at batch size 256. Pre-production TRT for A100, uses batch size 94 and INT8 with sparsity

# ACCELERATING HPC



All results are measured

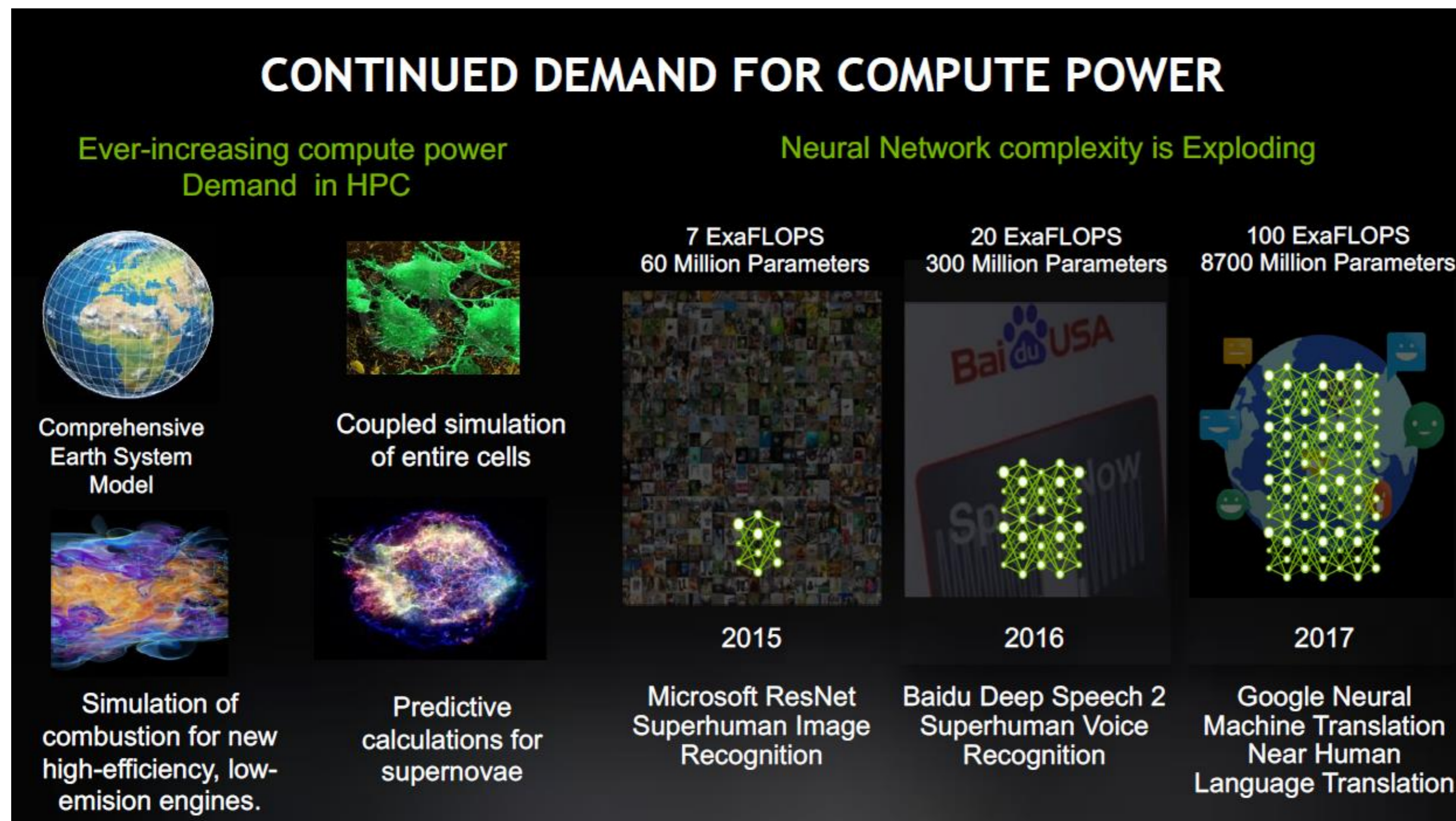
Except BerkeleyGW, V100 used is single V100 SXM2. A100 used is single A100 SXM4

More apps detail: AMBER based on PME-Cellulose, GROMACS with STMV (h-bond), LAMMPS with Atomic Fluid LJ-2.5, NAMD with v3.0a1 STMV\_NVE

Chroma with szsc21\_24\_128, FUN3D with dpw, RTM with Isotropic Radius 4 1024<sup>3</sup>, SPECFEM3D with Cartesian four material model

BerkeleyGW based on Chi Sum and uses 8xV100 in DGX-1, vs 8xA100 in DGX A100

# Neutuchající poptávka po výpočetním výkonu



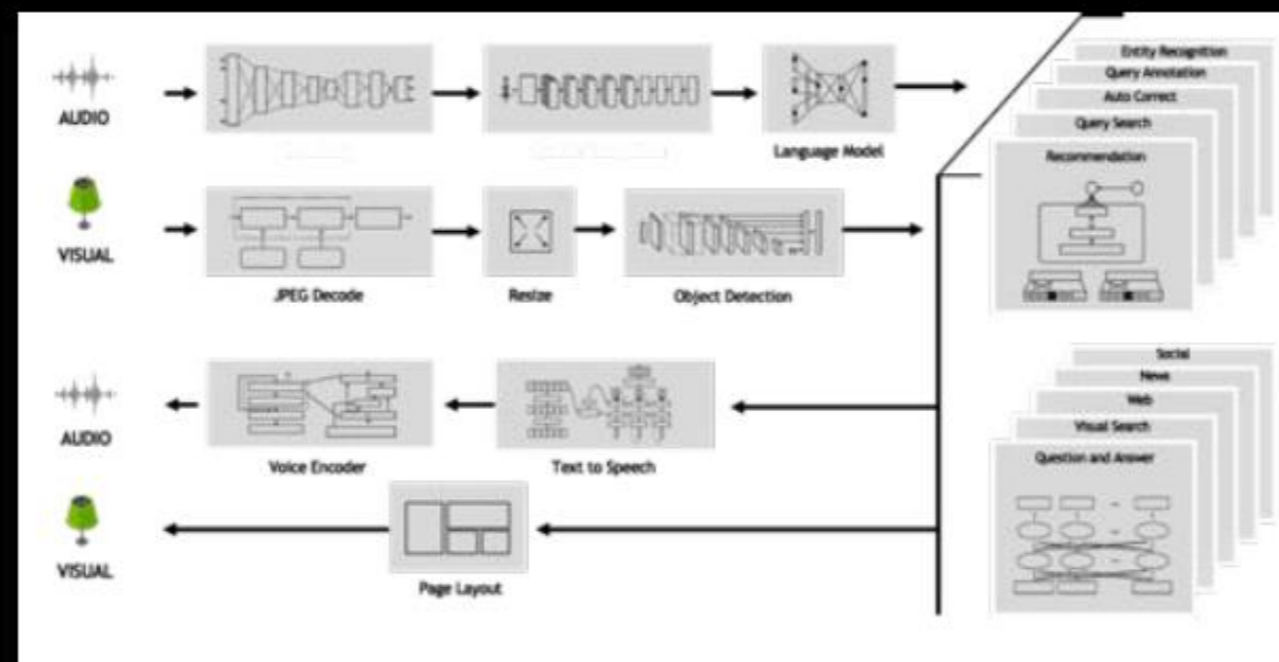


# Natural Language Processing (NLP)

Komplexní soubor AI kontejnerů, nízká latence

“What are the different types of lighting for a living room?”

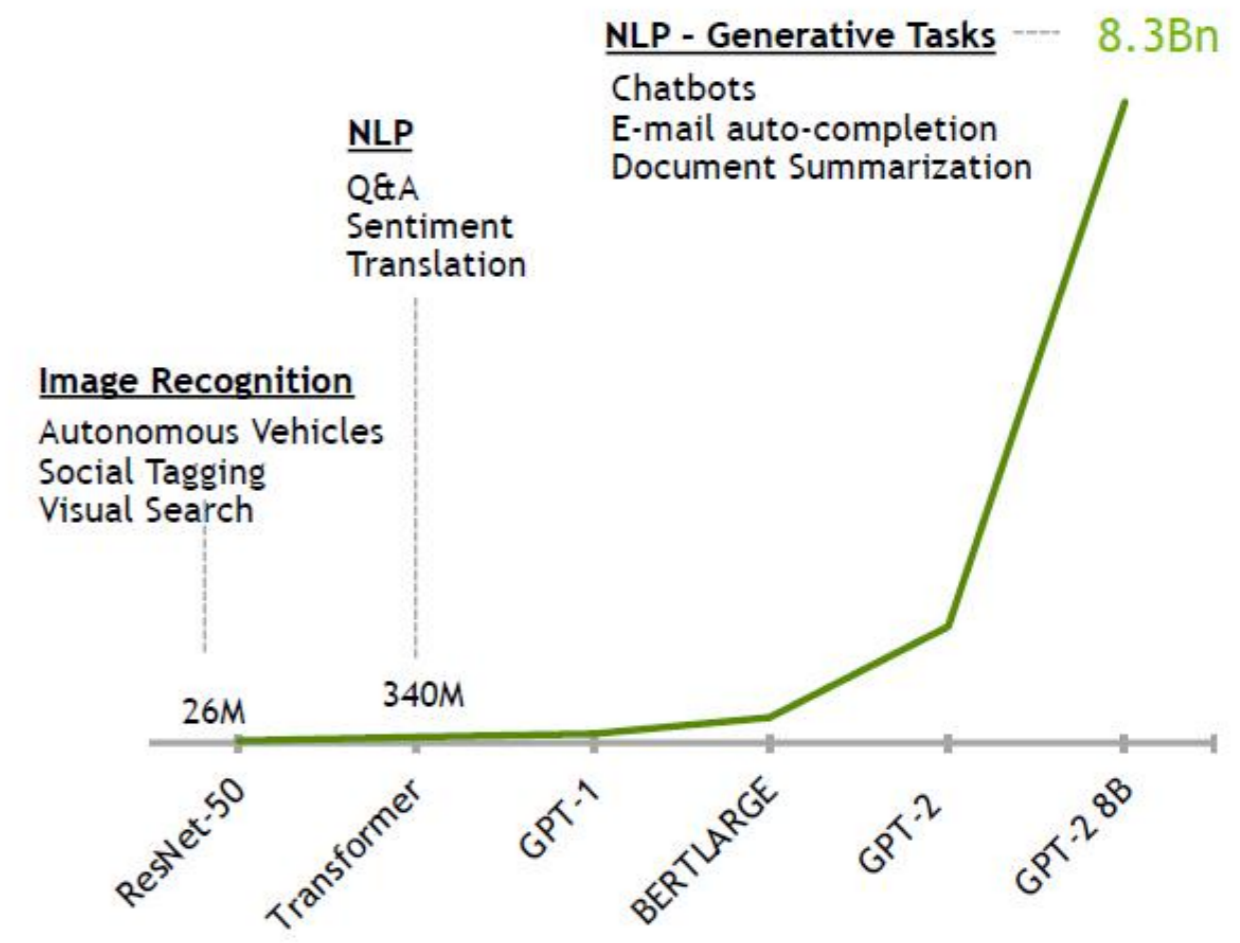
“Ambient, task and accent”



20-30 containers end-to-end  
RNN, CNN, MLP in INT8, FP16, FP32  
Latency <300ms

<https://github.com/NVIDIA/Megatron-LM>

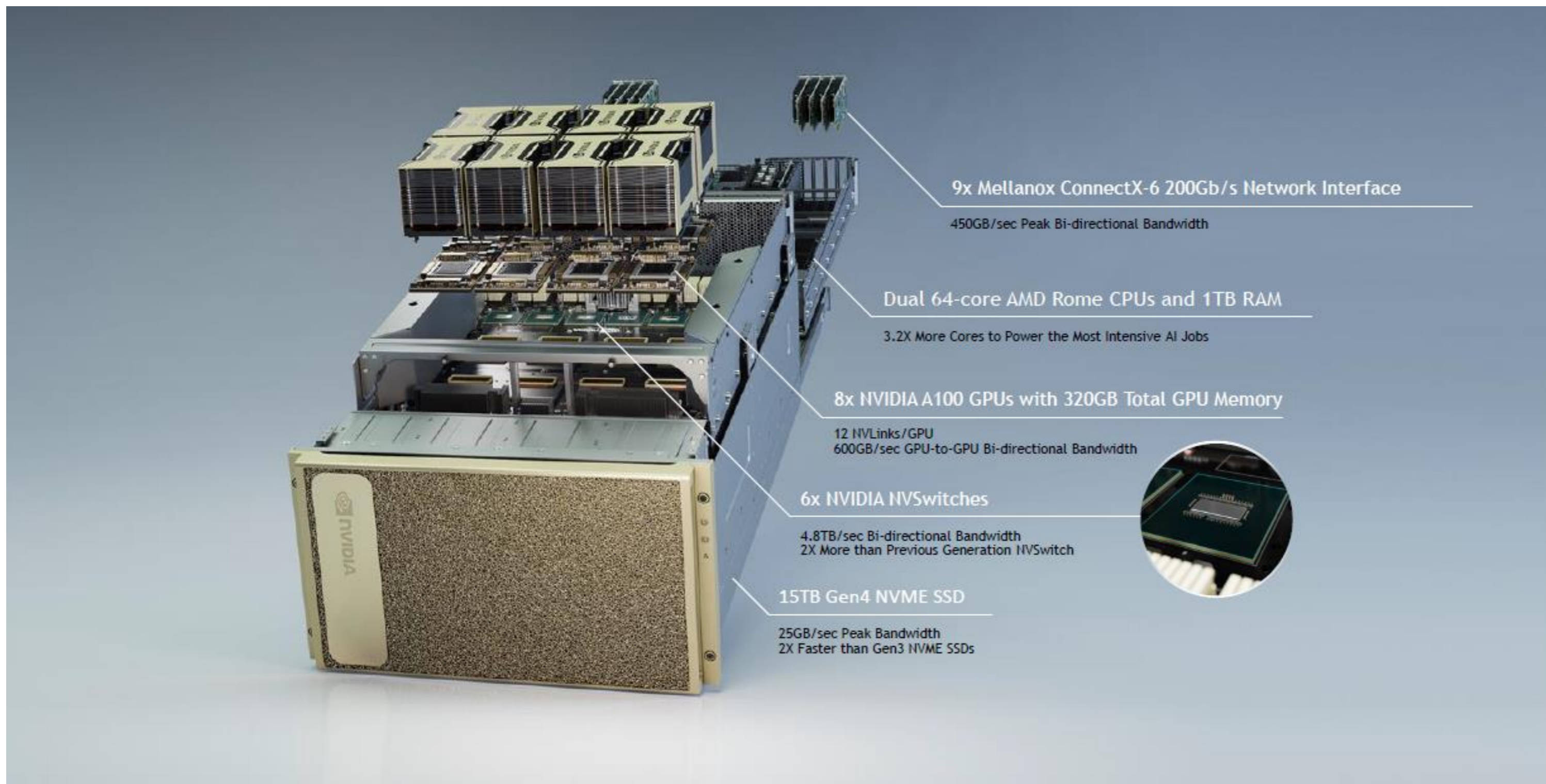
## Number of Parameters by Network



# Porovnání NVIDIA GPU akceleratorů

PARAMETRY	Novinka							
	GEFORCE RTX 2080TI	TITAN RTX	QUADRO RTX 5000	QUADRO RTX 6000 / 8000	TESLA T4	TESLA V100 / V100S PCIE	TESLA V100 SXM2	A100
Architektura	Turing	Turing	Turing	Turing	Turing	Volta	Volta	Ampere
# CUDA jader	4 352	4 608	3 072	4 608	2 560	5 120	5 120	6 912
# Tensor jader	544	576	384	576	320	640	640	432
FP64 (TFlops)	0,4	0,5	0,4	0,5	0,25	7 / 8,2	7,8	9,7 / 19,5
FP32 (TFlops)	13,4	16,3	11,2	16,3	8,1	14 / 16,4	15,7	19,5
Tensor (TFlops)	107,6	130	89,2	130	65	112 / 130	125	312
GPU paměť	11 GB	24 GB	16 GB	24 / 48 GB	16 GB	32 GB	32 GB	40 GB
Paměti	GDDR6	GDDR6	GDDR6	GDDR6	GDDR6	HBM2	HBM2	HBM2
Propustnost paměti	616 GB / s	672 GB / s	448 GB / s	624 GB / s	300 GB / s	900 GB/s / 1 134 GB/s	900 GB / s	1,5 TB/s
ECC paměti	není	není	ECC	ECC	ECC	ECC	ECC	ECC
Max. příkon	250 W	280 W	265 W	295 W	70 W	250 W	300 W	400 W
Provedení	PCIe gen3	PCIe gen3	PCIe gen3	PCIe gen3	PCIe gen3	PCIe gen3	SXM2	SXM4
Pro datacentra	Ne	Ne	Ano	Ano	Ano	Ano	Ano	Ano
Oznámení	2018	2018	2018	2018	2018	2017 / 2019	2017	2020
Listová cena	1 249 USD	2 499 USD	2 809 USD	5 949 / 8 269 USD	2 909 USD	11 839 USD	11 839 USD	součást systému

# NVIDIA DGX A100



9x Mellanox ConnectX-6 200Gb/s Network Interface

450GB/sec Peak Bi-directional Bandwidth

Dual 64-core AMD Rome CPUs and 1TB RAM

3.2X More Cores to Power the Most Intensive AI Jobs

8x NVIDIA A100 GPUs with 320GB Total GPU Memory

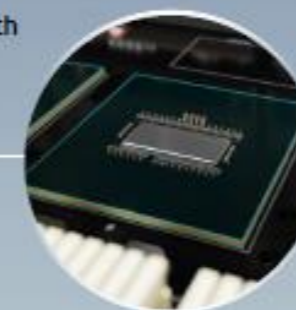
12 NVLinks/GPU  
600GB/sec GPU-to-GPU Bi-directional Bandwidth

6x NVIDIA NVSwitches

4.8TB/sec Bi-directional Bandwidth  
2X More than Previous Generation NVSwitch

15TB Gen4 NVME SSD

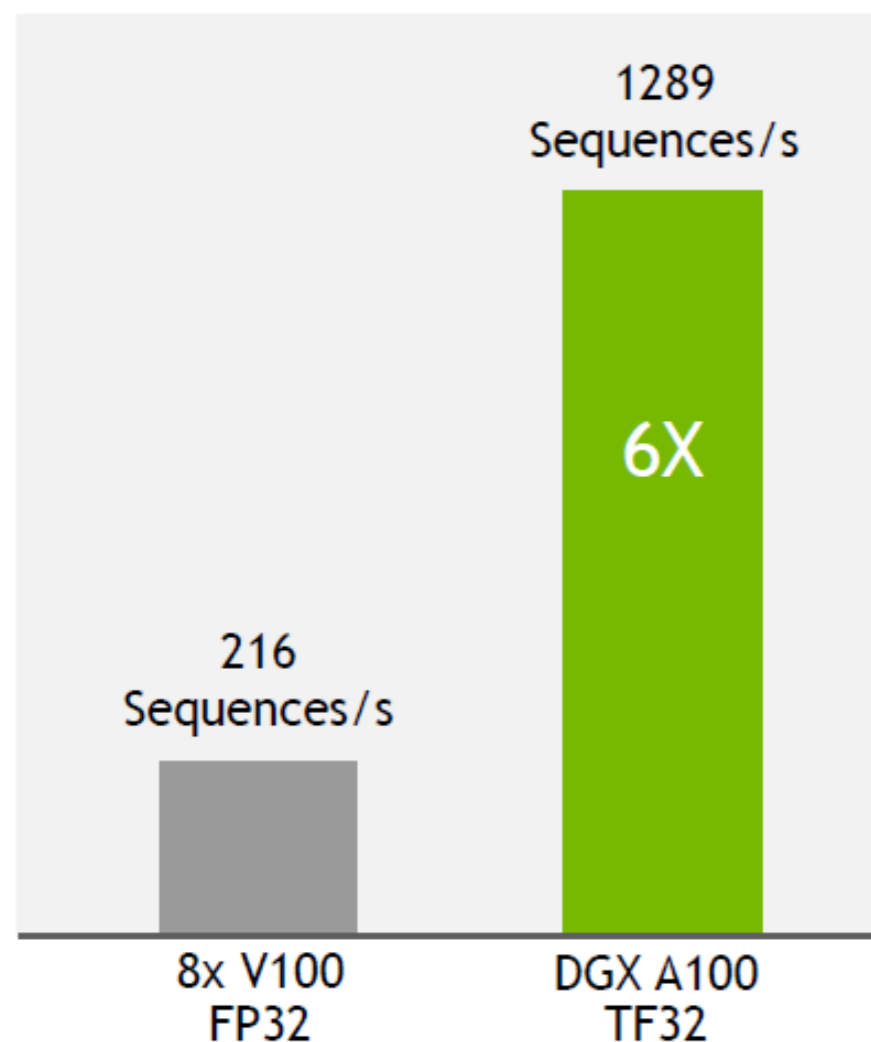
25GB/sec Peak Bandwidth  
2X Faster than Gen3 NVME SSDs



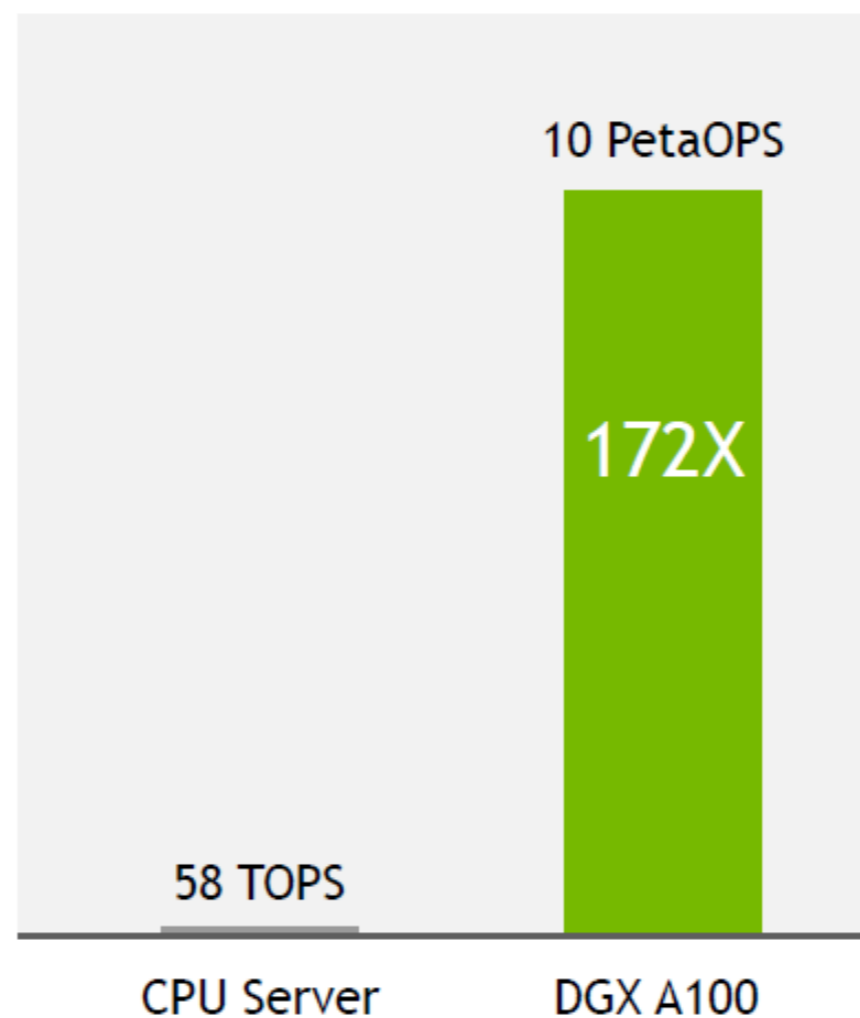
# NVIDIA DGX systémy

Parametr	DGX A100	DGX-2	DGX-1	DGX Station
<b>GPUs</b>	8× NVIDIA A100 40GB	16× NVIDIA Tesla V100 32GB	8× NVIDIA Tesla V100 32GB	4× NVIDIA Tesla V100 32GB
<b>Výkon (tensor operace)</b>	5 PetaFLOPS	2 PetaFLOPS	1 PetaFLOPS	0,5 PetaFLOPS
<b>GPU paměť</b>	320 GB celkem	512 GB celkem	256 GB celkem	128 GB celkem
<b>CPU</b>	2× AMD 7742, 2.7 GHz 64c	2× Intel 8168, 2.7 GHz 24c	2× Intel E5-2698 v4 2.2GHz 20c	Intel E5-2698 v4 2.2GHz 20c
<b>NVIDIA CUDA cores</b>	55 296	81 920	40 960	20 480
<b>NVIDIA Tensor cores</b>	3 456 (64-bit)	10 240	5 120	2 560
<b>Propojení GPU karet</b>	NVSwitch, non-blocking 2,4TB/s	NVSwitch, non-blocking, 2,4TB/s	NVLink, hypercube topologie	NVLink
<b>RAM</b>	1 TB	1,5 TB	512 GB	256 GB
<b>HDD</b>	2× 1.92TB NVME SSD 4× 3.84TB NVME SSD	2× 960GB NVME SSD 8× 3.84TB NVME SSD	4× 1,92TB SSD	4× 1,92TB SSD
<b>Network</b>	1× 200GbE dual-port Eth. 8× 200Gb HDR Infiniband	2× 10/25GbE, 8× 100Gb EDR Infiniband/Ethernet	2× 10GbE, 4× 100Gb EDR Infiniband/Ethernet	2× 10GbE
<b>Maximální příkon</b>	6,6 kW	10 kW	3 500 W	1 500 W
<b>Provedení</b>	rack, 6U	rack, 10U	rack, 3U	tower, vodní chlazení CPU a GPU

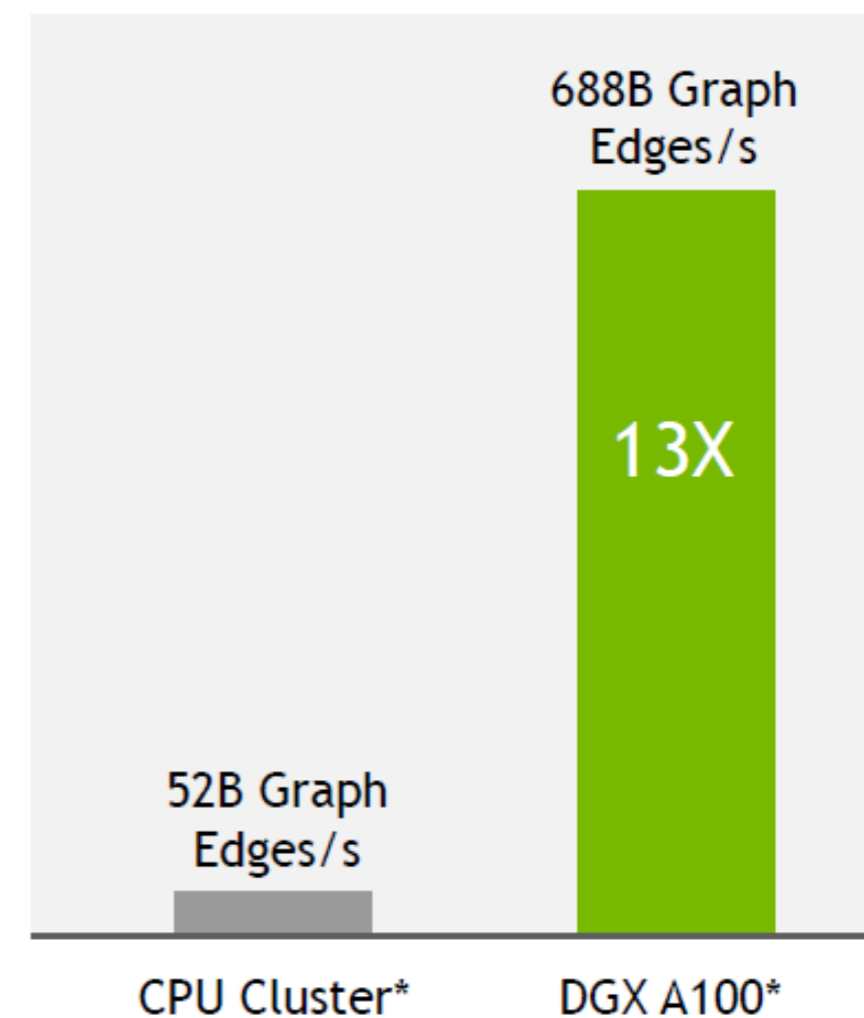
# DGX A100 PERFORMANCE



**Training**  
NLP: BERT-Large



**Inference**  
Peak Compute



**Analytics**  
PageRank

BERT Pre-Training Throughput using PyTorch including (2/3)Phase 1 and (1/3)Phase 2 | Phase 1 Seq Len = 128, Phase 2 Seq Len = 512  
 V100: DGX-1 Server with 8x V100 using FP32 precision  
 DGX A100: DGX A100 with 8x A100 using TF32 precision

CPU Server: 2x Intel Platinum 8280 using INT8  
 DGX A100: DGX A100 with 8x A100 using INT8 with Structural Sparsity

3000x CPU Servers vs. 4x DGX A100  
 Published Common Crawl Data Set:  
 128B Edges, 2.6TB Graph

# Porovnání cen

DGX A100 – 199 000 USD + support

DGX Station – 49 000 USD (původně 79 000 USD) + support

DGX-1 – 119 000 USD (původně 139 000 USD) + support

DGX-2 – 399 000 USD + support



# Dostupnost NVIDIA DGX A100: The Trucks Are Rolling In!

Filled With DGX A100s for YOU To Buy.... 😊



# Další novinky v AI SW z GTC 2020

## AI Software Announcements

- **NVIDIA Jarvis** - application framework for multimodal conversational AI services that delivers real-time performance on GPUs
- **NVIDIA NeMo** - open-source toolkit for developing state-of-the-art conversational AI models
- **NVIDIA® TensorRT™ 7.1** - high-performance SDK for DL inference. Tuned for Ampere GPUs and smashes earlier BERT inference perf with new INT8 optimizations
- **NVIDIA Merlin** - framework for building high-performance, deep learning-based recommender systems

**Dostupnost:** Jarvis – Early Access | NeMo – již dostupný | TensorRT 7.1 – v nejbližších dnech | NVIDIA Merlin – již dostupný



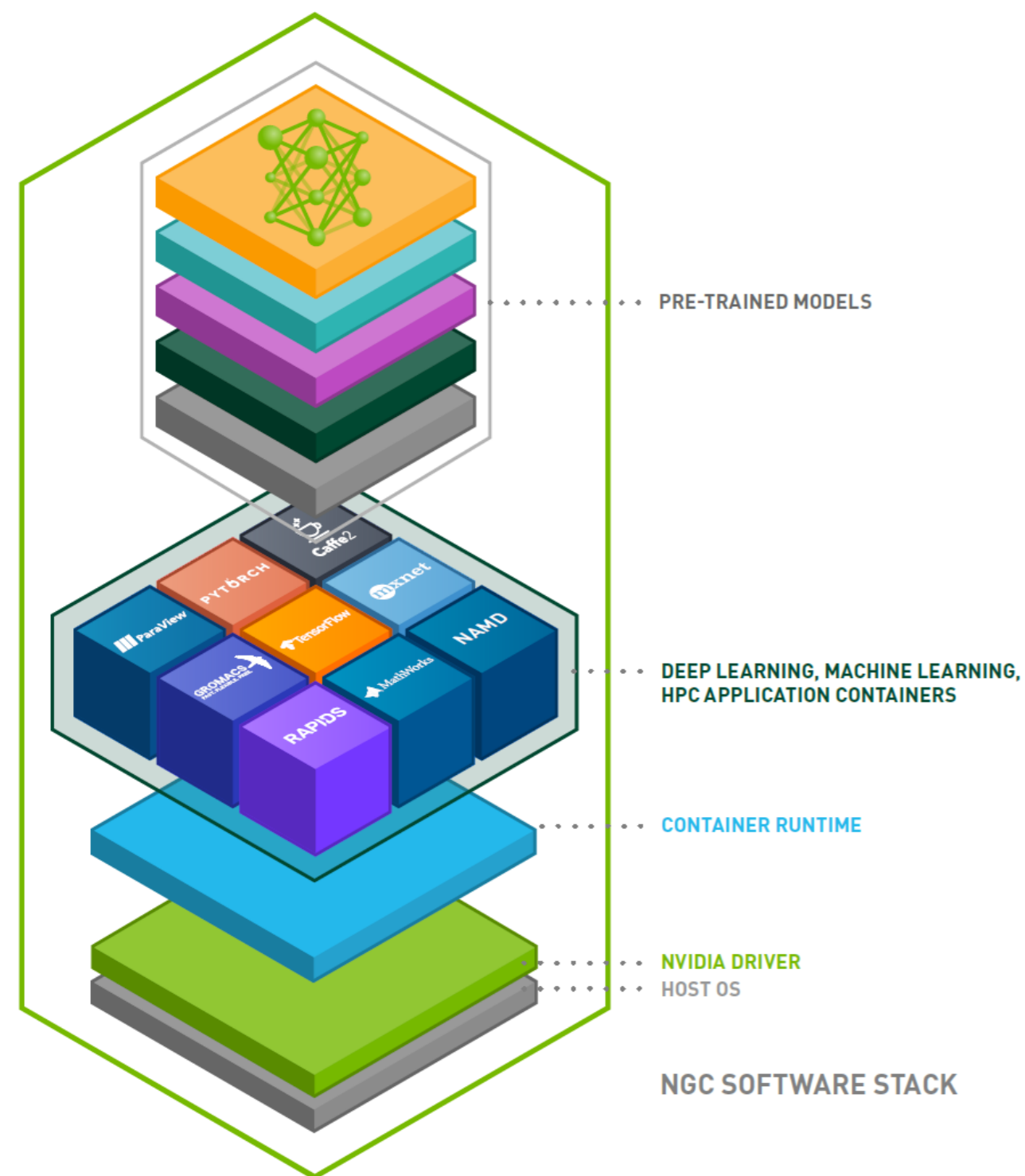
<https://youtu.be/bOf2S7OzFEg>



# NVIDIA GPU Cloud (NGC)

- AI a HPC kontejnery
- Předtrénované modely
- Skripty pro trénování modelů
- Workflow

volně dostupné, optimalizované pro NVIDIA GPU



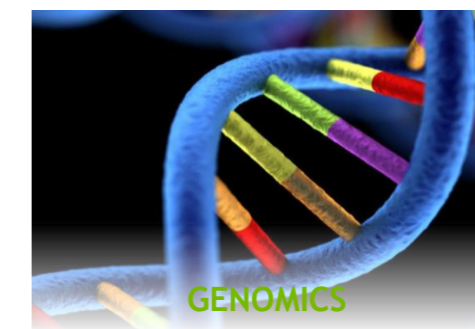
TensorFlow | PyTorch | more



NAMD | GROMACS | more



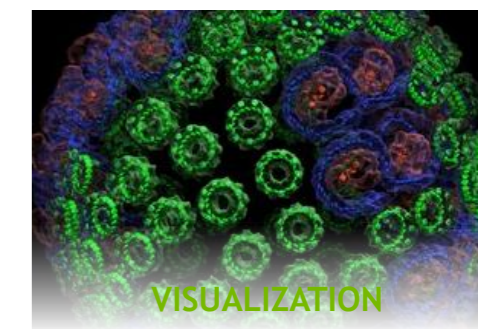
RAPIDS | H2O | more



Parabricks



TensorRT | DeepStream | more



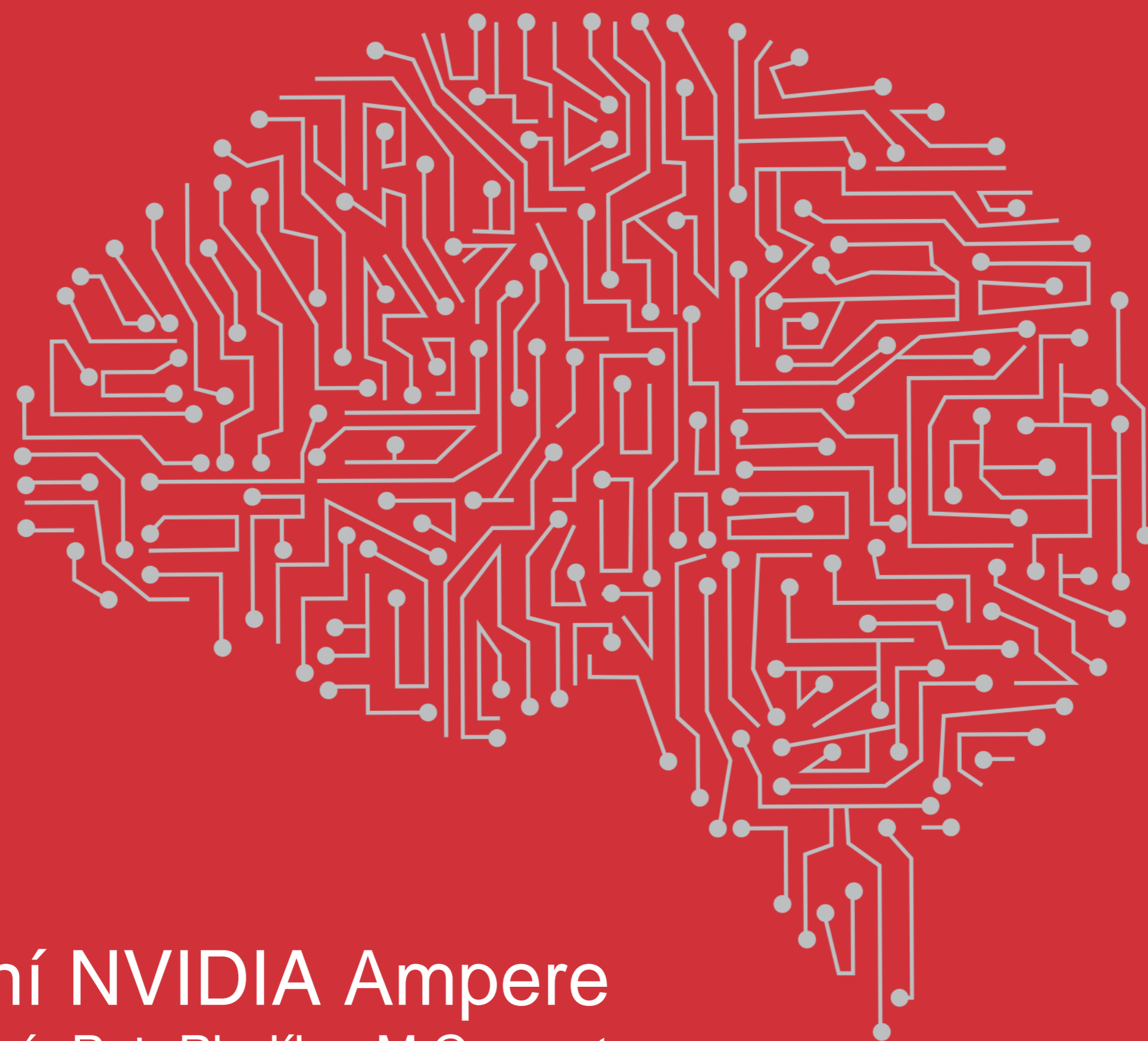
ParaView | Index | more

# NVIDIA GPU Cloud (NGC)

The screenshot displays the NVIDIA NGC Accelerated Software interface. At the top, it says "NVIDIA NGC | ACCELERATED SOFTWARE" and "Welcome Guest". Below this is a navigation bar with tabs for "ALL CONTENT TYPES", "CONTAINERS", "MODELS", "MODEL SCRIPTS", and "HELM CHARTS". The "CONTAINERS" tab is selected. A search bar "Search containers" and a sort dropdown "Sort: Last Modified" are visible. The main content area shows a grid of container cards:

- LAMMPS Container:** Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) is a software application designed for molecular dynamics simulations.
- CUDA Container:** CUDA is a parallel computing platform and programming model that enables dramatic increases in computing performance by harnessing the power of the NVIDIA GPUs.
- RAPIDS Container:** The RAPIDS suite of software libraries gives you the freedom to execute end-to-end data science and analytics pipelines entirely on GPUs.
- NAMD Container:** NAMD is a parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems. NAMD uses the popular molecular graphi...
- Triton Inference Server Container:** Triton Inference Server provides a data center inference solution optimized for NVIDIA GPUs. It maximizes inference utilization and performance on GPUs via ...
- TensorRT Container:** NVIDIA TensorRT is a C++ library that facilitates high-performance inference on NVIDIA graphics processing units (GPUs). TensorRT takes a trained network, which ...
- TensorFlow Container:** TensorFlow is an open-source software library for high-performance numerical computation. Its flexible architecture allows easy deployment of computation a...
- DIGITS Container:** The NVIDIA Deep Learning GPU Training System (DIGITS) puts the power of deep learning into the hands of engineers and data scientists.
- PyTorch Container:** PyTorch is a GPU accelerated tensor computational framework with a Python front end. Functionality can be easily extended with common Python libraries s...
- nvidia/rapidsai:** MXNet is a deep learning framework that allows you to mix the flavors of symbolic programming and imperative programming to maximize efficiency and ...

At the bottom left, it says "NGC Version: 2.30.2".



# Představení NVIDIA Ampere

Kamila Jeřábková, Petr Plodík – M Computers

