Multi-GPU akcelerace
technologie pro budoucnost simulací

Prezentující: Kamila Jeřábková, M Computers
kamila.jerabkova@mcomputers.cz

NVIDIA

M computers®

O NÁS

- ✓ Jsme česká technologická firma
- ✓ Specialisté na superpočítače a IT infrastrukturu

SUPERPOČÍTAČE

HPC software, cloud a služby
Machine learning
HPC pro výzkum
Benchmarky

IT INFRASTRUKTURA

Ukládání dat a zálohování
Servery včetně řešení na míru
Aplikační řešení
Networking

- ✓ Elite partner **NVIDIA**, Platinum partner **Lenovo** DCG i PCG, Gold partner **IBM**, HPE, Intel, AMD, NetApp, Gigabyte, Supermicro..
- ✓ Úzce spolupracujeme se vzdělávacími a vědeckými institucemi
- ✓ Dodáváme v celém regionu CEE

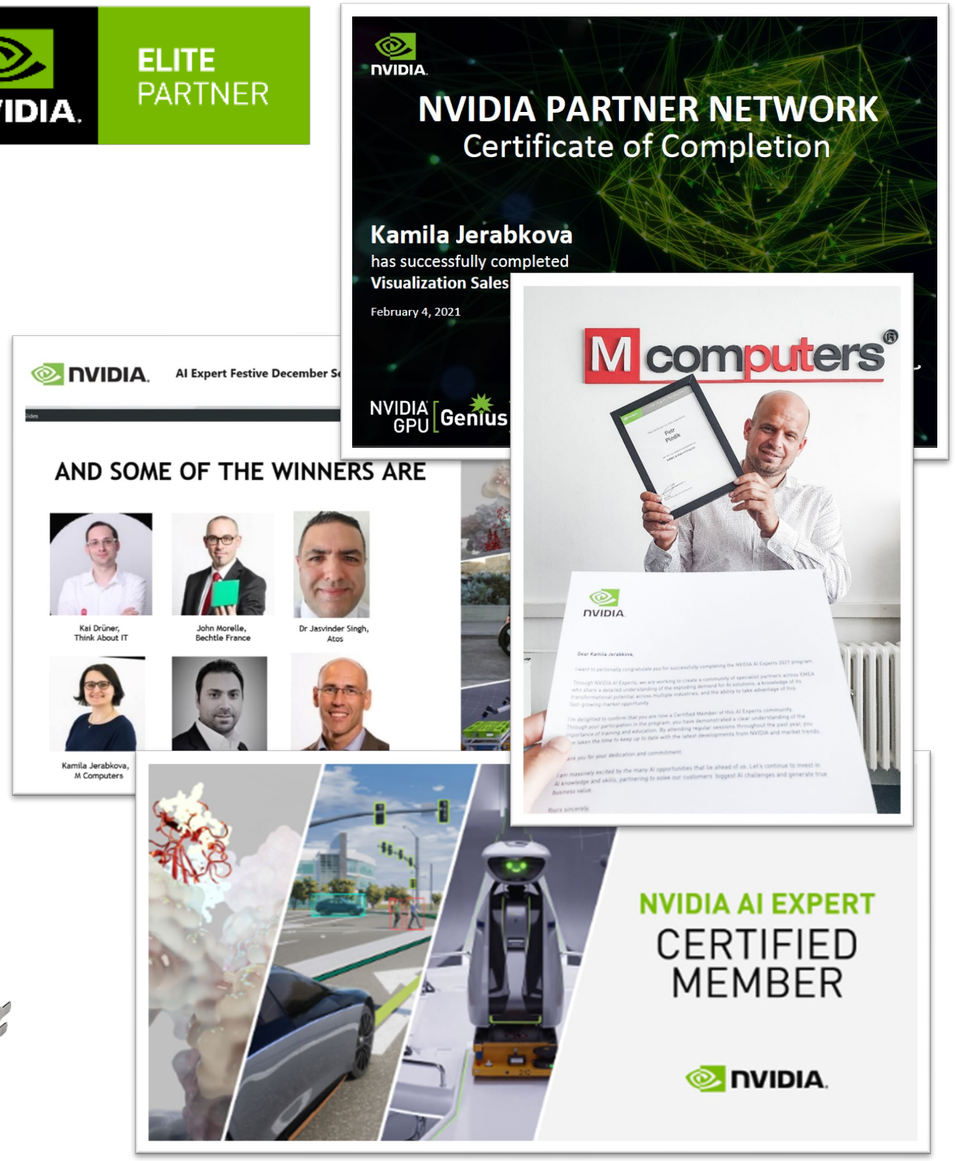
ROK ZALOŽENÍ FIRMY	OBRAT ZA ROK 2022 (v mil. Kč)		
2002	48	415	3
	AKTUÁLNÍ POČET ZAMĚSTNANCŮ		POČET POBOČEK V ČR

NVIDIA ELITE PARTNER

- ✓ NVIDIA certifikace:
  Compute DGX | Visualization (RTX) | Virtualization (vGPU) | Networking

- ✓ Slevy a promo nabídky pro vysoké školy, výzkumné instituce a start-upy
  - Inception promo RTX 6000 Ada za 50%
  - EDU promo (roční licence **Omniverse za 1 USD**)

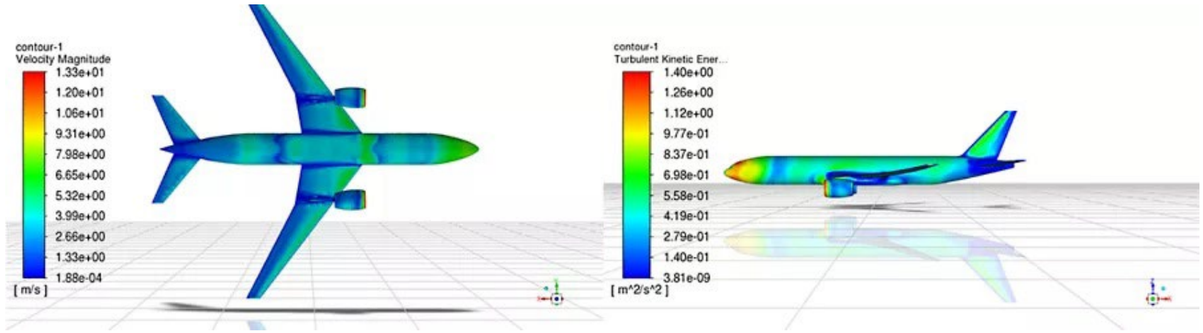
- ✓ Demo pool nejnovějších NVIDIA technologií
  - NVIDIA DGX Station A100 (4x A100)
  - Akcelerátory (**H100**, A100, A40, A30, **A2**, **A16**)
  - **vGPU licence** (vApps, vPC, vWS, vCS)
  - **AI Enterprise licence**

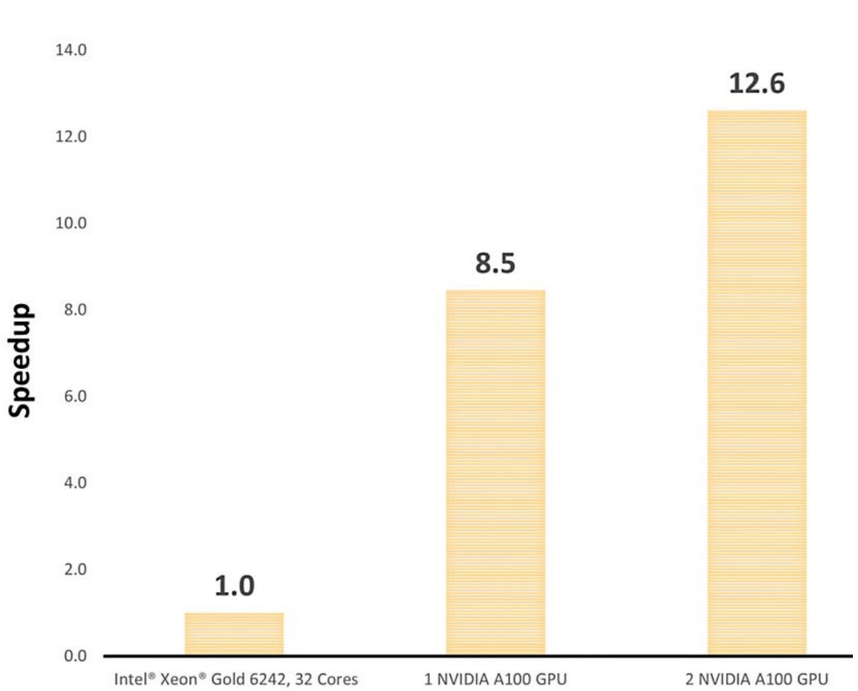


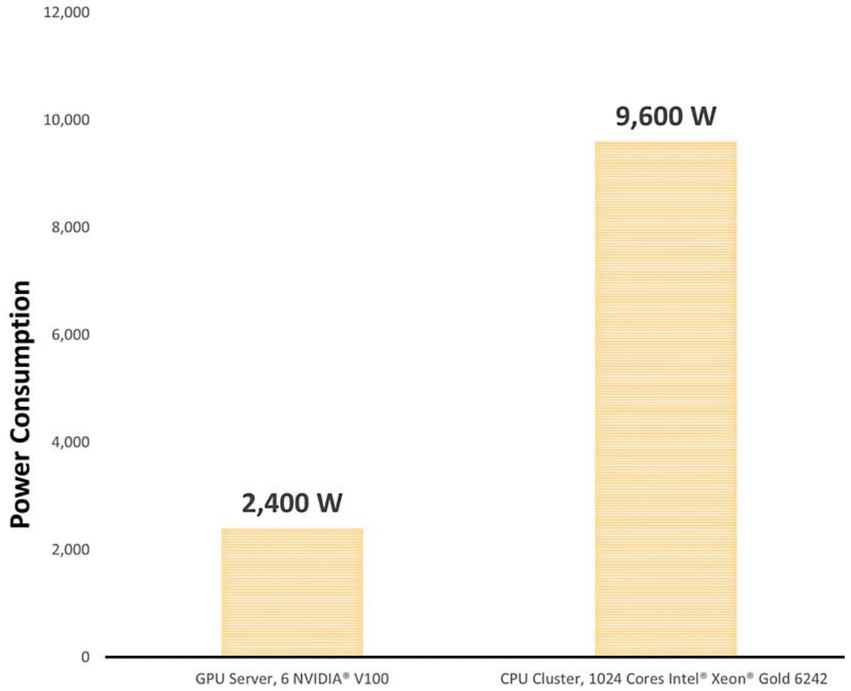
Akcelerace výpočtů pomocí GPU

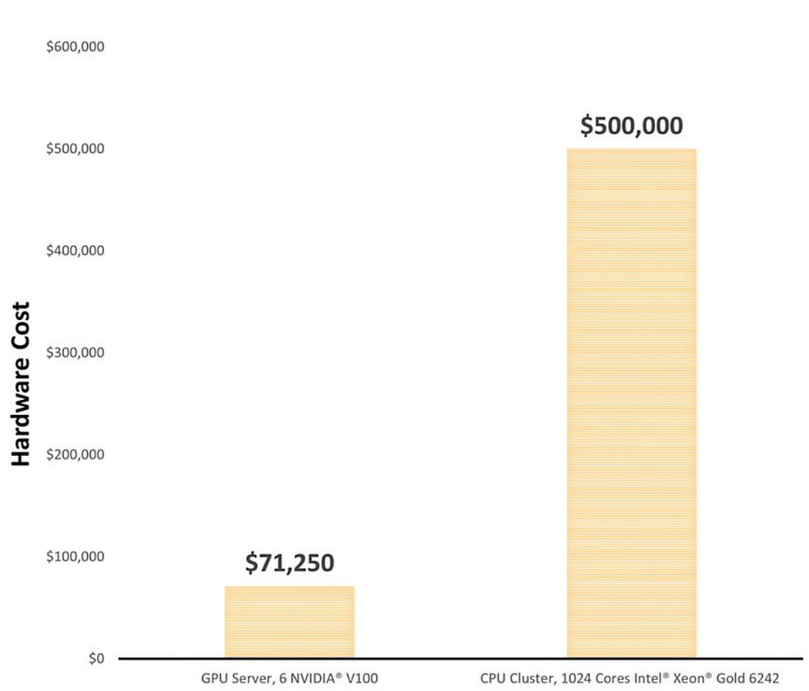
Výhody provádění simulací na GPU vs. CPU

- ✓ Násobně vyšší výkon
- ✓ Nižší spotřeba el. energie a menší nároky na chlazení
- ✓ Snížení nákladů na HW a využití prostoru





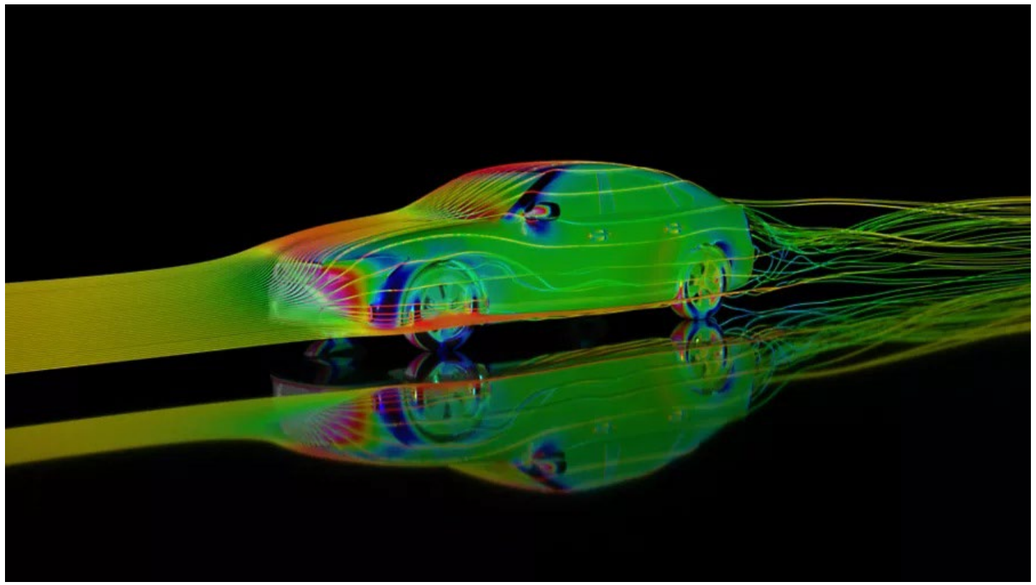




Akcelerace výpočtů pomocí multi-GPU

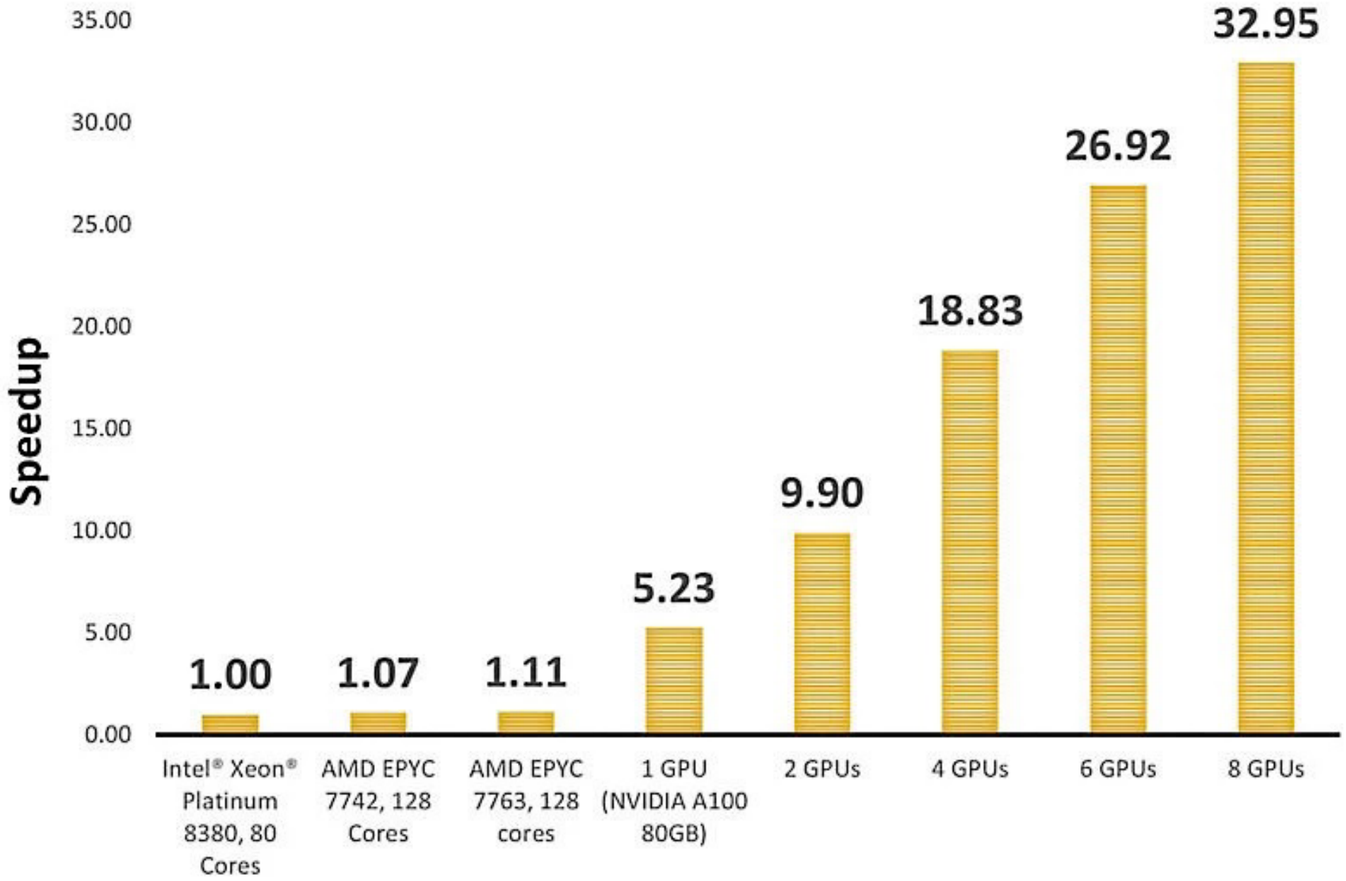
NVIDIA GPU a Ansys Fluent multi-GPU solver

- ✓ Zrychlení výpočtů
- ✓ Zpracování extrémně velkých modelů
- ✓ Až 640 GB grafické paměti v jednom serveru



Srovnání výkonu pomocí benchmarku *DrivAer model (300 milionů částic)* na CPU a GPU konfiguracích s využitím Ansys Fluent

„NVIDIA A100 GPU > 5x Intel Xeon Platinum 8380 (80 jader)“



Grafické karty pro akceleraci výpočtů

NVIDIA A100 80GB

Nejvýkonnější akcelerátor generace Ampere. Na těchto kartách je možné urychlit mnoho masivně paralelních úloh a vědeckých (HPC) aplikací nebo efektivně řešit algoritmy umělé inteligence (AI).

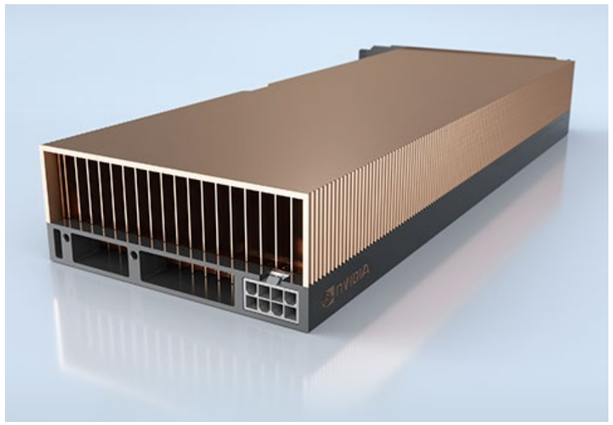
- **80 GB** paměti HBM2 s vysokoou propustností až **2 GB/s**
- **6 912** CUDA jader
- **432** Tensor jader 3. generace
- FP32: **19,5 TFLOPS**
- **SXM4** (10x rychlejší než PCIe4) **| PCIe Gen 4** (2x vyšší datová propustnost než PCIe Gen 3)

NVIDIA H100 80GB

Nejnovější a nejvýkonnější akcelerátor generace NVIDIA Hopper. Oproti předchozí generaci Ampere (A100) dosahuje až trojnásobného výkonu ve výpočtech s jednoduchou přesností (FP32).

- **80 GB** paměti HBM3 s propustností až **3,35 TB/s**
- **16 896** CUDA jader
- **528** Tensor jader 4. generace
- FP32: **67 TFLOPS**
- **SXM5 | PCIe Gen 5** (2x vyšší datová propustnost než PCIe Gen 4)

Grafické karty pro vizualizaci a akceleraci výpočtů



NVIDIA A40 a RTX A6000

Profesionální karta určená pro vizualizaci, strojové učení a datovou analýzu a momentálně také nejvýkonnější procesor pro grafické výpočty.

- nástupce karet **Quadro RTX 6000/8000**
- **10 752 CUDA** jader, **336 Tensor** jader
- FP32: **37,4** TFLOPS
- grafika, renderování, simulace, VR, 3D design a AI výpočetní výkon
- čip **GA102** (RTX série 3080/3090)
- **48 GB** grafické paměti GDDR6 ECC (až 96 GB pomocí NVLinku)
- propustnost paměti až **768 GB/s**
- **PCIe Gen 4** (2x více než PCIe Gen 3)
- max. příkon **300W**
- podpora virtuálních pracovních stanic **vWS** (vGPU) i virtuálních serverů (vCS)

NVIDIA L40 a RTX 6000 Ada

Akcelerátor postavený na generaci *Ada Lovelace*, nástupce karet A40 / RTX A6000 a momentálně nejvýkonnější karta pro simulace a grafické výpočty.

- nástupce karet **NVIDIA A40** / **RTX A6000**
- **18 176 CUDA** jader, **568 Tensor** jader
- FP32 **95,42** TFLOPS
- nejnovější generace **RT**, **Tensor**ových a **CUDA** jader pro grafiku, renderování, simulace, VR, 3D design a AI výpočetní výkon
- čip **AD102** (RTX 4090)
- **48 GB** grafické paměti GDDR6 ECC
- propustnost paměti až **864 GB/s**
- **PCIe Gen 4** (2x více než PCIe Gen 3)
- max. příkon **300W**
- Podpora vGPU i NVAIE

Ve dvou variantách:
RTX A6000 / 6000 Ada: aktivní chlazení vhodné do profesionálních pracovních stanic
A40 / L40: pasivní chlazení pro osazení do serverů v datových centrech

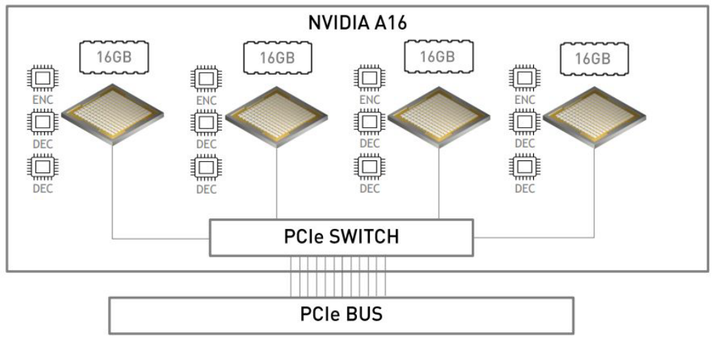
Grafické karty pro vizualizaci a akceleraci výpočtů

NVIDIA A16

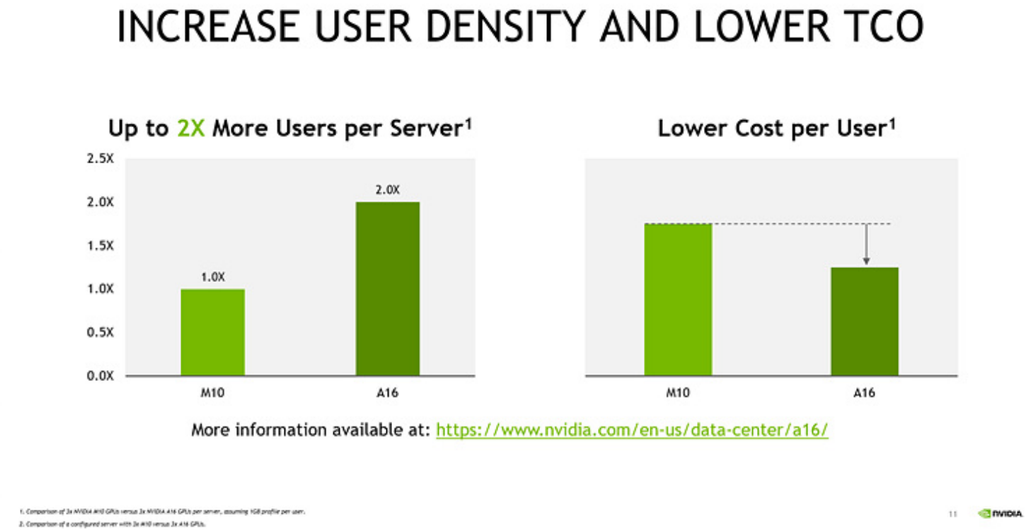
A16 v sobě kombinuje výkon 4 grafických čipů na jednom boardu s pamětí 4x 16GB GDDR6.
To znamená vysokou hustotu výpočetního výkonu s nízkými nároky na prostor a se skvělým poměrem cena / výkon. NVIDIA A16 v kombinaci se softwarem NVIDIA vGPU tvoří součást platformy EGX pro profesionální vizualizaci a přináší multimediální virtuální desktopy (**VDI**) umožňující vzdálenou práci odkudkoli. Podporuje až 64 souběžně pracujících uživatelů na jediné kartě ve dvouslotovém provedení.

- **64 GB** (4x 16GB) paměti **GDDR6** s podporou ECC
- 4x čipy GA107 na jedné desce
- propustnost paměti až **4x 200 GB/s**
- 4x 1 280 **CUDA** jader a 4x 40 **Tensor** jader
- **PCIe Gen 4** x16
- pasivní chlazení, dual-slot
- TDP **250W** (235W Max-Q)

NVIDIA A16 ARCHITECTURE



Dostupná k zapůjčení v našem DEMO labu



Grafické karty pro vizualizaci a akceleraci výpočtů

NVIDIA RTX A5000 / A5500 (Ada)

NVIDIA RTX A5000 nahrazuje předchozí generaci ***Quadro RTX 5000***. Ve srovnání s ní nabízí až 2,5 násobný nárůst výkonu ve výpočtech s jednoduchou přesností (FP32). RTX A5000 přináší dokonalou rovnováhu mezi výkonem, spotřebou a spolehlivostí pro náročné výpočty a je ideální pro využití v profesionálních pracovních stanicích.

- **24 GB** paměti GDDR6 s podporou ECC
- propustnost paměti až **768 GB/s**
- **8192 – 10240** CUDA jader
- FP32: **27,7 – 34,1 TFLOPS**
- **PCIe Gen 4** (2x více než PCIe Gen 3)
- aktivní chlazení, dual-slot

NVIDIA RTX A4000 / A4500 (Ada)

NVIDIA RTX A4000 je nejvýkonnější jednoslotová grafická karta nabízející skvělý výkon ve vykreslování světla v reálném čase (real-time ray tracing), v akcelerovaných AI výpočtech a profesionálním grafickém renderování.

- **16 – 20 GB** paměti GDDR6 s podporou ECC
- propustnost paměti až **448 – 640 GB/s**
- **6144 – 7168** CUDA jader
- FP32: **19,2 – 23,6 TFLOPS**
- **PCIe Gen 4** (2x více než PCIe Gen 3)
- aktivní chlazení, single-slot

GPU řešení pro akcelerované výpočty

GPU akcelerátory

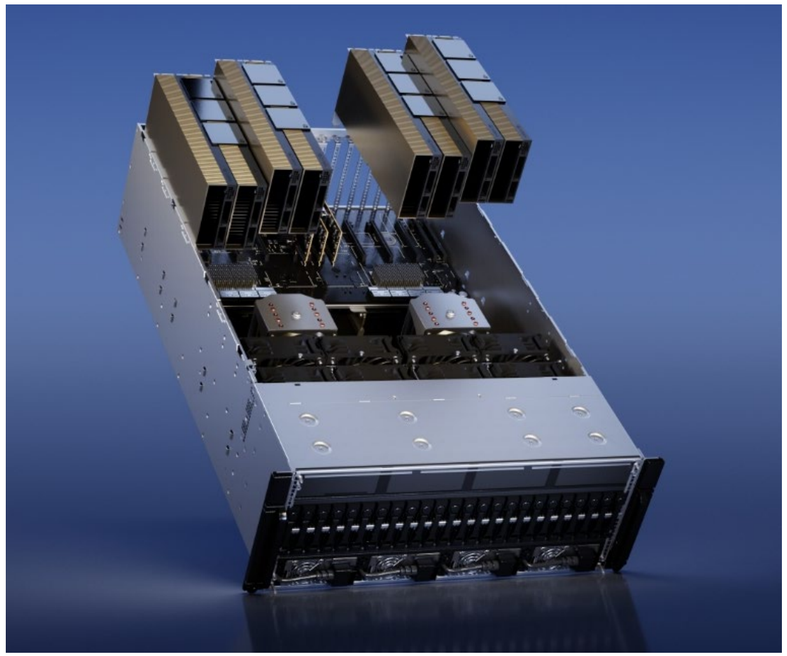
- Servery OEM výrobců s PCIe GPU akcelerátory
- GPU propojeny NVLink bridge / PCIe

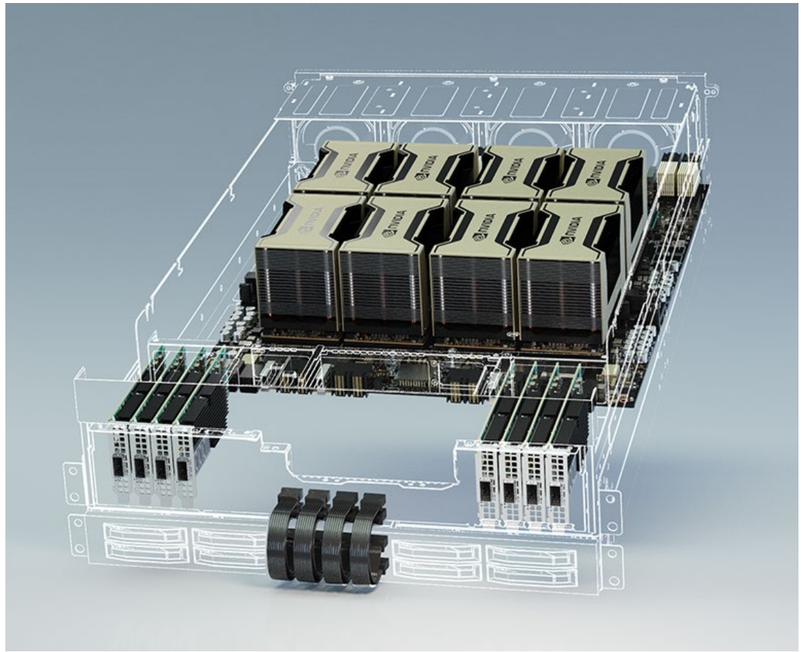
HGX servery

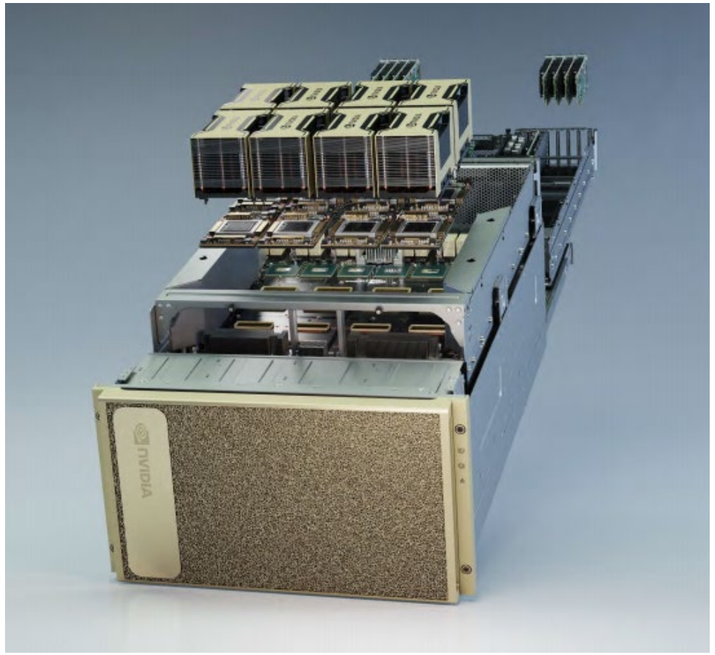
- Servery OEM výrobců s GPU boardy
- GPU propojeny NVLink technologií
- Bez NVIDIA SW a supportu

DGX systémy

- Vyladěné NVIDIA systémy s GPU boardy
- kompletní SW vrstvoa (OS, Docker, napojení na NGC)





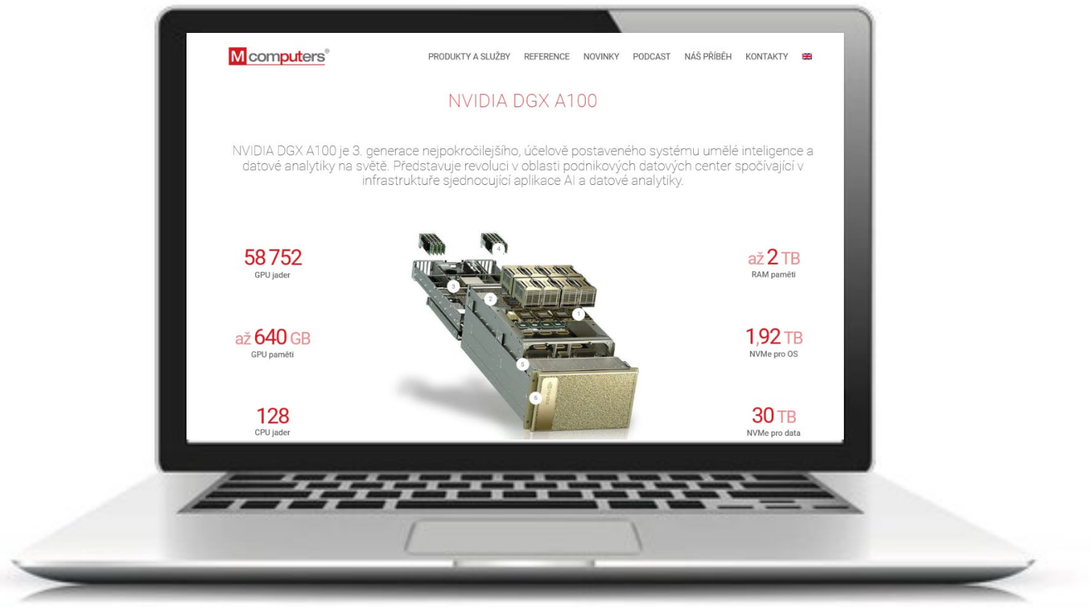


NVIDIA produkty přehledně

www.mcomputers.cz/nvidia

NVIDIA – kompletní informace přehledně a na jednom místě v čj / aj

- ✓ aktuality a nová produktová oznámení
- ✓ přehled NVIDIA produktů
- ✓ specifikace, datasheety, benchmarky
- ✓ srovnání všech Nvidia karet
- ✓ referenční architektury
- ✓ záznamy webinářů a prezentace ke stažení + registrace
- ✓ reference a příklady úspěšných nasazení



POROVNÁNÍ NVIDIA GPU PRO DATOVÁ CENTRA

PARAMETR	NVIDIA A2	NVIDIA A16	NVIDIA A40	NVIDIA A30	NVIDIA A10	NVIDIA A100 SXM4 | PCIE	DGX STATION A100	DGX A100
Architektura karty	Ampere	Ampere	Ampere	Ampere	Ampere	Ampere	Ampere	Ampere
Čip karty	GA107	GA107	GA102	GA100	GA102	GA100	GA100	GA100
# CUDA jader	1 280	4x 1 280	10 752	6 912	9 216	6 912	27 648	55 296
# Tensor jader	40	4x 40	336	224	288	432	1 728	3 456
FP64 (TFlops)	0,07	0,271	1,179	5,2	0,97	9,7	38,8	77,6
FP64 Tensor (TFlops)	–	–	–	10,3	–	19,5	78	156

POROVNÁNÍ NVIDIA KARET PRO VIZUALIZACI

PARAMETR	RTX 3080	RTX 3090	RTX A6000	RTX A5000	RTX A4500	RTX A4000	RTX A2000
Architektura	Ampere	Ampere	Ampere	Ampere	Ampere	Ampere	Ampere
Čip karty	GA102	GA102	GA102	GA102	GA102	GA104	GA106
# CUDA jader	8704	10 496	10 752	8 192	7 168	6 144	3 328
# Tensor jader	272	328	336	256	224	192	104
FP64 (TFlops)	0,47	0,56	1,25	0,87	0,739	0,6	0,124
FP32 (TFlops)	29,8	35.6	40	27,7	23,65	19,2	8
FP16 Tensor (TFlops)	119/238*	142/284*	309,7*	222,2*	189,2	153,4*	63,9*