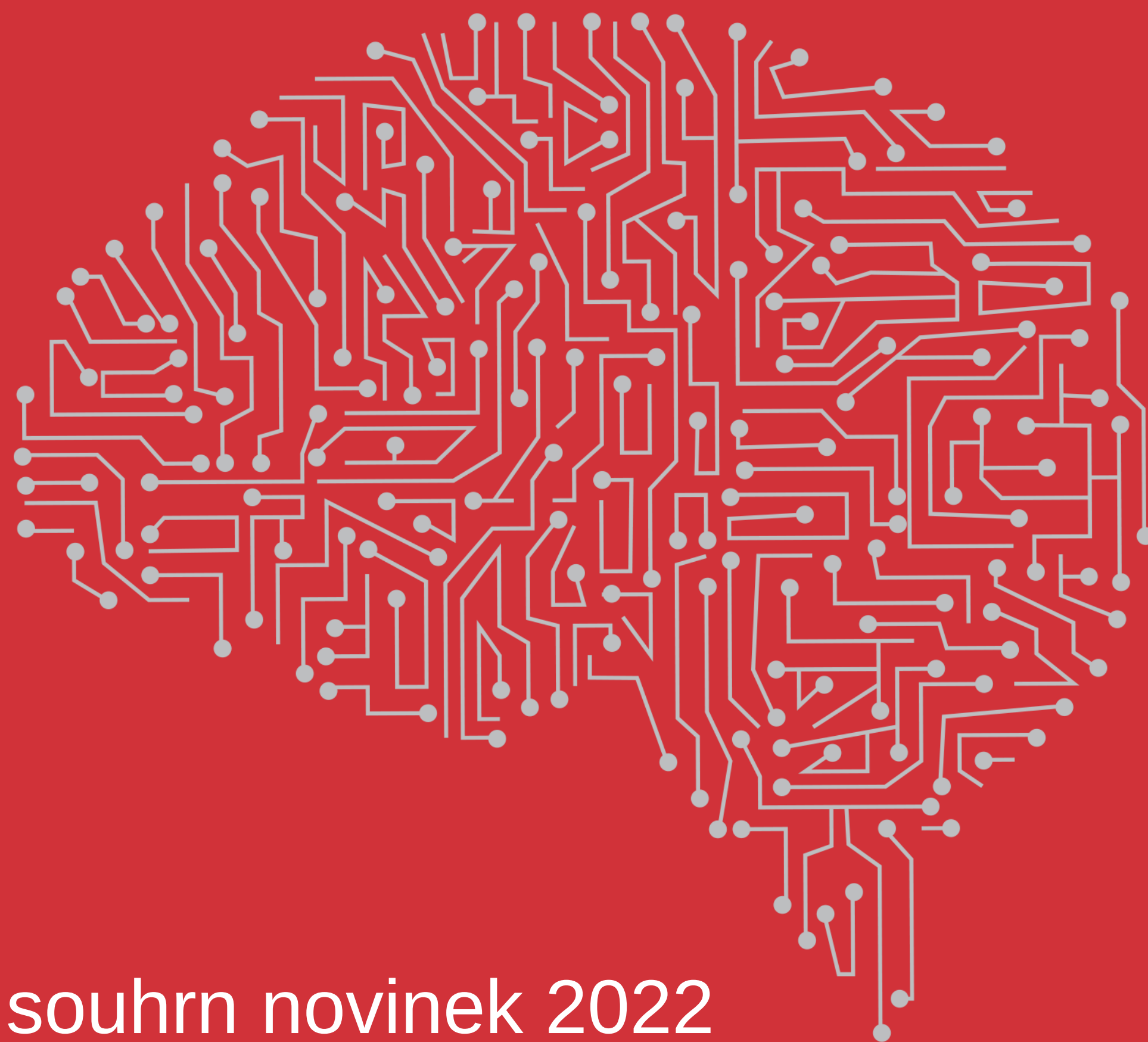




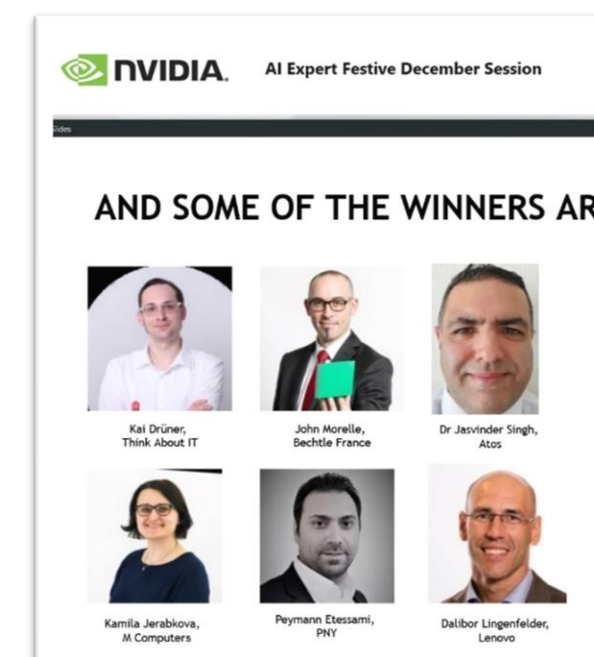
NVIDIA – souhrn novinek 2022

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



NVIDIA ELITE PARTNER

- ✓ Jsme první NVIDIA ELITE partner ve střední a východní Evropě
- ✓ NVIDIA certifikace:
Compute DGX | Visualization (RTX) | Virtualization (vGPU) | Networking
- ✓ Nabízíme slevy pro vysoké školy, výzkumné instituce a start-upy
- ✓ Demo pool nejnovějších NVIDIA technologií
 - NVIDIA DGX Station A100 (4x A100)
 - Akcelerátory (**H100**, A100, A40, A10, A30, **A2**, **A16**)
 - vGPU licence (vApps, vPC, vWS, vCS)
 - **AI Enterprise licence**



Aktuální portfolio NVIDIA, řešení pro každou oblast (HW + SW)



Frameworky pro vývoj aplikací
- rozpoznávání řeči, kybernetická bezpečnost, doporučování obsahu, medicínská analýza obrazu, smart cities, autonomní řízení, robotika a další..

Kompletní softwarová vrstva pro instalaci, optimalizaci, management a sdílení akcelerovalých systémů
- Base Command, AI Enterprise, vGPU, Parabricks...

Hardwarová vrstva zahrnuje nejen GPU akcelerátory (H100, A100, RTX...), ale i CPU (Grace), DPU (BlueField), networking (switche, síťové karty NIC), edge zařízení, a zároveň i extrémně výkonné výpočetní systémy DGX (referenční architektura Super POD), či OEM platformy HGX a OVX.

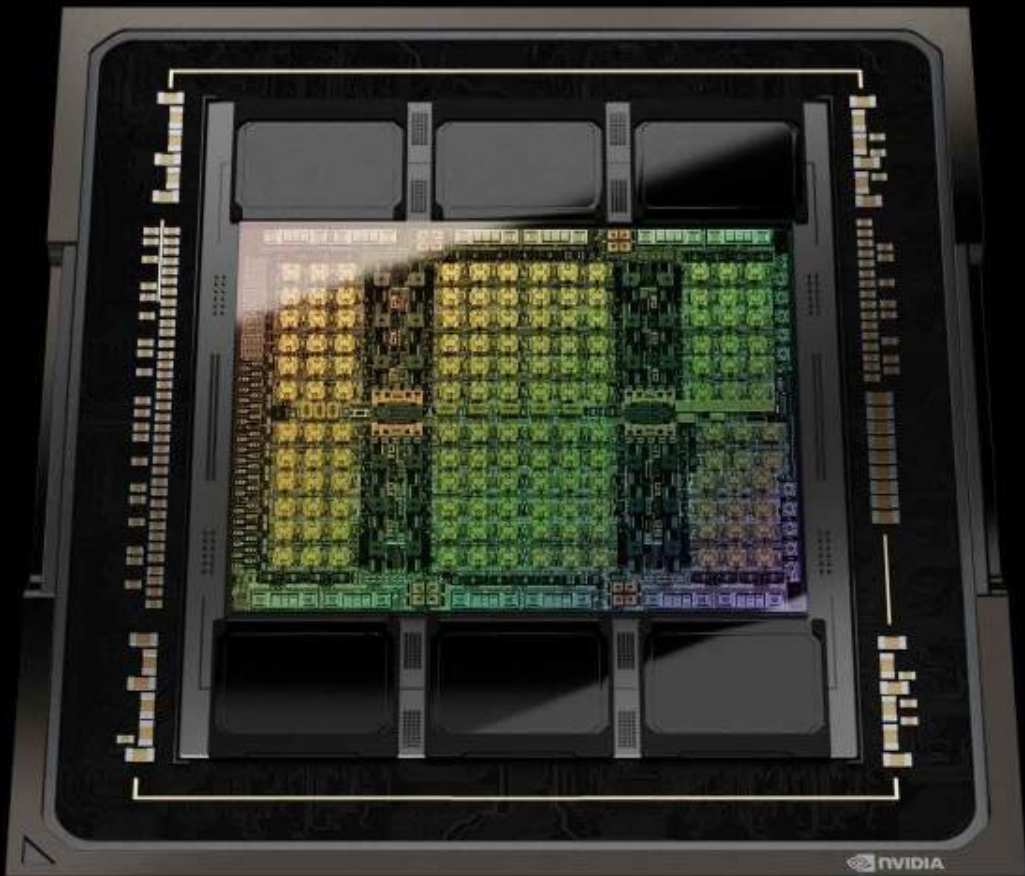
HW: Architektura Hopper, akcelerátor NVIDIA H100

NVIDIA H100 80GB HBM3

- nejvýkonnější akcelerátor současnosti
- 4nm technologie výroby
- 80 mld tranzistorů na 814 mm²
- propustnost paměti HBM3 = 3 TB/s (HBM2 u A100 měla “jen” 2TB/s, PCIe 1,5TB/s)
- 50 MB L2 cache
- 4. generace NVLinku (900GB/s)
- podpora PCIe 5.0 (128GB/s)
- MIG (Multi-Instance GPU) 2. generace

Dostupné varianty:

- **H100 SXM5** (700W)
- **H100 PCIe** Gen 5 (350W a 65% výkonu SXM verze)
- **H100 CNX** (ConnectX-7 NIC)



(TFLOPS)	H100 SXM5	H100 PCIe	A100 SXM4
FP64	34	26	9,7
FP64 Tensor	67	51	19,5
FP32	67	51	19,5
TF32	500/1000*	400/800*	156/312*
FP16	134	102	78
FP16 Tensor	1000/2000*	800/1600*	312/624*
FP8 Tensor	2000/4000*	1600/3200*	---

*výkon s funkcí sparsity

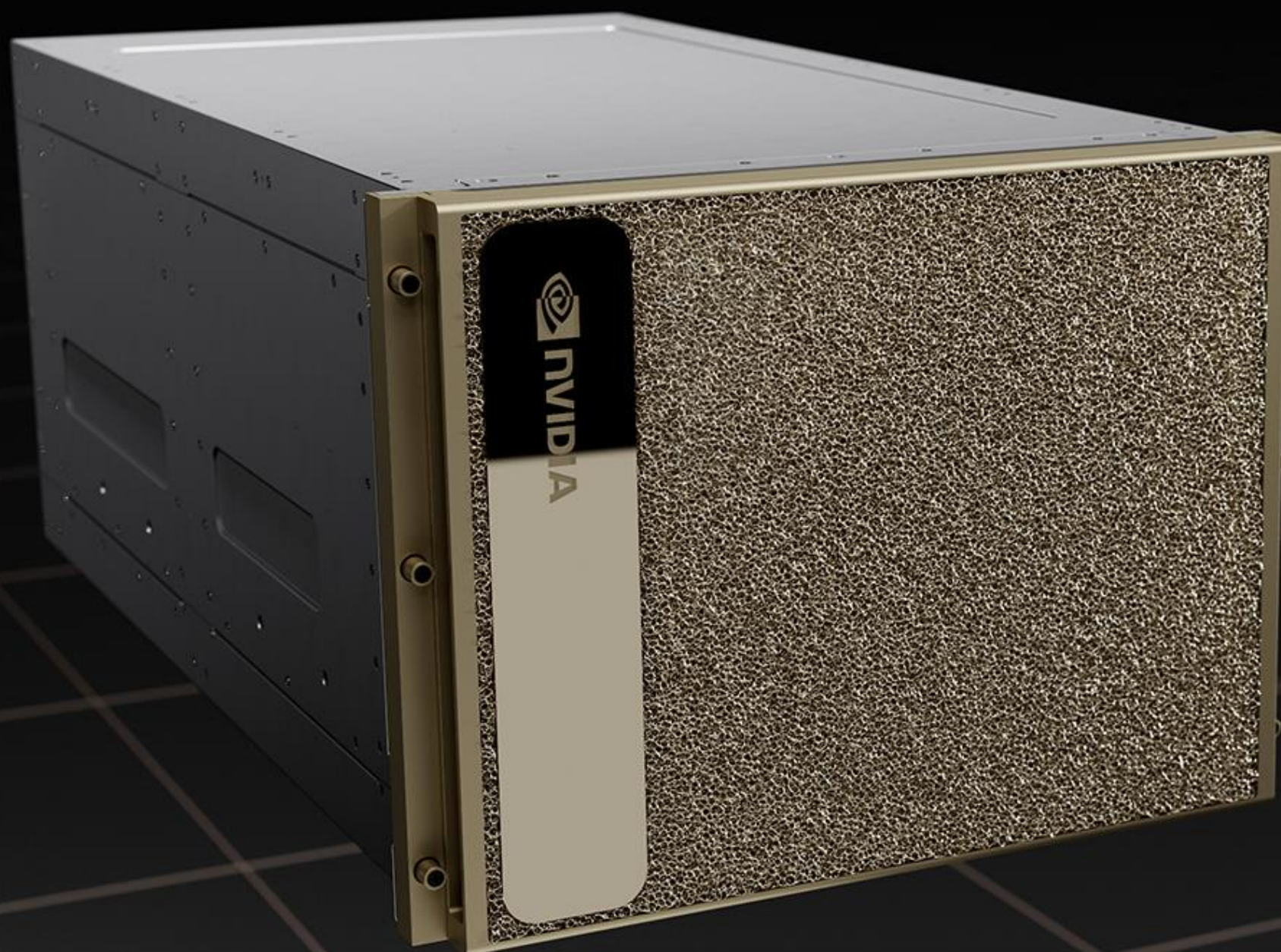
	SXM	PCIe
Tensor jader (4. gen):	528	456
CUDA jader:	16 896	14 592

NVIDIA H100

80 Billion Transistors
TSMC 4N Process
4.9 TB/s Bandwidth

FP8	4,000 TFLOPS	6X
FP16	2,000 TFLOPS	3X
TF32	1,000 TFLOPS	3X
FP64/FP32	60 TFLOPS	3X

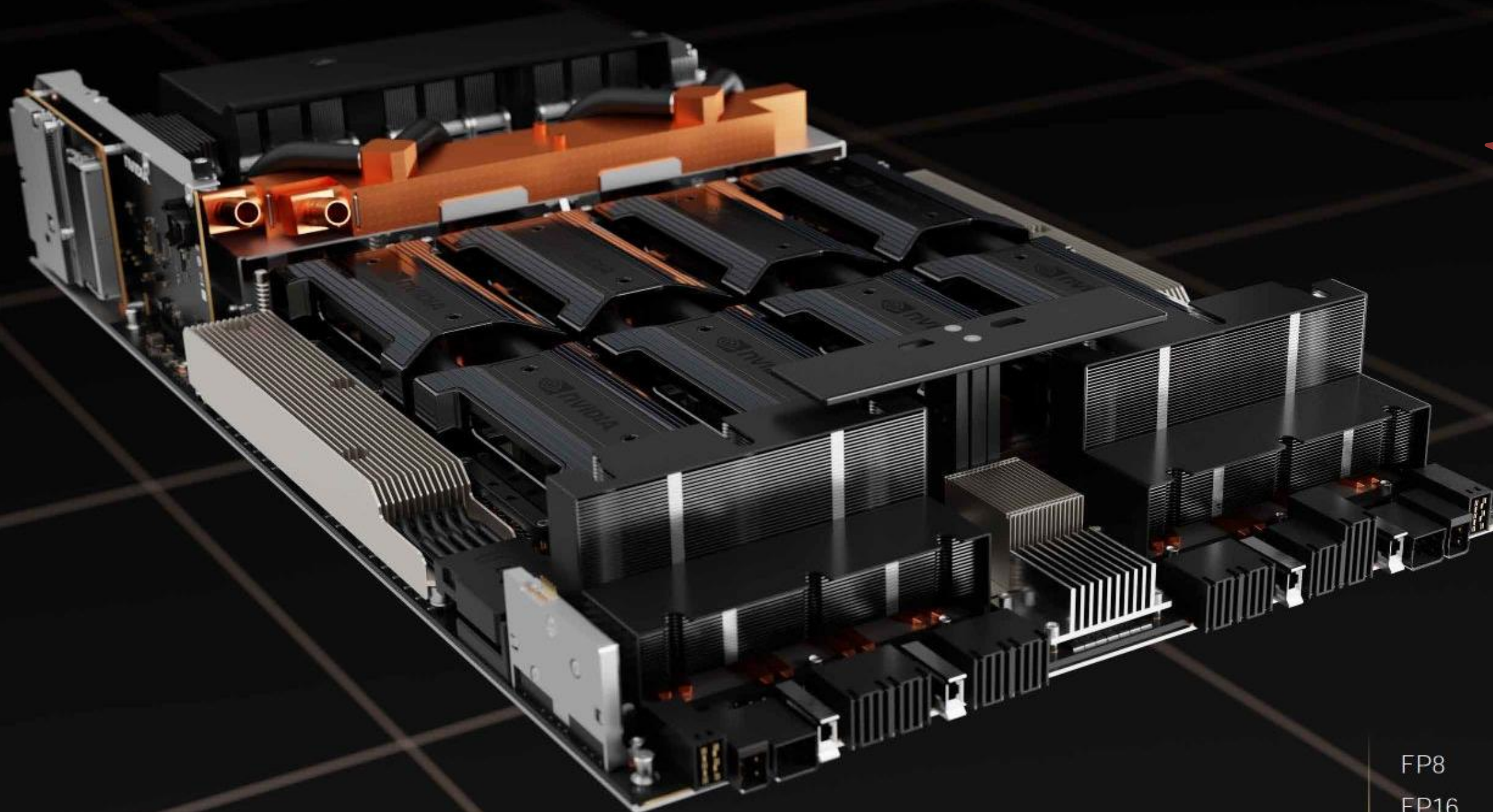
NVIDIA DGX H100, superpočítač pro HPC a AI



- 8x GPU **H100 80GB**
 - 640 GB HBM3 paměti s propustností až 24 TB/s
 - 640 miliard tranzistorů
 - 32 petaFLOPS výkonu pro AI výpočty
- 2x CPU Intel Sapphire Rapids Xeon (48 cores?)
- 2TB DDR5 RAM
- 4x NVSwitch (7,2 TB/s)
- OS: 2x 1.9TB NVMe M.2
- DATA: 8x 3.84TB NVMe U.2
- 8x single-port ConnectX-7 (400Gb/s InfiniBand/Ethernet)
- 2x dual-port ConnectX-7 (400Gb/s InfiniBand/Ethernet)
- 10Gb/s onboard NIC (RJ45) (volitelně 50Gb/s Eth)
- Management port (BMC) (RJ45)
- DGX Display Adapter card (4 DisplayPort, 4K)
- Spotřeba max. 10,2 kW
- Velikost 8U

NVIDIA HGX H100, platforma dostupná přes OEM výrobce

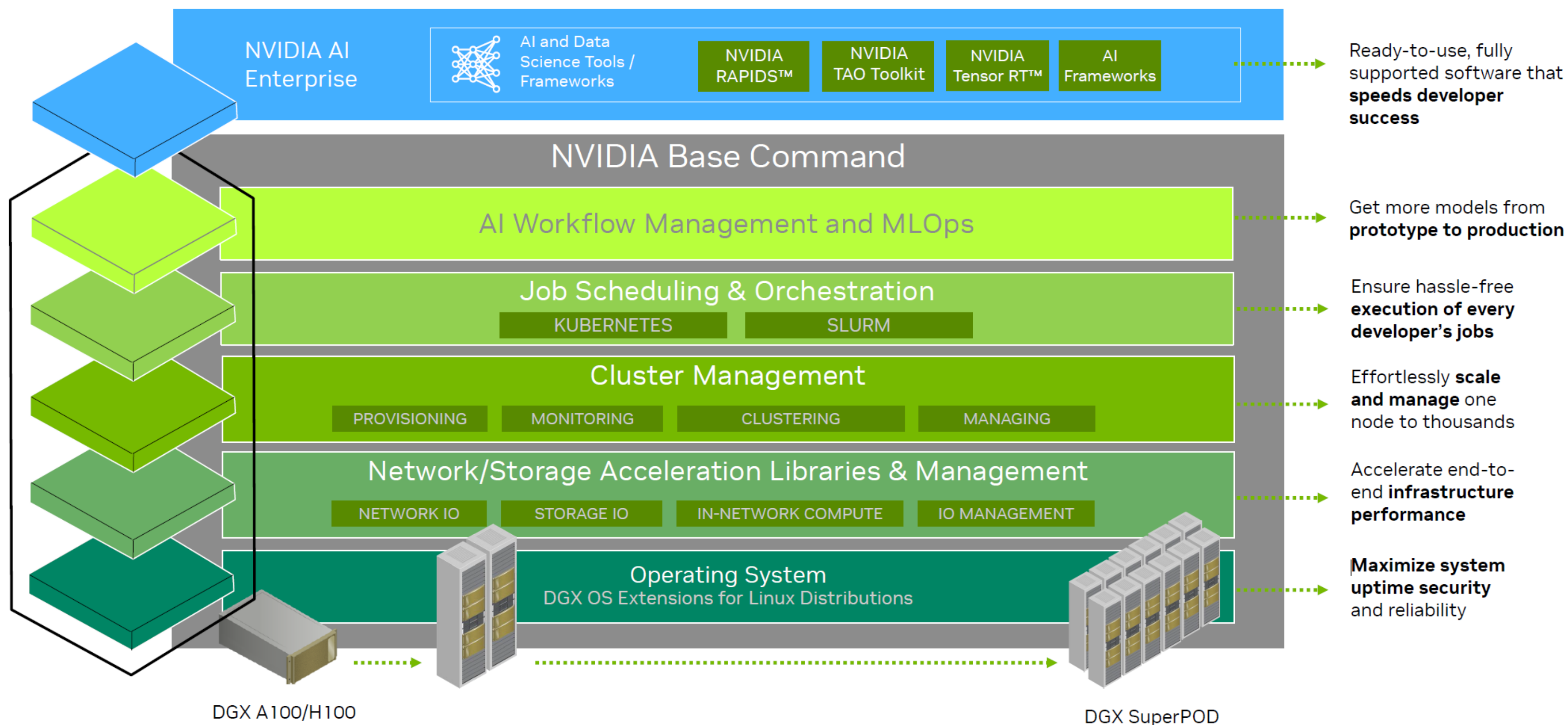
Dostupné u
OEM výrobců
2023



INTRODUCING HGX H100

FP8	32,000 TFLOPS	6X
FP16	16,000 TFLOPS	3X
FP64	480 TFLOPS	3X
In-Network Compute	3.6 TFLOPS	∞
Bisection Bandwidth	3.6 TB/s	1.5X

NVIDIA Base Command + AI Enterprise, SW vrstva pro DGX systémy

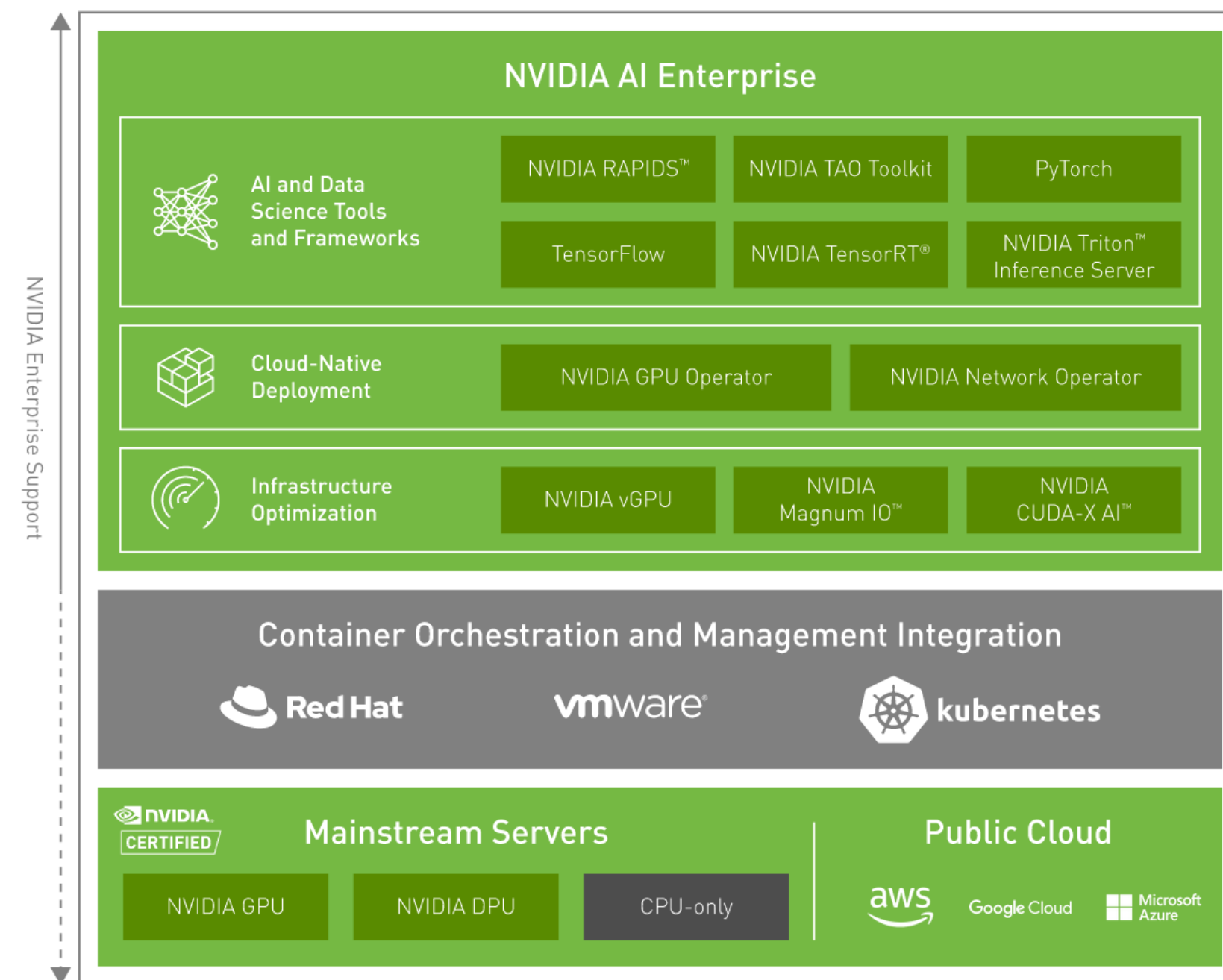


NVIDIA AI Enterprise, SW pro certifikované systémy

mcomputers.cz/nvidia/ai-enterprise/

Komplexní sada softwaru pro umělou inteligenci a analýzu dat, která je optimalizovaná a certifikovaná pro provoz na VMware vSphere a Red Hat OpenShift společně s DGX a certifikovanými OEM systémy (NVIDIA-Certified Systems). Zahrnuje technologie pro rychlé nasazení, správu a škálování AI výpočtů v hybridním cloudu.

- ✓ licencováno per CPU socket
- ✓ sdílení výkonných GPU akceleratorů více virtuálními servery
- ✓ zprovoznění vyladěného AI prostředí v řádu hodin
- ✓ podpora od NVIDIA na celý software stack včetně open-source knihoven, frameworků
- ✓ napojení na NVIDIA GPU katalog (NGC)
- ✓ podporované akcelerátory:
 - compute: HGX A100, A100, A30, V100s/V100, nově A100X, A30X a A2
 - mixed workloads: A40, A10, A16, T4, RTX A6000, RTX A5000
- ✓ nově certifikováno pro provoz Red Hat OpenShift jako bare metal nebo s VMware
- ✓ bare metal provoz na Red Hat Enterprise Linux 8.4 a Ubuntu 20.04 LTS
- ✓ nově podpora nasazení u CPU-only serverů, bez nutnosti osazení GPU
- ✓ podpora nasazení v cloudu (AWS, Azure, Google)



NVIDIA vGPU – virtualizace grafického výkonu

mcomputers.cz/nvidia/nvidia-vgpu/

Software NVIDIA Virtual GPU (vGPU) umožňuje virtualizaci a rozdělení GPU výkonu tak, aby byla jedna GPU karta sdílena mezi více virtuálními stroji. Zároveň lze i více GPU přidělit jedinému virtuálnímu stroji, což otevírá možnosti zpracování i extrémně výpočetně náročných úloh.

✓ GPU výkon bez omezení

Výkon virtualizovaných GPU není degradován a dosahuje téměř stejných hodnot jako u fyzických GPU karet.

✓ Snadné nastavení

S využitím běžných nástrojů pro řízení datových center je správa a monitoring virtualizovaných GPU snadnější a bezpečnější.

✓ Optimální využití zdrojů

Kombinujte výkon několika GPU karet, nebo je naopak rozdělte na více menších instancí, přesně dle vašich potřeb.

✓ Měňte výkon dle potřeby

Snadno přidávejte, odebírejte či měňte profily, podle toho, jak se momentálně vyvíjí vaše potřeby grafického výkonu.

Rozdělení VIRTUAL GPU licencí

✓ Virtual Applications (vApps)

vhodný pro akceleraci streamování aplikací s řešeními RDSH, včetně Citrix Virtual Apps a VMware Horizon vApps

✓ Virtual PC (vPC)

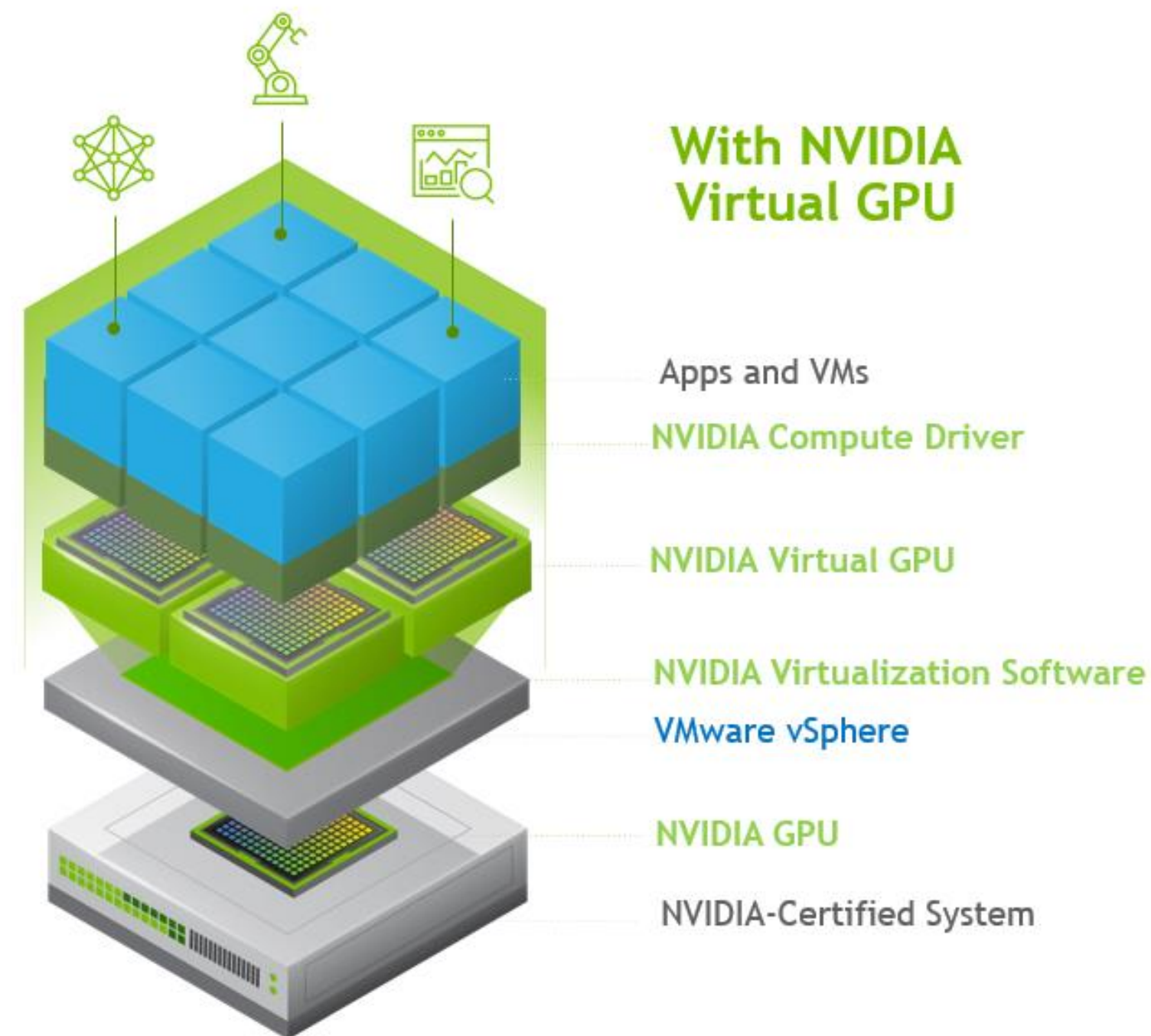
akcelerace výkonu pro virtuální desktopy, základní funkcionality, velikost profilu max. 2 GB

✓ Virtual Workstation (vWS)

akcelerace výkonu pro virtuální pracovní stanice, více funkcionalit a větší velikost profilů

✓ Virtual Compute Server (vCS)

akcelerace výpočtů virtualizovaných serverů, vhodné pro výpočetně náročné úlohy jako je trénování umělé inteligence, hluboké učení a datová věda



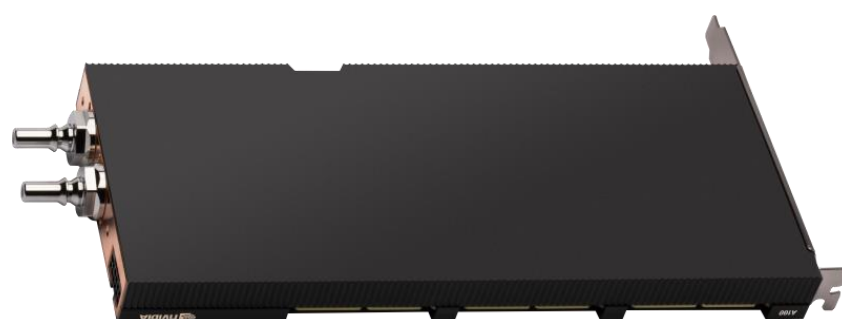
Grafické karty pro akceleraci výpočtů



NVIDIA A100 80GB

Nejvýkonnější akcelérátor 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 vysokou 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)

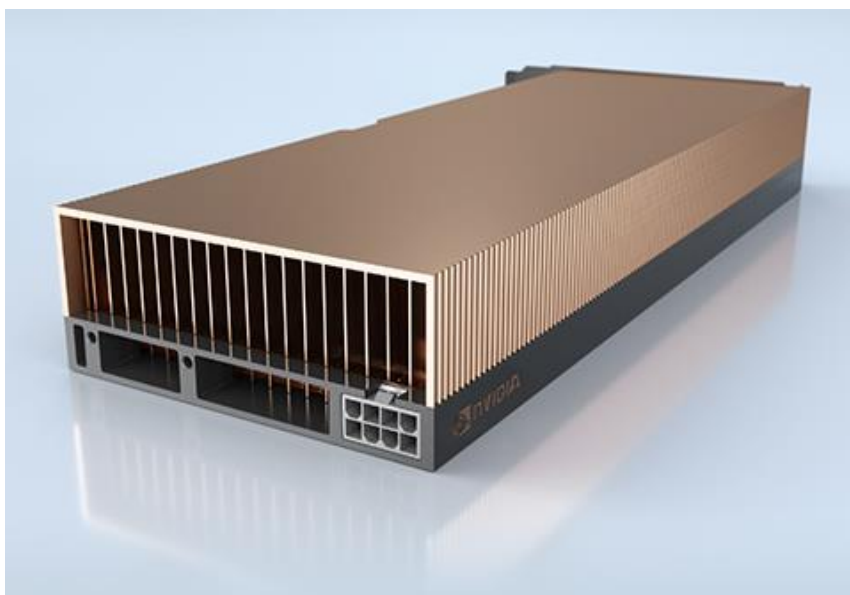


NVIDIA H100 80GB

Nejnovější a nejvýkonnější akcelérátor 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 RTX A6000 a NVIDIA A40 (generace Ampere)

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

Nabízí 3 nebo 4 display porty pro zapojení více širokoúhlých ultraHD obrazovek zároveň.

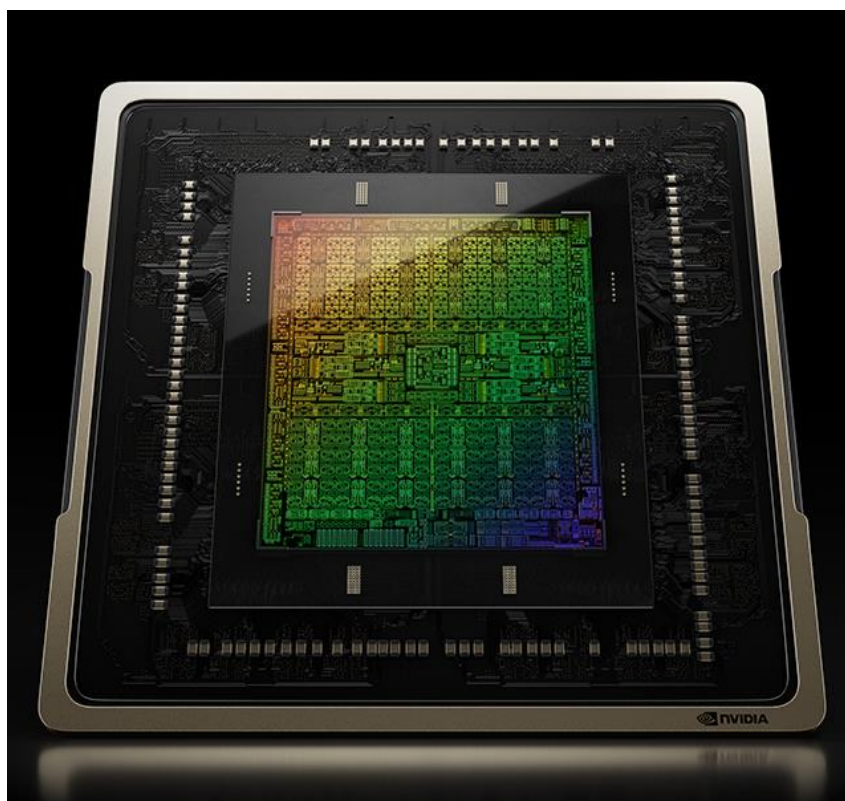
- nástupce karet **Quadro RTX 6000/8000**
- **10 752 CUDA** jader, **336 Tensor** jader
- FP32: **37,4 TFLOPS**
- grafika, renderování, simulace, VR, 3D design a AI výpočetní výkon
- čip **GA102** (RTX série 3080/3090)
- **48 GB** grafické paměti GDDR6 ECC (až 96 GB pomocí NVLinku)
- propustnost paměti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

Ve dvou variantách:

RTX A6000: aktivní chlazení vhodné do profesionálních pracovních stanic

A40: pasivní chlazení pro osazení do serverů v datových centrech (A40)

Grafické karty pro vizualizaci a akceleraci výpočtů



NVIDIA L40 a RTX 6000 Ada (2023)

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

Nabízí 4 display porty a podporu Quadro Sync pro propojení více karet a zobrazování na více širokoúhlých ultraHD obrazovkách zároveň.

- nástupce karet **NVIDIA A40 / RTX A6000**
- **18 176 CUDA** jader, **568 Tensor** jader
- FP32 **95,42 TFLOPS**
- nejnovější generace **RT**, **Tensorových** 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 virtuálních pracovních stanic **vWS** (vGPU)

Grafické karty pro vizualizaci a akceleraci výpočtů



NVIDIA RTX A5000 / A5500

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

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

Reference ČR

České vysoké učení technické v Praze

Posílení interaktivní práce na RCI klastru



ČESKÉ
VYSOKÉ
UČENÍ
TECHNICKÉ
V PRAZE

9x NVIDIA DGX Station A100 160GB

- 4x NVIDIA A100 SXM4 40 GB
- 1x AMD Epyc 7742 64 cores 3.6GHz
- 515 GB DDR4 RAM
- 1x 1.92 TB NVMe (OS)
- 1x 3.84 TB NVMe (Data)
- 4x Mini DisplayPort
- 4 GB GPU paměti
- 2x 10 Gb/s Ethernet
- 1x 1 GbE (BMC management)

1x NVIDIA DGX Station A100 320GB

- 4x NVIDIA A100 SXM4 40 GB
- 1x AMD Epyc 7742 64 cores 3.6GHz
- ...



Reference ČR

Masarykova univerzita



MUNI

NVIDIA H100 je nejnovější a nejvýkonnější akcelerátor generace Hopper.

2x NVIDIA H100 80GB PCIe

- ✓ nejvýkonnější akcelerátor současnosti
- ✓ 4nm technologie výroby
- ✓ 80 mld tranzistorů na 814 mm²
- ✓ propustnost paměti HBM3 = 3 TB/s
- ✓ 50 MB L2 cache
- ✓ 4. generace NVLinku (900GB/s)
- ✓ podpora PCIe 5.0 (128GB/s)
- ✓ 310 W
- ✓ MIG (Multi-Instance GPU) 2. generace

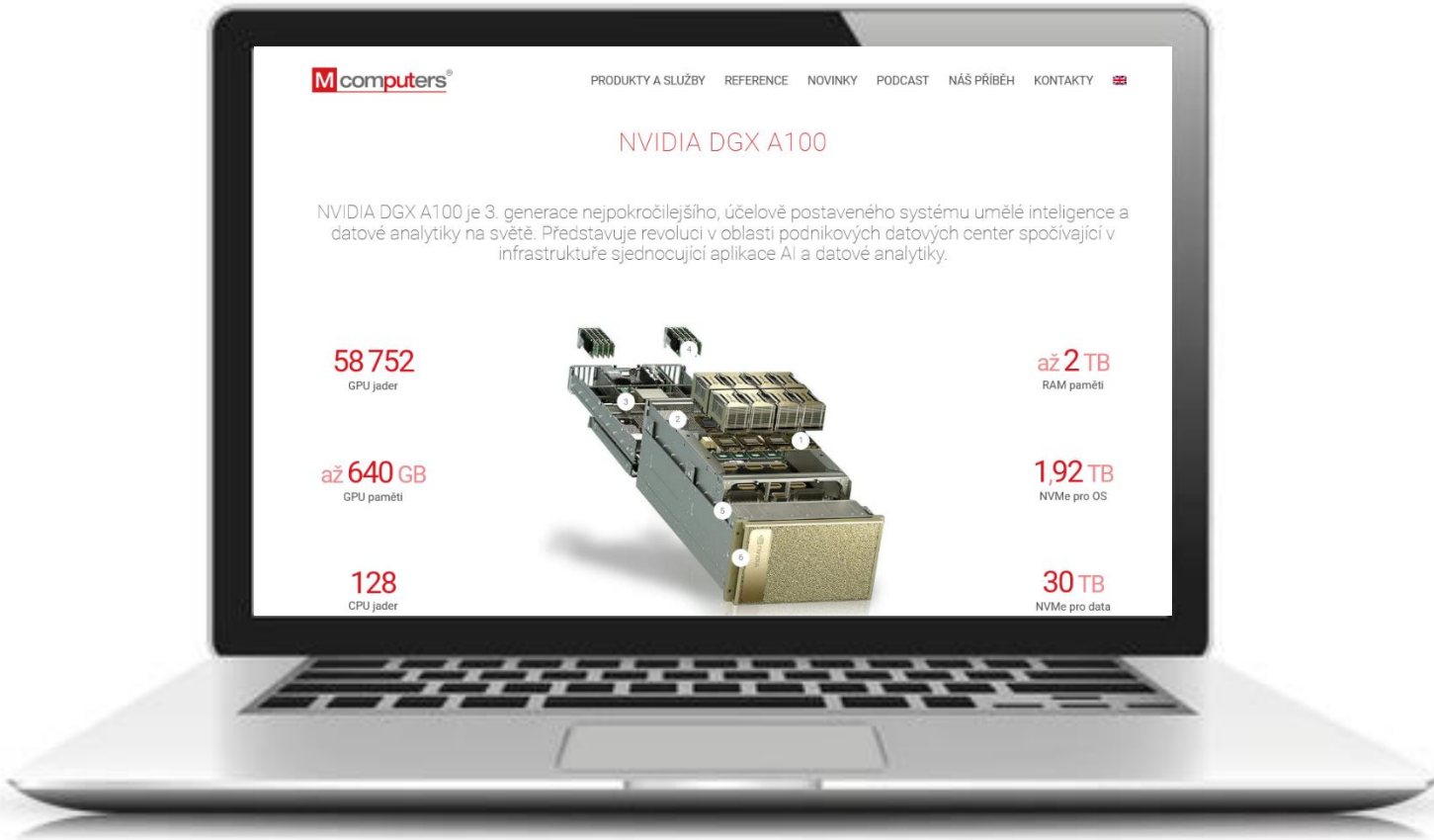


NVIDIA produkty přehledně

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 intensity	—	—	—	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*