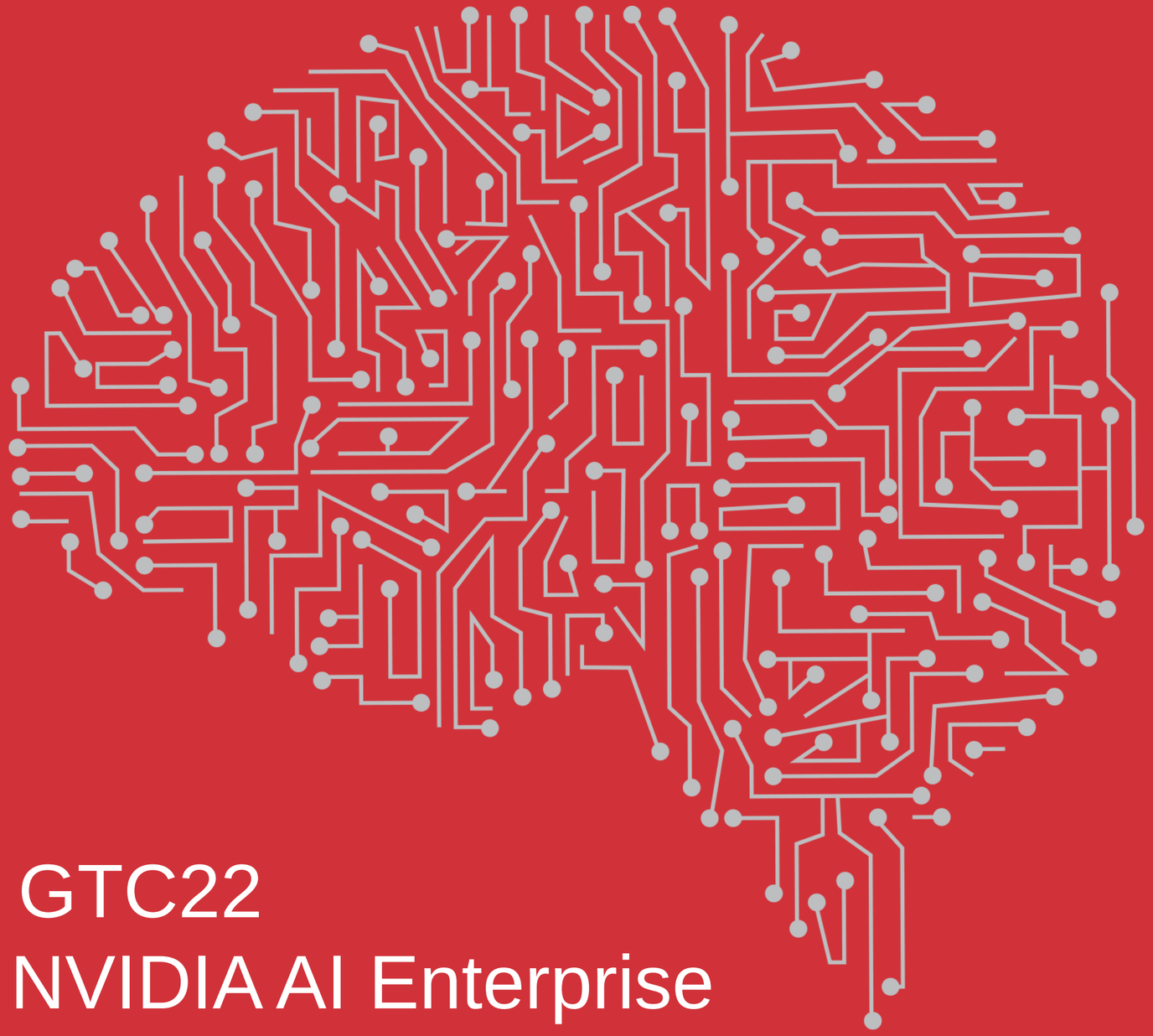


Novinky z GTC22
Lenovo & NVIDIA AI Enterprise

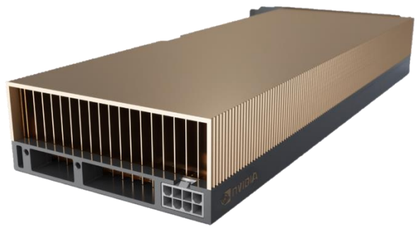
Prezentující:
Kamila Jeřábková
Petr Plodík

Lenovo NVIDIA M computers®

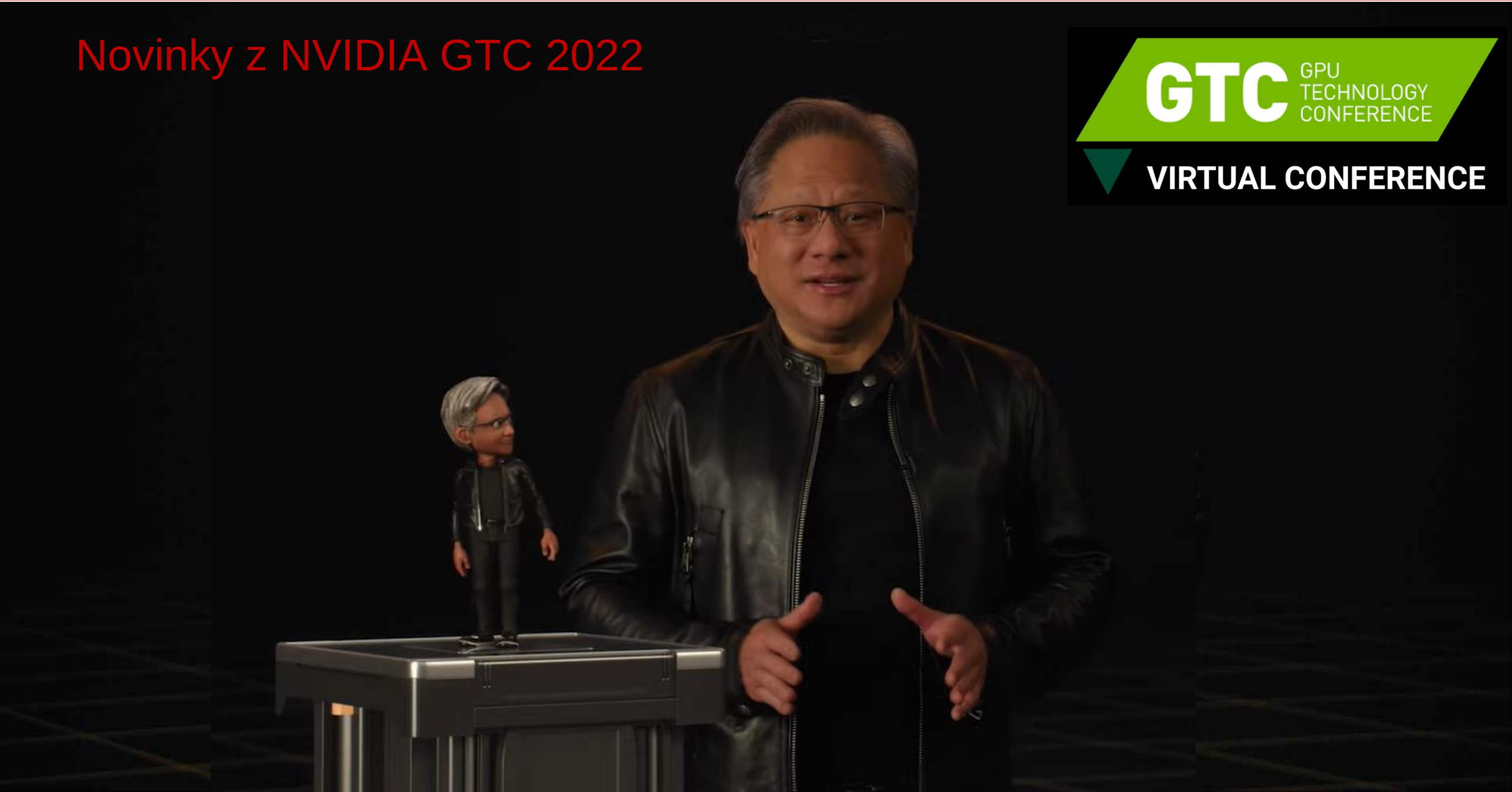
NVIDIA ELITE PARTNER

- ✓ Jsme NVIDIA ELITE partner pro region střední a východní Evropy
- ✓ NVIDIA certifikace:
  Compute DGX | Visualization (RTX) | Virtualization (vGPU)
- ✓ Nabízíme slevy pro vysoké školy, výzkumné instituce a start-upy
- ✓ Demo pool nejnovějších NVIDIA technologií
  NVIDIA DGX Station (4x V100)
  NVIDIA DGX Station A100 (4x A100)
  Akcelerátory V100, A100, A40, A10, A30, **A2**, **A16**
  vGPU licence (vApp, vPC, vWS, vCS), **AI Enterprise**

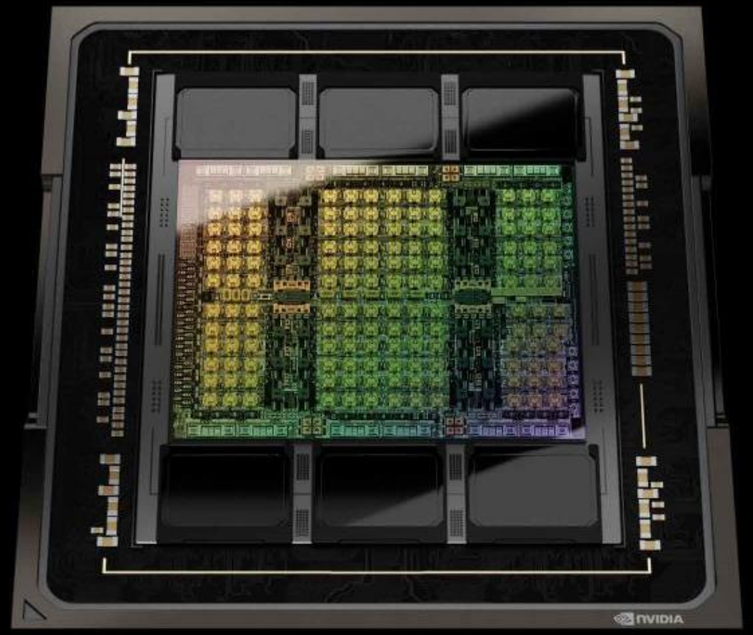
NEW

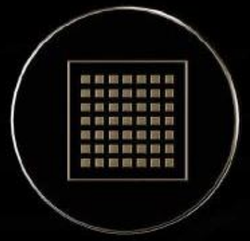


Novinky z NVIDIA GTC 2022

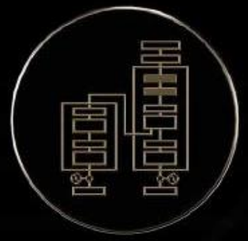


Architektura HOPPER

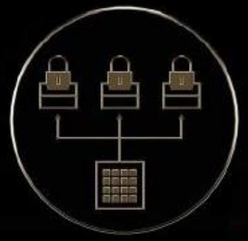




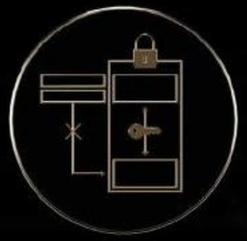
WORLD'S MOST ADVANCED CHIP
80 Billion Transistors



TRANSFORMER ENGINE
6X Transformer Performance



2ND GENERATION MULTI-INSTANCE GPU
7X Secure Tenants



CONFIDENTIAL COMPUTING
Secure Data and AI Models In-Use

4TH GENERATION NVLINK
7X PCIe Gen5

DPX INSTRUCTIONS
7X Dynamic Programming Performance

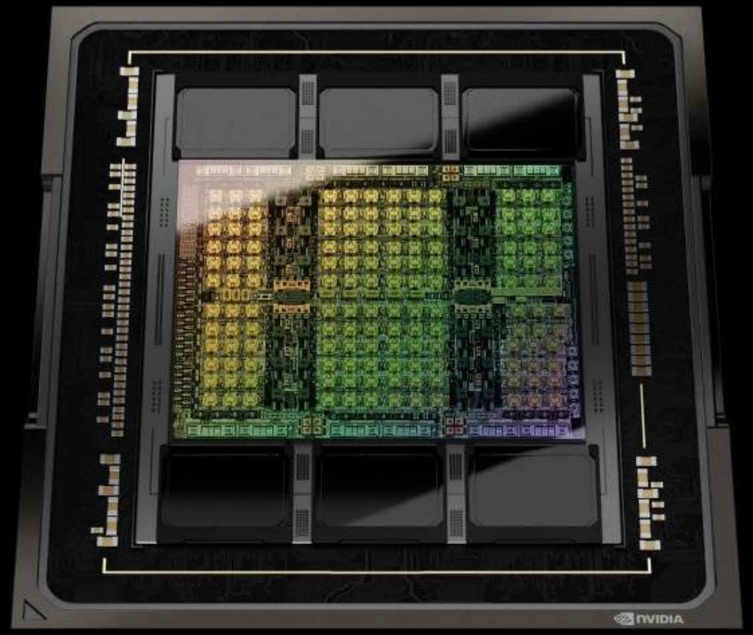
Akcelerátor NVIDIA H100

NVIDIA H100 80GB HBM3

- 4nm technologie výroby
- 80 mld tranzistorů na 814 mm^2
- propustnost paměti HBM3 = 3 TB/s
  (HBM2 u A100 měla “jen” 2TB/s, PCIe 1,5TB/s)
- 4. generace NVLinku (900GB/s)
- podpora PCIe 5.0 (128GB/s)

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	30	24	9,7
FP64 Tensor	60	48	19,5
FP32	60	48	19,5
TF32	500/1000*	400/800*	156/312*
FP16	120	96	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

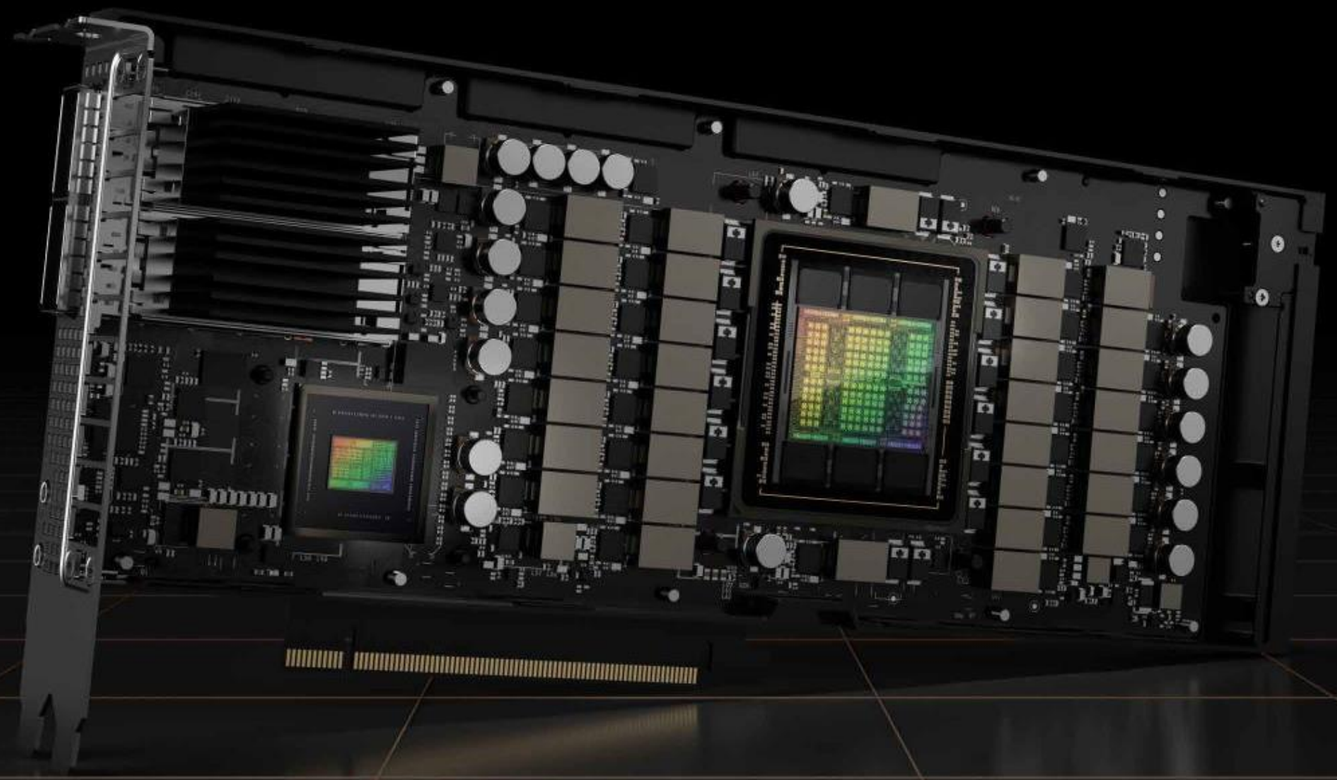
NVIDIA H100 CNX

ANNOUNCING
H100 CNX

Converged H100 and CX-7 SmartNIC

7 Multi-GPU Instances

PCIe Gen5



NVIDIA DGX H100

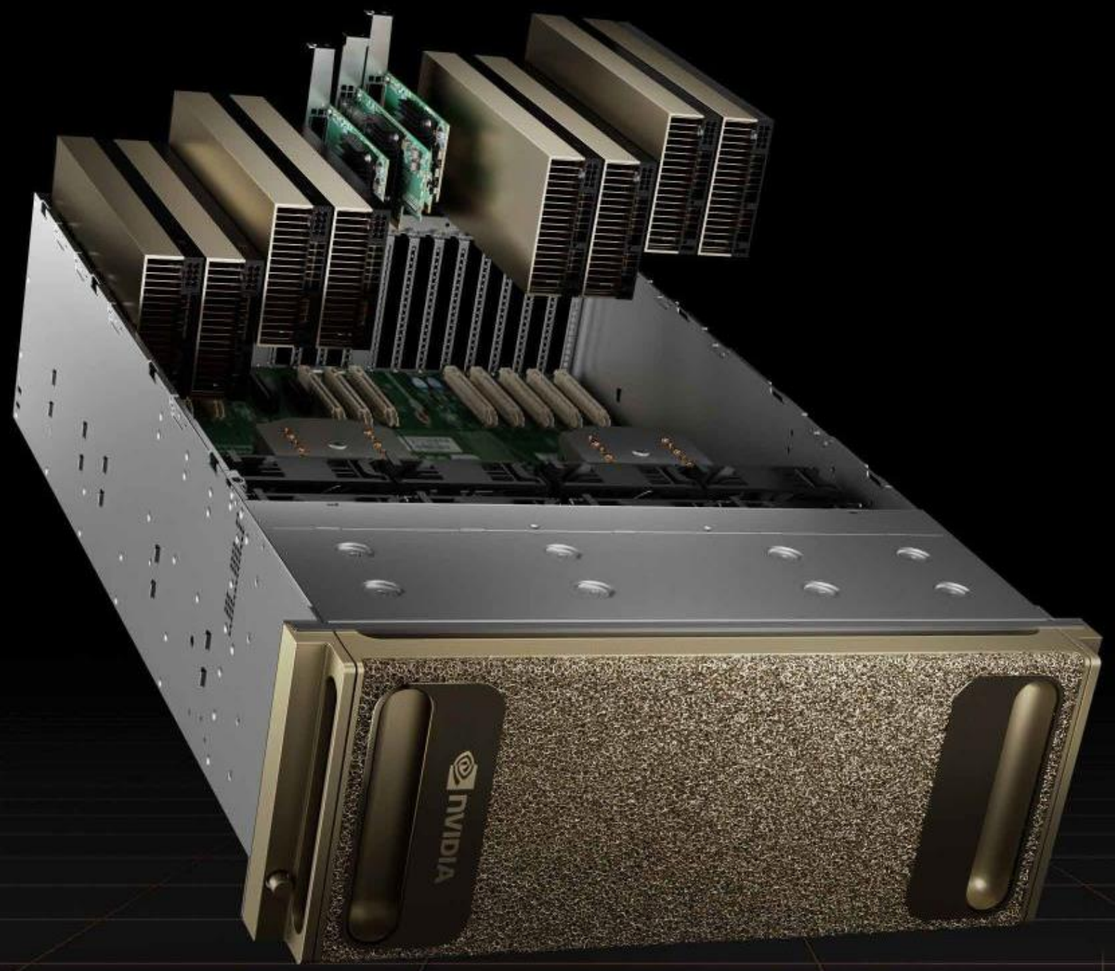
INTRODUCING
DGX H100

8 H100 GPUs | 640 Billion Transistors
32 PFLOPS of AI Performance
640 GB of HBM3 Memory
24 TB/s of Memory Bandwidth

- CPU - dual x86
- 2TB RAM (DDR5?)
- 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 BlueField-3 DPU VPI (400Gb/s IB/Eth)
- 10Gb/s onboard NIC (RJ45) (volitelně 50Gb/s Eth)
- Management (BMC) (RJ45)
- 2x NVIDIA BlueField-3 DPU BMC (RJ45)
- Spotřeba max. 10,2kW

AI performance is sparse FP8.

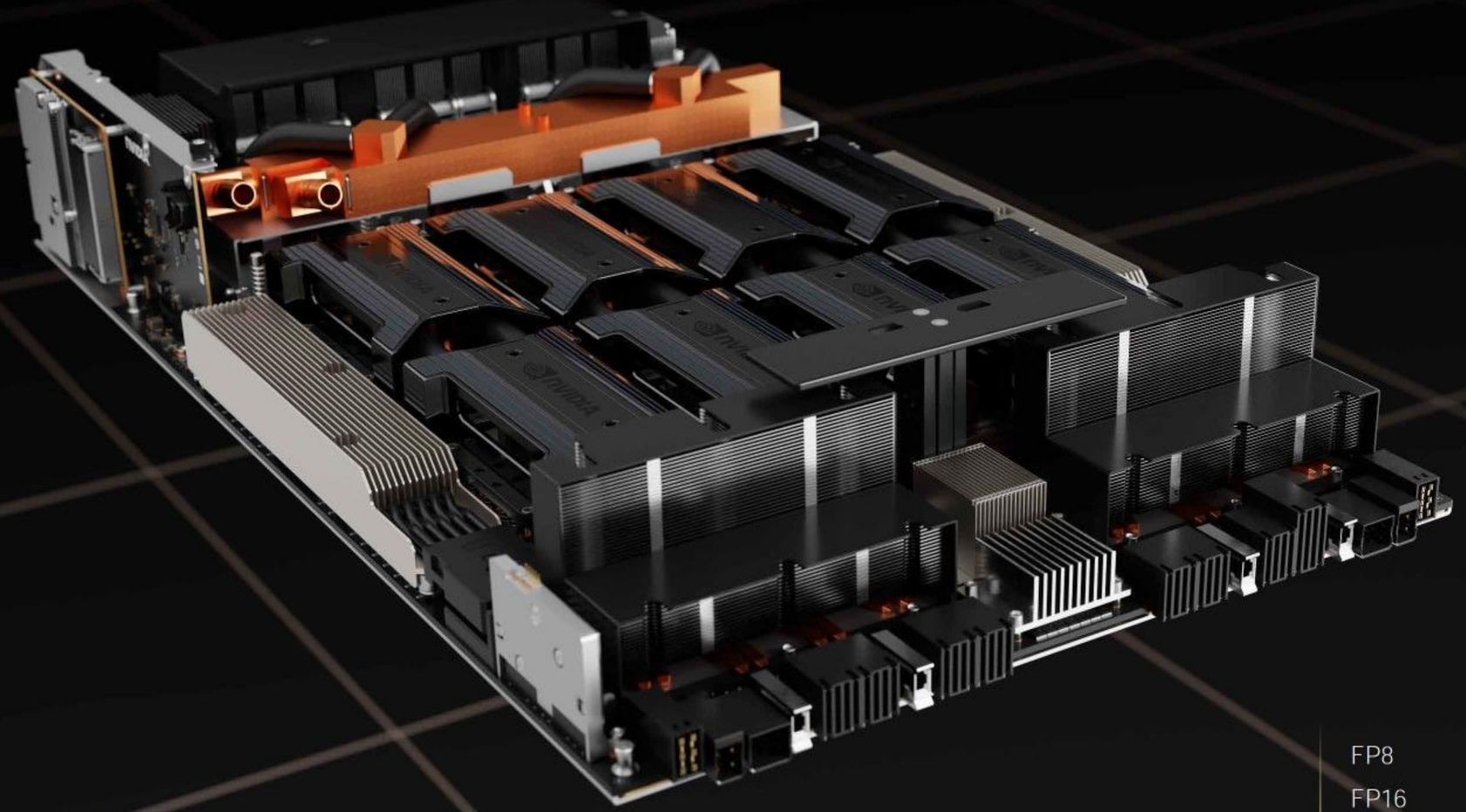
NVIDIA OVX



NVIDIA OVX SERVER

- 8 NVIDIA A40 GPUs
- 3 NVIDIA ConnectX-6 200Gbps NICs
- 2 Intel Ice Lake 8362 CPUs
- 1 TB System Memory and 16 TB NVMe Storage

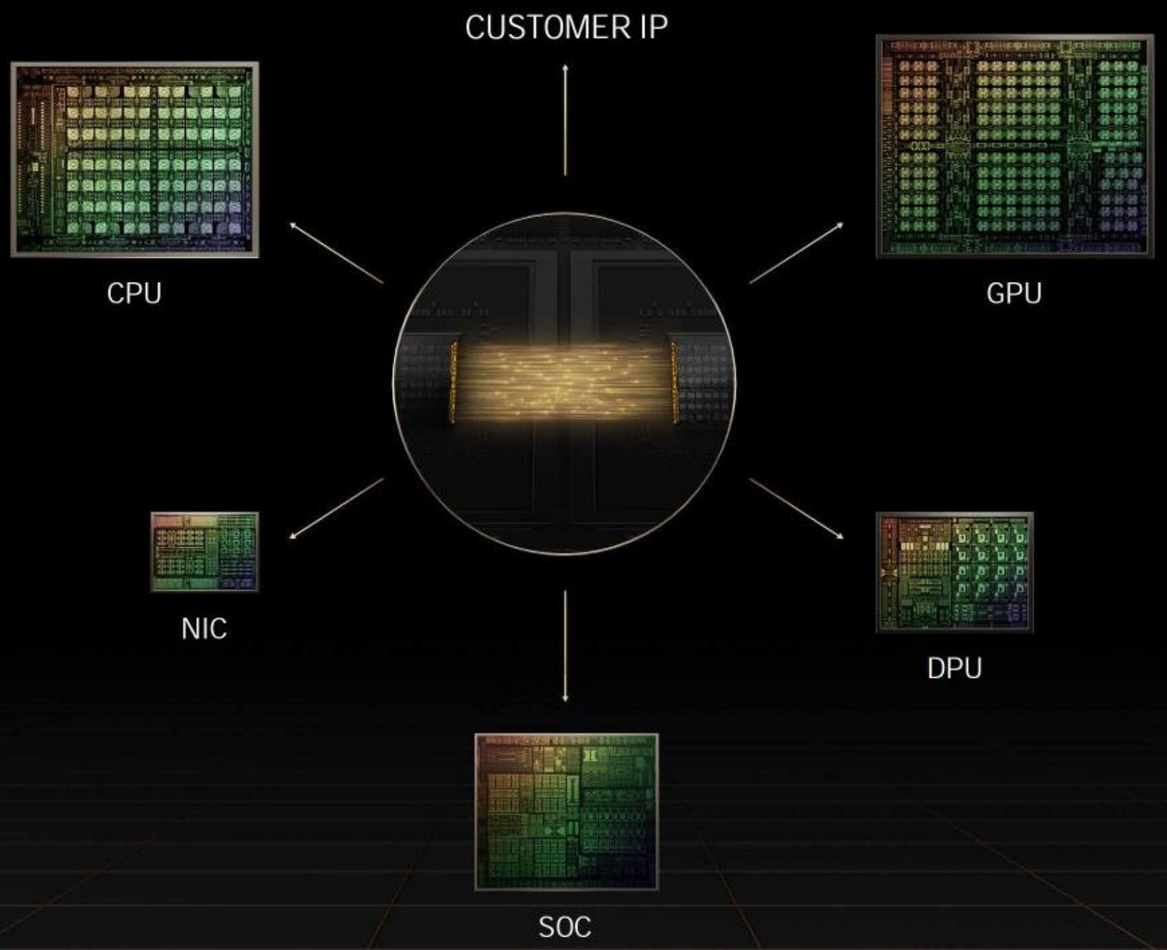
NVIDIA HGX H100



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 NVLink



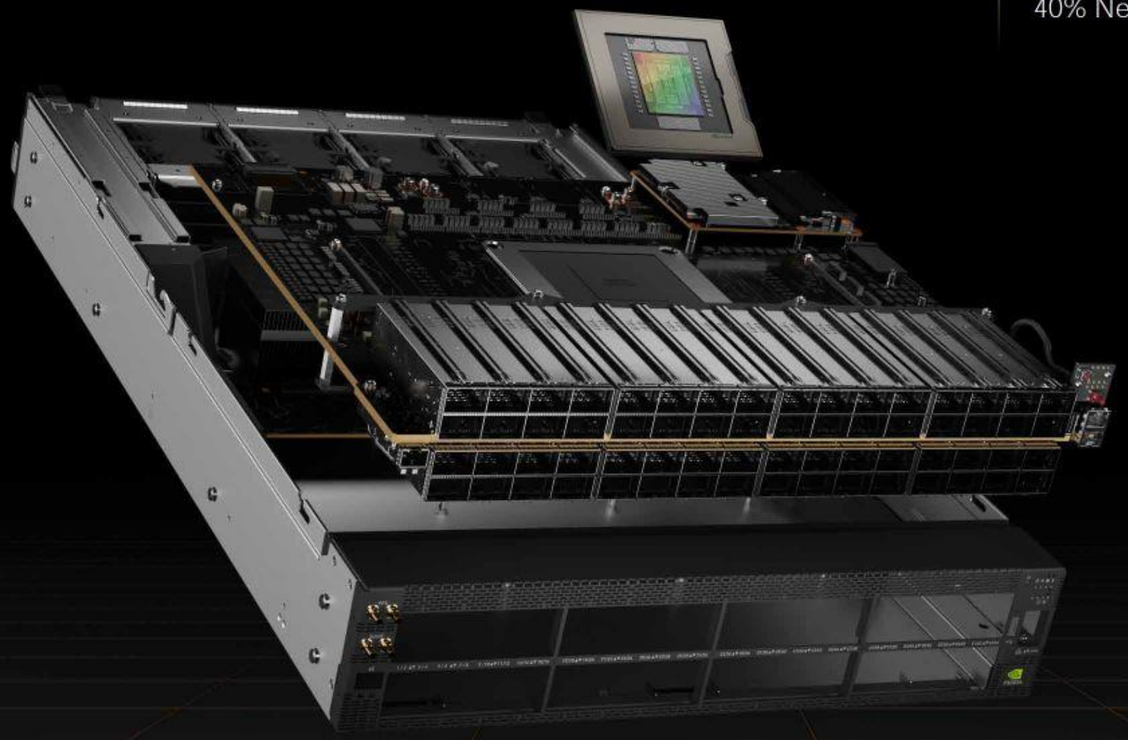
NVLINK OPENS A WORLD OF SEMI-CUSTOM CHIPS AND SYSTEMS

NVIDIA NVLink

NVLINK C2C OFFERS MULTITUDE OF SYSTEM ARCHITECTURES

NVIDIA SPECTRUM-4

SPECTRUM-4 400G
ETHERNET SWITCH



100 Billion Transistors – TSMC 4N Process

51.2T Bandwidth | 12.8T Line Rate Crypto | 100G SerDes

128 Ports of 400GbE

4X Higher Switching Throughput

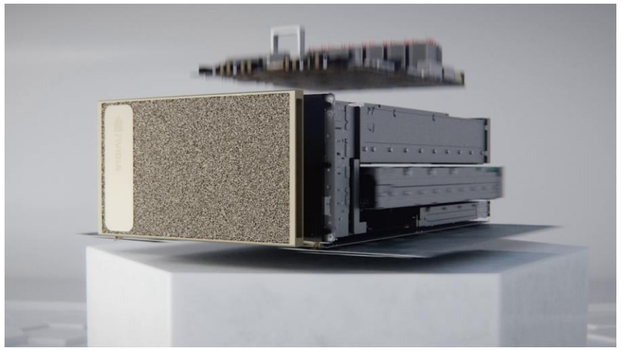
40% Network Power Savings

Nově oznámené produkty

NVIDIA DGX A100 refresh

Rozšíření možností konektivity pro systémy NVIDIA DGX A100 o ConnectX-7 (InfiniBand nebo Ethernet) síťové adaptéry.

- ✓ **ConnectX-7 InfiniBand** adaptéry pro propojení DGX systémů do clusterů
- ✓ **ConnectX-7 VPI** (InfiniBand nebo Ethernet) adapréry pro storage a networking



NVIDIA RTX A5500

Ideální pro provoz vysoce výkonných virtuálních pracovních stanic (vWS) s profesionálními vizualizačními aplikacemi nebo v kombinaci se softwarem NVIDIA Virtual PC (vPC) pro multimediální virtuální desktopy.

- ✓ **24 GB** paměti GDDR6 s podporou ECC
- ✓ propustnost GPU paměti **768 GB/s**
- ✓ **10 240 CUDA** jader a **320 Tensor** jader
- ✓ **80 RT** jader s výkonem **66,6 TFLOPS**
- ✓ podpora **PCIe Gen 4**
- ✓ grafický výstup **4x DisplayPort**
- ✓ aktivní dual-slot karta se spotřebou **230W**
- ✓ výkon **34,1 TFLOPS** (FP32) a až **272,8 TFLOPS** (TF32) s využitím sparsity

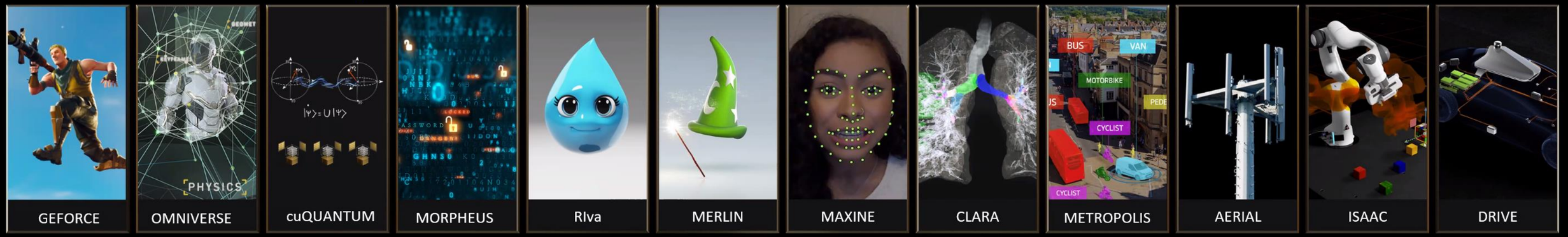
NVIDIA Jetson AGX Orin

Jednodeskový počítač pro vývoj pokročilých AI robotů a dalších autonomních strojů v různých odvětvích, jako je průmyslová výroba, chytrá města, retail, zdravotnictví, internet věcí, atp.

- ✓ **275 TOPS** pro AI výpočty
- ✓ až **32GB DRAM** paměť
- ✓ 12-core Arm Cortex-A78AE 64-bit CPU
- ✓ **204.8GB/s** propustnost paměti
- ✓ Jetson AGX Orin Developer Kit - $1999
- ✓ Jetson AGX Orin 64GB - $1599*
- ✓ Jetson AGX Orin 32GB - $899*
- ✓ Jetson Orin NX 16GB - $599*
- ✓ Jetson Orin NX 8GB - $399*
- ✓ dostupnost Q4 2022, developer kit již nyní

NVIDIA PLATFORM

Comprehensive E2E Hardware and Software Stack



GEFORCE | OMNIVERSE | cuQUANTUM | MORPHEUS | RIva | MERLIN | MAXINE | CLARA | METROPOLIS | AERIAL | ISAAC | DRIVE

APPLICATION FRAMEWORKS

PLATFORM SOFTWARE

NVIDIA AI

NVIDIA GPU CLOUD REGISTRY

cuIO	cuDF	cuDNN	cuML	cuGraph	TAO	TensorRT	Triton

NVIDIA CUDA, DOCA, vGPU, Magnum IO

RTX | DGX | EGX | AGX

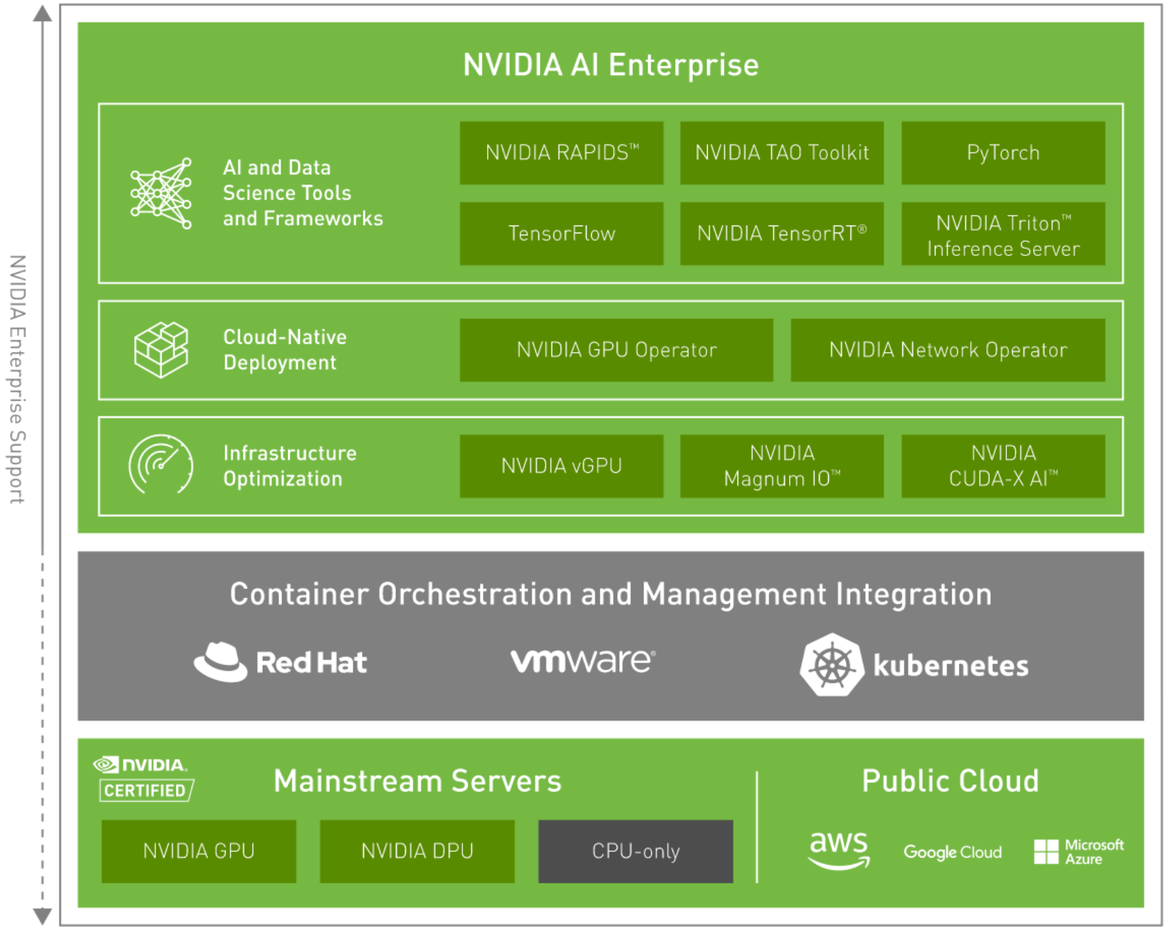
CHIPS & SYSTEMS

NVIDIA AI Enterprise 2.0

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 certifikovanými systémy (NVIDIA-Certified Systems). Zahrnuje technologie pro rychlé nasazení, správu a škálování AI výpočtů v hybridním cloudu.

- ✓ licencováno per CPU socket
- ✓ podporované akcelerátory:
  - compute: HGX A100, A100, A30, V100s/V100, nově A100X, A30X a A2
  - mixed workloads: A40, A10, A16, T4, RTX A6000, RTX A5000
- ✓ certifikováno pro provoz na VMware vSphere
- ✓ 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 AI Enterprise

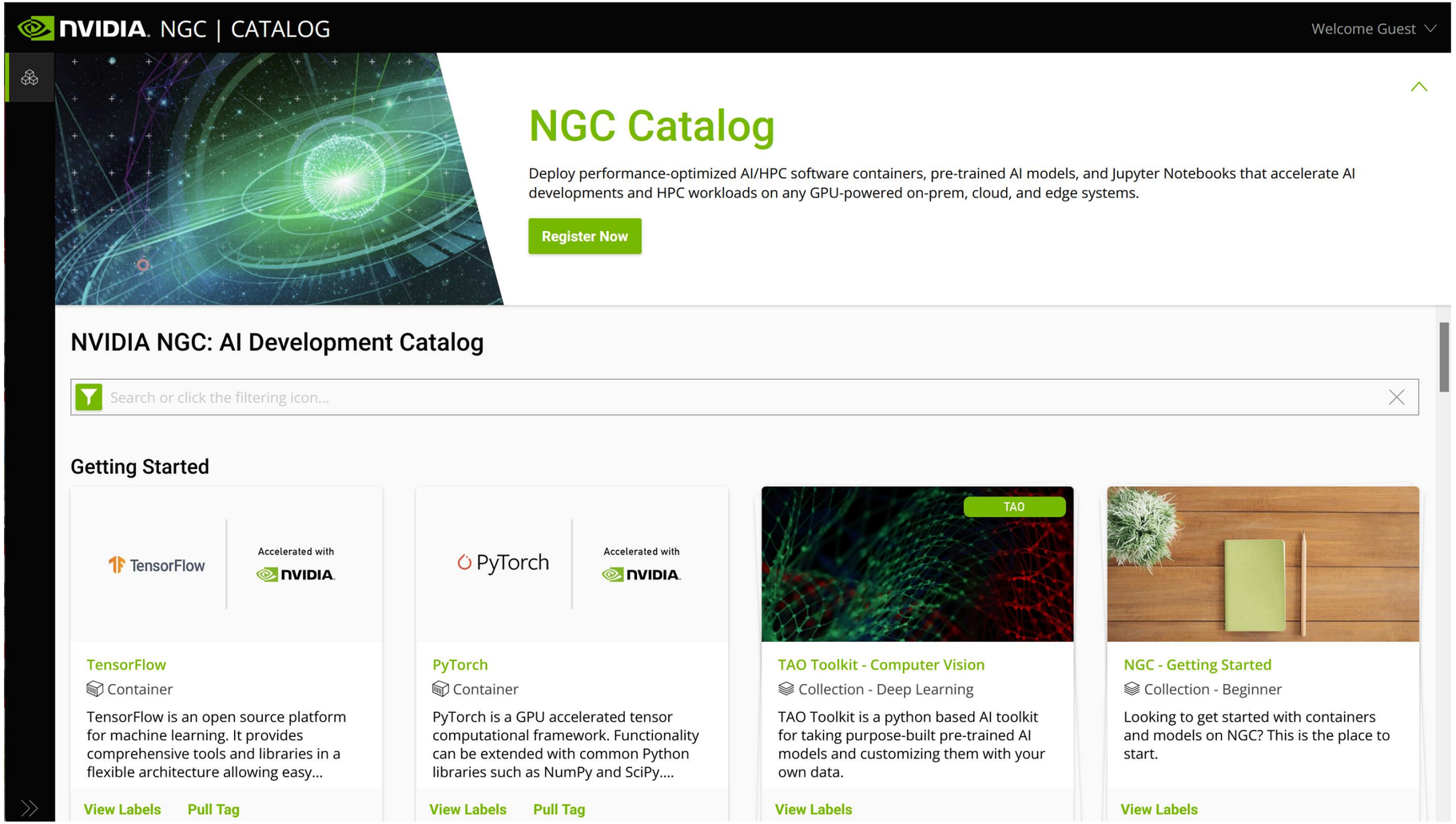
Prostředí pro nasazení HPC & AI aplikací na platformě VMware
Sdílení výkonných GPU akcelerátorů více virtuálními servery
Inspirace NVIDIA DGX systémy – AI appliance
Zprovoznění vyladěného AI prostředí v řádu hodin
Podpora od NVIDIA na celý software stack včetně open-source knihoven, frameworků
Napojení na NVIDIA GPU katalog (NGC)

https://docs.nvidia.com/ai-enterprise/deployment-guide/dg-software.html
https://www.nvidia.com/en-us/data-center/products/ai-enterprise/

NVIDIA AI ENTERPRISE LICENSE TYPE	SUGGESTED PRICE	Support Services
Subscription, 1-year term	$2,000 / CPU socket	Business Standard Included +$450 for Business Critical
Subscription, 3-year term	$6,000 / CPU socket	Business Standard Included +$1,350 for Business Critical
Subscription, 5-year term	$8,000 / CPU socket	Business Standard Included +$2,250 for Business Critical
Perpetual License + 1-Year Support	$4,494 / CPU socket	Business Standard Included +$450 for Business Critical
Perpetual License + 3-Year Support	$6,292 / CPU socket	Business Standard Included +$1,350 for Business Critical
Perpetual License + 5-year support services	$8,090 / CPU socket	Business Standard Included +$2,250 for Business Critical

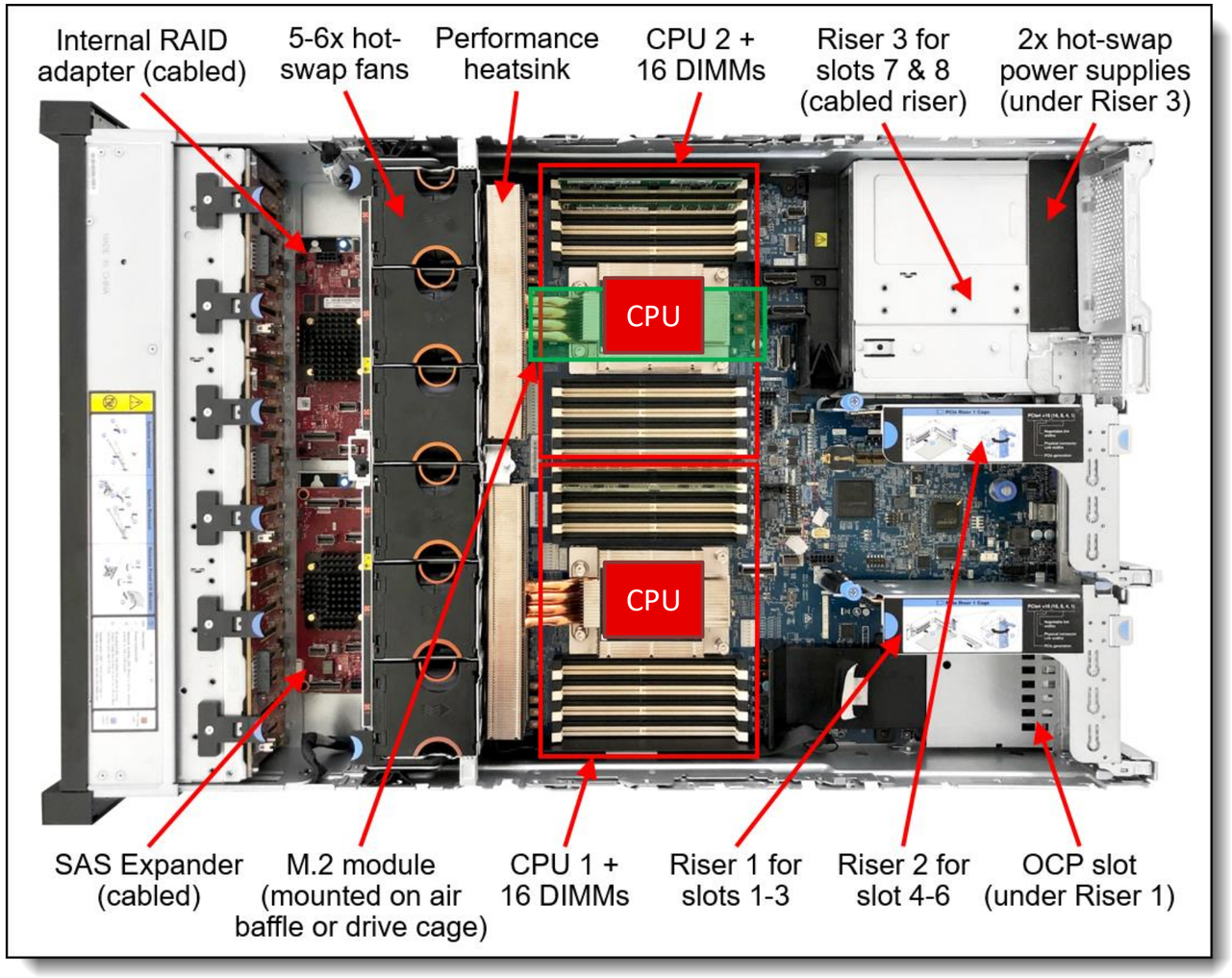
NVIDIA AI ENTERPRISE OFFERING	SUGGESTED EDU PRICE
Subscription, 1-Year Term	$500 / CPU socket
Subscription, 3-Year Term	$1,500 / CPU socket
Subscription, 5-Year Term	$2,000 / CPU socket
Perpetual License + 1 year support	$1,124 / CPU socket
Perpetual License + 3 years support	$1,573 / CPU socket
Perpetual License + 5 years support	$2,022 / CPU socket

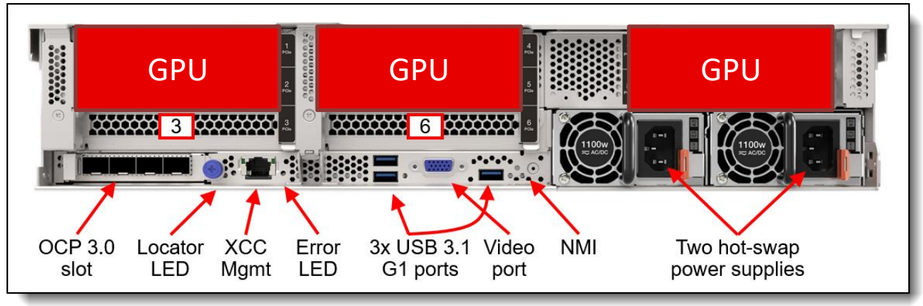
NVIDIA GPU katalog (NGC)



https://catalog.ngc.nvidia.com/

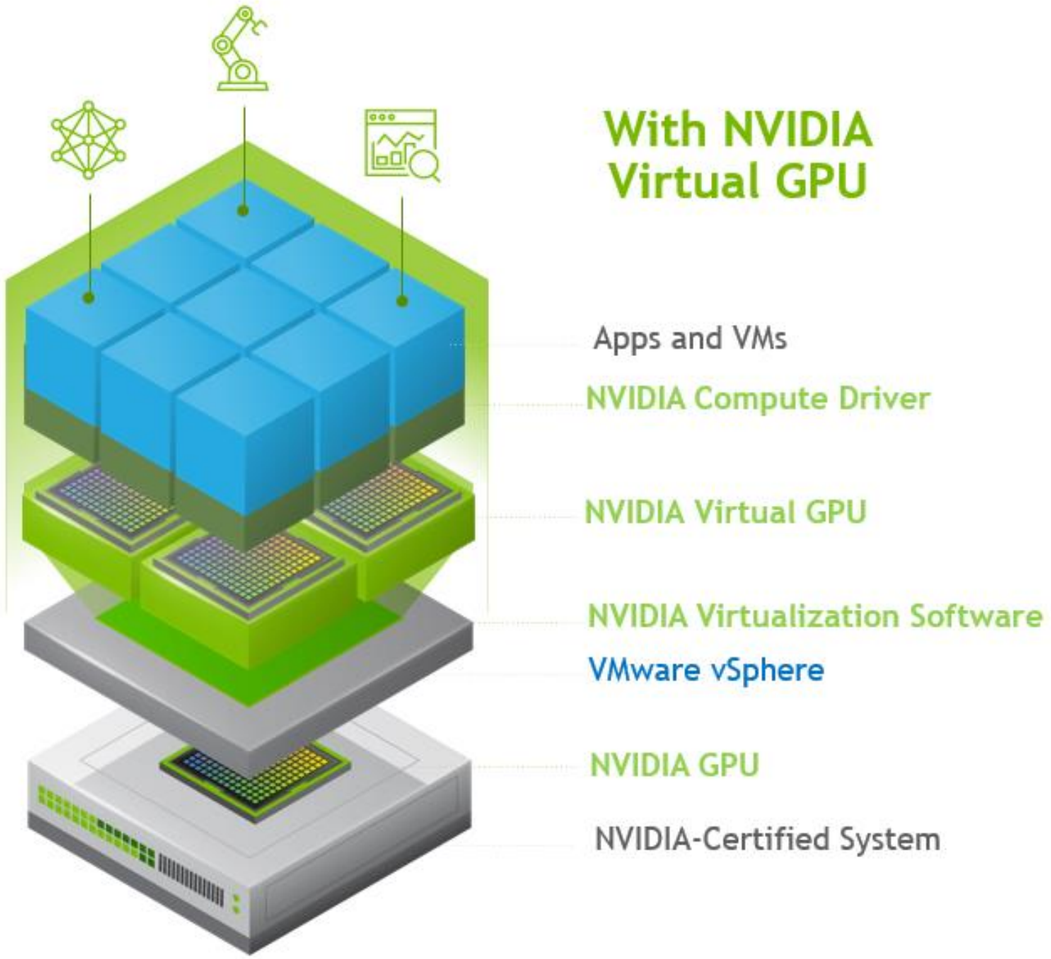
Lenovo ThinkSystem SR665 – dual-socket, AMD EPYC, 3 DW GPUs





Lenovo

NVIDIA vGPU



Concurrent User License Annual Subscription Pricing	
License	Annual Subscription Pricing
NVIDIA Virtual Applications	$10 per CCU subscription
NVIDIA Virtual PC	$50 per CCU subscription
	4 year subscription - $43.75 per CCU
	5 year subscription - $40 per CCU
NVIDIA RTX Virtual Workstation	$250 per CCU subscription
	4 year subscription - $206.25 per CCU
	5 year subscription - $180 per CCU

<< Cena za jednoho připojeného uživatele za rok, již vč. supportu

<< Při objednávce předplatného na více let, je cena na rok nižší

Enterprise Perpetual Licensing plus SUMS Pricing	
License	Annual Licensing plus SUMS Pricing
NVIDIA Virtual Applications	$20 per CCU perpetual license
	$5 SUMS per year
NVIDIA Virtual PC	$100 per CCU perpetual license
	$25 SUMS per year
NVIDIA RTX Virtual Workstation	$450 per CCU perpetual license
	$100 SUMS per year

<< Cena za jednoho připojeného uživatele za neomezenou dobu, je nutné přikoupit support (SUMS) na 5 let

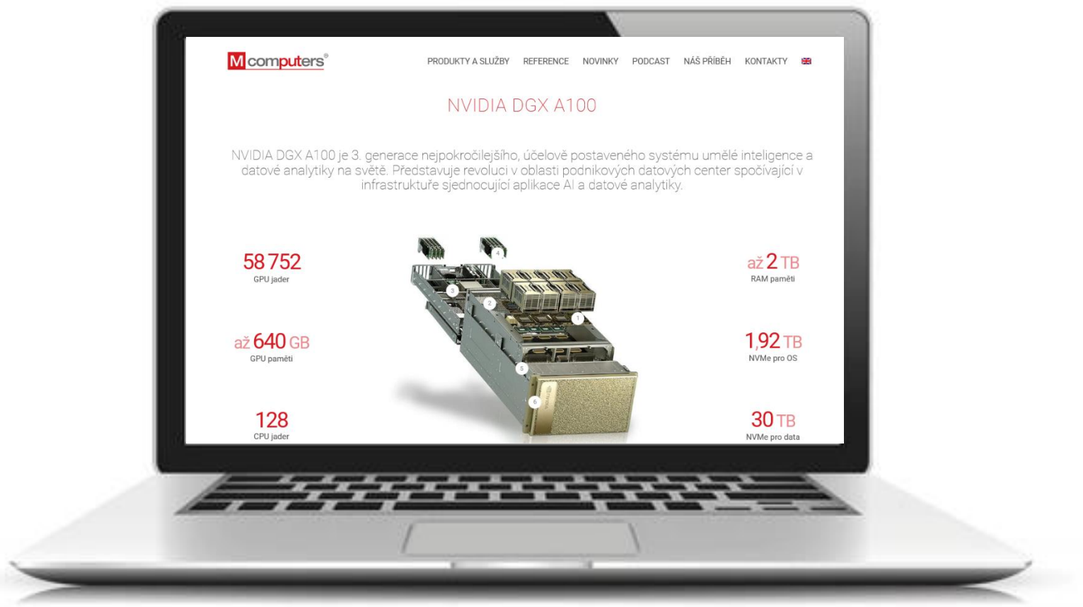
Min. objednávka SUMS je 5 let = 500 USD za uživatele

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*