NVIDIA – novinky a trendy

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

NVIDIA ELITE PARTNER

- ✓ NVIDIA kompetence:
  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)

- ✓ NVIDIA technologie pro testování
  - NVIDIA DGX Station A100 (4x A100)
  - Akcelerátory (H100, A100, A40, A30, A2, A16, L4, L40)
  - vGPU licence (vApps, vPC, vWS, vCS)
  - AI Enterprise licence

Kompletní 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,..

Kompletní softwarová vrstva pro instalaci, optimalizaci, management a sdílení akcelerovaných 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.

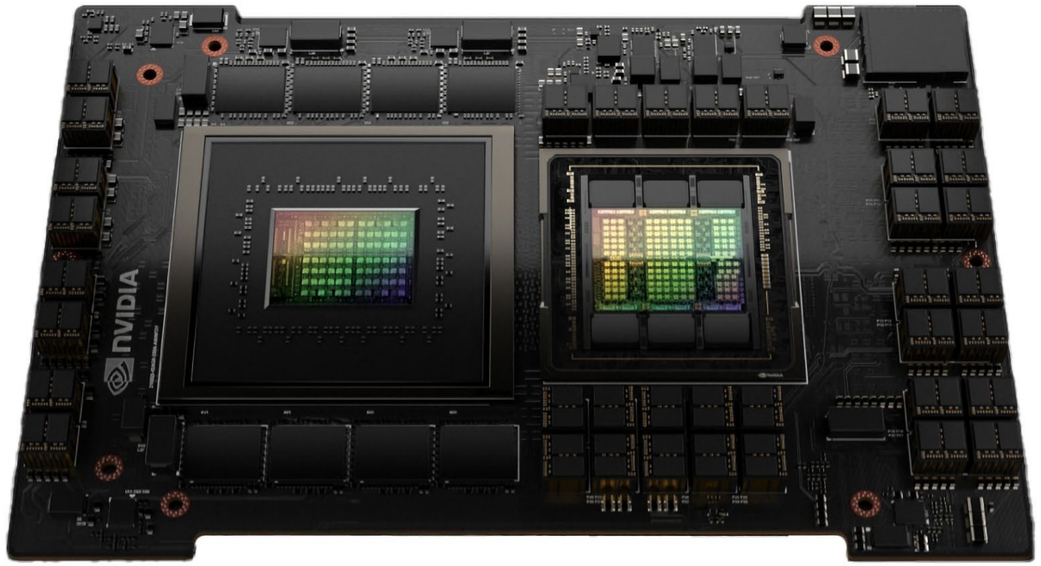
Superchip Grace

NVIDIA Grace Superchip

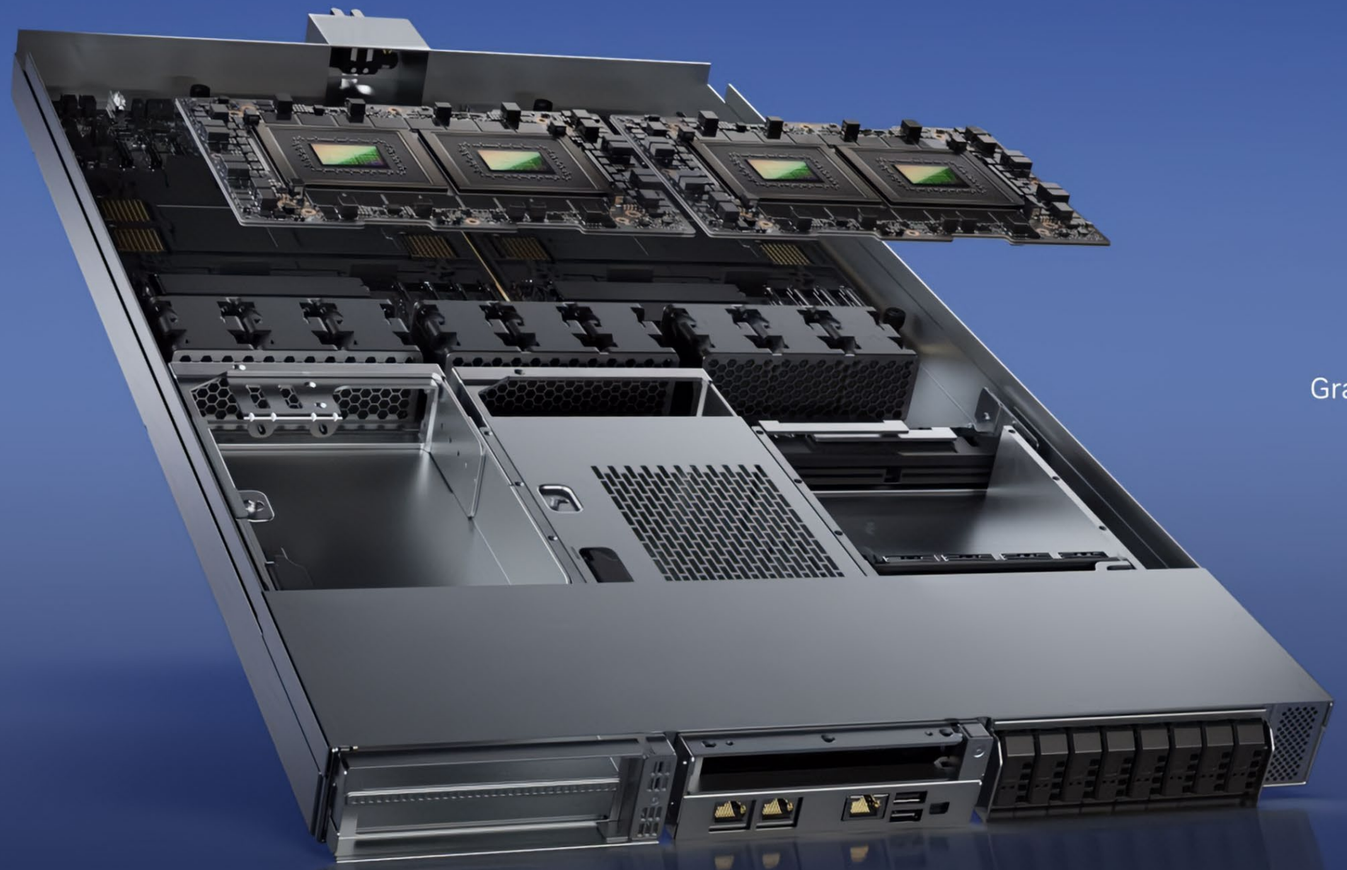
Superchip Grace jsou dva čipy ARM procesoru Grace na jedné desce. Oba CPU čipy jsou propojeny vysokorychlostním C2C (Chip-to-Chip) NVLinkem. Tato technologie umožňuje propojení různých čipů Nvidia, tedy nejenom CPU+CPU, ale i např. CPU+GPU, CPU+GPU+NIC

- ✓ High-performance **CPU** pro **HPC** a **cloud** computing
- ✓ Super chip design s **144 Arm v9** CPU jádry (2x72 jader)
- ✓ LPDDR5x s ECC paměť s celkovou propustností **1TB/s**
- ✓ **SPECrate2017_int_base** přes **740** (přibližný teoretický výkon)
- ✓ **FP64** až **7,1** TFLOPs
- ✓ **900 GB/s** C2C interface, 7x rychlejší než PCIe Gen 5
- ✓ 2x vyšší výkon per watt ve srovnání s nynějšími CPU (**500W TDP** with memory)
- ✓ Podpora všech NVIDIA softwarových stacků a platforem (RTX, HPC, AI, Omniverse)

	NVIDIA Grace CPU Superchip architecture
Core architecture	Neoverse V2 Cores: Armv9 with 4x128b SVE2
Core count	144
Cache	L1: 64 KB I-cache + 64 KB D-cache per core L2: 1 MB per core L3: 234 MB per superchip
Memory technology	LPDDR5X with ECC, co-packaged
Raw memory BW	Up to 1 TB/s
Memory size	Up to 960 GB
FP64 peak	7.1 TFLOPS
PCI Express	8x PCIe Gen 5 x16 interfaces; option to bifurcate Total 1 TB/s PCIe bandwidth. Additional low-speed PCIe connectivity for management.
Power	500 W TDP with memory, 12 V supply



Grace CPU Superchip 1U Server



Grace CPU Superchip 1U Server

NVIDIA GRACE
VS
NEXT-GEN X86

	Performance	Energy Efficiency	Data Center Throughput
Microservices	1.3X	1.7X	2X
Big Data	1.2X		

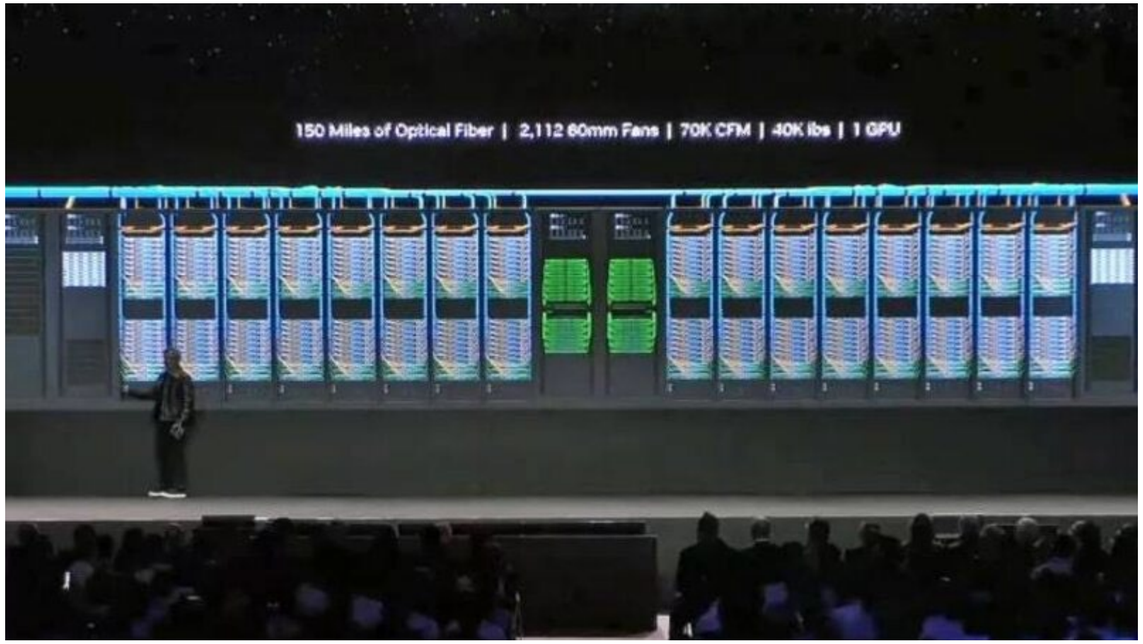
ASUS Atos GIGABYTE Hewlett Packard Enterprise

QCT Supermicro wistron zt Systems

DGX GH200 AI Supercomputer

NVIDIA GH200

- ✓ **256** NVIDIA Grace Hopper Superchipů navzájem propojených NVLinkem
- ✓ celková sdílená GPU paměť **144 TB**
- ✓ **900 GB/s** celková propustnost (GPU-to-GPU)
- ✓ 1 exaFLOPS výkon pro AI výpočty (FP8)
- ✓ NVIDIA Base Command, NVIDIA AI Enterprise software
- ✓ Referenční architektura



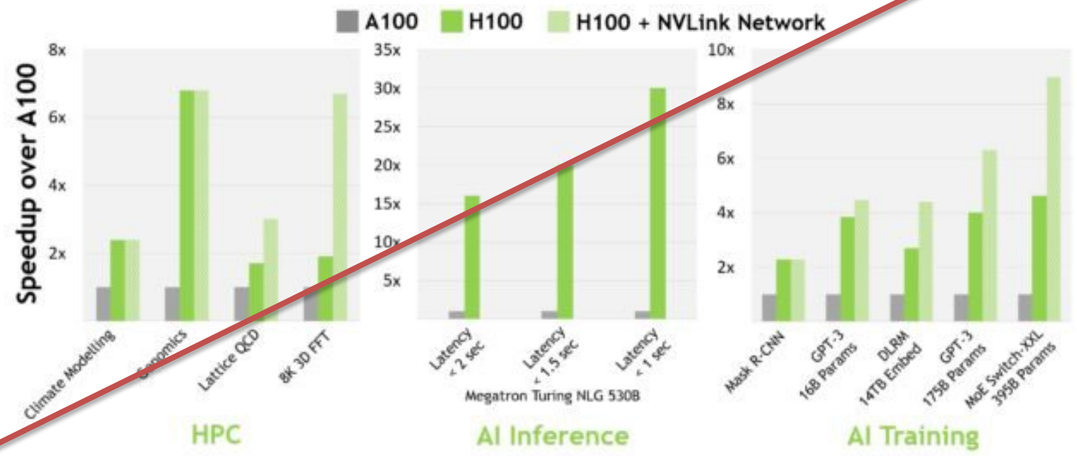
	NVIDIA DGX GH200
CPU and GPU	256x NVIDIA Grace Hopper Superchips
CPU Cores	18,432 Arm Neoverse V2 Cores with SVE2 4X 128b
GPU Memory	GPU Memory 144TB
Performance	Performance 1 exaFLOPS
Networking	256x OSFP single-port ConnectX®-7 VPI with 400Gb/s InfiniBand 256x dual-port BlueField®-3 VPI with 200Gb/s InfiniBand and Ethernet 24x Quantum-2 QM9700 InfiniBand Switches 20x Spectrum™ SN2201 Ethernet Switches 22x Spectrum SN3700 Ethernet Switches
NVIDIA NVLink Switch System	NVIDIA NVLink Switch System 96x L1 NVIDIA NVLink Switches 36x L2 NVIDIA NVLink Switches
Management Network	Host baseboard management controller (BMC) with RJ45
Software	NVIDIA AI Enterprise (optimized AI software) NVIDIA Base Command (orchestration, scheduling, and cluster management) DGX OS / Ubuntu / Red Hat Enterprise Linux / Rocky (operating system)
Support	Comes with three-year business-standard hardware and software support

Architektura Hopper

NVIDIA H100 NVL 94 GB HBM3e

Optimalizovaný dual akcelerátor pro Transformer modely a pro trénování LLM (Large Language) modelů (např. GPT)

- **188 GB** paměti HBM3e (2x 94 GB propojených NVLinkem)
- propustnost paměti **7,8 TB/s** (2×3,9 TB/s)
  (A100 měla “jen” 2TB/s, PCIe 1,5TB/s)
- 6 aktivních čipů HBM3
- podpora **PCIe 5.0** (128GB/s)
- 14x 12GB **MIG** profilů
- konfigurovatelné TDP **2× 350-400 W**



Form Factor	H100 SXM	H100 PCIe	H100 NVL[2]
FP64	34 teraFLOPS	26 teraFLOPS	68 teraFLOPs
FP64 Tensor Core	67 teraFLOPS	51 teraFLOPS	134 teraFLOPs
FP32	67 teraFLOPS	51 teraFLOPS	134 teraFLOPs
TF32 Tensor Core	989 teraFLOPS	756teraFLOPS	1,979 teraFLOPs
BFLOAT16 Tensor Core	1,979 teraFLOPS	1,513 teraFLOPS	3,958 teraFLOPs
FP16 Tensor Core	1,979 teraFLOPS	1,513 teraFLOPS	3,958 teraFLOPs
FP8 Tensor Core	3,958 teraFLOPS	3,026 teraFLOPS	7,916 teraFLOPs
INT8 Tensor Core	3,958 TOPS	3,026 TOPS	7,916 TOPS
GPU memory	80GB	80GB	188GB
GPU memory bandwidth	3.35TB/s	2TB/s	7.8TB/s
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG	14 NVDEC 14 JPEG
Max thermal design power (TDP)	Up to 700W (configurable)	300-350W (configurable)	2x 350-400W (configurable)
Multi-Instance GPUs	Up to 7 MIGS @ 10GB each		Up to 14 MIGS @ 12GB each
Form factor	SXM	PCIe Dual-slot air-cooled	2x PCIe Dual-slot air-cooled
Interconnect	NVLink: 900GB/s PCIe Gen5: 128GB/s	NVLink: 600GB/s PCIe Gen5: 128GB/s	NVLink: 600GB/s PCIe Gen5: 128GB/s
Server options	NVIDIA HGX H100 Partner and NVIDIA-Certified Systems™ with 4 or 8 GPUs NVIDIA DGX H100 with 8 GPUs	Partner and NVIDIA-Certified Systems with 1–8 GPUs	Partner and NVIDIA-Certified Systems with 2-4 pairs
NVIDIA AI Enterprise	Add-on	Included	Add-on

Architektura Ada Lovelace

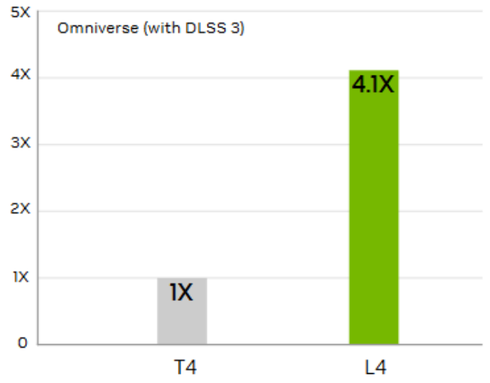
NVIDIA L4

Univerzální serverový akcelerátor s nízkým TDP vhodný pro akceleraci videa, grafiku a inference AI.

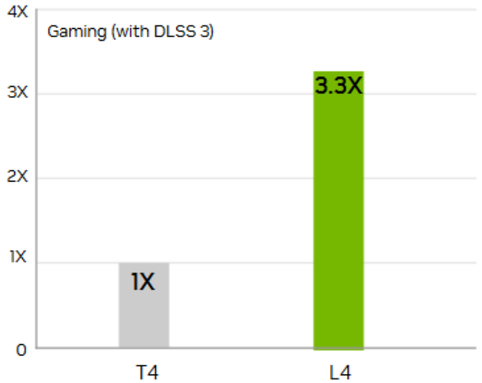
- **24 GB** GPU paměti GDDR6X
- propustnost paměti **300 GB/s**
- podpora **PCIe 4.0** (64GB/s)
- konfigurovatelné TDP **40-72 W**
- kompaktní single slotové low-profile provedení
- podpora virtualizace (vGPU)

FP32	30.3 teraFLOPs
TF32 Tensor Core	120 teraFLOPS*
FP16 Tensor Core	242 teraFLOPS*
BFLOAT16 Tensor Core	242 teraFLOPS*
FP8 Tensor Core	485 teraFLOPs*
INT8 Tensor Core	485 TOPs*
GPU memory	24GB
GPU memory bandwidth	300GB/s
NVENC | NVDEC | JPEG decoders	2 | 4 | 4
Max thermal design power (TDP)	72W
Form factor	1-slot low-profile, PCIe
Interconnect	PCIe Gen4 x16 64GB/s

Over 4X Higher Real-Time Rendering Performance



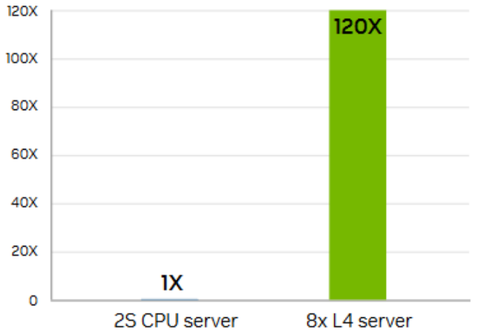
Over 3X Higher Ray-Tracing Performance



L4 Visual Computing Performance

Measured performance:
Real-time rendering: NVIDIA Omniverse™ performance for real-time rendering at 1080p and 4K with NVIDIA Deep Learning Super Sampling (DLSS) 3.
Ray tracing: Gaming performance geomean for AAA titles supporting ray tracing and DLSS 3.

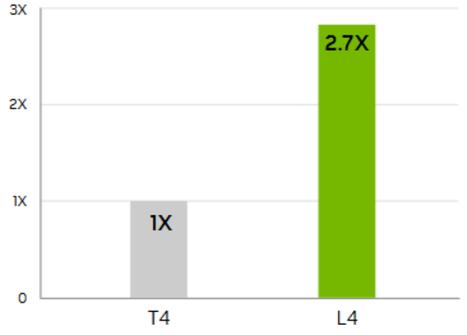
Up to 120X Higher AI Video Performance



L4 AI Video Performance

Measured performance: 8x L4 vs 2S Intel 8362 CPU server comparison, end-to-end video pipeline with CV-CUDA® decode, preprocessing, inference (SegFormer), postprocessing, encode, NVIDIA® TensorRT™ 8.6 vs CPU-only pipeline using OpenCV 4.7, PyTorch inference.

2.7X More Generative AI Performance



L4 Generative AI Performance

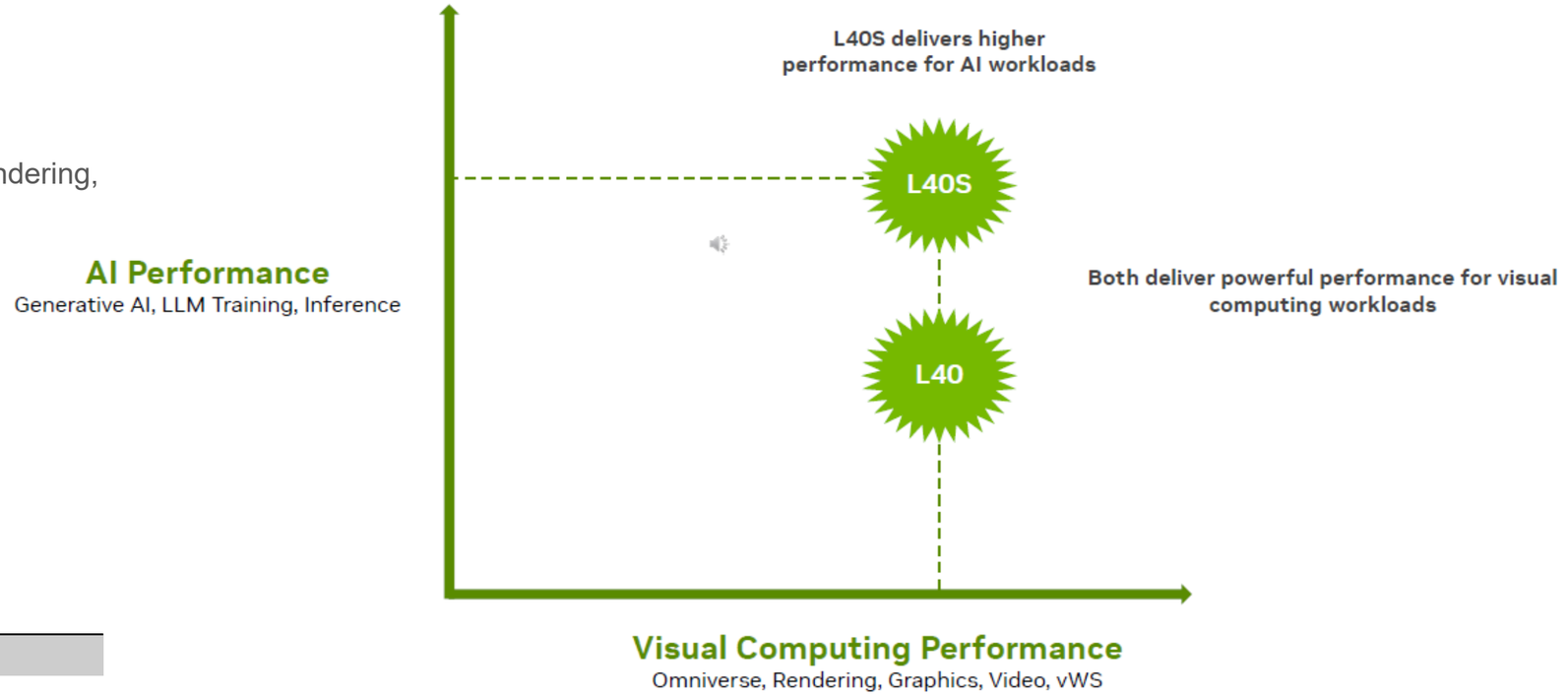
Projected performance: L4 vs T4 image generation, 512x512 stable diffusion v2.1, FP16, TensorRT 8.5.2.

Architektura Ada Lovelace

NVIDIA L40S

Nejvýkonnější univerzální karta pro grafické zpracování, 3D rendering, simulace, virtualizaci, AI výpočty a FP32 HPC.

- **48 GB** GPU paměti GDDR6
- propustnost paměti **864 GB/s**
- podpora **PCIe 4.0** x16 (64GB/s)
- TDP **350 W**



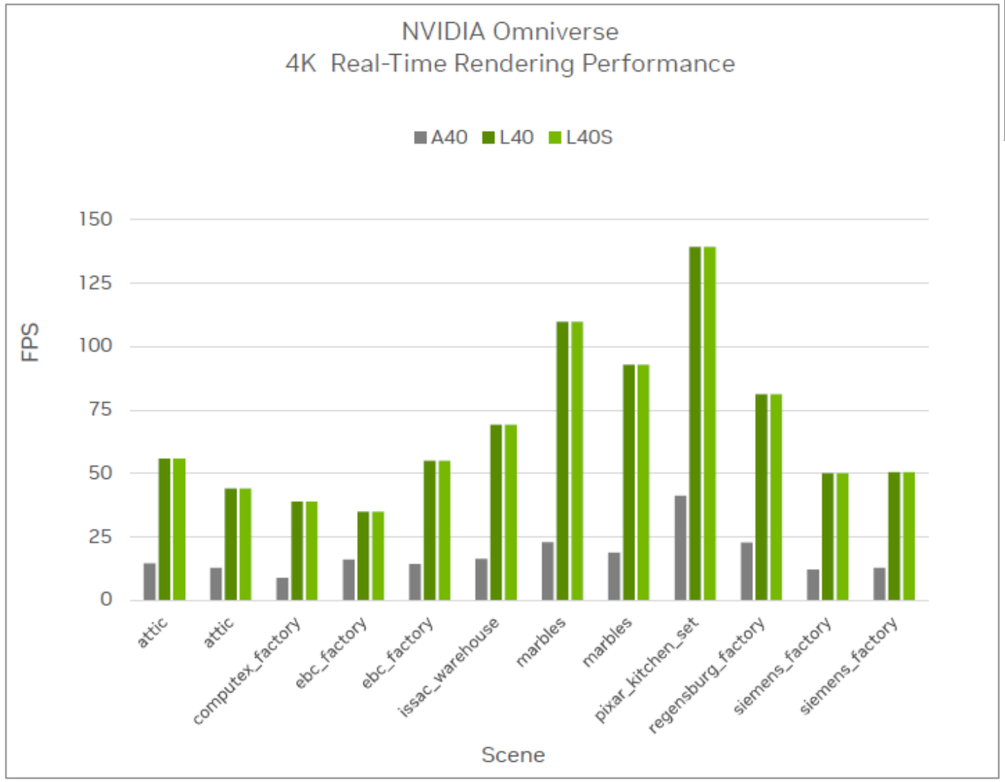
Specializovaná na AI výpočty (FP16) a inference (FP8)

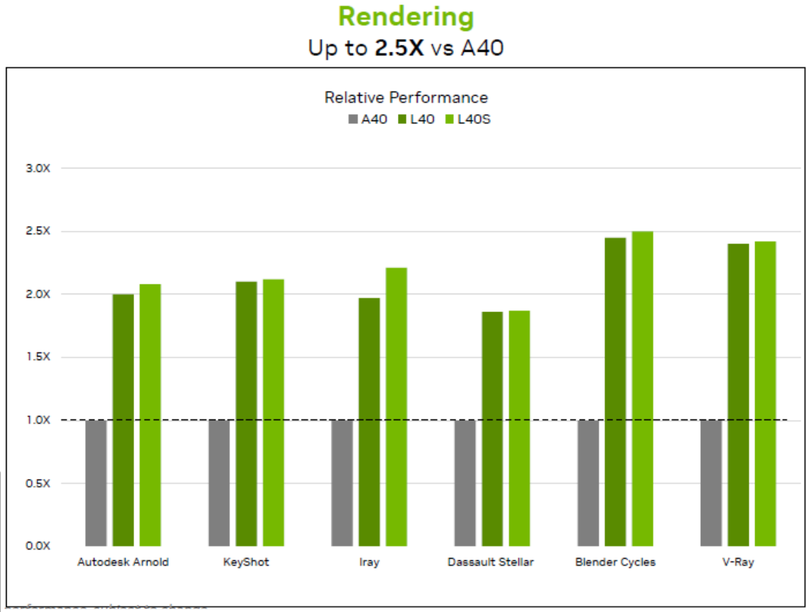
FP32	91,6 teraFLOPs
TF32 Tensor Core	366 teraFLOPS*
FP16 Tensor Core	733 teraFLOPS*
FP8 Tensor Core	1466 teraFLOPs*
GPU memory	48 GB
GPU memory bandwidth	864GB/s
NVENC | NVDEC | JPEG decoders	3 | 3 | 4
Max thermal design power (TDP)	350W
Form factor	2-slot FHFL, PCIe gen4
Interconnect	PCIe Gen4 x16 64GB/s

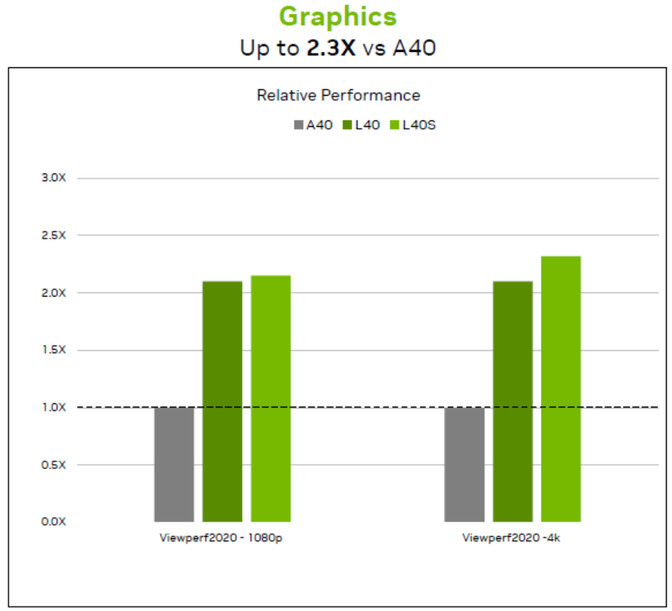
NVIDIA L40S

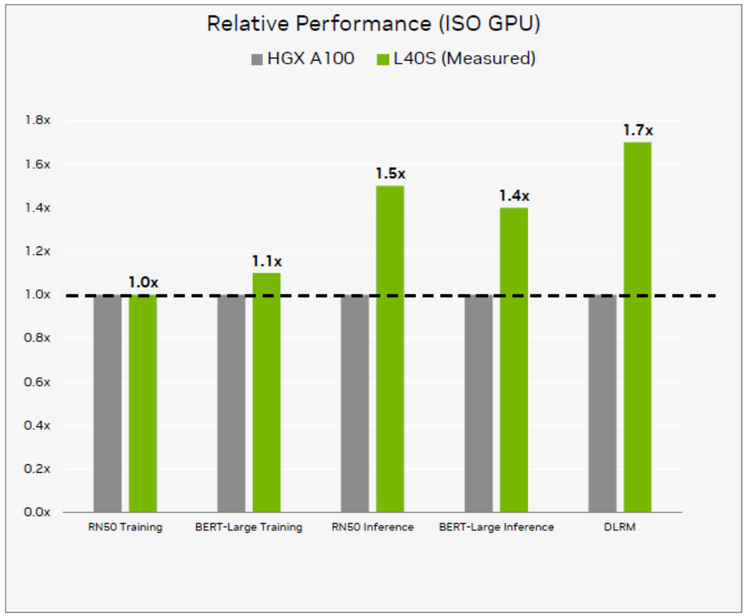
NVIDIA L40S vs. L40

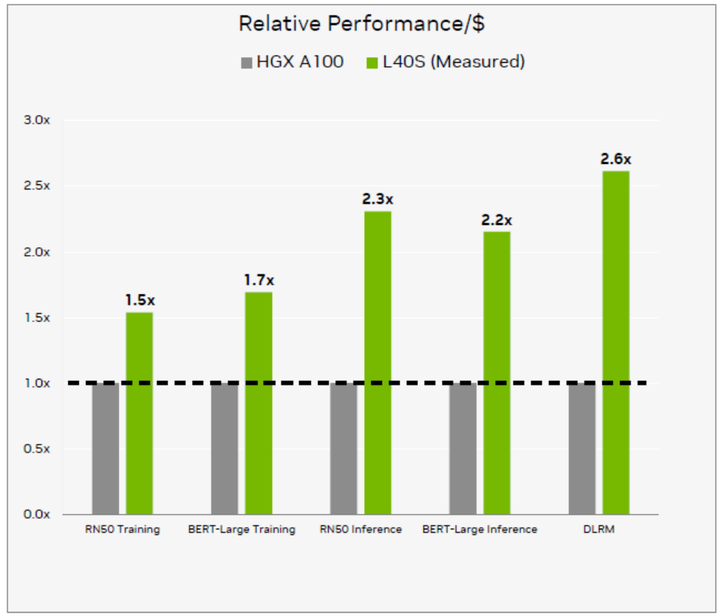
- ✓ Pro grafické operace, renderování a simulace poskytují stejný výkon, pro AI výpočty a FP32 HPC je L40S výkonnější.



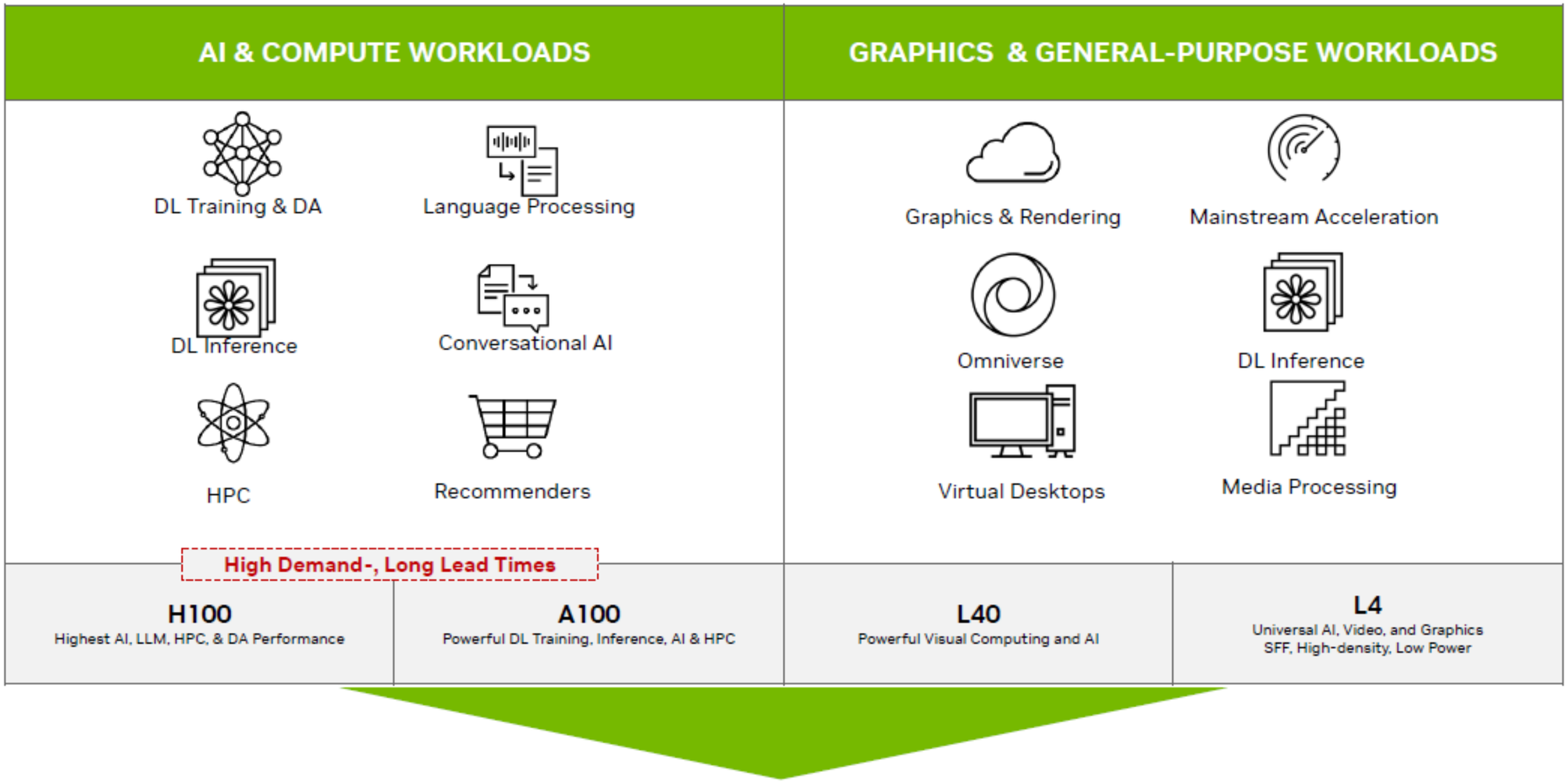








Pozicování a dostupnosti



NVIDIA L40S

The Most Powerful Universal Data Center GPU for AI and Graphics

Srovnání L40S s dalšími kartami

	NVIDIA L40	NVIDIA L40S	NVIDIA A40	A100 80GB SXM
GPU Architecture	**NVIDIA Ada Lovelace**		NVIDIA Ampere	
FP64	N/A	N/A	N/A	9.7 TFLOPS
FP32	90.5 TFLOPS	91.6 TFLOPS	37.4 TFLOPS	19.5 TFLOPS
RT Core	209 TFLOPS	212 TFLOPS	73.1 TFLOPS	N/A
TF32 Tensor Core*	181 TFLOPS	366 TFLOPS	150 TFLOPS	312 TFLOPS
FP16/BF16 Tensor Core*	362 TFLOPS	733 TFLOPS	299 TFLOPS	624 TFLOPS
FP8 Tensor Core*	744 TFLOPS	1466 TFLOPS	N/A	N/A
INT8 Tensor Core*	744 TOPS	1466 TOPS	598 TOPS	1248 TOPS
GPU Memory	48 GB GDDR6	48 GB GDDR6	48 GB GDDR6	80 GB HBM2e
GPU Memory Bandwidth	864 GB/s	864 GB/s	696 GB/s	2039 GB/s
L2 Cache	96 MB	96 MB	40 MB	40 MB
Media Engines	3 NVENC (+AV1) 3 NVDEC 4 NVJPEG	3 NVENC (+AV1) 3 NVDEC 4 NVJPEG	1 NVENC 2 NVDEC	0 NVENC 5 NVDEC 5 NVJPEG
Power	300W	Up to 350 W	300W	Up to 400 W
Form Factor	2-slot FHFL	2-slot FHFL	2-slot FHFL	8-way HGX
Interconnect	PCIe Gen4 x16: 64 GB/s	PCIe Gen4 x16: 64 GB/s	PCIe Gen4 x16: 64GB/s	PCIe Gen4 x16: 64 GB/s

Architektura Ada Lovelace

	Architecture	NVIDIA CUDA Cores	NVIDIA RT Cores	NVIDIA Tensor Cores	Peak Single-Precision Perf (TFLOPS)	Peak RT Core Perf (TFLOPS)	Peak Tensor Perf (TFLOPS)	GPU Memory (GDDR6)	Memory Interface (bits)	Peak Memory Bandwidth (GB/s)	Max Power Consumption (TGP)8	Graphics Bus	Display Technology	NVENC | NVDEC	NVIDIA Max-Q Technologies	GPUDirect® for Video	VR-Ready
RTX 5000 Ada	Ada Lovelace	9728	76	304	42.6	98.5	681.8	16 GB, ECC	256	576	175	PCI Express 4.0 x16, x8	DP 1.4a, HDMI 2.1	2x (8th Gen) | 2x (5th Gen)	Yes	Yes	Yes
RTX 4000 Ada	Ada Lovelace	7424	58	232	33.6	77.8	538.0	12 GB, ECC	192	432	175	PCI Express 4.0 x16, x8	DP 1.4a, HDMI 2.1	2x (8th Gen) | 2x (5th Gen)	Yes	Yes	Yes
RTX 4000 SFF Ada	Ada Lovelace	6144	48	192	19,2	44,3	306,8	20 GB, ECC	160	280	70	PCI Express 4.0 x16, x8	4x Mini DisplayPort 1.4a	2x (8th Gen) | 2x (5th Gen)	Yes	Yes	Yes
RTX 3500 Ada	Ada Lovelace	5120	40	160	23.0	53.3	368.6	12 GB, ECC	192	432	140	PCI Express 4.0 x16, x8	DP 1.4a, HDMI 2.1	2x (8th Gen) | 1x (5th Gen)	Yes	No	Yes
RTX 3000 Ada	Ada Lovelace	4608	36	144	19.9	46.0	318.6	8GB, ECC	128	256	140	PCI Express 4.0 x8, x4	DP 1.4a, HDMI 2.1	1x (8th Gen) | 1x (5th Gen)	Yes	No	Yes
RTX 2000 Ada	Ada Lovelace	3072	24	96	14.5	33.5	231.6	8GB	128	256	140	PCI Express 4.0 x8, x4	DP 1.4a, HDMI 2.1	1x (8th Gen) | 1x (5th Gen)	Yes	No	Yes
RTX A1000 6GB	Ampere	2560	20	80	9.3	18.2	74.6	6GB	96	168	95	PCI Express 4.0 x8, x4	DP 1.4a, HDMI 2.1	1x (7th Gen) | 2x (5th Gen)	Yes	No	Yes, 60W +
RTX A500	Ampere	2048	16	64	7.0	13.7	56.0	4GB	64	112	60	PCI Express 4.0 x8, x4	DP 1.4a, HDMI 2.1	1x (7th Gen) | 1x (5th Gen)	Yes	No	No

NVIDIA vGPU – virtualizace grafického výkonu

mcomputers.cz/nvidia/nvidia-vgpu/

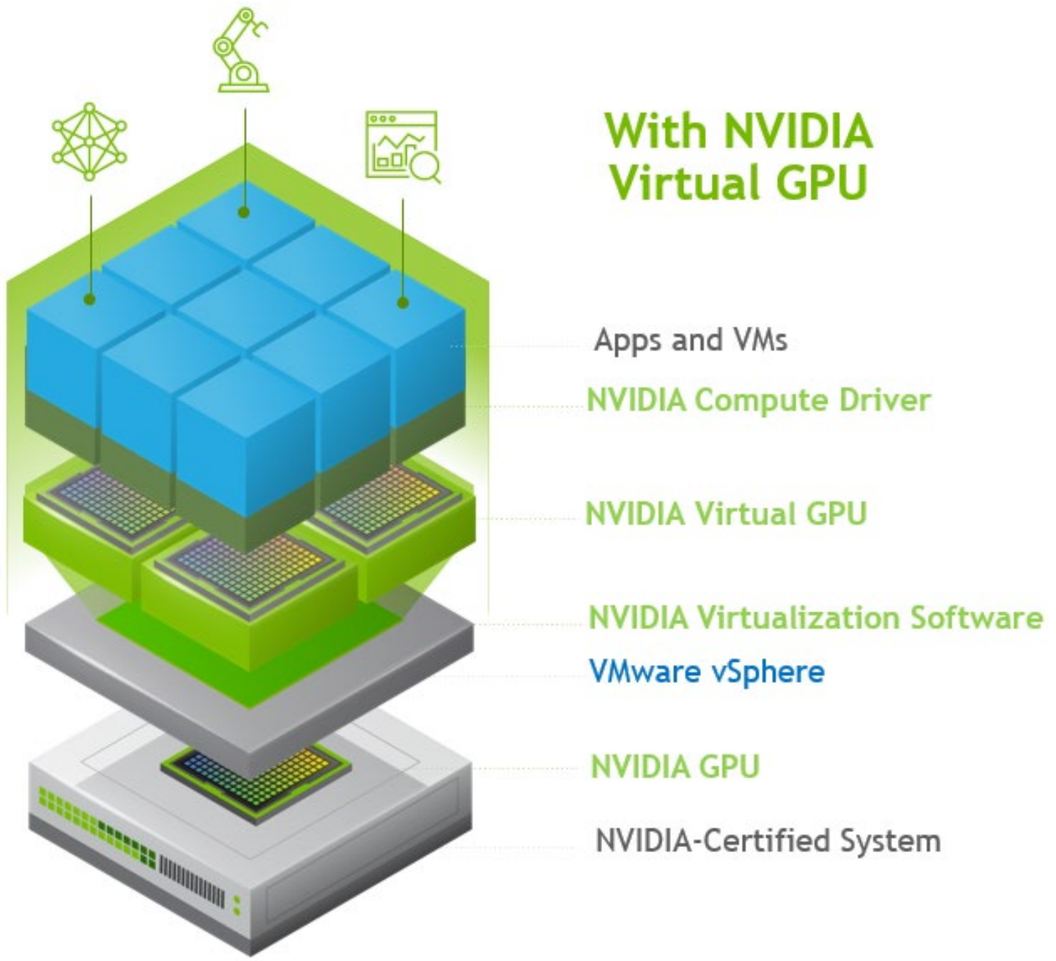
Software NVIDIA Virtual GPU (vGPU) umožňuje virtualizaci a rozdělení GPU výkonu tak, aby byla jedna GPU karta sdílena mezi více virtuálními stroji. Zároveň lze i více GPU přidělit jedinému virtuálnímu stroji, což otevírá možnosti zpracování i extrémně výpočetně náročných úloh.

✓ **GPU výkon bez omezení**
Výkon virtualizovaných GPU není degradován a dosahuje téměř stejných hodnot jako u fyzických GPU karet.

✓ **Optimální využití zdrojů**
Využijte výkon několika GPU karet, nebo je naopak rozdělte na více menších instancí, přesně dle vašich potřeb.

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

✓ **Měňte výkon dle potřeby**
Snadno přidávejte, odebírejte či měňte profily, podle toho, jak se momentálně vyvíjí vaše potřeby grafického výkonu.



Rozdělení VIRTUAL GPU (vGPU) 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
.

NVIDIA Omniverse

mcomputers.cz/nvidia/nvidia-omniverse/

NVIDIA Omniverse Enterprise je jednoduše rozšiřitelná platforma pro spolupráci v oblastech 3D designu, akcelerovaných realistických simulací (digital twin) nebo umělé inteligence. NVIDIA Omniverse je ve své podstatě virtuální svět, ve kterém mohou spolupracovat jednolivci nebo týmy pracující v různých aplikacích.

Licence Omniverse Enterprise:

- ✓ **NUCLEUS** – sdílení, prohlížení i editace projektů
  základní licence umožňující spolupráci mezi uživateli v reálném čase a využití pouze aplikací 3. stran
- ✓ **CREATOR** – kompozice scény, simulaci a vykreslování.
  licence umožňující využít dostupné Nvidia aplikace i aplikace 3. stran
- ✓ **REVIEWER** – pouze nahlížení do projektů
  společně s licencí Nucleus lze komentovat, měnit textury a mírně editovat

Možnost propojení s aplikacemi:

- ✓ Autodesk 3ds Max, Adobe Substance 3D Painter, SketchUp 3D, Rhino & Associates, Autodesk Maya, Epic Games Unreal Engine, Visual Components, Graphisoft Archicad, a mnoho dalších…

Lze vyzkoušet 30-denní Trial verzi
Aktuálně běží promo nabídka pro **EDU – licence za 1 USD**

	NVIDIA Omniverse for Individuals	NVIDIA Omniverse Enterprise
Collaboration	Between multiple apps and one other user	Between multiple apps and licensed Creators
Licensing	Free for individuals	Subscription Licenses, Annual, and Multi-Year
Support	Public forums, Training Videos	NVIDIA Enterprise Support
Nucleus	Nucleus Workstation only	Enterprise Nucleus Server Nucleus Workstation
Connectors	Use of all Connectors, including beta	Use of all production Connectors
Apps	All Omniverse apps, including beta	Omniverse Create and Code (license per Creator) Omniverse View (license per user) *Use of Kit-based apps require Creator license*
Batch Microservices	On up to 2 GPUs	On up to 64 GPUs per Creator subscription

	Term			
Subscription	1 Year	3 Years	4 Years	5 Years
Omniverse Enterprise Nucleus Subscription per Named User	$1,000	$3,000	$4,000	$5,000
Omniverse Enterprise Creator Subscription per Floating User (CCU)	$2,000	$6,000	$8,000	$10,000
Omniverse Enterprise Reviewer Subscription per Floating User (CCU)	$100	$300	$400	$500

NVIDIA AI Enterprise 4.0

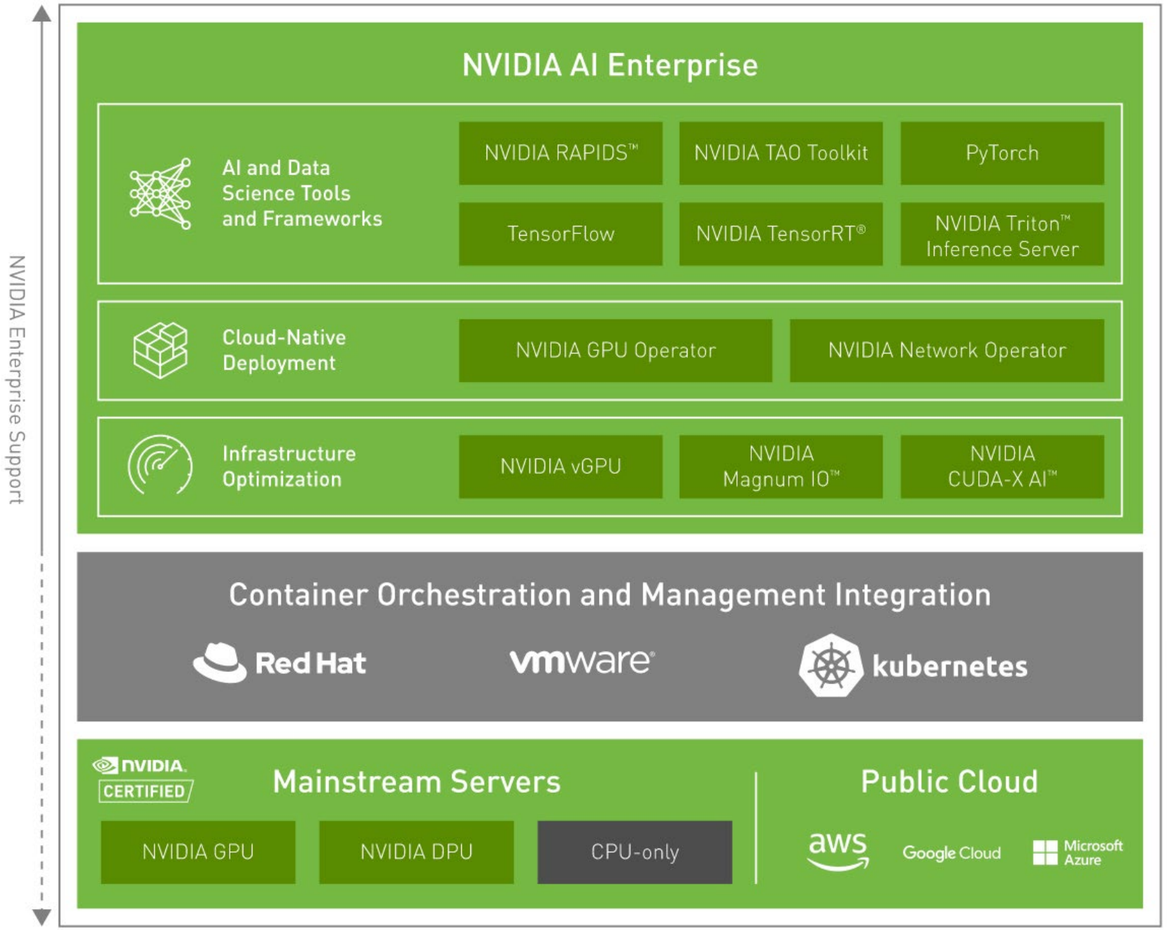
mcomputers.cz/nvidia/ai-enterprise/

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

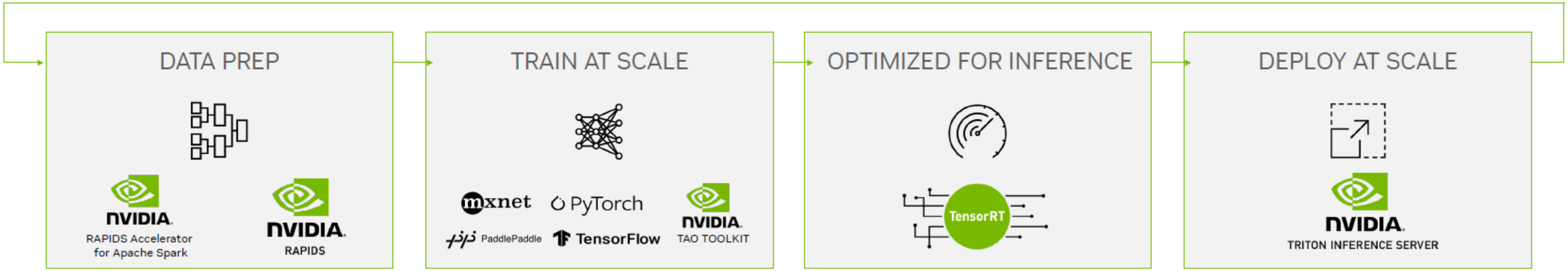
- ✓ licencováno per **GPU** (max. 20 VM per GPU)
- ✓ sdílení výkonných GPU akcelerátorů více virtuálními servery
- ✓ snadné zprovoznění vyladěného AI prostředí v řádu hodin a jeho škálování
- ✓ podpora na celý software stack včetně open-source knihoven, frameworků a aplikací
- ✓ napojení na NVIDIA GPU katalog (**NGC**)
- ✓ **100+** frameworků, předtrénovaných modelů a referenčních aplikací pro AI
- ✓ nově obsahuje **AI Workflows**

doporučenéné akcelerátory: H100, A100, A30
podporované akcelerátory:
compute: H100, A100, A30, V100s/V100, A100X, A30X a A2
mixed workloads: A40, L40, A10, A16, T4, L4, RTX 6000 Ada, RTX A6000/A5000

- ✓ 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
- ✓ nasazení v cloudu (AWS, MS Azure, Google, Oracle)
- ✓ spouštění kontejnerů (Kubernetes, VMware Tanzu, HPE Ezmeral,..)



NVIDIA AI Enterprise a AI Workflows



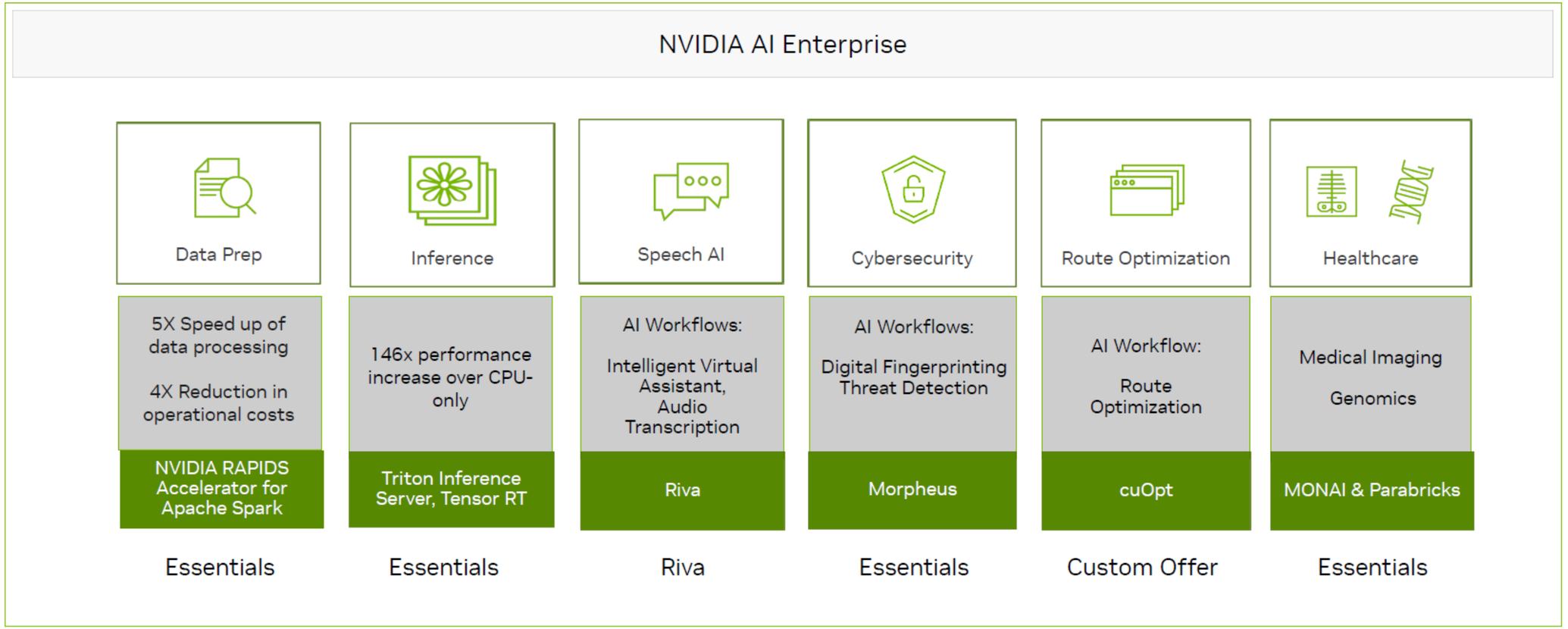
May

Intelligent Virtual Assistant	Audio Transcription	Digital Fingerprinting Threat Detection	Next Item Prediction	Route Optimization	Inference	Generative AI
Engaging contact center assistance 24/7 for lower operational costs	World-class, accurate transcripts based on GPU-optimized models	Cybersecurity threat detection and alert prioritization to identify and act faster	Personalized product recommendations for increased customer engagement and retention	Vehicle and robot routing optimization to reduce travel times and fuel costs	At-scale inference and MLOps for data science teams	TBD

NVIDIA AI Workflows

NVIDIA AI Workflows jsou předtrénované referenční aplikace, které umožňují rychlou a snadnou implementaci AI do podnikových systémů.
AI Workflows jsou nově součástí softwarové sady AI Enterprise.

- ✓ **Essentials** - standardní licence NVAIE
- ✓ **RIVA** – licence pro Riva Framework a jeho Workflows (Audio Transcription, Intelligent Virtual Assistant)
- ✓ **Custom offers** – licence pro Workflows, které nejsou v ceníku)

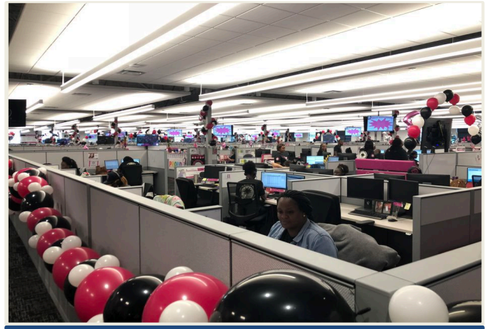


NVIDIA AI Enterprise využívají již tisíce firem po celém světě

A Top 10 Finance Company

Accuracy & 5X Speed Improvement

- ASR for Customer Experience Center
  - Virtual Assistant
  - Offline Transcriptions



T-Mobile

~2X Accuracy & 10X Speed Improvement

- ASR for Customer Support
  - Agent Assist
  - Real-Time Analytics

RingCentral

2X Accuracy Increase with No Accent & Noisy Environments Issues

- ASR for Video Call Transcripts
  - Real-Time Transcriptions
  - Meeting Summarizations

SOTA WER for Noisy Environments & Expressive Synthetic Voices

- ASR & TTS for Consumer Application
  - 100M Monthly Active Users
  - Integrated in Over 1000 Lenses

Northwestern Medicine® ARTISIGHT

Human-Like Vocalization for Patient Announcement System

- TTS for Patient Announcement
  - A Top 10 US Best Hospital

AI² LABS · AVAYA · botpress · Computacenter · DATA MONSTERS · Deloitte. · Dialoga Group · Floatbot® · Hewlett Packard Enterprise · interactions

InstaDeep™ · kore.ai · LEXISTEMS. · MALAMUTE · MINERVA · moneypenny · ncs// · pendulum · plaBook · quantiphi

Read · replicant · siminsights · SmartCow An AI Engineering Company · SVA · softserve · TARTEEL AI · VectorVentures

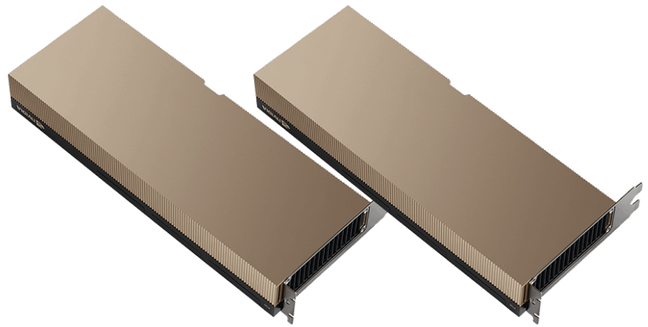
Reference ČR

Masarykova univerzita

NVIDIA H100 je nejvýkonnější akcelerátor generace Hopper.

2x NVIDIA H100 80GB PCIe

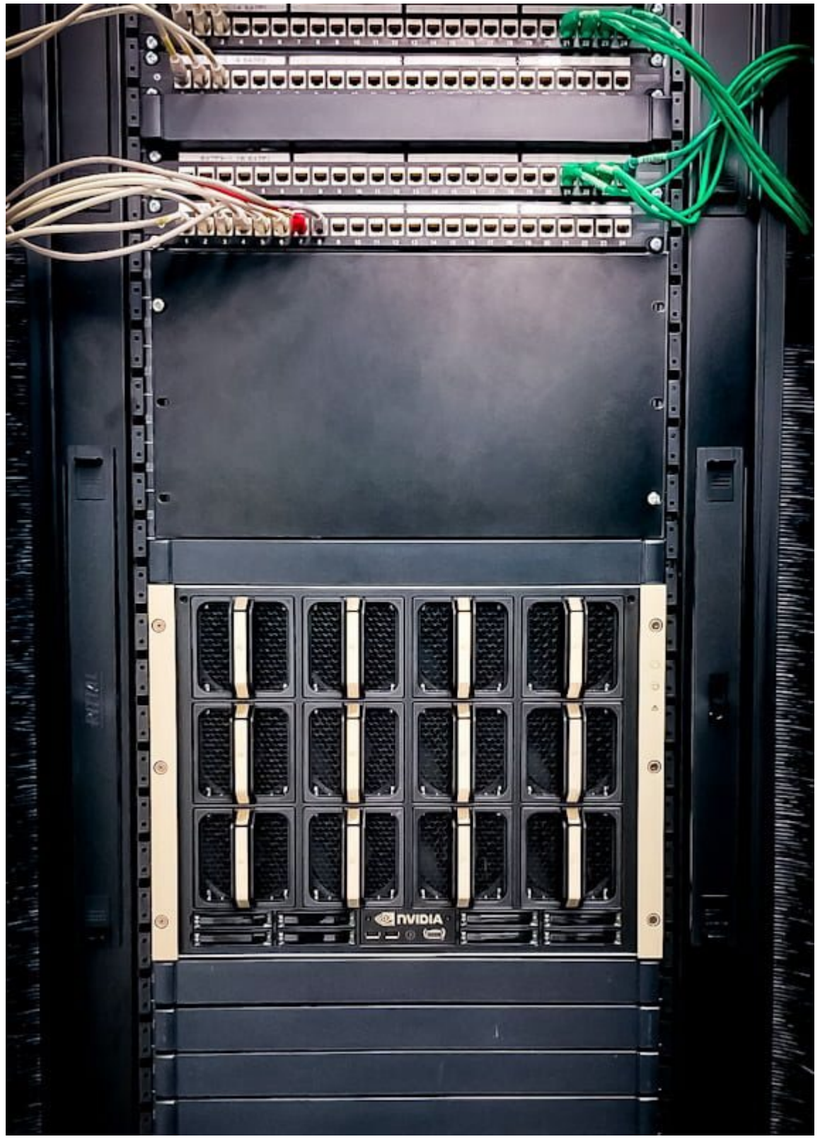
- ✓ nejvýkonnější akcelerátor současnosti
- ✓ 4nm technologie výroby
- ✓ 80 mld tranzistorů na 814 mm2
- ✓ propustnost paměti HBM3 = 3 TB/s
- ✓ 50 MB L2 cache
- ✓ 4. generace NVLinku (900GB/s)
- ✓ podpora PCIe 5.0 (128GB/s)
- ✓ 310 W
- ✓ MIG (Multi-Instance GPU) 2. generace



NVIDIA DGX H100 je nejnovější a nejvýkonnější sériově vyráběný systém pro výpočty AI.

1x NVIDIA DGX H100 640GB

- ✓ 8x GPU H100 80GB
- ✓ 640 GB HBM3 paměti s propustností až 24 TB/s
- ✓ 2x CPU Intel Xeon Platinum 8480C (112 jader)
- ✓ 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 IB/Eth)
- ✓ 2x dual-port ConnectX-7 (400Gb/s IB/Eth)
- ✓ Management port (BMC) (RJ45)
- ✓ 10Gb/s onboard NIC (RJ45) (až 100Gb/s Eth)
- ✓ DGX Display Adapter card (4 DisplayPort, 4K)
- ✓ Spotřeba max. 10,2 kW
- ✓ Velikost 8U



Reference

TECHNICAL UNIVERSITY OF KOŠICE

Technická univerzita v Košicích (SVK)

1x NVIDIA DGX A100 320GB

- ✓ 8× NVIDIA A100 SXM4 40 GB
- ✓ 2× AMD Epyc 7742 64 cores 3.6GHz
- ✓ 1 TB DDR4 RAM
- ✓ 2× 1.92 TB NVMe (OS)
- ✓ 4× 3.84 TB NVMe (15 TB Data)
- ✓ 8x 200 Gb/s HDR InfiniBand
- ✓ 2x 200 Gb/s Ethernet
- ✓ Velikost 6U

University of Ljubljana
Faculty of Computer and Information Science

Univerzita Ljubljana (SL)

1x NVIDIA DGX H100 640GB

- ✓ 8x GPU H100 80GB
- ✓ 640 GB HBM3 paměti s propustností až 24 TB/s
- ✓ 2x CPU Intel Xeon Platinum 8480C (112 jader)
- ✓ 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 IB/Eth)
- ✓ 2x dual-port ConnectX-7 (400Gb/s IB/Eth)
- ✓ Management port (BMC) (RJ45)
- ✓ 10Gb/s onboard NIC (RJ45) (až 100Gb/s Eth)
- ✓ DGX Display Adapter card (4 DisplayPort, 4K)
- ✓ Spotřeba max. 10,2 kW
- ✓ Velikost 8U

Reference

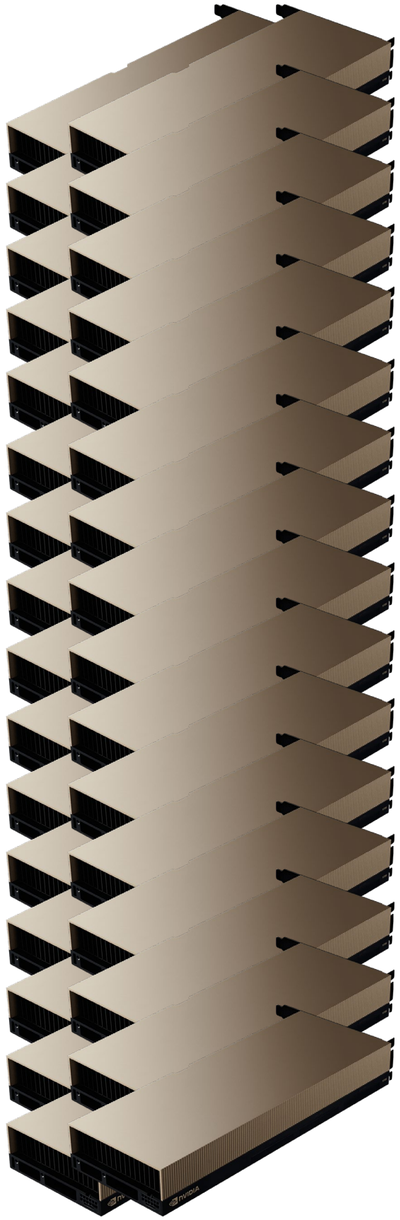
EUROPEAN UNION
SATELLITE CENTRE

Analysis for decision making

European Union Satelite Center (ES)

1x NVIDIA DGX A100 320GB

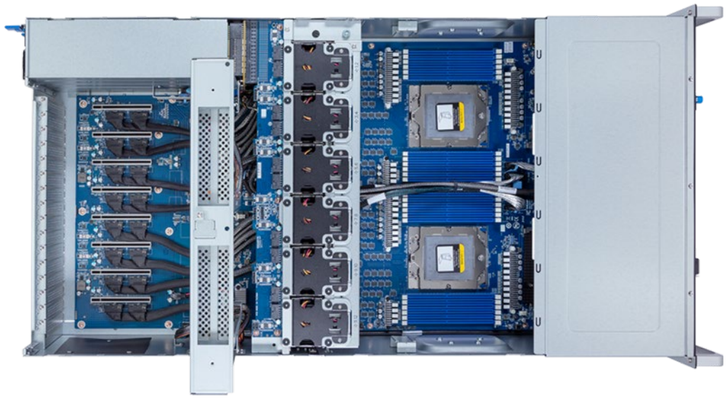
- ✓ 8× NVIDIA A100 SXM4 40 GB
- ✓ 2× AMD Epyc 7742 64 cores 3.6GHz
- ✓ 1 TB DDR4 RAM
- ✓ 2× 1.92 TB NVMe (OS)
- ✓ 4× 3.84 TB NVMe (15 TB Data)
- ✓ 8x 200 Gb/s HDR InfiniBand
- ✓ 2x 200 Gb/s Ethernet
- ✓ Velikost 6U

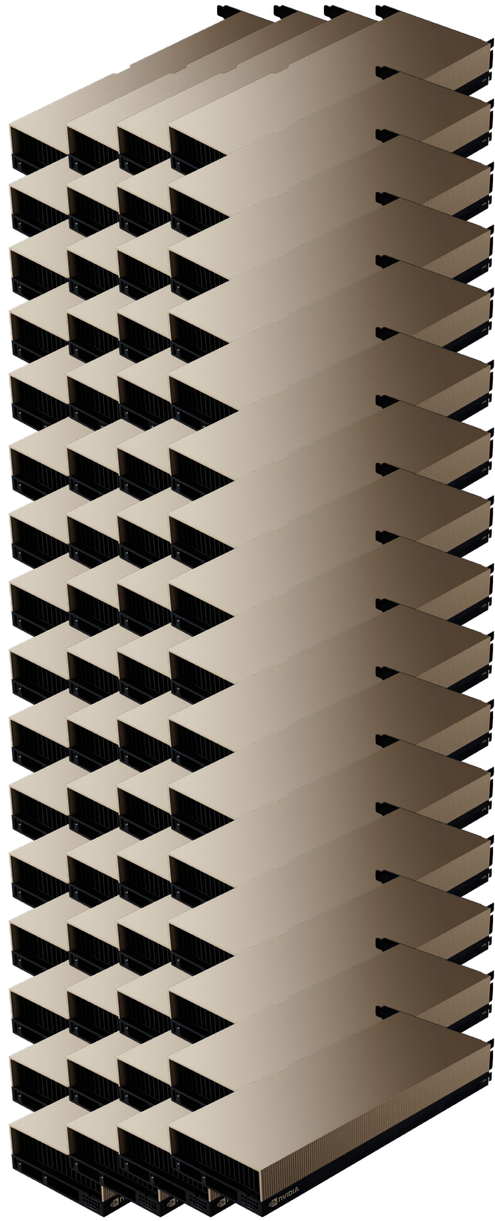


********** (CZ)

96x GPU H100 80GB PCIe

- ✓ **8x H100 80GB** PCIe + NVLink bridge
- ✓ **7 680 GB** HBM4 GPU paměti
- ✓ **2x CPU** AMD Epyc Genoa 9354 (64 jader)
- ✓ **1,5 TB** DDR5 4800MHz RAM
- ✓ Velikost 4U



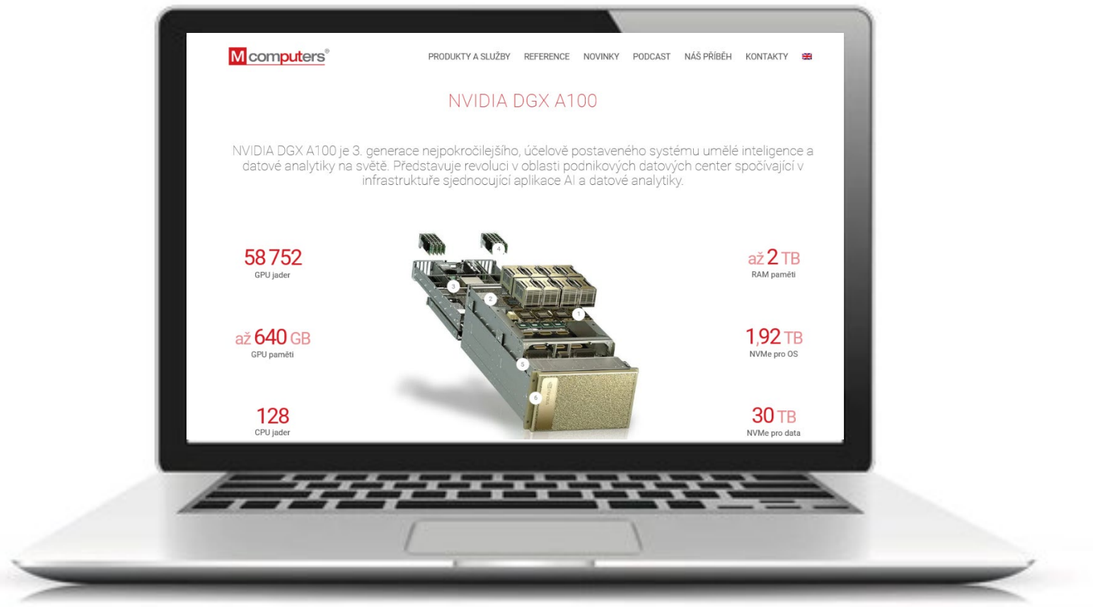


NVIDIA produkty přehledně

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*