mcomputers

NVIDIA ELITE PARTNER · TECHNICKÝ UPDATE

NVIDIA AI referenční architektury

Enterprise RA pro AI továrny

Petr Plodík

REFERENČNÍ ARCHITEKTURY · ÚVOD

Co je NVIDIA referenční architektura

Ověřený a opakovatelný návrh AI továrny, který spojuje **výpočet, síť, úložiště a software**. Zkracuje nasazení, snižuje integrační riziko a umožňuje předvídatelný růst.

1 Server (node)

NVIDIA-Certified Systems od OEM výrobců – např. HPE ProLiant, HPE Compute XD

2 Cluster

Enterprise RA – jak z certifikovaných nodů postavit cluster: síť, škálování, úložiště

3 Software

Referenční návrhy AI ekosystému – AI Factory Validated Design, sizing a deployment guidy

Jak vzniká: od vlastní AI továrny NVIDIA po řešení partnera

1 Develop	2 Establish	3 Endorse	4 Scale
NVIDIA staví a provozuje vlastní AI infrastrukturu (DGX, interní AI factory)	Zkušenosti zapíše do certifikace nodů, clusterových RA a softwarových návrhů	Design Review Board NVIDIA posoudí návrh OEM partnera	Partner nabízí schválený návrh jako vlastní AI Factory – např. HPE
Schválení (endorsement) DRB:	**Infrastructure · povinné**	**Spectrum-X · doporučené**	**Síťová logika · doporučené**

Zdroj: NVIDIA Enterprise RA white paper – Introducing NVIDIA Reference Architectures

REFERENČNÍ ARCHITEKTURY · NCP A ENTERPRISE

NCP vs. Enterprise RA: stejná DNA, jiné měřítko

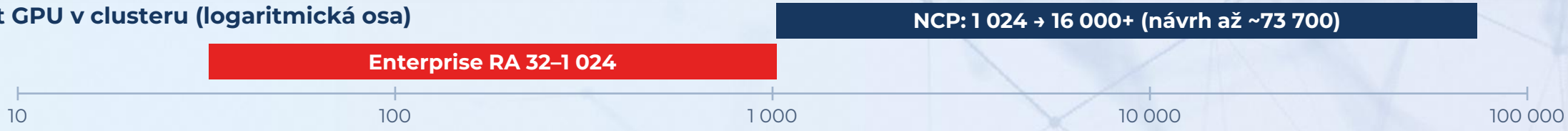
NVIDIA Cloud Partner (NCP) RA

Pro koho	cloudoví poskytovatelé, GPU-as-a-service, tvůrci foundation modelů
Měřítko	od 128 nodů po 16 000+ GPU; NCP design: až 64 PODů × 1 152 GPU
Tenanti	single i multi-tenant (IaaS, CaaS, PaaS)
Síť	InfiniBand Quantum i Ethernet Spectrum-X
Nasazení	převážně NVIDIA professional services

Enterprise RA

Pro koho	podniky budující vlastní AI továrnu on-premise
Měřítko	32–1 024 GPU: 4–128 nodů nebo 1–8 racků
Tenanti	single-tenant, více uživatelů a týmů
Síť	Ethernet Spectrum-X, BlueField-3, ConnectX
Nasazení	NVIDIA partneři

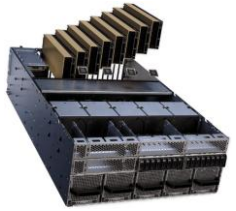
Počet GPU v clusteru (logaritmická osa)



Zdroje: NVIDIA Enterprise RA white paper; NVIDIA DSX – NCP Reference Guide (~73 700 = 64 PODů × 1 152 GPU)

REFERENČNÍ ARCHITEKTURY · RODINY

Tři rodiny AI továren

	RTX PRO AI Factory vzduch · PCIe	HGX AI Factory vzduch · SXM + NVLink	NVL72 AI Factory kapalina · rack-scale
			
Node (C-G-N-B)	2-8-5-200 (+ 2-4-3-200, 2-X-1-NA)	2-8-9-800	2-4-5-800 (compute tray)
GPU	RTX PRO 6000 / 4500 Blackwell Server Edition	HGX B300 – 8× Blackwell Ultra SXM	GB300 NVL72 – 72 GPU v NVLink doméně
Měřítko	4–32 nodů · 32–256 GPU	4–128 nodů · 32–1 024 GPU	1–8 racků · 72–576 GPU
Typické úlohy	agentní inference, physical AI, vizuální computing, HPC FP32	trénink, fine-tuning, distribuovaná inference, analytika	trénink bilionových modelů, MoE, real-time reasoning
Tvrzení NVIDIA	vstupní bod do AI v běžném DC	až 15× vyšší propustnost tokenů vs. HGX H100	až 50× vyšší výstup AI továrny vs. Hopper
HPE	**ProLiant DL380a Gen12**	**Compute XD690**	**GB300 NVL72 by HPE**

Zdroj: NVIDIA Enterprise RA white paper – Introducing, Design Tenets. Tvrzení NVIDIA jsou horní meze výrobce.
Obrázky: NVIDIA

ÚVOD · GPU PLATFORMY

GPU: generační skok v paměti a propojení

Parametr	H200 (Hopper)	B300 (Blackwell Ultra)	Rubin	Rubin vs. Hopper
Tranzistory	80 mld. (1 čip)	208 mld. (2 čipy)	336 mld. (2 čipy)	**4,2×**
Paměť GPU	141 GB HBM3e	až 288 GB HBM3e	288 GB HBM4	**2×**
Propustnost paměti	4,8 TB/s	8 TB/s	22 TB/s	**4,6×**
NVFP4 inference (dense)	– (FP8: 2 PFLOPS)	13–15 PFLOPS	50 PFLOPS	**25× (vs. FP8)**
FP64 / FP64 Tensor	34 / 67 TFLOPS	1,2–1,3 TFLOPS	[doplnit]	**–**
Transformer Engine	1. generace	2. generace	3. generace	**+2 generace**
Softmax (B200 = 1×)	–	2×	2× FP32 / 4× BF16	**–**
NVLink na GPU	NVLink 4 · 0,9 TB/s	NVLink 5 · 1,8 TB/s	NVLink 6 · 3,6 TB/s	**4×**
CPU ↔ GPU (NVLink-C2C)	900 GB/s (GH200)	900 GB/s (Grace)	1,8 TB/s (Vera)	**2×**
Max. TDP GPU	až 700 W	1 100–1 400 W	[doplnit]	**–**
Důvěrné výpočty	TEE	TEE	TEE-I/O	**TEE → TEE-I/O**
Rok uvedení	2024	2025	2026	**+2 roky**

Zdroje: NVIDIA Data Center Portfolio; NVIDIA H200 datasheet; NVIDIA Technical Blog (7/2026); NVIDIA GTC 2026. Rubin vs. Hopper = poměr Rubin / H200. B300 dle provedení HGX/GB.

ÚVOD · GPU PLATFORMY

GPU pro AI továrny: od PCIe karty po rack

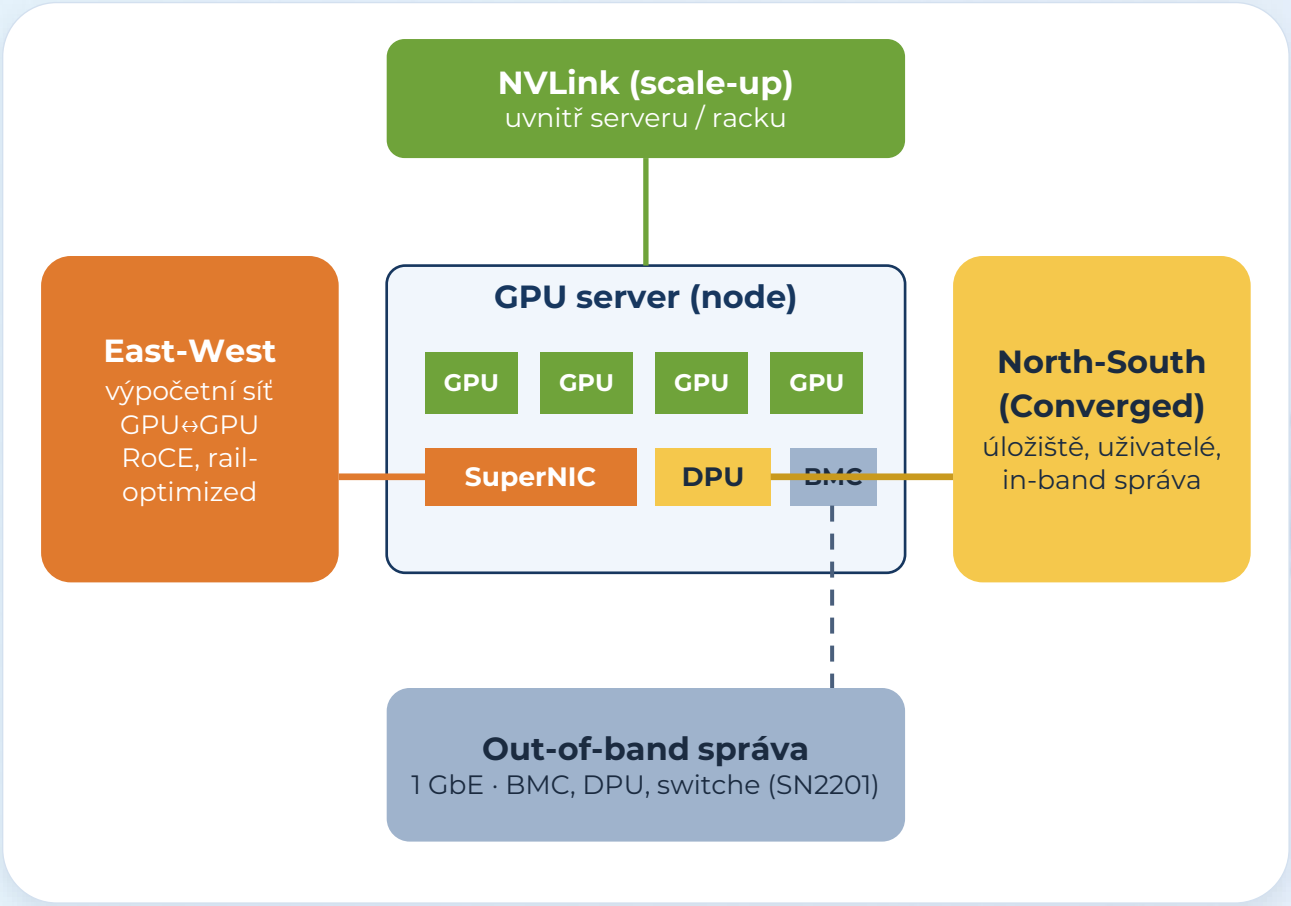
	PCIe · RTX PRO AI Factory			Desktop	HGX AI Factory		NVL72 AI Factory	
Parametr (na 1 GPU)	**H200 NVL**	**RTX PRO 4500**	**RTX PRO 6000**	**GB300 (DGX Station)**	**B300 NVL8**	**Rubin NVL8**	**GB300 NVL72**	**VR NVL72**
Architektura	Hopper	Blackwell	Blackwell	Blackwell Ultra	Blackwell Ultra	Rubin	Blackwell Ultra	Rubin
Formát	PCIe, 2 sloty	PCIe, 1 slot	PCIe, 2 sloty	superchip GPU + Grace	SXM, deska 8 GPU	SXM, deska 8 GPU	rack, 72 GPU	rack, 72 GPU
Paměť GPU	141 GB HBM3e	32 GB GDDR7	96 GB GDDR7	252 GB HBM3e + 496 GB LPDDR5X	288 GB HBM3e	~256 GB HBM4[2]	288 GB HBM3e	288 GB HBM4
Propustnost paměti	4,8 TB/s	0,8 TB/s	1,6 TB/s	7,1 TB/s	8 TB/s	**~16 TB/s[2]**	8 TB/s	**19,2 TB/s[2]**
FP4 se sparsitou	– (FP8: 3,3 PF)	1,6 PF[1]	4 PF[1]	20 PF	18 PF	**50 PF**	20 PF	**50 PF**
FP4 dense	–	n/a	n/a	15 PF	13,5 PF	**35 PF**	15 PF	**35 PF**
FP64	30 TF	n/a	n/a	n/a	~1,25 TF	**33 TF**	~1,4 TF	**33 TF**
GPU ↔ GPU	NVLink bridge 0,9 TB/s (2–4 GPU)	jen PCIe Gen5	jen PCIe Gen5	1 GPU · C2C 900 GB/s k CPU	NVLink 5 · 1,8 TB/s (8 GPU)	**NVLink 6 · 3 TB/s[2] (8 GPU)**	NVLink 5 · 1,8 TB/s (72 GPU)	**NVLink 6 · 3 TB/s[2] (72 GPU)**
Příkon · chlazení	600 W · vzduch	165 W · vzduch	600 W · vzduch / kapalina	1 600 W celý systém	~1 100 W · vzduch / kapalina	n/a · kapalina	1 400 W · kapalina	n/a · kapalina
Referenční node	RTX PRO 2-8-5-200 (volitelně)	RTX PRO 2-8-5-200	RTX PRO 2-8-5-200	vývoj u stolu	HGX 2-8-9-800	nová generace HGX	NVL72 2-4-5-800	nová generace NVL72
Rok uvedení	2024	2026	2025	2026	2025	2026	2025	2026

[1] NVIDIA neuvádí dense/sparse. [2] Specifikace Rubinu upravené NVIDIA 9/2026; NVL8 přepočteno z 2 TB a 130 TB/s na 8 GPU. n/a = neuvádí.

Zdroje: NVIDIA produktové stránky a datasheety (H200, RTX PRO, DGX Station, HGX, NVL72), stav 11/2026

NÁVRHOVÉ PRINCIPY · SÍŤ

Sítě AI továrny: North-South a East-West provoz



Volby návrhu

- **Sloučená vs. oddělená** síť North-South a East-West – malé clustery mohou sdílet jednu fabric.
- **Single-plane vs. dual-plane:** dual-plane rozdělí 800G port GPU na 2× 400G do dvou nezávislých fabric; zátěž rozkládá NCCL, výpadek roviny jen sníží propustnost.
- **Single-plane** je levnější, ale má o 50 % nižší propustnost GPU.
- **Jen inference:** síť East-West lze vynechat – cluster pak ale neumí trénink ani fine-tuning.
- **Switche:** Spectrum-4 SN5600 / SN5610 (64× 800G) pro data, SN2201 (48× 1G) pro OOB.

Zdroje: NVIDIA Reference Architectures for Enterprise AI Factories (white paper) – Design Tenets; NVIDIA HGX AI Factory Enterprise RA – Physical Topologies

NÁVRHOVÉ PRINCIPY · NOMENKLATURA

Jak číst označení C-G-N-B

2	–	8	–	5	–	200
C		**G**		**N**		**B**
CPU sockety v serveru		GPU v serveru		síťové adaptéry celkem (vč. 1× North-South)		Gb/s na GPU v síti East-West

Příklady – každý čtvereček je jeden síťový adaptér

2-8-5-200	**RTX PRO**		**200 Gb/s**	1× BlueField-3 B3220 + 4× B3140H SuperNIC (1 na 2 GPU)
2-8-9-800	**HGX B300**		**800 Gb/s**	1× BlueField-3 + 8× ConnectX-8 SuperNIC (1 na GPU)
2-4-5-800	**NVL72 tray**		**800 Gb/s**	2× Grace + 4 GPU: 1× BlueField-3 B3240 + 4× ConnectX-8

North-South (DPU) East-West (SuperNIC)

Zdroj: NVIDIA Reference Architectures for Enterprise AI Factories (white paper) – Design Tenets, Deep Dive. Starší HGX RA (H100/H200/B200) používá 2-8-9-400.

NÁVRHOVÉ PRINCIPY · REFERENCE CONFIGURATIONS

Referenční konfigurace na jednom místě

Rodina	Konfigurace	CPU	GPU	Síťové adaptéry	East-West / GPU	Adaptéry (E-W / N-S)	Podporovaná GPU
RTX PRO	2-8-5-200	2	8	5: 1 N-S + 4 E-W	200 Gb/s	4× BF-3 B3140H / 1× BF-3 B3220	RTX PRO 6000 / 4500 BSE
RTX PRO	2-4-3-200	2	4	3: 1 N-S + 2 E-W	200 Gb/s	2× BF-3 B3140H / 1× BF-3 B3220	RTX PRO 6000 / 4500 BSE
RTX PRO	2-2-3-400	2	2	3: 1 N-S + 2 E-W	400 Gb/s	2× BF-3 B3140H / 1× BF-3 B3220	RTX PRO 6000 / 4500 BSE
RTX PRO	2-X-1-NA	2	2 / 4 / 8	1: jen N-S	–	– / 1× BF-3 B3220	RTX PRO 6000 / 4500 BSE
HGX	2-8-9-800	2	8	9: 1 N-S + 8 E-W	800 Gb/s	8× ConnectX-8 / 1× BF-3	HGX B300
NVL72	2-4-5-800	2 Grace	4 / tray	5: 1 N-S + 4 E-W	800 Gb/s	4× ConnectX-8 / 1× BF-3 B3240	GB300 NVL72

Zvýrazněné = základní (foundational) konfigurace každé rodiny.

VOLUME
2-4-3-200 a 2-2-3-400: levnější varianty se 2 nebo 4 GPU; u vybraných konfigurací i s NVIDIA Vera CPU.

EXTREME VOLUME
2-X-1-NA: jen síť North-South, bez E-W – pro úlohy, které se vejdou do jednoho GPU nebo nodu.

VOLITELNĚ
U 2-8-5-200 lze ve vybraných konfiguracích použít i H200 NVL.

Zdroj: NVIDIA Enterprise RA white paper – Design Tenets (Table 2), Deep Dive. BSE = Blackwell Server Edition.

NÁVRHOVÉ PRINCIPY · SCALABLE UNIT

Scalable Unit: stavební kostka AI továrny

Scalable Unit (SU) je opakovatelný blok, podle kterého se navrhuje topologie sítě, napájení, chlazení, rozmístění racků i plán nasazení.

- **RTX PRO a HGX:** 1 SU = 4 nody, růst po 4 nodech
- **NVL72:** 1 SU = celý rack (18 compute trayů)

HGX: od 1 SU do 32 SU (1 dlaždice = 4 nody = 32 GPU)



1 SU 4 SU 8 SU 16 SU 32 SU

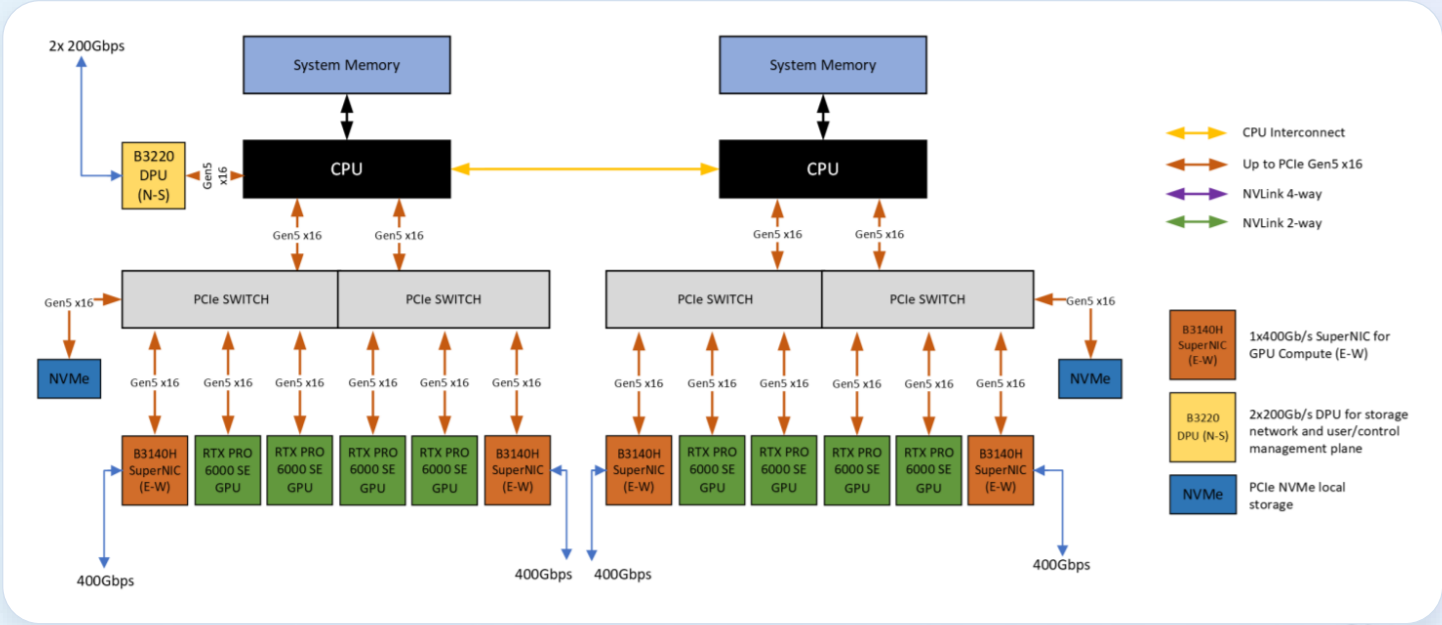
Rodina	Scalable Unit	Minimum	Přírůstek	Maximum
RTX PRO	4 nody (32 GPU)	1 SU / 4 nody	+4 nody	8 SU · 32 nodů · 256 GPU
HGX	4 nody (32 GPU)	1 SU / 4 nody	+4 nody	32 SU · 128 nodů · 1 024 GPU
NVL72	1 rack (72 GPU)	1 rack	+1 rack	8 racků · 576 GPU

Kdy oddělit sítě: RTX PRO do 16 nodů sloučená síť North-South + East-West, nad 16 nodů oddělené. HGX: od 16 nodů doporučeno oddělit, 32 SU = spine-leaf. NVL72: vždy oddělená North-South a dual-plane East-West.

Zdroj: NVIDIA Reference Architectures for Enterprise AI Factories (white paper) – Design Tenets (Scalable Unit), Deep Dive

DO HLOUBKY · RTX PRO AI FACTORY

RTX PRO: node 2-8-5-200





Např.: **HPE ProLiant DL380a Gen12**

Varianty: **2-4-3-200** a **2-2-3-400** (levnější, 2–4 GPU), **2-X-1-NA** (bez sítě East-West).

Konfigurace serveru

- **8× RTX PRO 6000 BSE** – 96 GB GDDR7, až 1,6 TB/s na GPU
- **PCIe Gen5 x16** pro každé GPU, přes PCIe switche
- **East-West:** 4× BlueField-3 B3140H, 1× 400G – 1 SuperNIC na 2 GPU
- **North-South:** 1× BlueField-3 B3220, 2× 200G
- **CPU:** AMD EPYC nebo Intel Xeon; min. 7 jader na GPU
- **RAM:** min. 128 GB na GPU (doporučeno ≥ 1 TB)

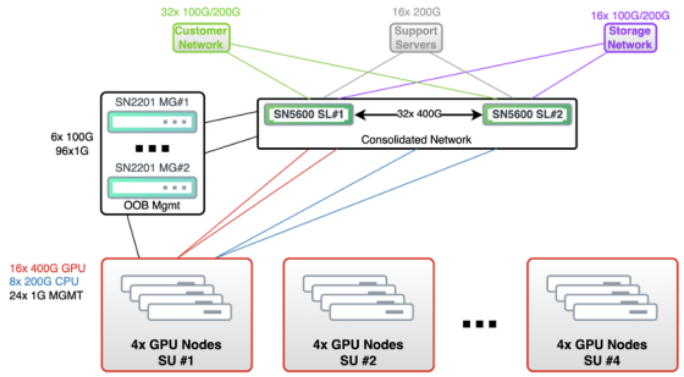
Velikost modelu

Inference až ~70B parametrů na GPU; nad ~40B v FP16 je pro fine-tuning potřeba paralelismus.

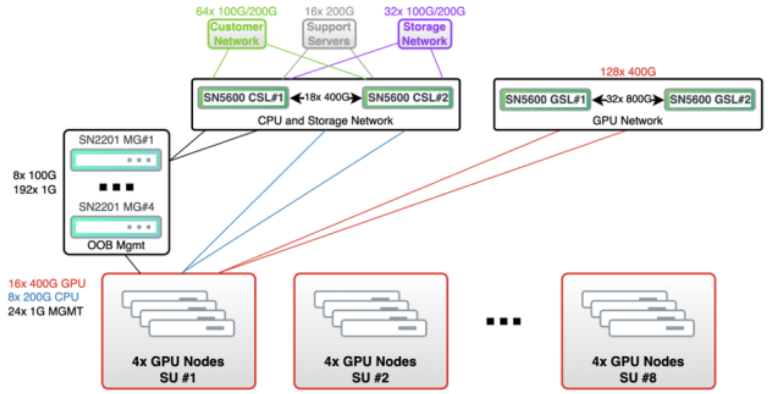
Zdroje: NVIDIA Enterprise RA white paper – Reference Architectures Deep Dive (Figure 1); NVIDIA RTX PRO AI Factory Enterprise RA. Foto: HPE

DO HLOUBKY · RTX PRO AI FACTORY

RTX PRO: škálování od 4 do 32 nodů



4 SU – 16 Nodes – 128 GPUs
Consolidated North-South / East-West Network



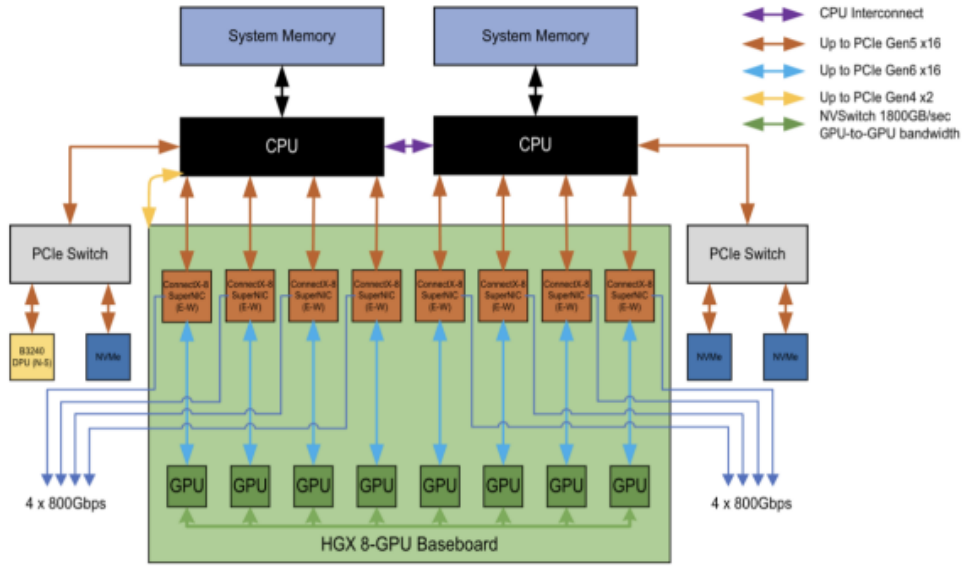
8 SU – 32 Nodes – 256 GPUs
Dedicated North-South / East-West Network

Nody	GPU	SU	SN5610 compute	SN5610 converged	SN2201 OOB	Na 1 SU (4 nody)
16	128	4	2	sloučeno s E-W	2	E-W 16× 400G = 6,4 Tb/s
32	256	8	2	2	4	N-S 8× 200G = 1,6 Tb/s · OOB 24× 1G

Zdroje: NVIDIA Enterprise RA white paper – Reference Architectures Deep Dive (Figure 4); NVIDIA RTX PRO AI Factory Enterprise RA – Network Logical Architecture

DO HLOUBKY · HGX AI FACTORY

HGX: node 2-8-9-800



Konfigurace serveru

- **HGX B300:** 8× Blackwell Ultra SXM, 288 GB HBM3e na GPU (2,3 TB na node)
- **NVLink 5:** 1,8 TB/s mezi GPU, 14,4 TB/s celkem
- **East-West:** 8× ConnectX-8 SuperNIC na desce, 800G na GPU (2× 400G)
- **North-South:** BlueField-3 B3240, 2× 400G – doporučeno kvůli offloadu KV cache
- **CPU:** 2 sockety, doporučeno 56 jader na socket; RAM ≥ 2 TB
- **Model:** čistá inference až ~120B parametrů na GPU



Např.: **HPE Compute XD690**
10U, 8× B300 SXM, vzduchem chlazený

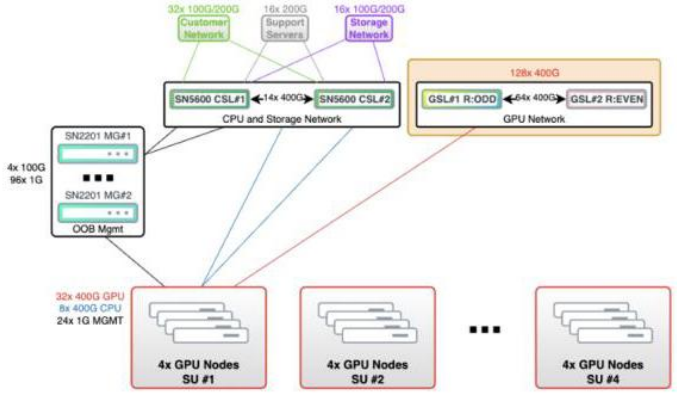
Propustnost East-West na GPU: 4× víc



Zdroje: NVIDIA Enterprise RA white paper – Reference Architectures Deep Dive (Figure 5); NVIDIA HGX AI Factory Enterprise RA. Foto: HPE

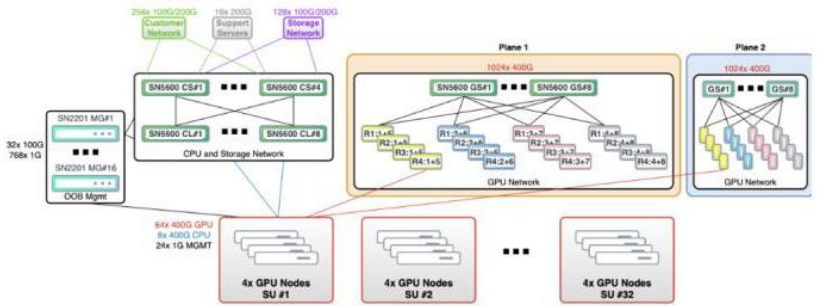
DO HLOUBKY · HGX AI FACTORY

HGX: škálování od 4 do 128 nodů



4 SU – 16 Node – 128 GPUs

Dedicated North-South | **Single Plane** or Dual Plane East-West



32 SU - 128 Nodes – 1024 GPUs

Dedicated North-South | Single Plane or **Dual Plane** East-West

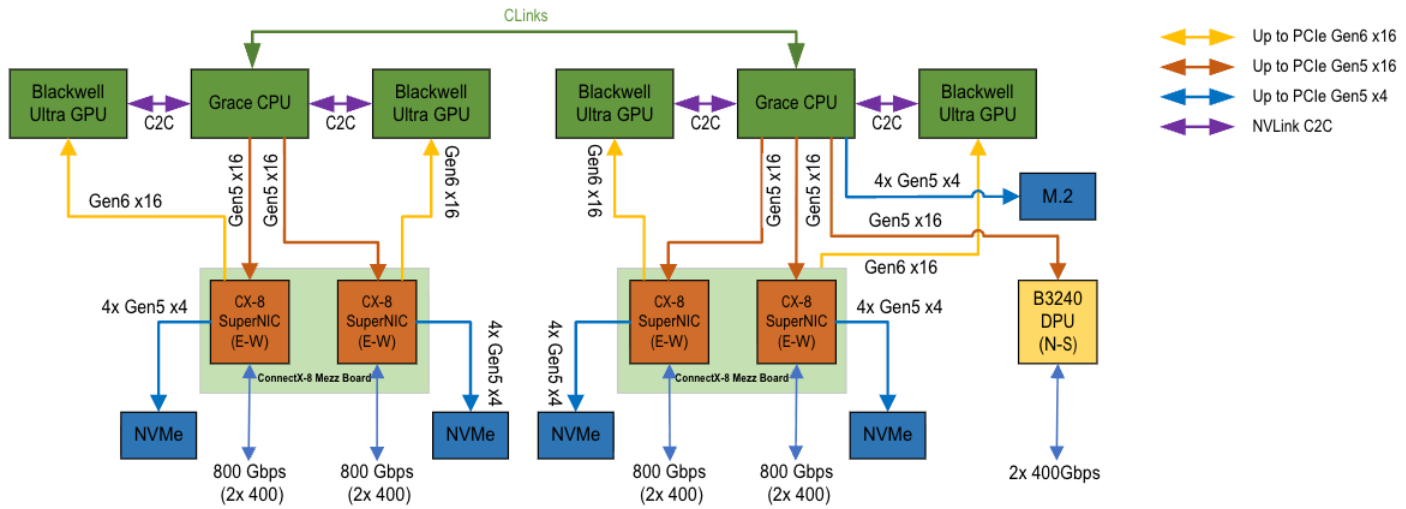
Nody	GPU	SU	Leaf	Spine	SN5600 compute	SN2201 OOB	E-W kabely mezi switchi
32	256	8	8	4	12	4	512
64	512	16	16	8	24	8	1 024
128	1 024	32	32	16	48	16	2 048

Dual-plane (doporučeno), obě roviny dohromady. Single-plane: −50 % propustnosti GPU.

Zdroje: NVIDIA Enterprise RA white paper – Reference Architectures Deep Dive (Figure 6); NVIDIA HGX AI Factory Enterprise RA – Tables 6 a 9

DO HLOUBKY · NVL72 AI FACTORY

NVL72: celý rack jako stavební kostka



Compute tray **2-4-5-800**: 2× Grace + 4× Blackwell Ultra (1 152 GB HBM3e), 4× ConnectX-8 (800G = 2× 400G), 1× BlueField-3 B3240.



SPRÁVA A SOFTWARE

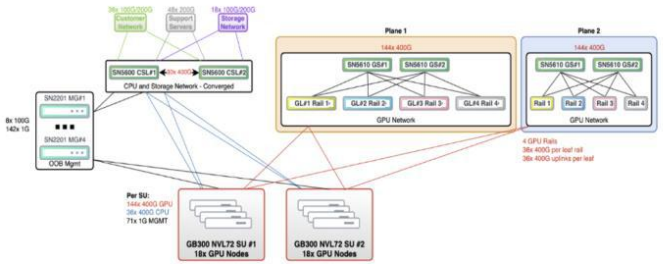
NVIDIA doporučuje Mission Control; NVLink oddíly spravuje Global Fabric Manager.

72	**GPU Blackwell Ultra**
36	**Grace CPU**
18 + 9	**compute + NVLink switch traye**
130 TB/s	**NVLink doména (1,8 TB/s na GPU)**
142 kW	**příkon racku GB300 NVL**
100 %	**kapalinové chlazení**

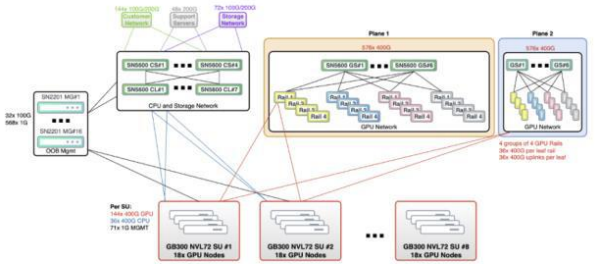
Zdroje: NVIDIA Enterprise RA white paper – Reference Architectures Deep Dive (Figure 7); NVIDIA NVL72 AI Factory Enterprise RA. Foto: HPE

DO HLOUBKY · NVL72 AI FACTORY

NVL72: škálování od 1 do 8 racků



2 SU — 36 Nodes – 144 GPUs
Dedicated North-South | Dual Plane East-West



8 SU — 144 Nodes – 576 GPUs
Dedicated North-South | Dual Plane East-West

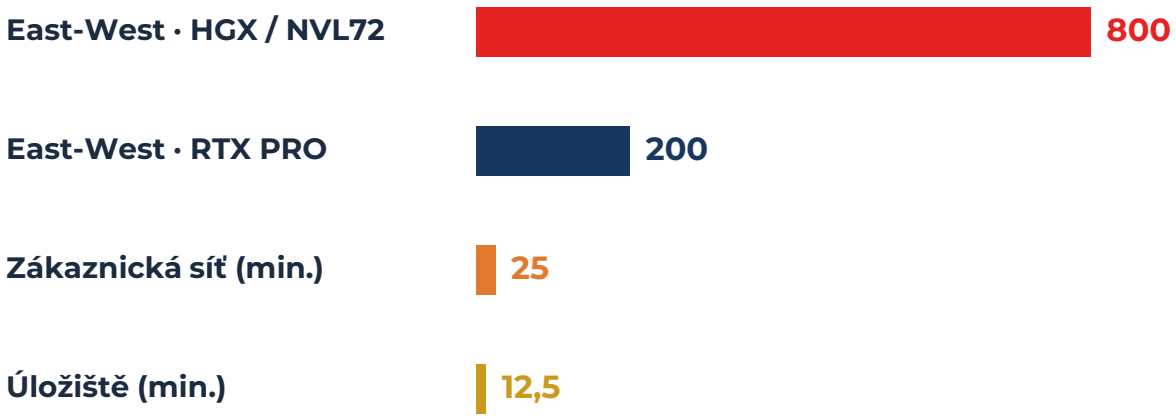
Racky	1	2	3	4	5	6	7	8
GPU	72	144	216	288	360	432	504	576
SN5600 výpočetní síť	12	12	24	24	32	32	44	44
SN5600 North-South	2	2	5	6	9	10	11	11
SN2201 OOB	2	4	6	8	10	12	14	16

Zdroje: NVIDIA Enterprise RA white paper – Reference Architectures Deep Dive (Figure 8); NVIDIA NVL72 AI Factory Enterprise RA – Table 11

DO HLOUBKY · ÚLOŽIŠTĚ A PROPUSTNOST

Pravidla dimenzování sítě a úložiště

Propustnost na jedno GPU (Gb/s)



PŘÍKLAD

256 GPU × 12,5 Gb/s = 3,2 Tb/s (cca 400 GB/s) do úložiště. North-South zvládne až 40 GB/s na node.

NVIDIA-Certified Storage

Program ověřuje úložiště pod AI zátěží a páruje je s Enterprise RA podle velikosti clusteru.

- **Foundation:** reference PCIe s RTX PRO 6000
- **Enterprise:** reference HGX B200

Lokální NVMe v nodu

Inference 1 TB, trénink 2 TB na CPU socket + 1 TB boot

HPE

Alletra Storage MP X10000 – první objektové úložiště s NVIDIA-Certified Storage (Foundation, ověřeno do 128 GPU), S3 over RDMA.

Zdroje: NVIDIA Enterprise RA (HGX, RTX PRO) – NVIDIA-Certified Storage; Enterprise AI Factory Design Guide; HPE GTC 2026

DO HLOUBKY · SOFTWARE

Softwarový stack Enterprise AI Factory

Aplikace a agenti	NVIDIA Blueprints: Enterprise RAG, AI-Q Research Agent, Video Search & Summarization · Agent Toolkit, OpenShell
Modely a inference	NVIDIA AI Enterprise: NIM mikroslužby, NeMo, Nemotron modely, Dynamo-Triton
Orchestrace	Run:ai (fronty, priority, frakční GPU) · Kubernetes · Slurm
Infrastrukturní SW	GPU Operator · Network Operator · NIM Operator · DOCA a DPF pro BlueField · NetQ
Správa clusteru	Base Command Manager (provisioning bare-metal) · Mission Control (NVL72)
Hardware	Enterprise RA: RTX PRO / HGX / NVL72 · Spectrum-X · NVIDIA-Certified Storage

Observabilita

- DCGM Exporter – metriky GPU
- Prometheus + Grafana – dashboardy GPU, NIM, K8s, síť
- NetQ – telemetrie switchů
- HPE OpsRamp – validováno pro Enterprise AI Factory

Zdroje: NVIDIA AI Enterprise Infrastructure Software; Enterprise AI Factory Design Guide; Observability Guide for Enterprise RAs

DO HLOUBKY · DIMENZOVÁNÍ

Kolik GPU potřebuje podnikový RAG chat

Velikost firmy	Souběžní uživatelé	LLM GPU	GPU celkem	Nody	Platforma
100–200 uživatelů	≤ 20	1	~3	1	RTX PRO 6000
500 uživatelů	30–50	1–2	~4–5	1	RTX PRO 6000
1 000 uživatelů	50–200	3–5	~6–10	1–2	RTX PRO 6000
5 000 uživatelů	500–1 000	16–30	~19–38	3–5	RTX PRO / H200 NVL
10 000 uživatelů	1 000–2 000	40–80	~47–92	6–12	RTX PRO / H200 NVL

Chat 128/128 tokenů, cílové TTFT ≤ 5 s (≤ 3 s nízká latence). Sumarizace potřebuje zhruba 2× víc LLM GPU.

SCALE 1X = 2,75 GPU

1 GPU LLM (Nemotron Super 49B) + 1 reranker + 0,5 embedding + 0,25 Milvus. Škáluje lineárně s účinností ~90 % až do 96X.

NIM NA 1 GPU

H100 NVL, Llama 3.1 8B FP8: ~140 souběžných uživatelů při TTFT ~1 s, ~3 000 tokenů/s. Run:ai s 0,5 GPU na model: 2× víc uživatelů.

Zdroje: NVIDIA Enterprise RAG Retrieval Scaling and Sizing Guide; NIM LLM with Run:ai Sizing Guide (Enterprise RA)

HPE AI TOVÁRNY · PORTFOLIO

NVIDIA AI Computing by HPE: tři typy AI továren

HPE Private Cloud AI
Pro podniky

- Hotové řešení na klíč
- 2. generace: ProLiant DL380a Gen12 + RTX PRO 6000
- Varianta i bez připojení (air-gapped)
- Rozšiřující racky pro víc GPU

HPE AI Factory at scale
Pro poskytovatele a tvůrce modelů

- Stovky až desítky tisíc GPU
- Bezpečný multi-tenant provoz
- Připraveno na NVIDIA NCP certifikaci
- NVIDIA Mission Control v 2026

HPE Sovereign AI Factory
Pro státní správu a regulované obory

- Data a provoz pod vlastní kontrolou
- Podpora NVIDIA AI Factory for Government
- Multi-tenant a air-gapped scénáře



HPE Private Cloud AI – AI továrna na klíč

SERVERY
ProLiant DL380a Gen12 · Compute XD690 · GB300 NVL72 by HPE

ÚLOŽIŠTĚ
Alletra Storage MP X10000 / B10000

SOFTWARE
NVIDIA AI Enterprise · Run:ai · OpsRamp · Morpheus · Data Fabric

SÍŤ
NVIDIA Spectrum-X

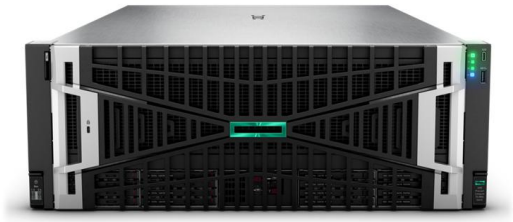
Zdroje: HPE – AI Factory with NVIDIA; tiskové zprávy HPE (06/2025, 10/2025, 03/2026). Foto: HPE

HPE AI TOVÁRNY · SERVERY

HPE servery v referenčních architekturách NVIDIA

ProLiant DL380a Gen12

RTX PRO AI Factory · 2-8-5-200



Formát	4U, 2× Intel Xeon 6
GPU	až 8× RTX PRO 6000 BSE
Chlazení	vzduch nebo DLC

SCHVÁLENO NVIDIA

Infrastructure ✓

16 a 32 nodů · 128 / 256 GPU

Compute XD690

HGX AI Factory · 2-8-9-800



Formát	10U, 2× Xeon 6 (64 jader)
GPU	HGX B300, 8× 288 GB
Příkon	~14,3 kW trvale, 16,4 kW špička

SCHVÁLENO NVIDIA

Infrastructure ✓ · Spectrum-X ✓

128 nodů · 1 024 GPU · dual-plane

GB300 NVL72 by HPE

NVL72 AI Factory · 2-4-5-800



Formát	48U rack, 72 GPU + 36 Grace
Chlazení	DLC, CDU 1,3 MW pro až 8 racků
Příkon	~132 kW nominálně, ~155 kW špička

REFERENČNÍ NÁVRH

dle NVL72 RA

1–8 racků · 72–576 GPU

Zdroje: NVIDIA Partner Endorsed Designs; HPE QuickSpecs DL380a Gen12, XD690, GB300 NVL72 by HPE. Foto: HPE

HPE AI TOVÁRNY · VÝBĚR

Jak vybrat AI továrnu: tři otázky

1 Máte kapalinové chlazení a víc než 100 kW na rack?

- ANO → **NVL72 AI Factory**
  GB300 NVL72 by HPE · největší modely
- NE ↓

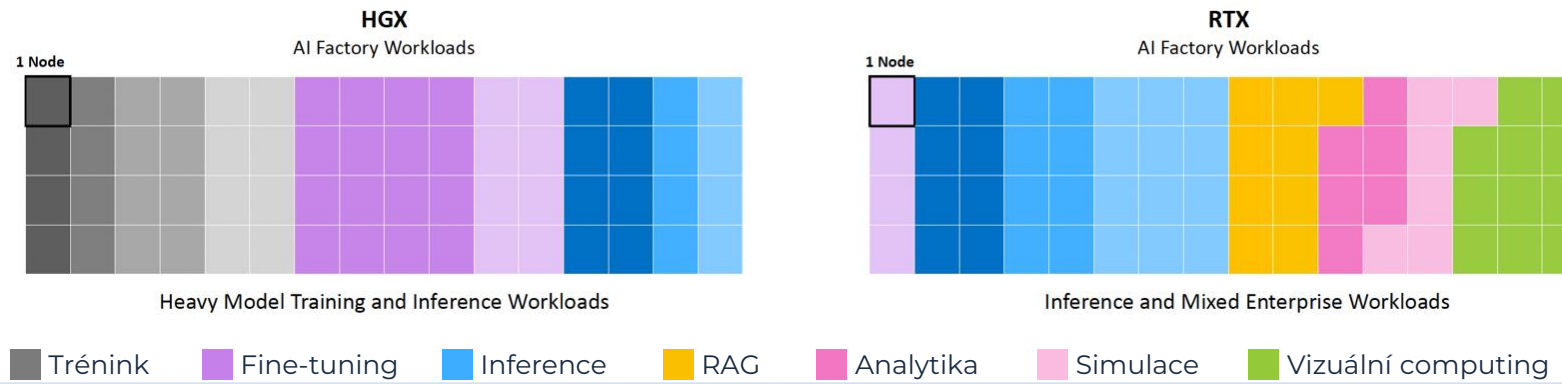
2 Potřebujete trénink nebo distribuovanou inferenci velkých modelů?

- ANO → **HGX AI Factory**
  HPE Compute XD690 · 800 Gb/s na GPU · ~14–16 kW/server
- NE ↓

3 Inference, agenti, vizuální computing, běžné DC pod 20 kW na rack?

- ANO → **RTX PRO AI Factory**
  HPE ProLiant DL380a Gen12 · PCIe · vzduch

Poznejte svůj workload – z něj vychází návrh AI továrny (1 čtvereček = 1 node)



A PAK VELIKOST

Začněte jednou Scalable Unit (4 servery = 32 GPU) a rosťte po 4 serverech, u NVL72 po racích. Rychlý start na klíč: HPE Private Cloud AI.

Zdroje: NVIDIA Enterprise RA white paper a dokumenty; HPE QuickSpecs; grafika workloadů: NVIDIA. Zjednodušený postup M Computers.

NVIDIA Extreme Co-Design Delivering X-Factors Every Year
From Chips to Racks to AI Factories



mcomputers

Děkuji

KONTAKTY

Petr Plodík · petr.plodik@mcomputers.cz