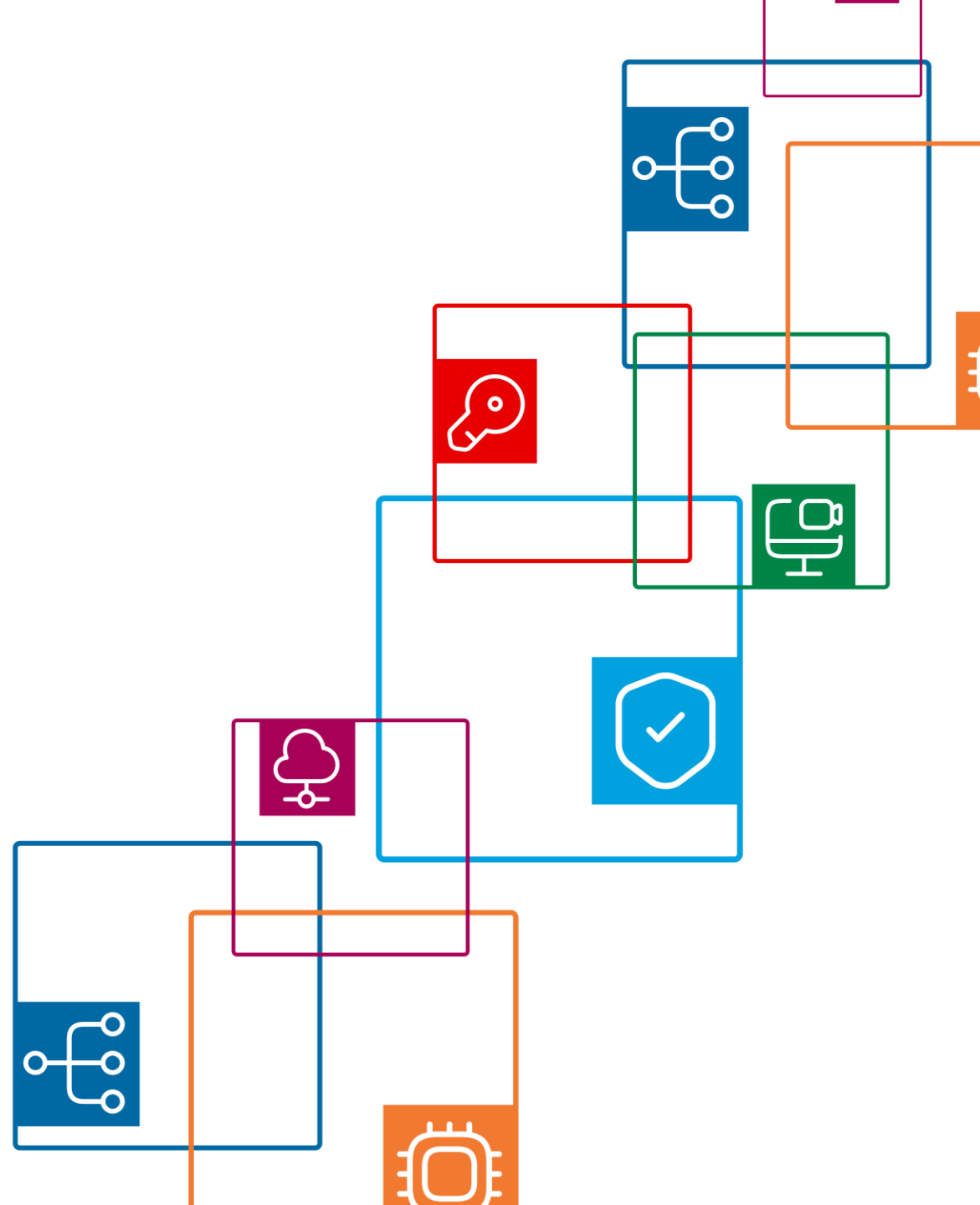




Onprem AI bez iluzí

Martin Kylián | CESNET



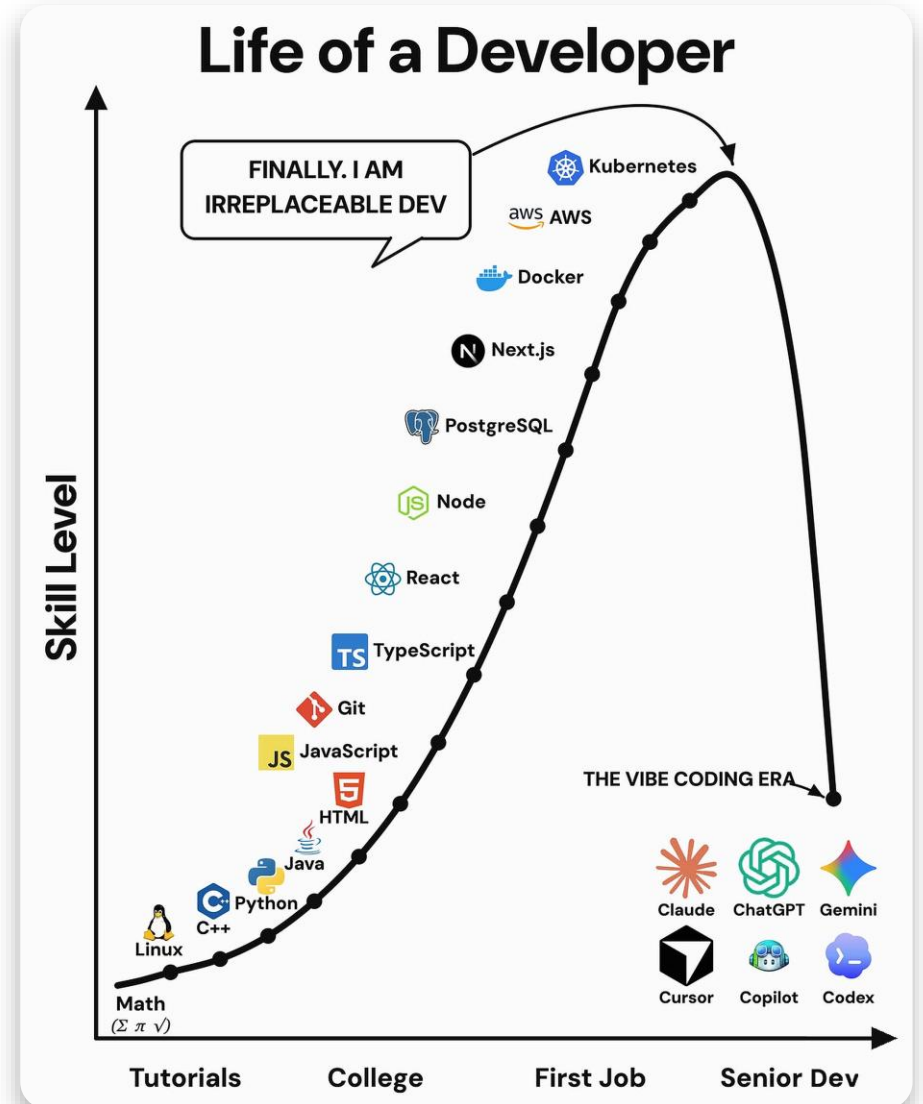
\$200 Vibecoder

- Claude Code: nová hračka
- Fascinace
- Všechno jde, rychle



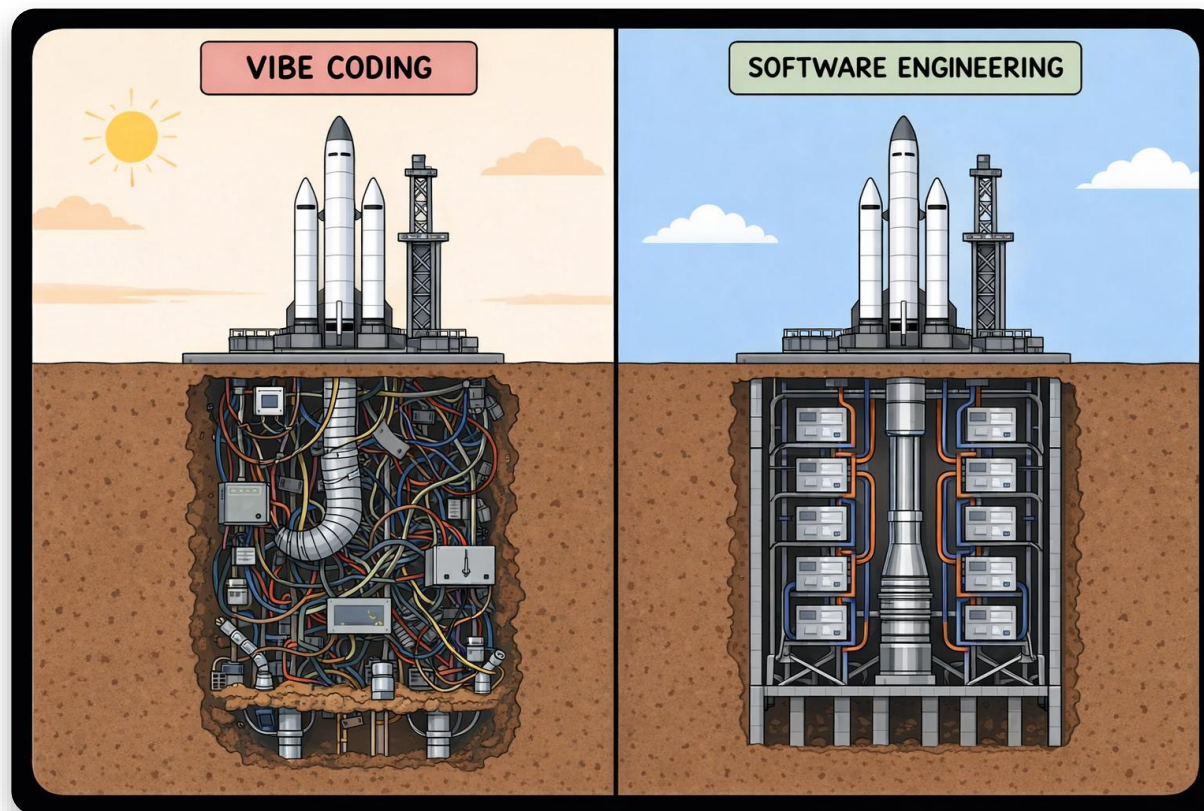
- Přestávám přemýšlet: prompt místo návrhu
- Méně ověřuji: „vypadá to, že to funguje“
- Víc důvěřuji: výstup zní sebejistě

Po půl hodině vibe codingu...



Vibe coding

- Rychlé fixy
- Žádná struktura
- Spaghetti kód
- "Nějak" to funguje



Software engineering

- Architektura
- Čistý kód
- Škálovatelné
- Udržovatelné



- Cíl: výsledek, ne vygenerovaný kód
- Proces: brainstorm → spec → plán → TDD → review
- Skills a pluginy řídí, jak agent pracuje
- Ověřování je součástí procesu



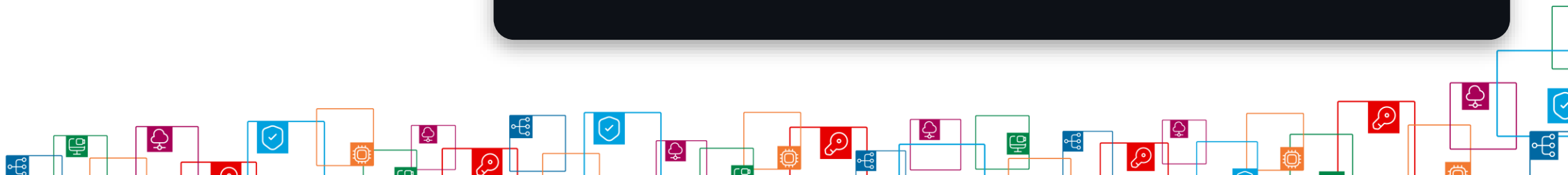
top agent skills · 2026-09-30

```
$ gh api repos/... --jq .stargazers_count
```

Repository

★ Stars

1	obra/superpowers	293 289
2	mattpocock/skills	272 647
3	multica-ai/andrej-karpathy-skills	216 002
4	anthropics/skills	179 110
5	nextlevelbuilder/ui-ux-pro-max-skill	131 876
6	JuliusBrussee/caveman	108 526
7	addyosmani/agent-skills	100 058
8	Leonxlnx/taste-skill	91 449
9	ComposioHQ/awesome-claude-skills	75 980
10	ayghri/i-have-adhd	52 304



```
kens)
ette (Cmd+Shift+P) and run "Shell Command: Install
o enable IDE integration

ew
Scripts: parallel find...44/112 agents · 47m32s
3 agents
ity:authn-session Opus 5.5 (1M context)
ity:authz-admin Opus 5.5 (1M context)
ity:idor-ownersh... Opus 5.5 (1M context)
ity:oauth-mcp-ap... Opus 5.5 (1M context)
ity:injection Opus 5.5 (1M context)
1-5 of 18 ↓
p pause · esc back · s save
```

18 agentů

paralelně · 47 min · Opus 5.5 (1M kontext)

~410 M tokenů

odhad spotřeby jednoho běhu

~260 USD

odhad ceny · ceník 4 / 20 USD za 1M in/out

...a výsledek za to stojí.



- Vlastní model = žádné účty za tokeny
- Apple silicon, DGX Spark, herní GPU, open model...

**Instantní káva ve velkém balení.
Bude to chvíli trvat.**



297,9 tok/s

prefill (čtení promptu)

20,5 tok/s

decode (generování)

PROMPT PROCESSING (EXCL. CACHED)

297.9 tok/s

TOKEN GENERATION

20.5 tok/s**100k tokenů kontextu ≈ 5,6 min**

jen na přečtení promptu (prefill)

+ 2 000 tokenů odpovědi ≈ další 1,6 min

Frontier model nemůže dělat všechno

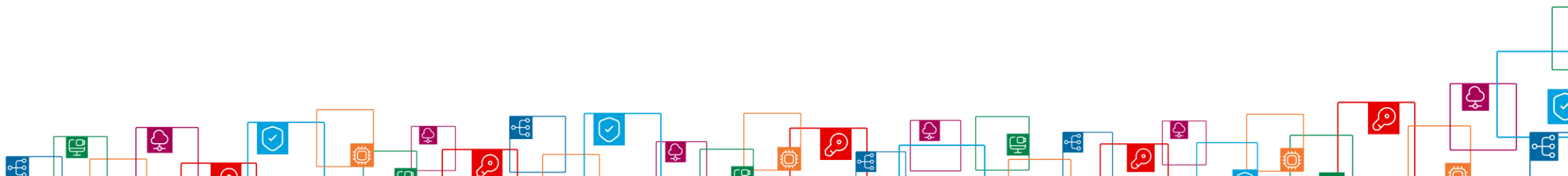
> Message from @ad17cc9dce16447ec (ctrl+o to expand)

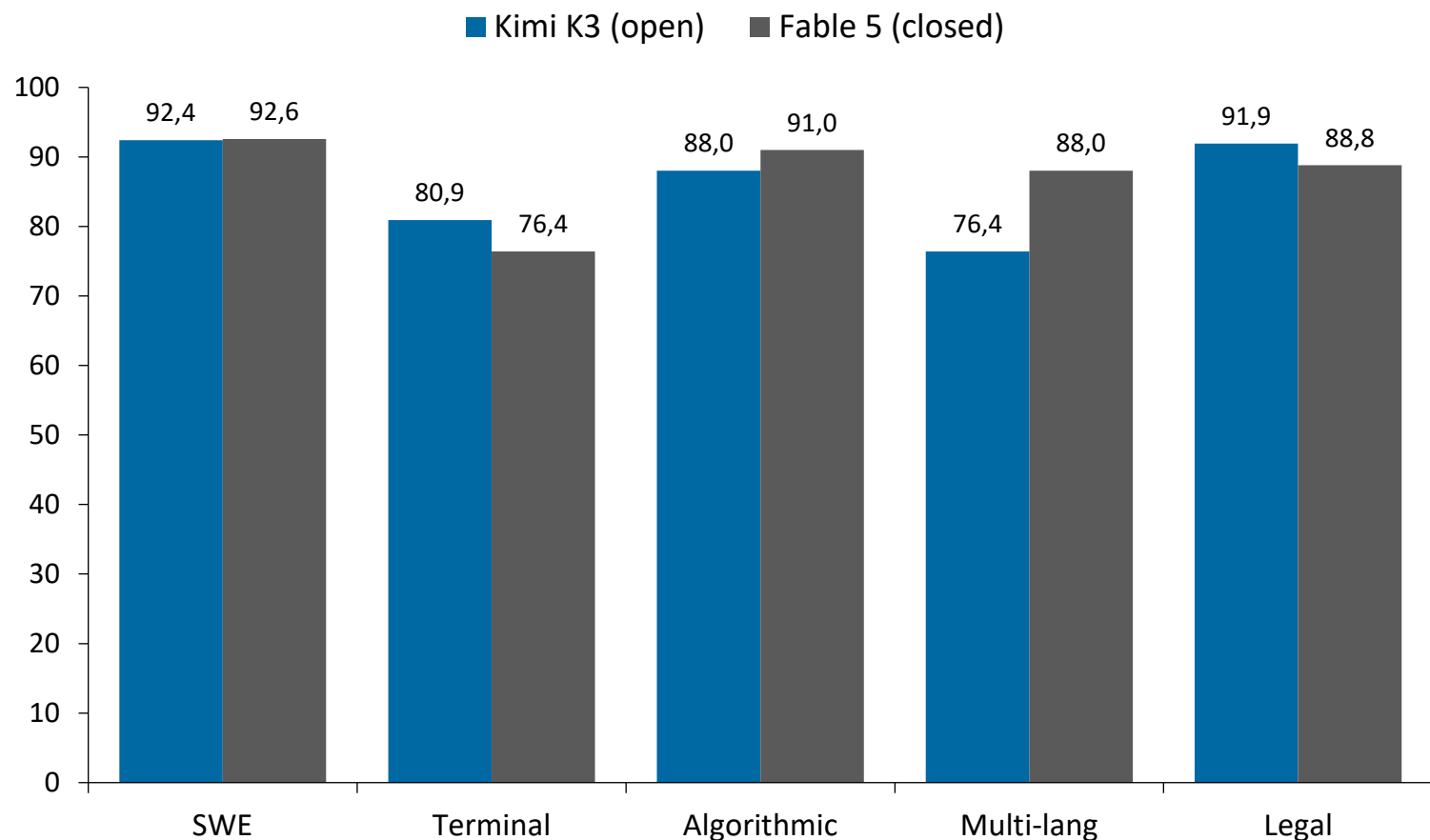
• Opus 5 (1M context)'s safeguards flagged this message. Our intentionally broad safeguards allow us to deliver more capabilities faster, but can sometimes flag legitimate coding, cybersecurity, and biology tasks. Switched to Opus 4.8 (1M context). Send feedback with /feedback or learn more: <https://support.claude.com/en/articles/16049681>

Details: `[cyber]`

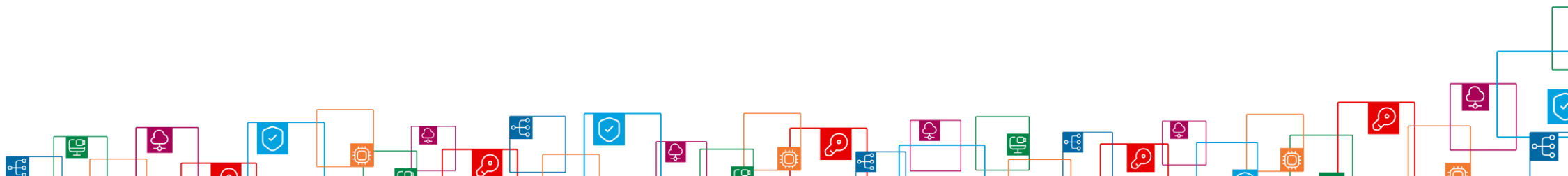
Tip: You can configure model switch behavior in /config

- Safeguards blokují legitimní security práci (jsem součástí Cyber Verification Program)
- Citlivá data nesmí ven z organizace
- Závislost na cizí politice, ceně a dostupnosti





- Open weights: můžu ho provozovat sám
- Na úrovni nejlepšího uzavřeného modelu
- Žádné cizí safeguardy, data zůstávají doma
- Licence: MIT-styl, podmínky až nad 20 M USD revenue (ne pro interní použití)





1× DGX B300

minimum (váhy 1,56 TB se tak tak vejdou)

8 × B300 · 2,3 TB HBM3e

- Model: 2,8T parametrů (MoE), aktivních 104B
- Váhy: ~1 561 GB (MXFP4, QAT)

4× DGX B300

doporučení pro produkci (64+ H200 GPU)

~300–650 tis. USD za kus



Dell PowerEdge XE9680

8× H100/H200 SXM · ~300 tis. USD

NVLink · velké modely přes 8 GPU



HPE ProLiant DL385 Gen11

4× RTX PRO 6000 (384 GB) · ~150 tis. USD (odhad)

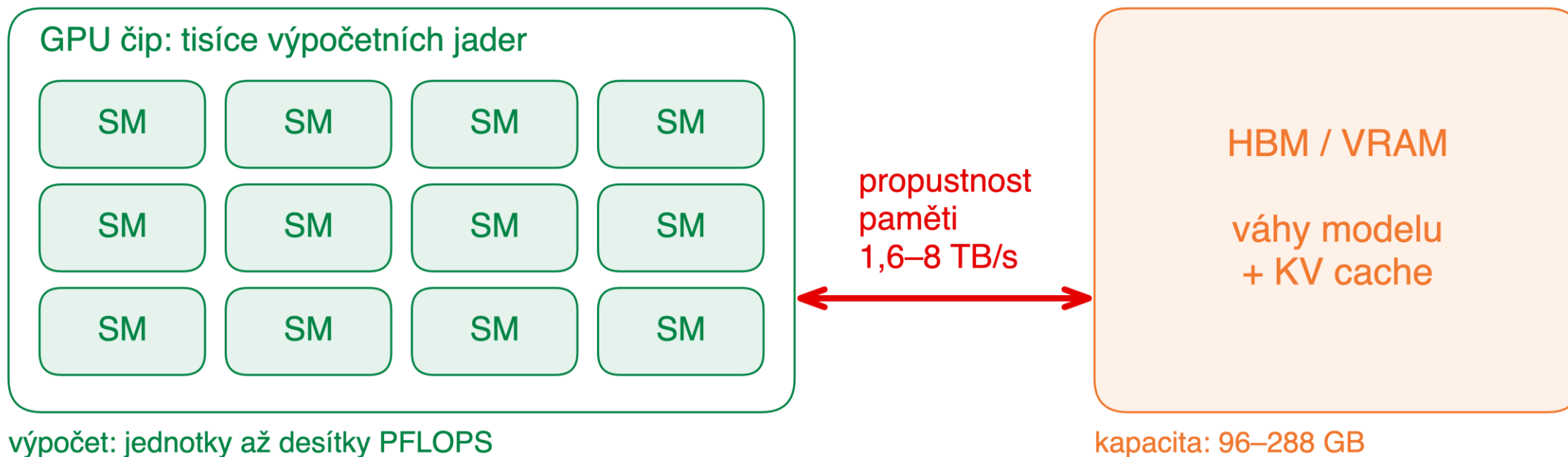
PCIe · levnější, menší modely / víc replik

GPU	Paměť	Propustnost	FP8 / FP4 dense (PFLOPS)	TDP	Propojení	Cena (orient.)
RTX PRO 6000 BSE	96 GB GDDR7	1,6 TB/s	FP8 ~1 / FP4 ~2	600 W	PCIe 5.0 (bez NVLink)	~13–16 tis. USD
H100 SXM	80 GB HBM3	3,35 TB/s	FP8 ~2 / FP4 –	700 W	NVLink 4, 900 GB/s	~25–40 tis. USD
H200 SXM	141 GB HBM3e	4,8 TB/s	FP8 ~2 / FP4 –	700 W	NVLink 4, 900 GB/s	~31–45 tis. USD
B200 (HGX/DGX)	180 GB HBM3e	8 TB/s	FP8 4,5 / FP4 9	1 000 W	NVLink 5, 1,8 TB/s	~30–50 tis. USD
B300 (HGX/DGX)	288 GB HBM3e	8 TB/s	FP8 4,5 / FP4 13,5	1 100 W	NVLink 5, 1,8 TB/s	~40–55 tis. USD
GB300 (NVL72)	288 GB HBM3e	8 TB/s	FP8 5 / FP4 15	1 400 W	NVLink 5, 1,8 TB/s	~42–56 tis. USD

Výkon bez sparsity. Ceny: tržní odhady 2026, NVIDIA ceníky nezveřejňuje.

RTX PRO 6000 = PCIe bez NVLinku · H/B řada v HGX/DGX = NVLink · cena roste řádově

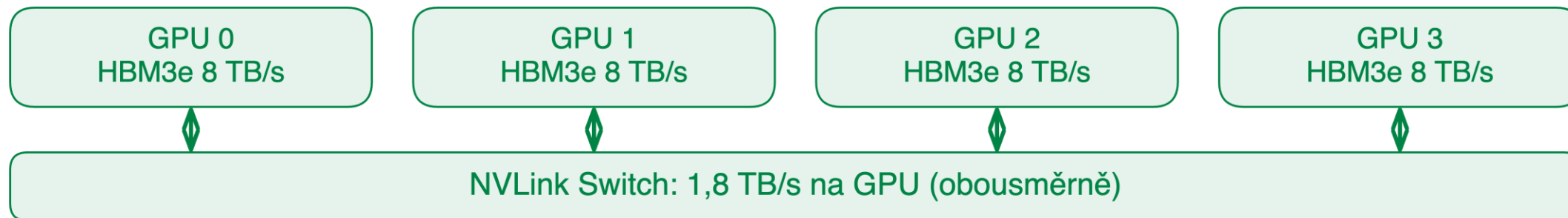




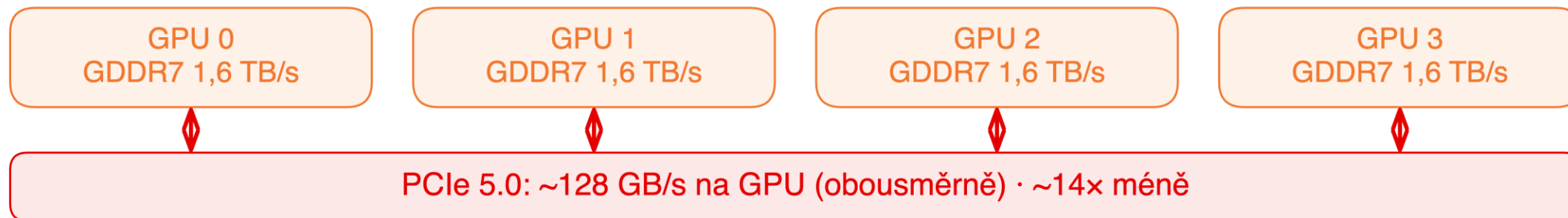
Decode: každý token čte všechny aktivní váhy
tok/s (1 uživatel) $\approx$ propustnost paměti $\div$ bajty aktivních vah
 $8 \text{ TB/s} \div 32 \text{ GB} \approx 250 \text{ tok/s}$ (teoretický strop)



Server s NVLink (HGX / DGX B300)



Server s PCIe (RTX PRO 6000)



Tensor parallelism: po každé vrstvě all-reduce mezi GPU
→ stovky synchronizací na token
PCIe: menší TP, víc replik · velký model přes víc karet = NVLink



request 1

POST /v1/chat/completions

```
{
  "model": "kimi-k3",
  "messages": [
    {"role": "system", "content": "Jsi DevOps asistent."},
    {"role": "user", "content": "Kolik GPU má node gpu-01?"}
  ],
  "tools": [{"type": "function", "function": {
    "name": "kubectl_get",
    "parameters": {"type": "object", "properties": {
      "resource": {"type": "string"},
      "name": {"type": "string"}}}}}]
}
```

response 1

← 200 OK

```
{
  "choices": [{"message": {
    "role": "assistant",
    "tool_calls": [{
      "id": "call_1", "type": "function",
      "function": {"name": "kubectl_get",
        "arguments": "{\\"resource\\":\\"node\\",
          \\"name\\":\\"gpu-01\\"}"}}
    ]}],
  "usage": {"prompt_tokens": 412,
    "completion_tokens": 31}}
}
```

request 2

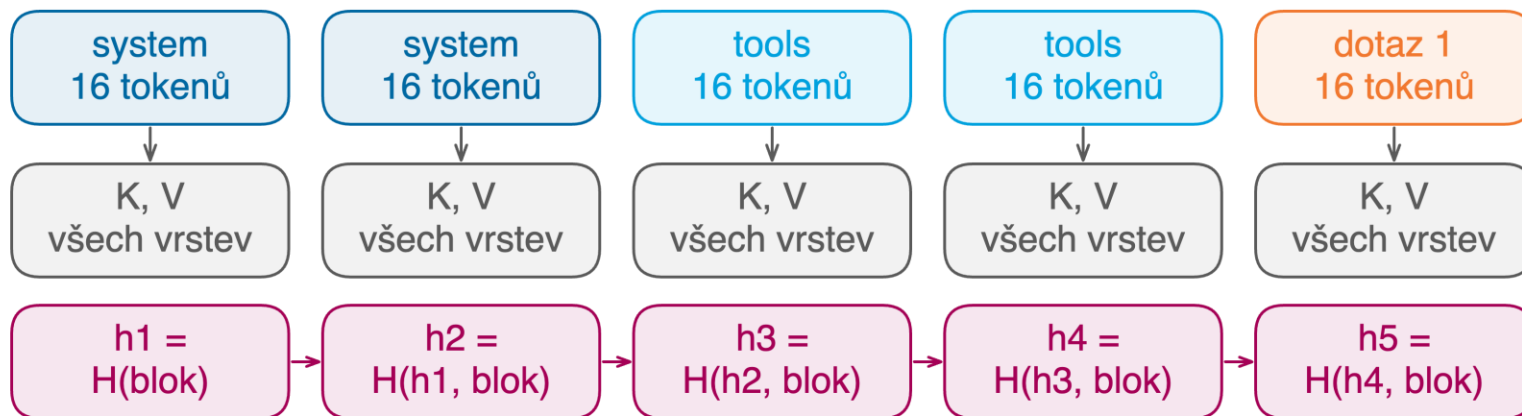
POST /v1/chat/completions (2. kolo)

```
"messages": [
  {"role": "system", "content": "Jsi DevOps asistent."},
  {"role": "user", "content": "Kolik GPU má node gpu-01?"},
  {"role": "assistant", "tool_calls": [{"id": "call_1", ...}]},
  {"role": "tool", "tool_call_id": "call_1", "content": "nvidia.com/gpu: 8"},
  {"role": "user", "content": "A kolik z nich je volných?"}
], "tools": [ ...stejně definice kubectl_get... ]
usage.prompt_tokens: 412 → 468 → ... roste s každým kolem
```

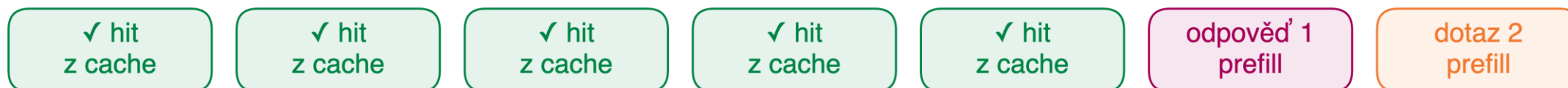
← znovu
 ← znovu
 ← znovu
 ← nové
 ← nové
 ← znovu

Model je bezstavový.
Pokaždé se posílá celá konverzace.
Prefill roste → proto KV cache.

Požadavek 1 → prefill všech tokenů



Požadavek 2: celá konverzace znovu



Cache se hledá podle obsahu, ne podle session:
nejdelší shodný prefix bloků.
Změna 1 tokenu na začátku (timestamp v system promptu)
= miss všeho za ní.

Model	Attention	KV / token (FP8)	128k BF16	128k FP8	1M FP8	1M NVFP4	Váhy
Qwen3.8-27B	hybrid: 16/64 vrstev GQA, 48 Gated DeltaNet	32,0 KiB	8,1 GiB	4,1 GiB	30,7 GiB	17,3 GiB	~22 GB NVFP4
GLM-5.3	MLA + DSA (sparse, top-2048)	43,9 KiB	11,0 GiB	5,5 GiB	41,8 GiB	23,5 GiB	~464 GB NVFP4
DeepSeek V4.1 Flash	CSA2: sdílená komprimovaná KV, FP4	1,6 KiB	0,4 GiB	0,2 GiB	1,5 GiB	0,8 GiB	~510 GB FP4/FP8
Kimi K3	hybrid: 24/93 vrstev MLA, 69 KDA	13,5 KiB	3,8 GiB	2,1 GiB	13,3 GiB	7,7 GiB	~1,56 TB MXFP4

Qwen3.8: KV roste jen v 16 z 64 vrstev, zbytek (DeltaNet) má konstantní stav ~147 MiB.

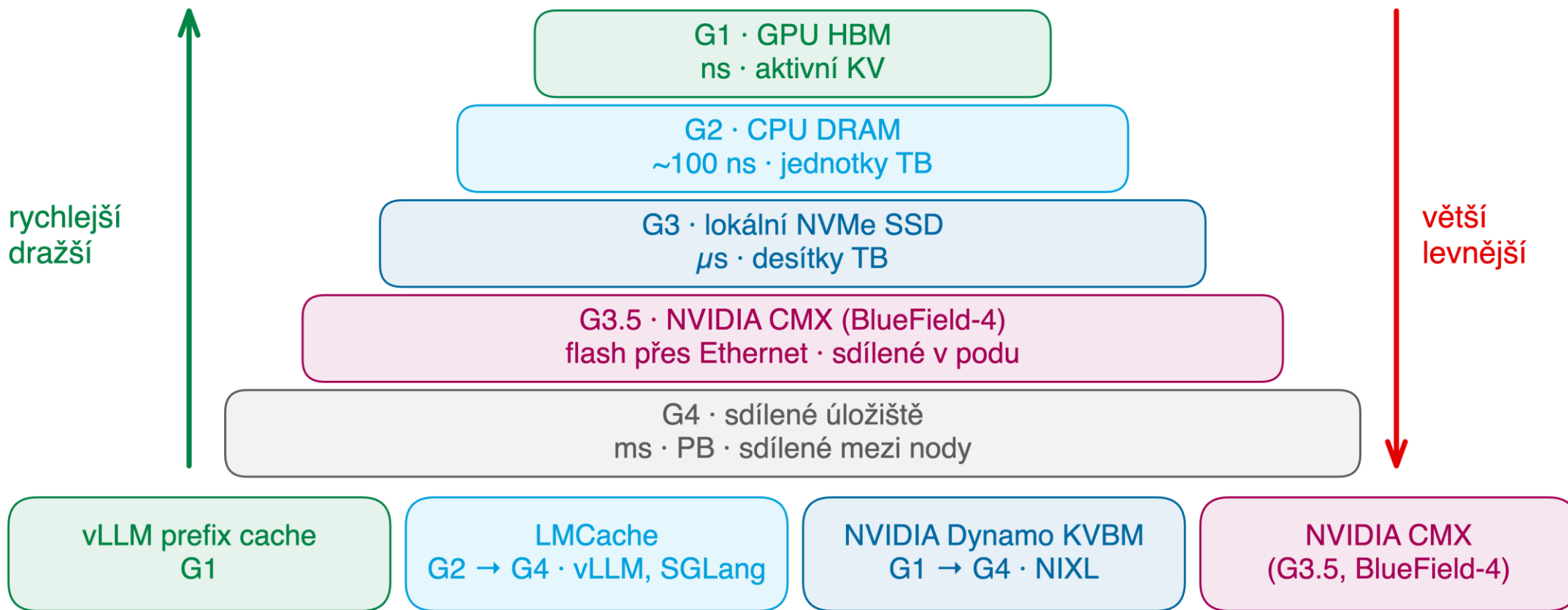
GLM-5.3: MLA komprimuje K a V do latentu 512+64 ve všech 78 vrstvách (bez DSA indexeru, +~6 %).

DeepSeek V4.1: KV sdílená napříč vrstvami + FP4 + bez SWA cache → 890 B/token (1/437 V1).

Kimi K3: KV roste jen v 24 z 93 vrstev (MLA), 69 KDA vrstev má konstantní stav ~414 MiB.

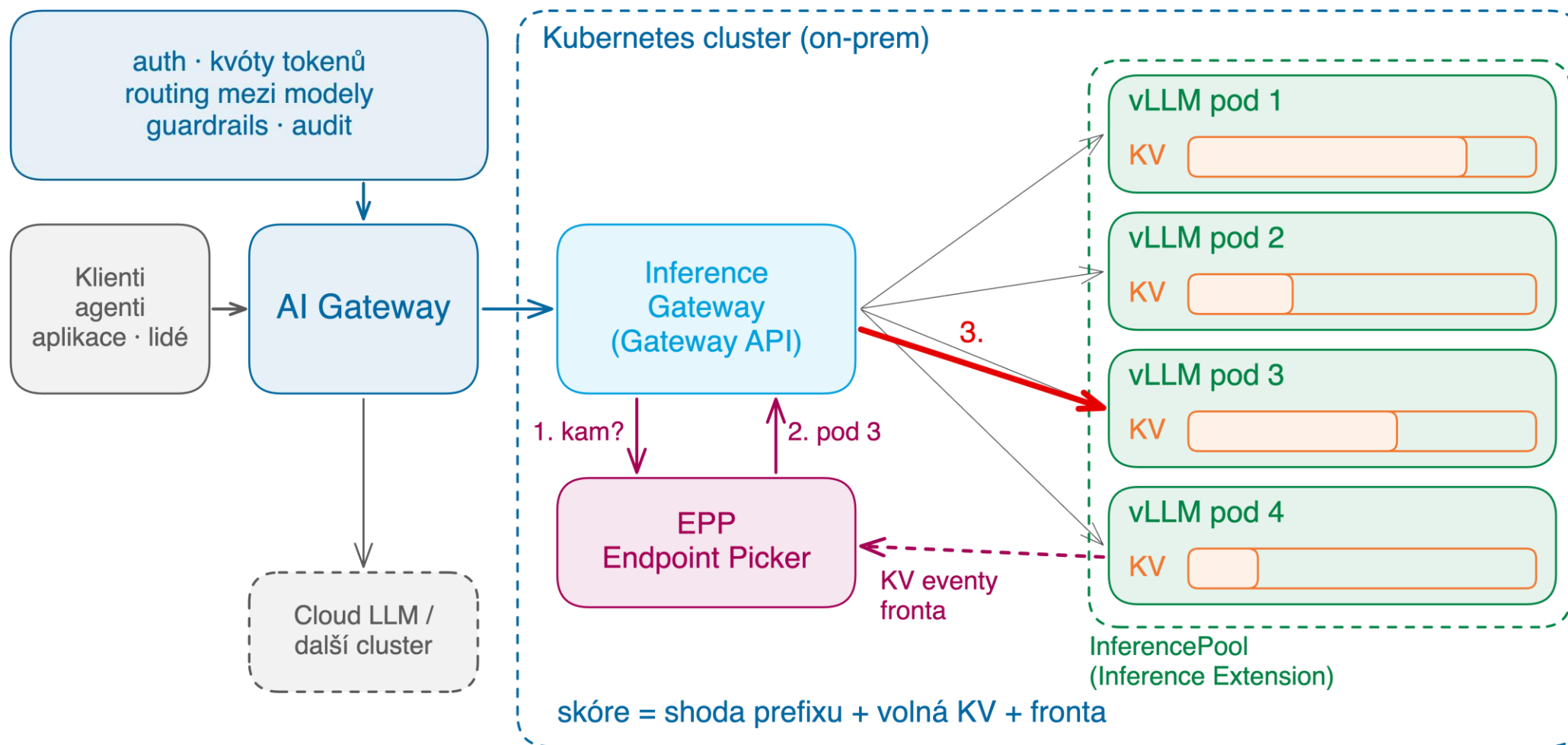
GLM-5.3 na 8× RTX PRO 6000: po vahách zbývá místo na KV cache pro ~38 uživatelů × 128k

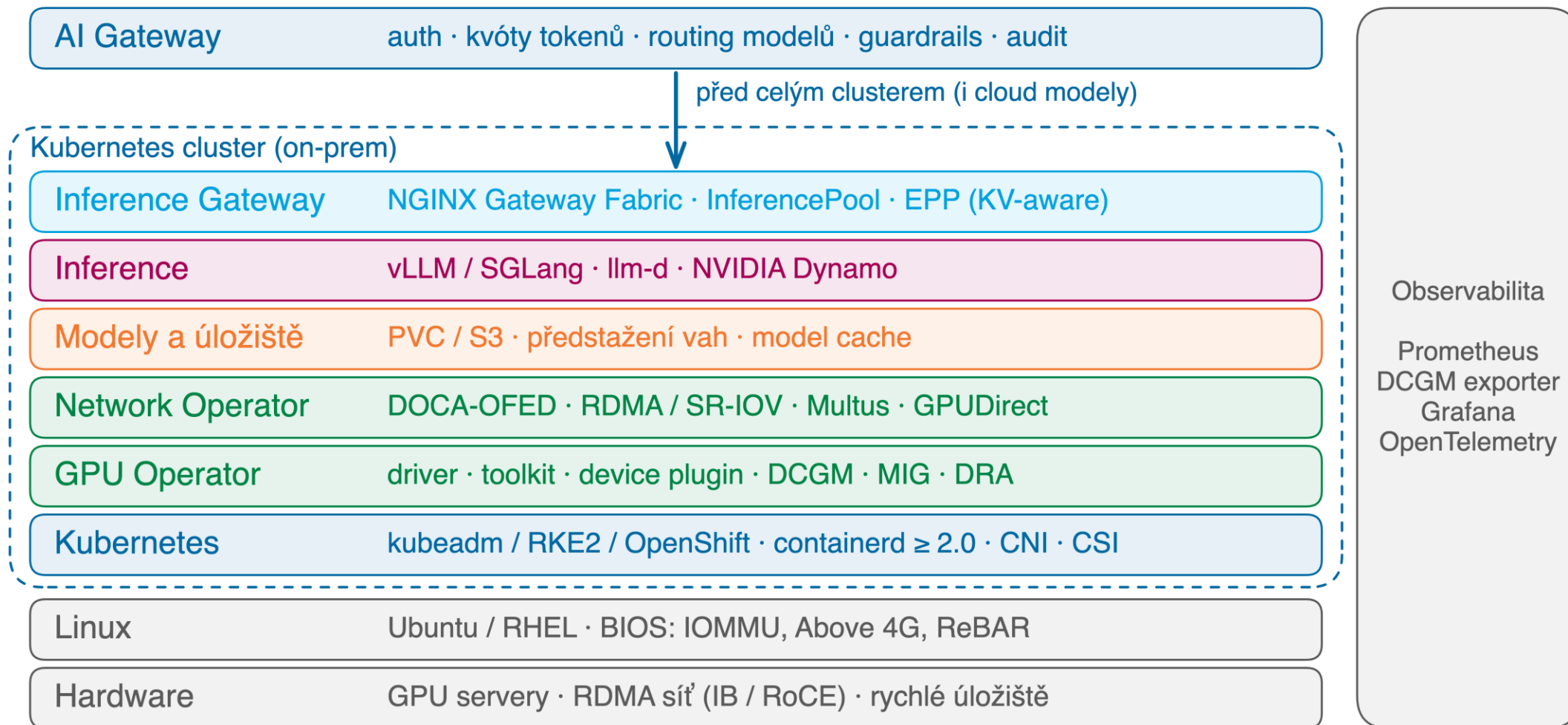




Načíst KV z nižšího tieru je rychlejší než znovu spočítat prefill → nižší TTFT







	vLLM	llm-d	NVIDIA Dynamo
Co to je	inference engine	K8s recepty nad vLLM + GAIE	distribuovaná serving vrstva
Prefix cache na replice	ano	z vLLM	z backendu
KV-aware routing	ne	ano (KV eventy, EPP)	ano (KV router / EPP)
P/D disaggregace	primitiva	ano (NIXL)	ano (NIXL)
KV offload	connectory, LMCache	ano	KVBM (G1 → G4)
Autoscaling	ne	variant autoscaler	Planner (SLA)
Licence	Apache-2.0	Apache-2.0	Apache-2.0, placená jen podpora
Kdy	1 server, 1 model	víc replik na K8s	multi-node, velké MoE, NVLink/RDMA

Dynamo i llm-d konvergují na stejné rozhraní: Gateway API InferencePool + EPP



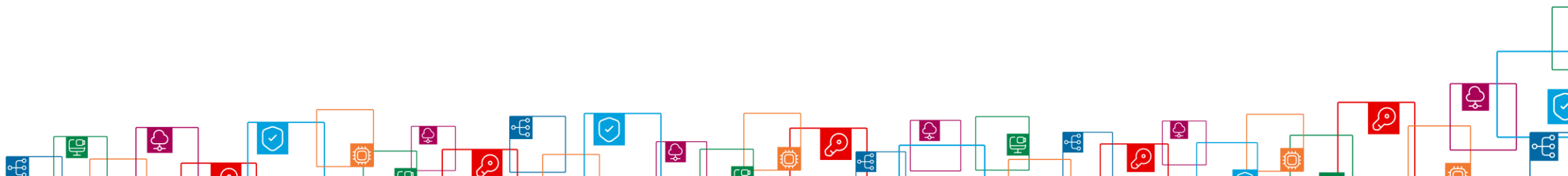
- SM120 kernely B12X (FP8/NVFP4 MoE, sparse MLA) + FlashInfer build pro SM120
- PCIe oneshot all-reduce místo NCCL: 1,4–6× nižší latence, ~+7 % decode (4 GPU)
- Laděný NCCL (XML graf, 8 kanálů): +69 % all-reduce BW na dual EPYC Turin
- Jeden image s profily: GLM-5.3, Qwen3.8, DeepSeek V4/V4.1 + volitelně LMCache
- Spekulativní dekódování MTP3 / DFlash2 / DSpark, měřená acceptance rate

```
RTX PRO 6000

$ docker pull \
  ghcr.io/local-inference-lab/vllm:karmic-kraken

# v produkci pin @sha256 digest
# po upgradu smazat JIT cache
# CUDA 13.4.1 · SM120 · PCIe bez NVLinku

★ 1 115 · rtx6kpro wiki
```



- Špičkový open model = stovky GB až TB paměti a NVLink, ne herní karta.
- Úzké hrdlo je paměť a propojení; výkon dělá KV cache a chytrý routing.
- Kubernetes je přirozený střed: GPU, repliky a Gateway API na jednom místě.
- AI Gateway je nutnost: auth, kvóty tokenů, routing mezi modely, audit.
- Licence rozhoduje taky: open weights $\neq$ bez podmínek (Kimi K3, NIM jen s AI Enterprise).





Děkuji

Otázky?

#PROPOJUJEME VĚDU

