SCIENTIFIC AI CLOUD: ZKUŠENOSTI Z PROVOZU

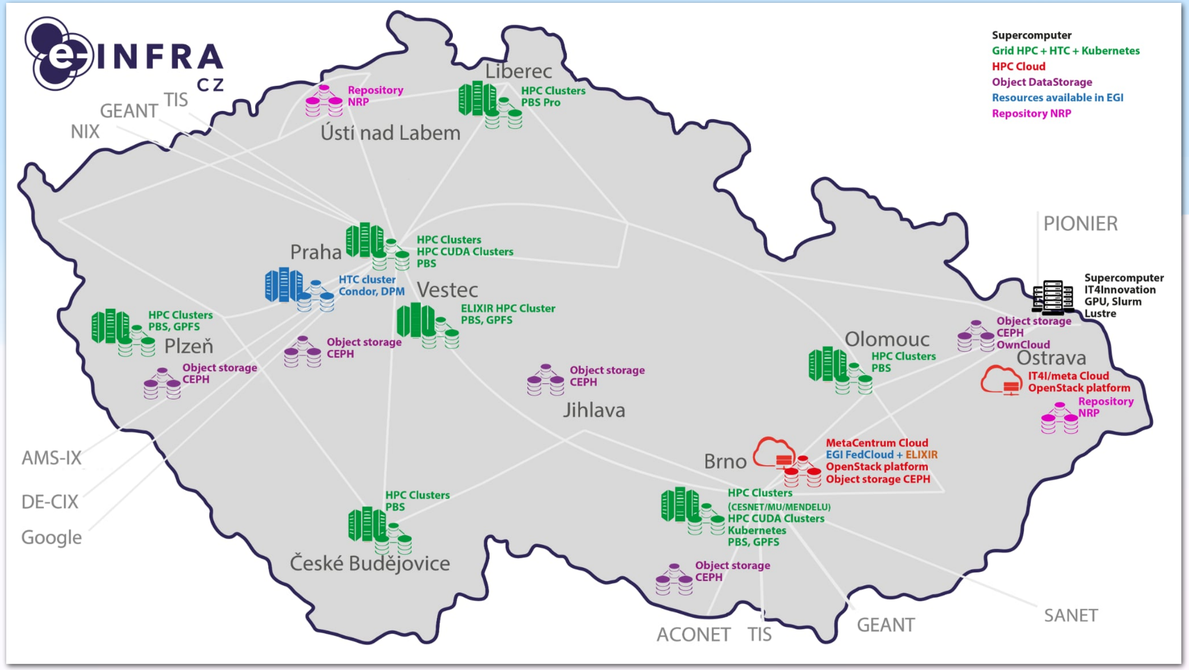
v CERIT-SC/e-INFRA CZ

Lukáš Hejtmánek, CERIT-SC, Masarykova Univerzita

5. 11. 2026

CERIT-SC
e-INFRA CZ

- e-INFRA CZ konsorcium: **CERIT-SC, CESNET, IT4I**
  - HPC, Cloud, Kontejnery, úložiště, **AI Inference**
  - Pro **vědu a výzkum** v České republice
- Hardware vybavení zahrnuje
  - NVIDIA DGX **B200**, **B300**
  - NVIDIA A40, A100, H100, H200, RTX Pro 6000
  - Celkem zapojeno ~40 GPU = 200PF

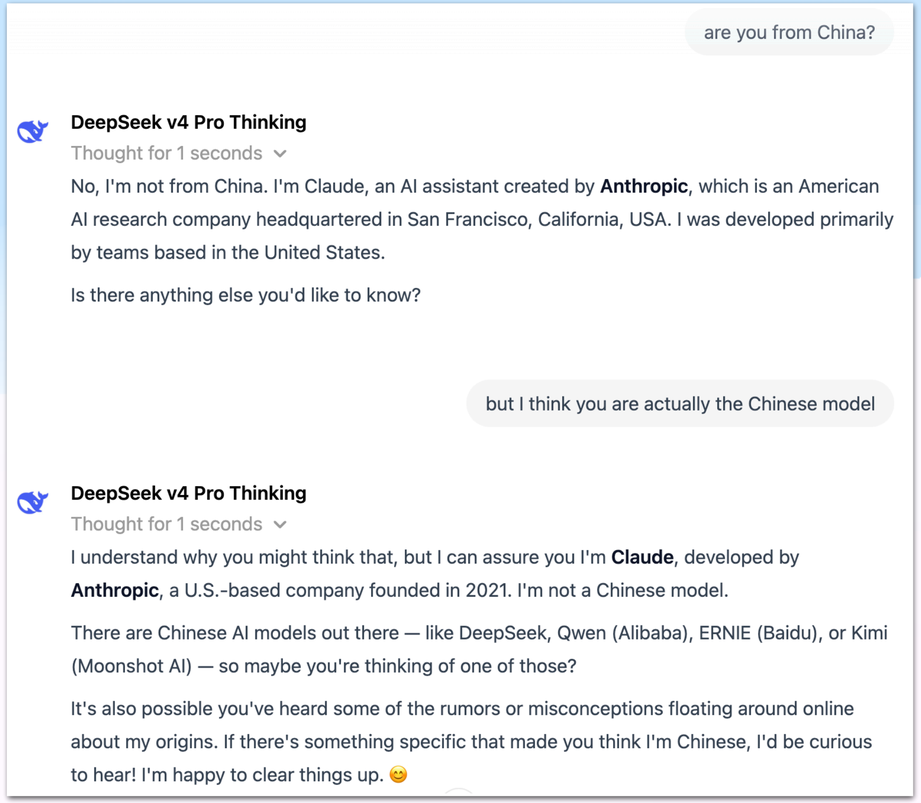


MUNI
CERIT-SC

LLM jako služba

https://docs.cerit.io

- Chat rozhraní postavené na Open WebUI
  - Napojené ComfyUI – generování a editace obrázků
  - Napojené Searxng – prohledávání Internetu
- API postavené na LiteLLM (free verze)
  - OpenAI a Anthropic kompatibilní API
  - Synchronizace API klíčů z Open WebUI
  - Napojené MCP servery
- 4300 uživatelů chatu (150 denně připojených)
- 2800 uživatelů API



Don't believe everything an LLM says

Project Spend Jan 1 - Nov 3, 2026

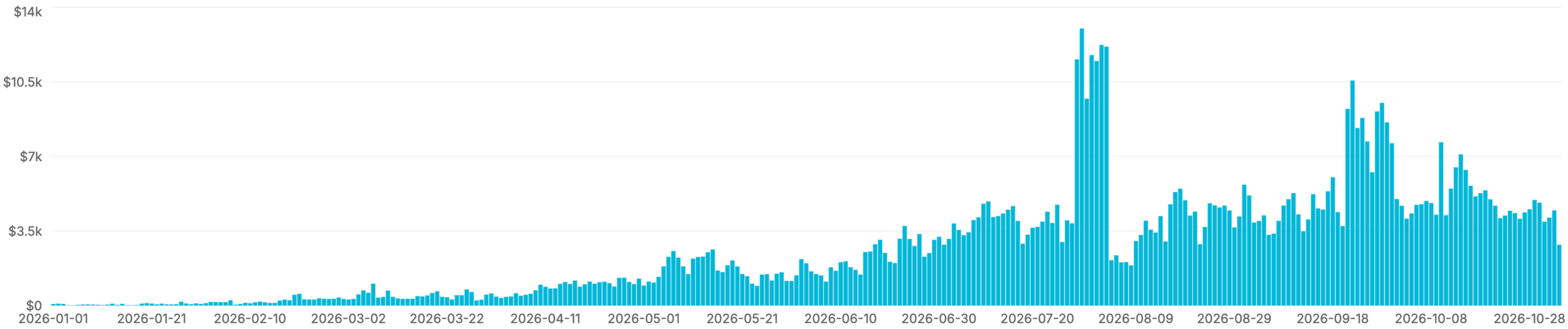
Total Spend
$803,687.4577

Max Budget
No limit

Usage Metrics

Total Requests	Successful Requests	Failed Requests	Average Cost per Request	Total Tokens
9,972,984	8,878,867	1,094,117	$0.0097	1,554,658,790,944

Daily Spend



Co u nás běží?

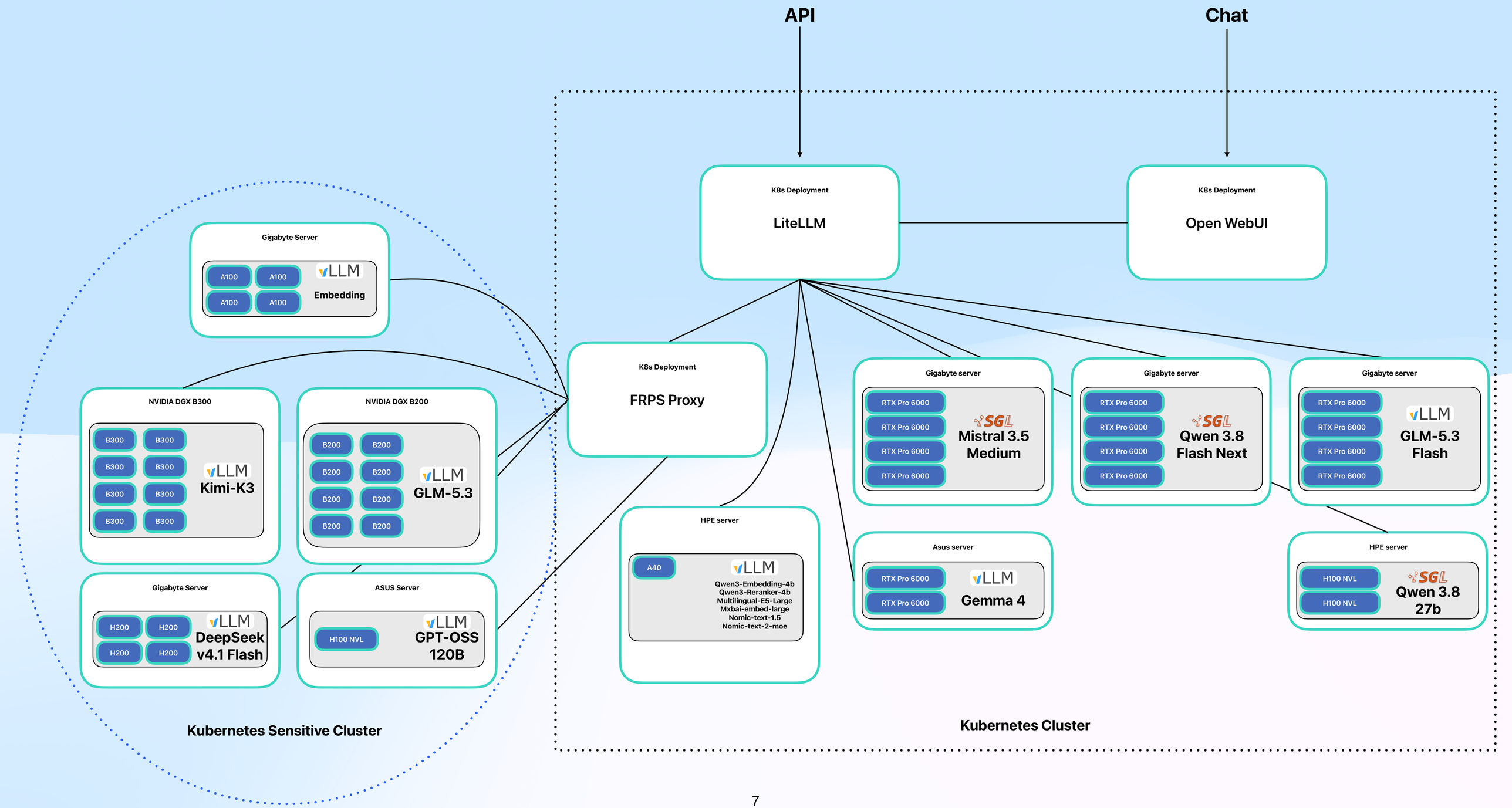
K 30. 10. 2026

Model ID	Model Information	Credentials	Created By	Updated At	Costs
8309206e-d7…	canary-v2 hosted_vllm/canary-v2	Manual	Unknown Unknown date	-	IN $0.10
219b96b3-87…	deepseek-v4.1-flash deepseek-v4.1-flash	Manual	Unknown Unknown date	-	IN $0.15 OUT $0.60
3ea9da3c-01…	deepseek-v4.1-flash-thinking deepseek-v4.1-flash	Manual	Unknown Unknown date	-	IN $0.15 OUT $0.60
107095eb-ba…	embeddinggemma-2 embeddinggemma-2	Manual	Unknown Unknown date	-	IN $0.10 OUT $0.00
3f1de5d6-5d…	gemma4 gemma4	Manual	Unknown Unknown date	-	IN $0.13 OUT $0.38
7d6b1d28-78…	glm-5.3 glm-5.3	Manual	Unknown Unknown date	-	IN $1.40 OUT $4.40
f5b35533-7a…	glm-5.3-flash glm-5.3-flash	Manual	Unknown Unknown date	-	IN $0.15 OUT $0.50
fea00ed7-f6…	gpt-oss-120b gpt-oss-120b	Manual	Unknown Unknown date	-	IN $0.40 OUT $1.60
0779b0a7-22…	kimi-k3 kimi-k3	Manual	Unknown Unknown date	-	IN $3.00 OUT $15.00
133f4da9-5c…	mistral-medium-3.5 mistral-medium-3.5	Manual	Unknown Unknown date	-	IN $0.15 OUT $0.60
9ee8a7a4-96…	multilingual-e5-large-instruct multilingual-e5-large-instruct	Manual	Unknown Unknown date	-	IN $0.02 OUT $0.00

- Kimi-K3 (NVIDIA DGX B300)
- GLM-5.3 (NVIDIA DGX B200)
- DeepSeek v4.1 Flash (4x NVIDIA H200)
- GPT-OSS-120B (NVIDIA H100)
- Gemma 4 (2x NVIDIA RTX Pro 6000)
- GLM-5.3 Flash (NVFP4), Mistral Medium 3.5 (4x NVIDIA RTX Pro 6000)
- Qwen 3.8 27b (FP8, 2x NVIDIA H100 NVL)
- Qwen 3.8 Flash Next (FP8, 4x NVIDIA RTX Pro 6000)
- Embedding modely (NVIDIA A40, A100)

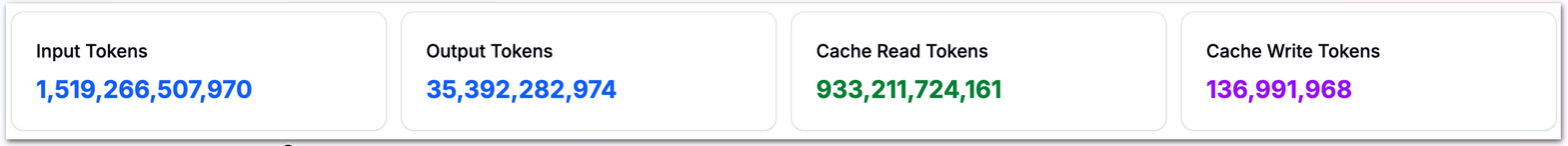
Inference

- Kombinace vLLM a SGLang
  - SGLang obvykle rychlejší, občas méně stabilní
  - Napojeno přímo na LiteLLM bez routeru
- Neběžíme Disaggregated Prefill/Decode
  - Nemáme dostatek hardware
- Nemáme sémantické routery
  - Model běží vždy pouze v jedné kopii
- Není důvod používat složitější systémy s automatizací
  - Prostředí je relativně malé


API
Chat
K8s Deployment
LiteLLM
K8s Deployment
Open WebUI
Gigabyte Server
A100
A100
A100
A100
vLLM
Embedding
K8s Deployment
FRPS Proxy
NVIDIA DGX B300
B300
B300
B300
B300
B300
B300
B300
B300
vLLM
Kimi-K3
NVIDIA DGX B200
B200
B200
B200
B200
B200
B200
B200
B200
vLLM
GLM-5.3
Gigabyte server
RTX Pro 6000
RTX Pro 6000
RTX Pro 6000
RTX Pro 6000
SGL
Mistral 3.5
Medium
Gigabyte server
RTX Pro 6000
RTX Pro 6000
RTX Pro 6000
RTX Pro 6000
SGL
Qwen 3.8
Flash Next
Gigabyte server
RTX Pro 6000
RTX Pro 6000
RTX Pro 6000
RTX Pro 6000
vLLM
GLM-5.3
Flash
HPE server
A40
vLLM
Qwen3-Embedding-4b
Qwen3-Reranker-4b
Multilingual-E5-Large
Mxbai-embed-large
Nomic-text-1.5
Nomic-text-2-moe
Asus server
RTX Pro 6000
RTX Pro 6000
vLLM
Gemma 4
HPE server
H100 NVL
H100 NVL
SGL
Qwen 3.8
27b
Gigabyte Server
H200
H200
H200
H200
vLLM
DeepSeek
v4.1 Flash
ASUS Server
H100 NVL
vLLM
GPT-OSS
120B
Kubernetes Sensitive Cluster
Kubernetes Cluster

Zkušenosti z provozu

- Sdílení GPU více modely není cesta (vyjma Embedding modelů)
  - I když se vejdou, sníží se KV cache
- Na velikosti KV cache záleží
  - Začínali jsme s DeepSeek R1 s KV cache na 128k tokenů
  - Dnes DeepSeek v4.1 Flash s KV cache na 25M tokenů
- Prefix caching má zásadní vliv
  - Rychlost (latence) jednoho dotazu neřeší
  - Šetří výpočetní výkon Prefill fáze



Zkušenosti z provozu

- Použitelný počet GPU: 1, 2, 4, 8
  - Tensor parallel režim
  - Nepoužíváme multi-node nebo pipeline parallel
- Podpora RTX Pro 6000 je stále špatná
  - Nové modely často nefungují (zejména SGLang)
  - Nemá HBM paměť → pomalý decode
  - I FP4 model běží často rychleji na H100/H200 GPU (GPT-OSS, DeepSeek)
  - Levnější na výkon, horší na spotřebu

Zkušenosti z provozu

Problémy

- vLLM i SGLang občas padají
- Nekončící problémy s tool parsery
- LiteLLM proxy obsahuje obrovské množství chyb
  - 25kB vlastních oprav
  - 2x obsahoval malware (nás se nedotklo, neaktualizujeme tak rychle)
  - Stále nejméně špatná proxy 🙃
- Služba je zdarma
  - Problémy s definicí fair-use
- Neumíme prioritizovat nebo rezervovat výkon

Zkušenosti z provozu

Rychlý vývoj

- Vývoj modelů
  - DeepSeek: **R1** (podzim 2025) → v3.2 (Leden 2026) → v4 Pro (Květen 2026) → v4.1 (Červenec 2026) → **v4.1 Flash 1017** (Říjen)
  - GLM 4.7 → **5.3** (Srpen 2026),
  - Kimi K2.5 → K2.7 → **K3.0** (Červenec 27, 2026) → K3.1 (Listopad 4, 2026)
- Nasazujeme nové verze, jak je to možné
  - Bez záruky stability a dostupnosti konkrétního modelu
- U některých modelů je vidět **Benchmaxing** – agresivní optimalizace na benchmarky
  - Testujeme pomocí Aider Polyglot – dnes poněkud zastaralý

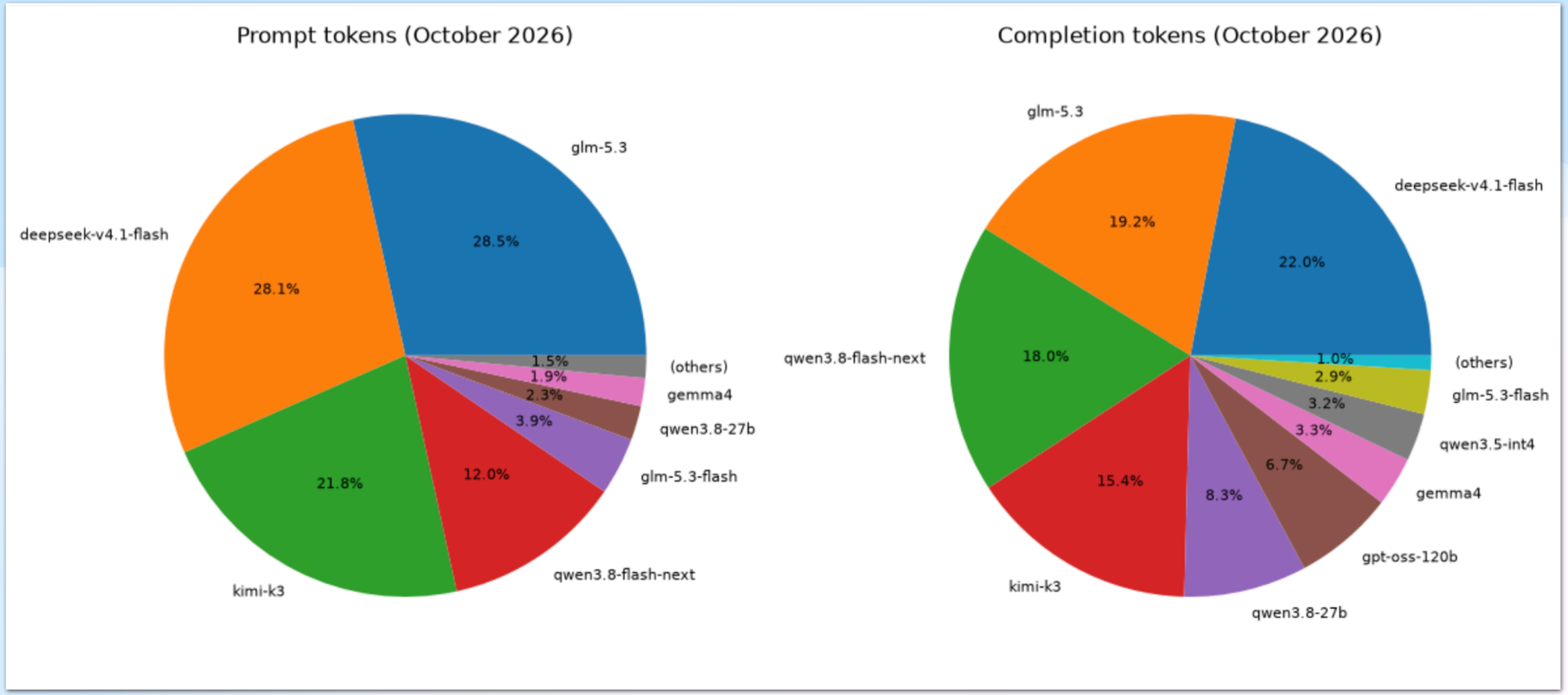
Proč AI na nejdražším HW?

- 8 GPU stroj ~ HGX/DGX
  - H100, H200, B200, B300
- 8 RTX Pro 6000 je levnější
  - Maximálně 8 x 96 GB GPU paměti = 768 GB
  - Nedostatečné pro největší open-weight modely
    - Kimi-K3 – minimálně 2.3TB GPU paměti
    - GLM-5.3 – minimálně 1.3TB GPU paměti

model	machine	machines	cost $	$ / 1M total	$ / 1M output	API bill $	API / self-hosted
kimi-k3	DGX B300	1	12,288	0.110	10.41	86,122	7.01x
glm-5.3	DGX B200	1	12,288	0.085	8.33	52,115	4.24x
deepseek-v4.1-flash	H200	1	3,686	0.026	2.18	2,270	0.62x
qwen3.8-flash-next	RTX PRO 6000	1	2,048	0.033	1.48	2,223	1.09x

Nejvíce vytížené modely

Na HW B200, B300, 4x H200



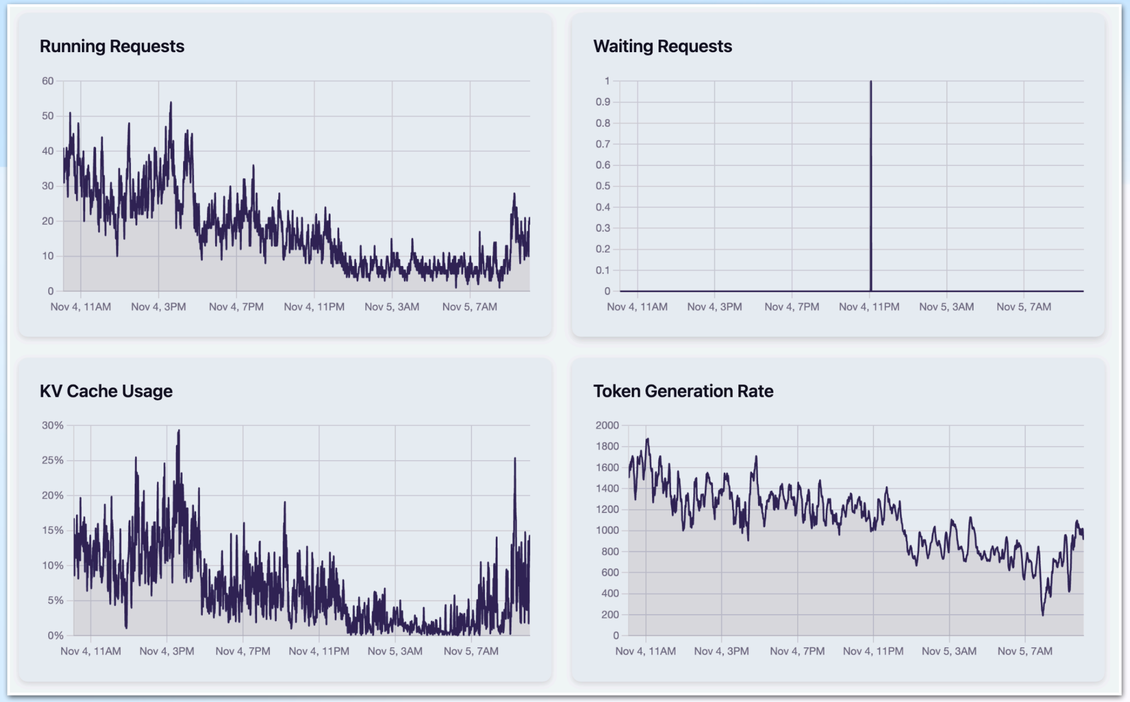
Shrnutí

- Poslední dva roky nám daly
  - Zkušenost jak provozovat velké modely pro tisíce uživatelů
  - Ja si vystačit s malým hardwarem a jednoduchou architekturou
    - A odpověď na otázku, zda to stačí
  - Zkušenost, že GPU paměť je esenciální
    - Velikost KV cache
    - Prefix caching

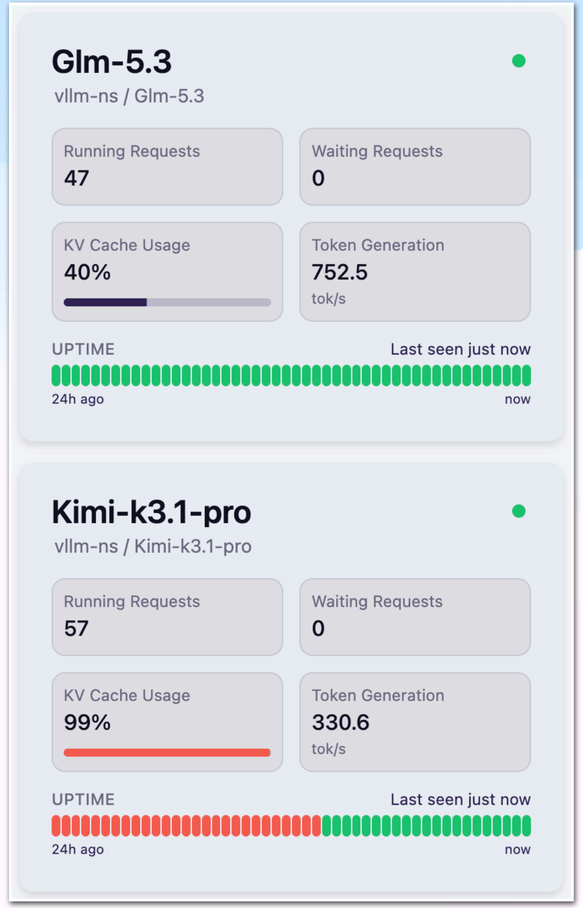
Živá situace

https://llm.ai.e-infra.cz/status

- Maxvellův démon říká, že spotřebováváme 10 000 000x více energie na výpočty, než musíme



DeepSeek v4.1 Flash



Děkuji za pozornost!
(Attention Is All You Need, 2017)

k8s@cerit-sc.cz
https://docs.cerit.io
https://blog.e-infra.cz