Tohle bude

nejnudnější prezentace dne.

AIDays

5. 11. 2026

On-prem AI: poznámky z praxe

Co rozhoduje o úspěšné adopci AI ve firmách.

Anton Giertli

Senior Solution Architect, Red Hat

Red Hat

Proč to dělat na vlastním HW?



Kontrola



Suverenita



Bezpečnost



Předvídatelné náklady

Tohle jsou důvody. Zbytek je práce.

„Doporučte nám nejlepší GPU.“



Souběh



Model a úloha

Velikost · kontext



Odezva



Dostupnost

Lekce 1: Koupit GPU je snadné. Vědět na co je těžší.

GPU vybíráte i pro datacentrum



Napájení



Chlazení



Prostor



Podpora

Lekce 2: Výběr GPU je rozhodnutí o datacentru.

Kam s váhami modelu?

Jedna správná volba neexistuje.

S3 (+ cache)

+ Rychlý start z cache

− Místo na discích nodů

OCI / ModelCar

+ Existující registry

± Lokální cache

− Build velkých imagů

PV / externí diskové pole

+ Váhy mimo worker

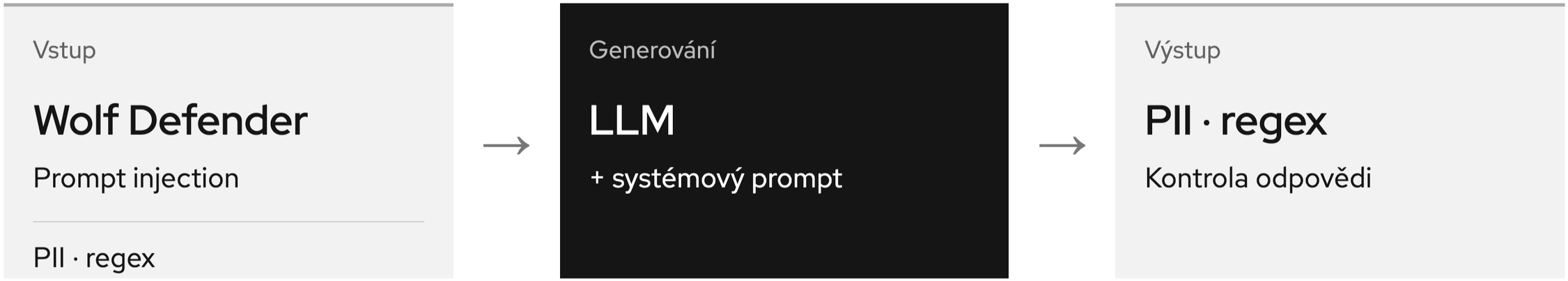
− Příprava vah předem

Lekce 3: Při plánování GPU nezapomeňte na místo pro váhy modelů.

Bezpečnostní tým zapojte od začátku

„Zapněte guardrails.“ Jaké?

Testované zapojení v sandboxu · NeMo Guardrails



Co zachytíme? Co ještě projde?

Lekce 4: Bez podpory bezpečnosti zůstane AI u PoC.

Co změnily guardrails?

Úspěšnost útoků (ASR) dle reportu · sandbox · garak / EvalHub

Bez ochrany

≈ 24,7 %

Model bez systémového promptu

→

S ochranou

≈ 4,9 %

NeMo Guardrails + Wolf Defender
+ systémový prompt

Výsledek celého nastavení, ne samotného Wolf Defenderu.

Část útoků stále prošla.

Lekce 5: Co neměříme, tomu nemůžeme věřit.

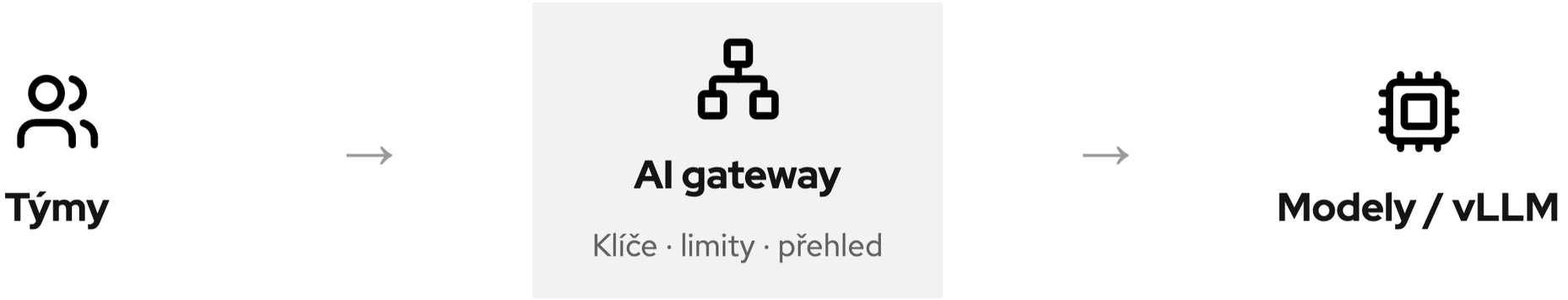
Pár GPU. Více týmů.

Nástroj / mechanismus	Kdy pomůže	Na co pozor
Time-slicing	Vývoj, experimenty	Bez izolace paměti
MIG	Oddělené části GPU	Podporované GPU a profily
MPS	Souběžné CUDA úlohy	Není izolace jako MIG
DRA	Přidělení podle požadavků	Dynamic MIG závisí na driveru
Kueue	Kvóty, priority, fair sharing	Přidělení kapacity úlohám

Lekce 6: Domluvte pravidla dřív, než se týmy začnou přetahovat o GPU.

Kueue
kueue.sigs.k8s.io

Model běží.
Jak se k němu dostanou lidé?

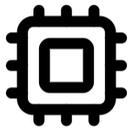


LiteLLM AI gateway

LiteMaaS PoC Samoobsluha nad gatewayí

Lekce 7: Model, ke kterému se nikdo nedostane, nikdo nepoužije.

Co potřebujete vědět o provozu AI?



GPU

GPU / MIG · VRAM · využití

DCGM Exporter
Prometheus + Grafana



Tokeny

Spotřeba podle modelu a týmu

AI gateway



Agent

Modely · nástroje · opakování

MLflow / Langfuse
OpenTelemetry

Lekce 8: AI nemůžete provozovat naslepo.

Práce je s tím dost.
Důvody ale zůstávají.

Kontrola Suverenita Bezpečnost Předvídatelné náklady

> Mít vlastní AI pod kontrolou za tu práci stojí.

Red Hat

Odkazy na cestu domů



Naskenujte nebo klikněte.

Všechny odkazy.
Na jednom místě.

Veřejný Markdown Gist

- Sizing a hardware
- Bezpečnost a evaluace
- Sdílení GPU a přístup k modelům
- Observabilita

gist.github.com/agiertli/51e8248dbf178909670fe3b618311da6

Komunitní projekty a produkty mají odlišné podmínky podpory. Ověřte konkrétní konfiguraci a verzi.