

Nasazení Red Hat OpenShift AI

Referenční architektura v air-gapped on-premise prostředích

Pavel Pokorný
DATERA

Co je cílovou službou

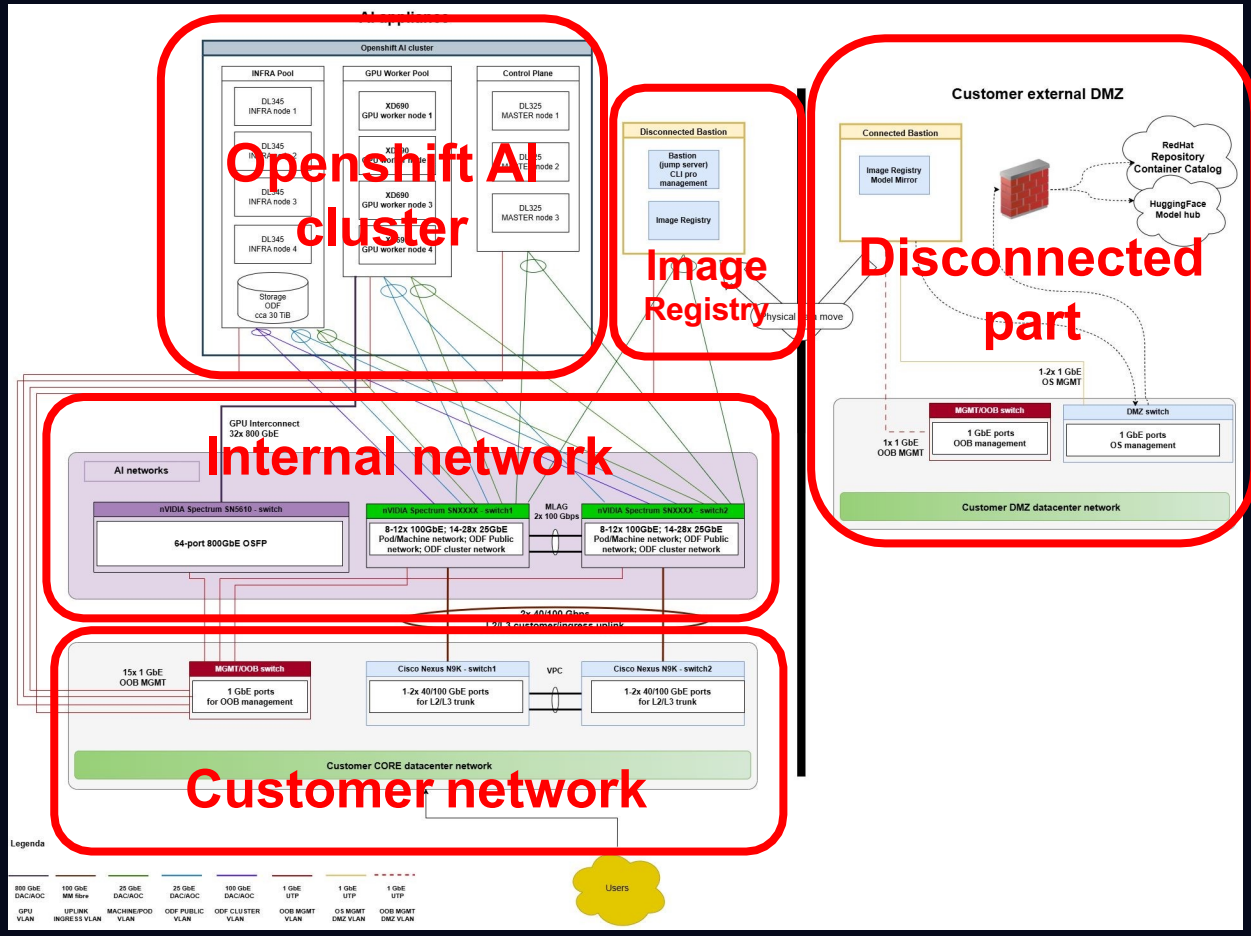
Definice cílové služby má vliv na architekturu řešení

Co je cílovou službou:

- **Model-as-a-Service:**
  - Nativní funkce RHOAI
    - vLLM provoz modelu
    - Katalog modelů
    - Registr modelů
    - Model Gateway
    - Distribuované LLM (llmd)
  - Důraz na volbu GPU
    - Podpora MIG (Multi-Instance GPU)
    - Využití Dynamic DRA (Dynamic Resource Allocation)
  - Zvážit nutnost “hosted” klastrů
- **AI datové projekty a Playground:**
  - Novější verze RHOAI
- **AI usecasy a obecný workload:**
  - Vhodná licenční politika
  - Oddělení pomocí “hosted” klastrů
  - Zesložitění síťového návrhu

Základy architektury

Popis základních bloků návrhu architektury



Infrastrukturní návrh

Doporučení návrhu architektury

HW požadavky a doporučení:

- Dedikovat Infra a Master nódy
- Oddělit INFRA / ODF a GPU nódy
  - Disky v GPU nodech pro image a ephemeral
- GPU nódy:
  - Volba GPU, uvažovat KV Cache
  - Sizing /var/lib/containers (images)
    - Nativní rezerva 15%
    - Počet a typ modelů
    - 1TB +
- Návrh a umístění bastion serverů
  - FW pravidla pro Bastion / Registry
  - Lze i bez konektivity s klastrem

Infrastrukturní návrh - sítě

Ukázkový návrh sítí

Možný návrh sítí a jejich role:

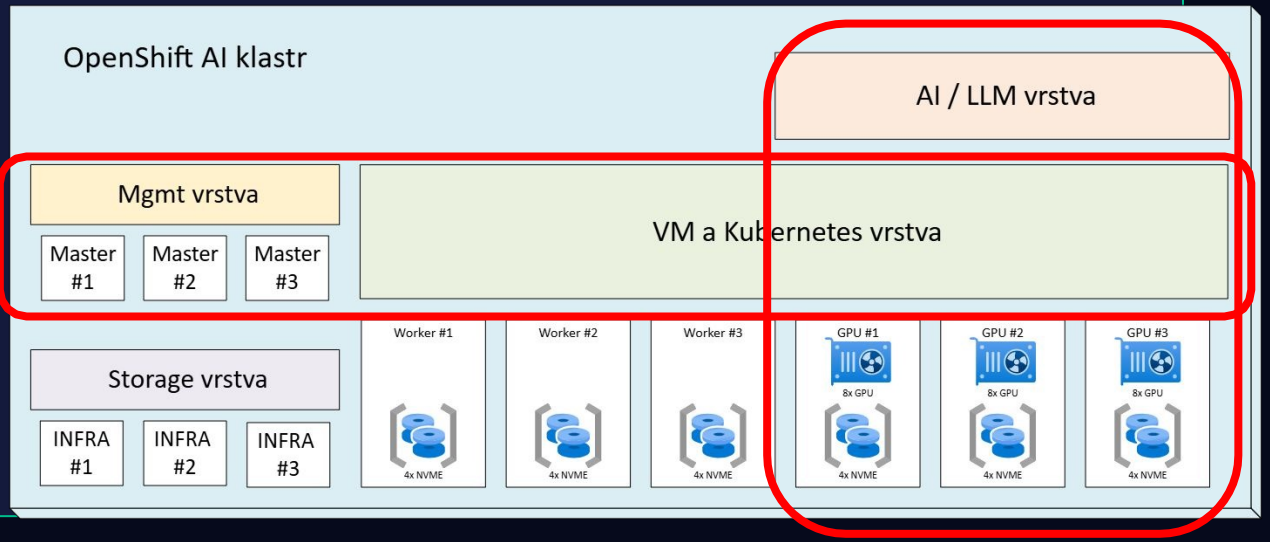
Pod / Machine	(25 GbE)	všechny nody	Pod komunikace
Client	(40 / 100 GbE)	switch uplink	přístup klientů k aplikacím
ODF Public	(25 GbE)	Infra + Worker nody	I/O aplikací vůči ODF
ODF Cluster	(100 GbE)	Infra nody	ODF interní repl. / recov.
GPU interconnect	(800 GbE)	Worker nody	GPU<->GPU komunikace
OOB	(1 GbE)	všechny technologie	

Licenční návrh

Více licenční možností, optimalizace dle usecasu

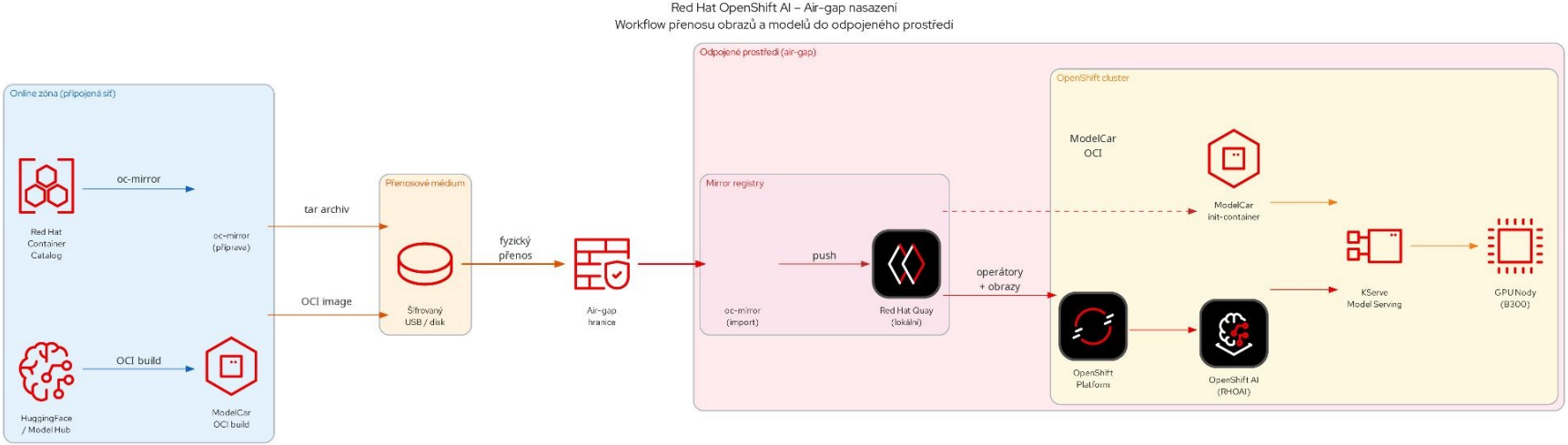
Nasazeni **“Čistě AI / LLM”** => licence **Red Hat AI Enterprise:**

- Pouze AI usecasy (trénování, tuning, inference a agenti)
- Bundle obsahuje:
  - Red Hat OpenShift AI
  - Red Hat OpenShift Container Platform
- Neobsahuje OCP Plus (ODF)
  - Možný add-on
  - Doplnění SKU komponent
- Per-node lic. model
- Neomezeně GPUs
- Bez omezení CPU jader
- Bez omezení soketů



Životní cyklus modelu v provozu

Workflow od importu po archiv



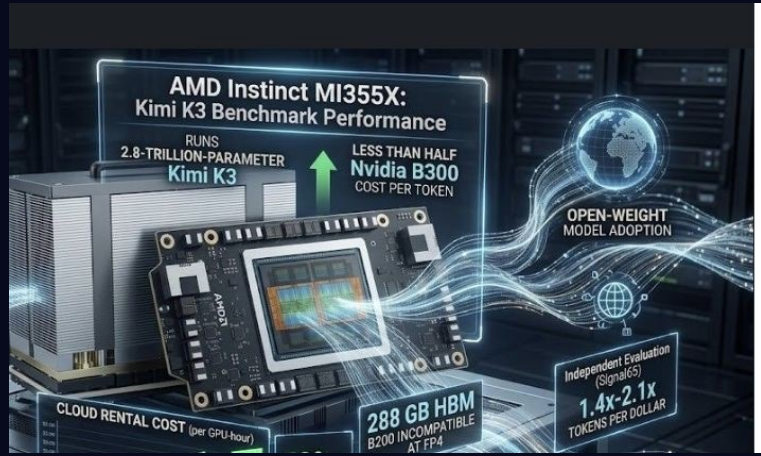
Podpora Multi-Vendor akceleratorů

Podpora NVIDIA & AMD

- **NVIDIA:** enterprise GPU včetně **B300 / GB300** Blackwell GPU
- **AMD:** Instinct MI325X, MI350X a MI355X GPU.

Intel a IBM akcelerátory

- **IBM Spyre**
- **Intel Gaudi**



Half the cost per token: AMD's MI355X challenges Nvidia on massive AI workloads

New benchmark data from SemiAnalysis's vendor-neutral InferenceX platform shows AMD's Instinct MI355X GPU running Moonshot AI's 2.8-trillion-parameter Kimi K3 model at less than half the cost of Nvidia's flagship B300. As models scale past trillion-parameter thresholds, the economics of AI inference are becoming as critical as raw throughput.

The Benchmark Breakdown:

📊 Cost Advantage: Cloud providers currently list MI355X instances starting around $2.50 per GPU-hour, compared to roughly $6.00 per GPU-hour for the B300. This lower hourly rate enables the MI355X to deliver Kimi K3 ...more

2

Half the cost per token: AMD's MI355X challenges Nvidia on massive AI workloads

Technické nuance - práce s parametry

Nutná znalost korektní parametrizace modelů

Parametry vLLM vs OpenShift:

- V některých případech (např. reranker) je nutné **specifikovat architekturu**
- Specifika parametrizace
  - vLLM: **--hf_overrides '{"architectures"**: ["Qwen3ForSequenceClassification"],"classifier_from_token": ["no", "yes"],"is_original_qwen3_reranker": true}'
  - RHOAI: **--hf_overrides={"architectures":** ["Qwen3ForSequenceClassification"],"classifier_from_token": ["no", "yes"],"is_original_qwen3_reranker": true}
  - Složitý troubleshooting:
    - `Supported tasks: ['generate']`
    - `Spravne: Supported tasks: ['token_classify', 'classify'] );`
    - **`vLLM načetlo reranker model jako model pro textovou inferenci`**



Děkujeme za pozornost

Pavel Pokorný

datera

linkedin.com/company/red-hat

facebook.com/redhat

youtube.com/@redhat

x.com/RedHat

Red Hat