Seznam LLM v produkci

Marek Šimůnek

Tech lead LLM týmu

Agenda

- Proč děláme vlastní Seznam LLM => SeLLMa => Šelma?
- Škálování inferencí
- Garance provozního trafficu
- Offline (Batchová) cesta

O mně

40,000,000,000 klíčů v databázi
300,000,000 stažení denně

- Big data
- MLOps
- LLM

1000 fyzických serverů
15 PB úložného prostoru
65 TB paměti

Platforma pro 5 výzkumnických týmů
25+ lidí
100+ modelů

Seznam LLM co umí česky

Lupa.cz » Seznam.cz chystá vlastní umělou inteligenci. V češtině už je o něco lepší než GPT-3.5

Seznam.cz chystá vlastní umělou inteligenci. V češtině už je o něco l než GPT-3.5

JAN SEDLÁK | 17. 1. 2024 | Doba čtení: **4 minuty**

9 NOVÝCH NÁZ

Lupa.cz » Jazykový model Seznamu se jmenuje Šelma, firma pracuje na zapojení AI do služeb

Jazykový model Seznamu se jmenuje Šelma, firma pracuje na zapojení AI do

Zprávy » Byznys » Byznys | Rozhovory » Vyhledávání na Seznamu se dramaticky změní, říká Pav...

Vyhledávání na Seznamu se dramaticky změní, říká Pavel Zima z vedení firmy

UMĚLÁ INTELIGENCE – 07. 3. 2024 – 4 min čtení

Nejrychlejších sedm měsíců života, říká žena, která vede vývoj generativní umělé inteligence Seznamu

Seznam.cz si jako česká internetová jednička nechce nechat příležitost v umělé inteligenci ujít. Investuje do ní desítky milionů a učí ji česky.

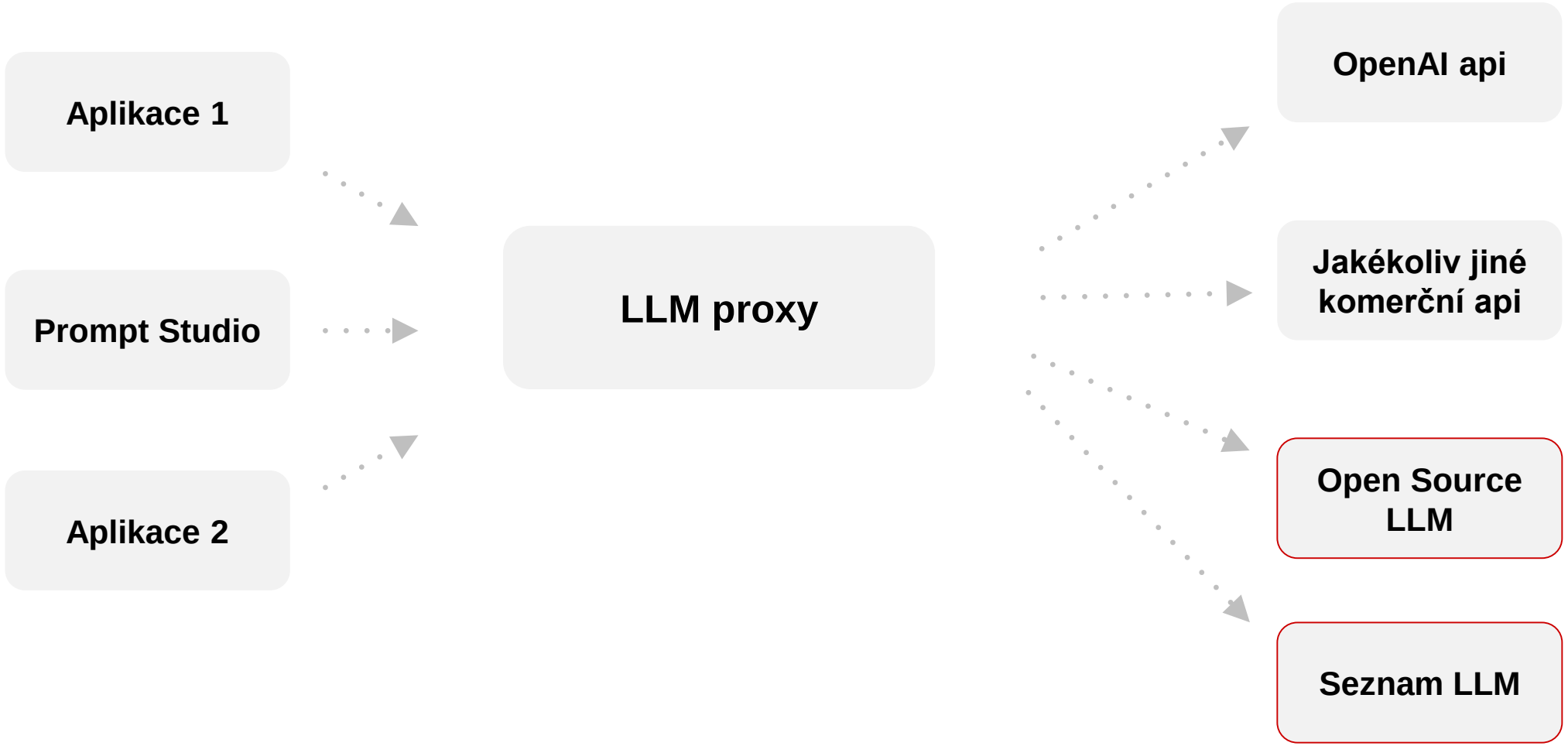
LLM v Seznamu

Výborná čeština modelu

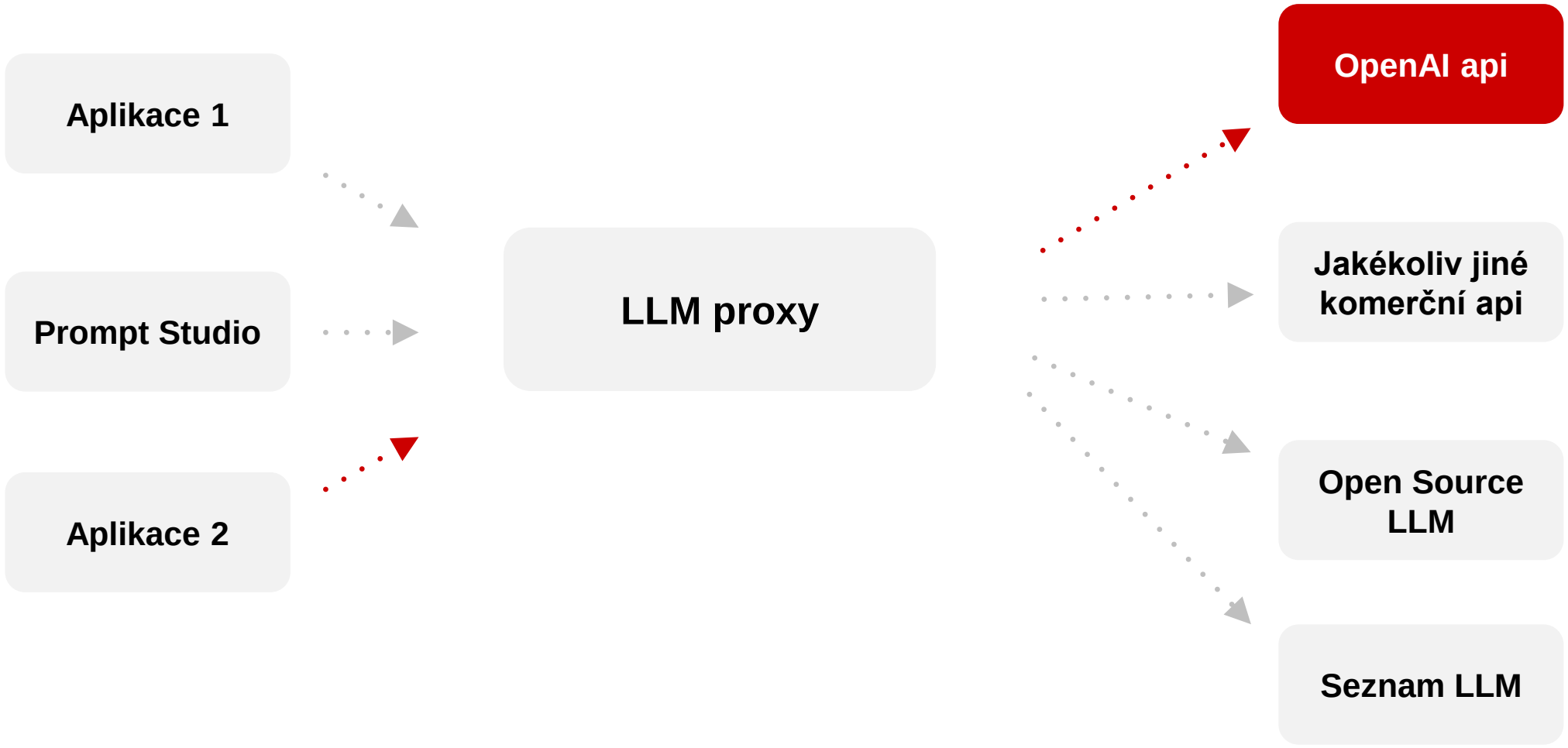
Provoz v naší režii

Bezpečnost pro uživatele

LLM proxy


Aplikace 1
Prompt Studio
Aplikace 2
LLM proxy
OpenAI api
Jakékoliv jiné komerční api
Open Source LLM
Seznam LLM

LLM proxy


Aplikace 1
Prompt Studio
Aplikace 2
LLM proxy
OpenAI api
Jakékoliv jiné komerční api
Open Source LLM
Seznam LLM

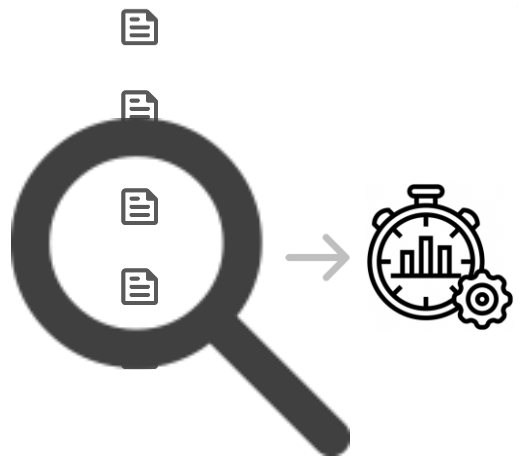
LLM proxy

Aplikace 1

Prompt Studio

Aplikace 2

LLM proxy



OpenAI api

Jakékoliv jiné komerční api

Open Source LLM

Seznam LLM

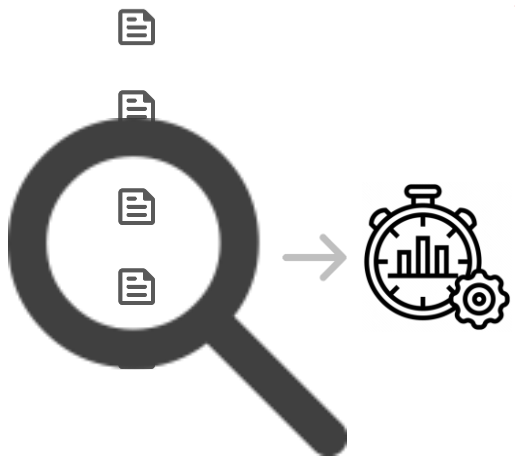
LLM proxy

Aplikace 1

Prompt Studio

Aplikace 2

LLM proxy



OpenAI api

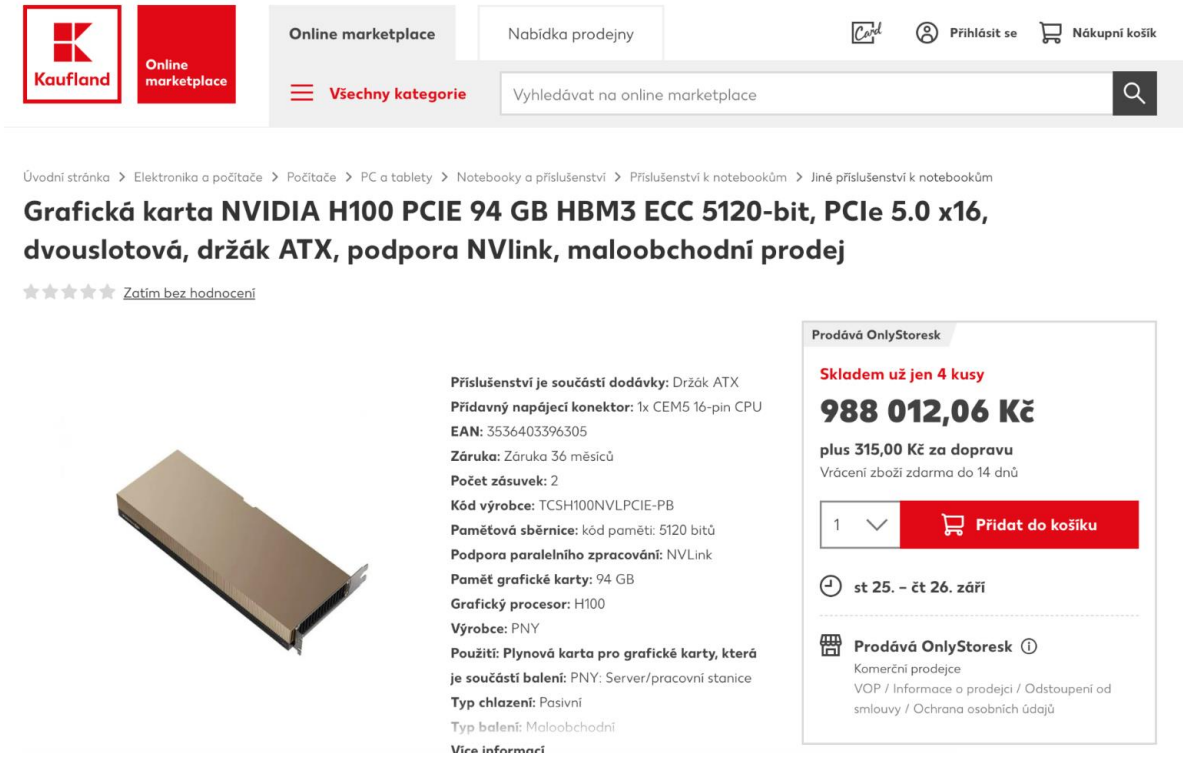
Jakékoliv jiné komerční api

Open Source LLM

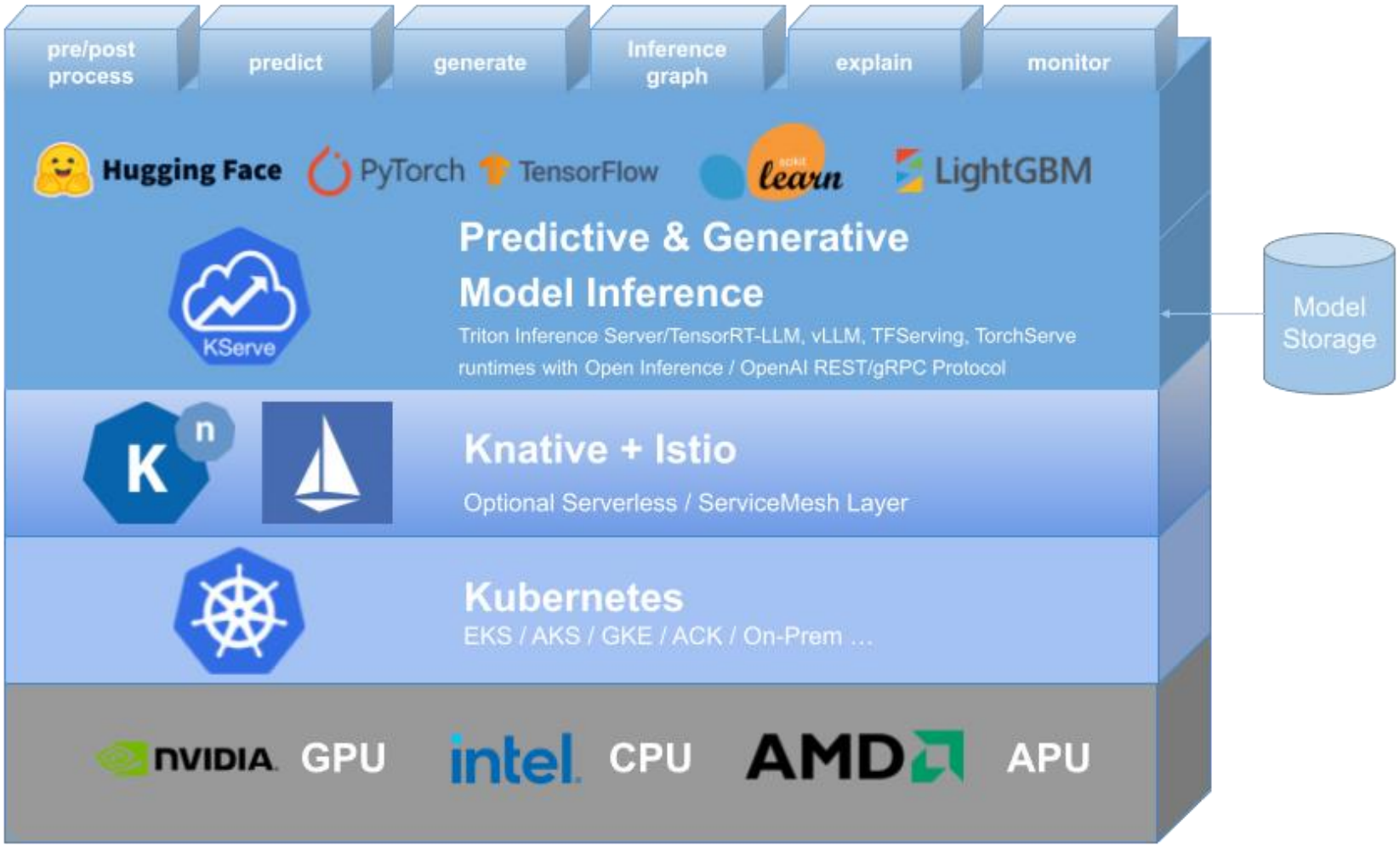
Seznam LLM

Požadavky na produkční řešení

- 1 GPU - cena cca 800k - 1M Kč
- proměnlivý traffic během dne
- 10 až 20 modelů v provozu (aktuálně 5)
- garance trafficu pro více uživatelů/aplikací
- umět dodržovat určité latence

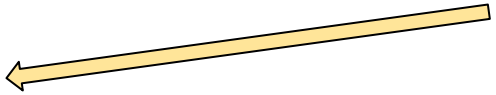


KServe

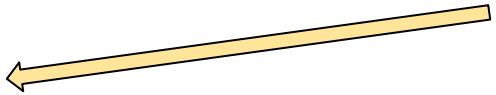

pre/post process
predict
generate
Inference graph
explain
monitor
Hugging Face
PyTorch
TensorFlow
learn
LightGBM
KServe
Predictive & Generative Model Inference
Triton Inference Server/TensorRT-LLM, vLLM, TFServing, TorchServe runtimes with Open Inference / OpenAI REST/gRPC Protocol
Model Storage
Knative + Istio
Optional Serverless / ServiceMesh Layer
Kubernetes
EKS / AKS / GKE / ACK / On-Prem …
NVIDIA GPU
intel CPU
AMD APU

KServe: jak řešíme škálování

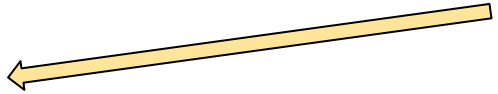
```
apiVersion: serving.kserve.io/v1beta1
kind: InferenceService
metadata:
  name: torch-service
spec:
  predictor:
    scaleTarget: 10
    scaleMetric: concurrency
    model:
      modelFormat:
        name: pytorch
      storageUri: pvc://some-persistent-volume-claim
```



metrika podle, které škálovat



jaký je formát modelu

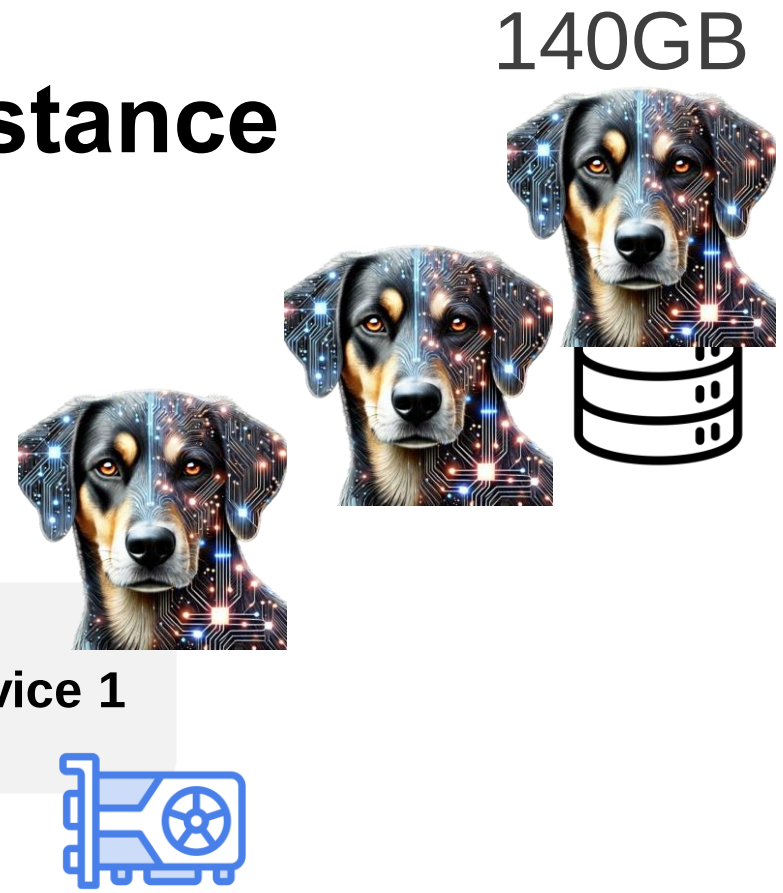


odkud model načíst

KServe: jak řešíme škálování

```
apiVersion: serving.kserve.io/v1beta1
kind: InferenceService
metadata:
  name: torch-service
spec:
  predictor:
    maxReplicas: 5
    minReplicas: 0
    scaleTarget: 10
    scaleMetric: concurrency
    model:
      modelFormat:
        name: pytorch
      storageUri: pvc://network-storage-gate-pvc
```

Nastartování nové instance



Doba naběhnutí:

- sellma-70B: **4-5min**
- sellma-7B: 1-2min

Applications use cases

Online

Chci odpověď co nejdřív

Offline

Mám hodně requestů, ale odpověď počká

Applications use cases

Online

Balancování mezi dobou odpovědí a celkovou propustností systému.

Offline

Nejjednodušší: Zaměření na maximální propustnost

Applications use cases

Online

Balancování mezi dobou odpovědí a celkovou propustností systému.

Offline

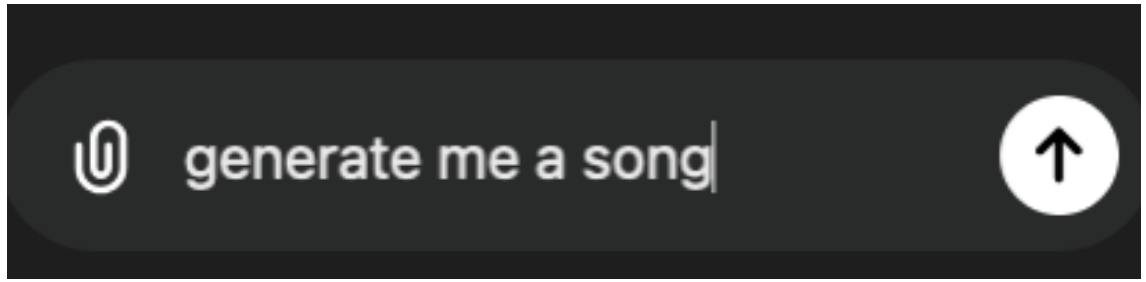
Nejjednodušší: Zaměření na maximální propustnost

Streaming

Sequential

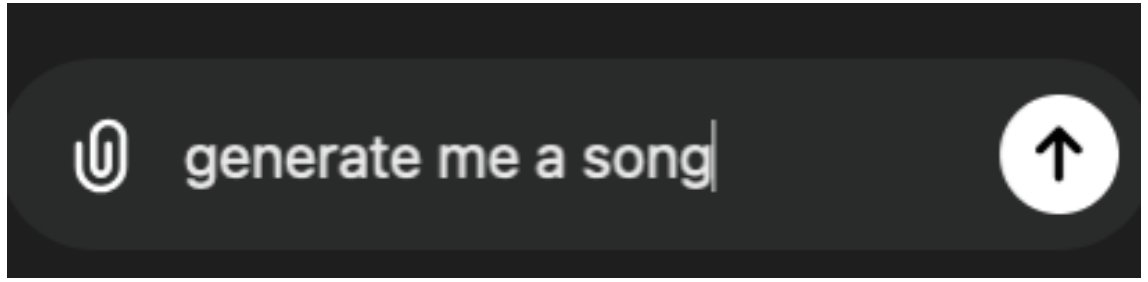
Vysvětlení inference pro streaming

Příklad:



Vysvětlení inference pro streaming

Příklad:



Prefill: 1.14s, 5 input tokens, 1 output token

`generate me a song`

Decoding: 1.62s, 33 output tokens

`Could you please provide me with some specifics`

Prefill, first-response	Decoding							
User: generate me a song. AI: *Could*	you	please	me	with	some	specifics	for	your

prompt

user prompt

Vysvětlení inference pro streaming

Prefill = time to first token (TTFT)
- Načtení celého promptu
- Čas od načtení promptu k vrácení prvního tokenu
- Závislé na počtu vstupních tokenů
- Limitovaný rychlostí výpočtu na GPUčku

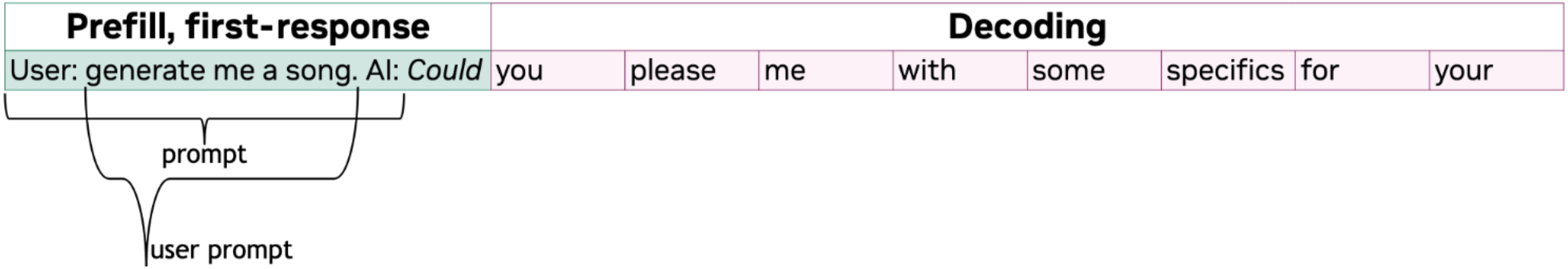
Decoding = inter-token latency (ITL)
- Generování výstupu token po tokenu
- ITL závisí jak na velikosti vstupu a tak i velikosti výstupu
- Limitovaný rychlostí GPU paměti

Prefill: 1.14s, 5 input tokens, 1 output token

`generate me a song`

Decoding: 1.62s, 33 output tokens

`Could you please provide me with some specifics`



Applications use cases

Online

Balancování mezi dobou odpovědí a celkovou propustností systému.

Streaming

metriky:
- Time to first token
- Inter token latency

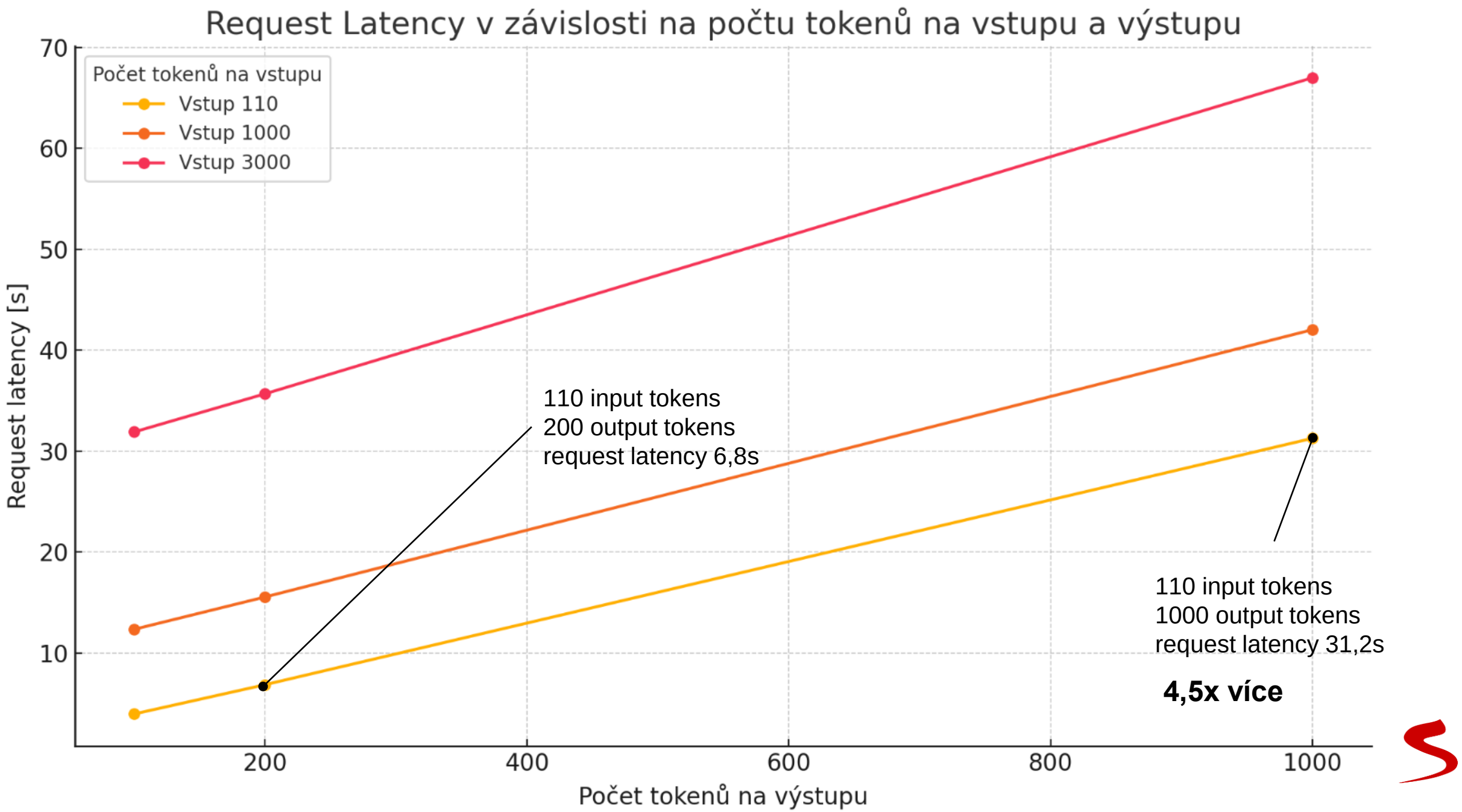
Sequential

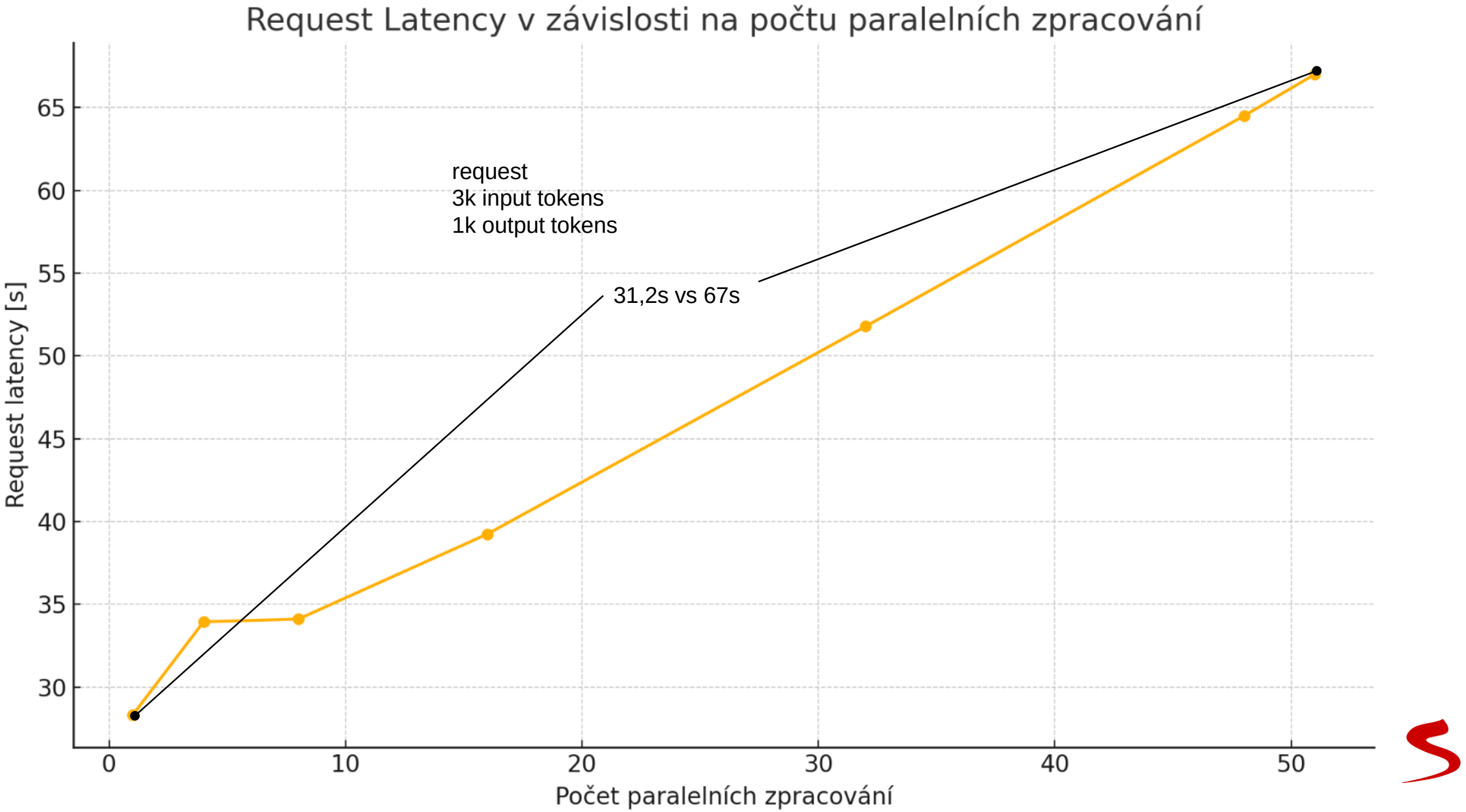
metriky:
- End to end latency (request latency)

Offline (Batch)

Nejjednodušší: Zaměření na maximální propustnost

metriky:
- request_throughput
- output_token_throughput


Request Latency v závislosti na počtu tokenů na vstupu a výstupu
Počet tokenů na vstupu
Vstup 110
Vstup 1000
Vstup 3000
Request latency [s]
70
60
50
40
30
20
10
200
400
600
800
1000
Počet tokenů na výstupu
110 input tokens
200 output tokens
request latency 6,8s
110 input tokens
1000 output tokens
request latency 31,2s
4,5x více


Request Latency v závislosti na počtu paralelních zpracování
request
3k input tokens
1k output tokens
31,2s vs 67s
Request latency [s]
65
60
55
50
45
40
35
30
0
10
20
30
40
50
Počet paralelních zpracování

Porovnání

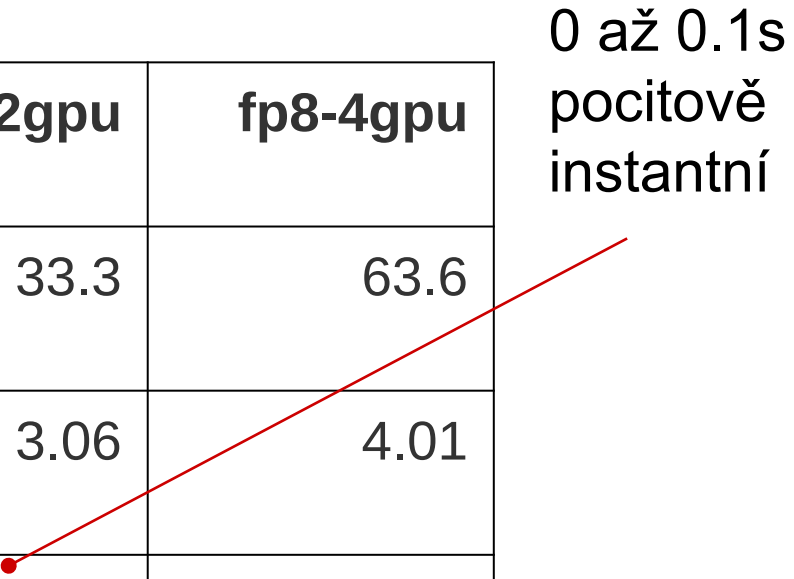
Sellma-70B, 1k input tokens, 1k output tokens, H100 PCIe

metric	fp16-4gpu	2x fp8-2gpu	fp8-4gpu
request_latency [s]	60.9	33.3	63.6
request_throughput [req/s]	2.13	3.06	4.01
time_to_first_token [s]	0.28	0.073	0.269
inter_token_latency [s]	0.05	0.028	0.054

Porovnání

Sellma-70B, 1k input tokens, 1k output tokens, H100 PCIe

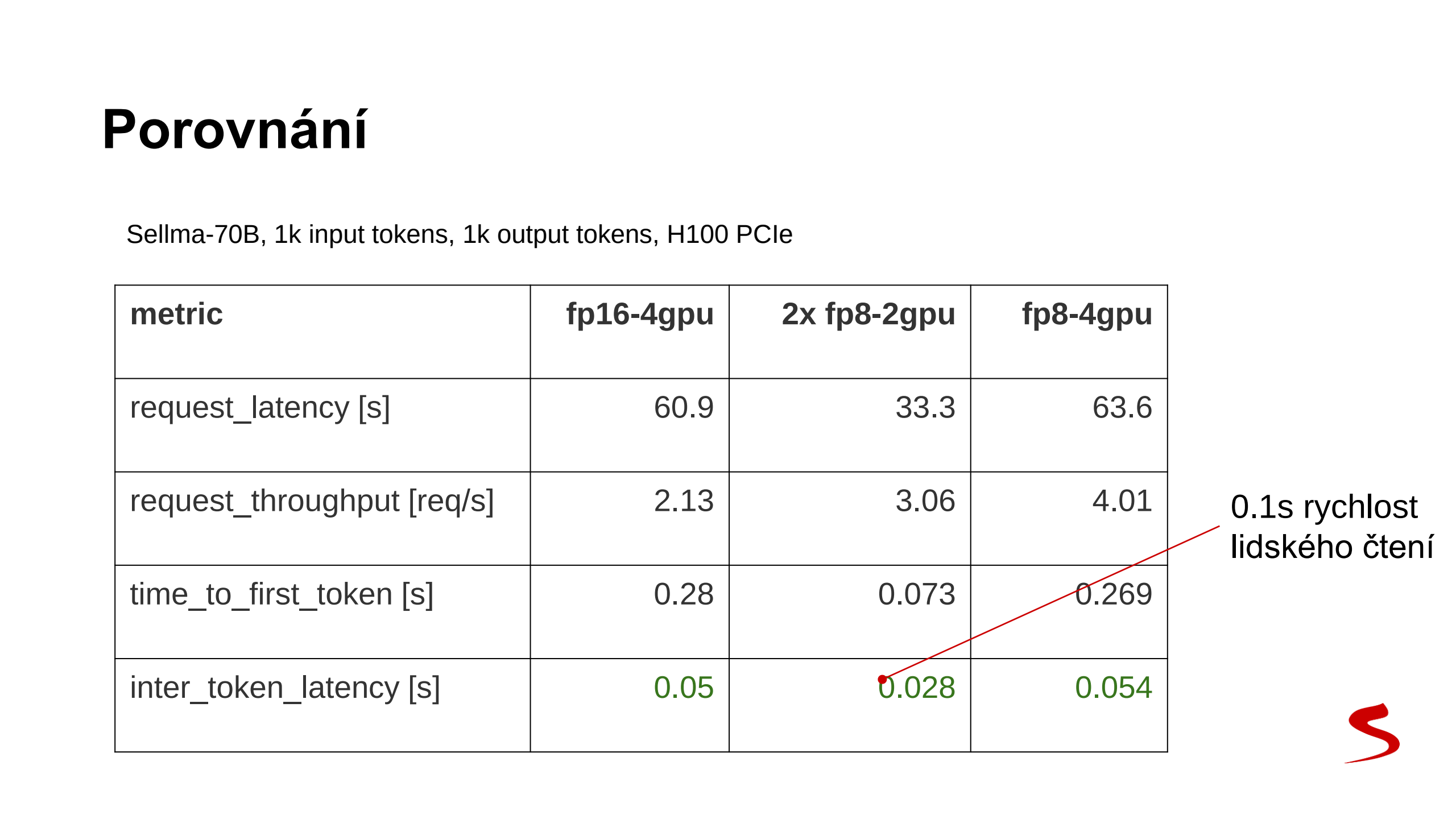
metric	fp16-4gpu	2x fp8-2gpu	fp8-4gpu
request_latency [s]	60.9	33.3	63.6
request_throughput [req/s]	2.13	3.06	4.01
time_to_first_token [s]	0.28	0.073	0.269
inter_token_latency [s]	0.05	0.028	0.054



Porovnání

Sellma-70B, 1k input tokens, 1k output tokens, H100 PCIe

metric	fp16-4gpu	2x fp8-2gpu	fp8-4gpu
request_latency [s]	60.9	33.3	63.6
request_throughput [req/s]	2.13	3.06	4.01
time_to_first_token [s]	0.28	0.073	0.269
inter_token_latency [s]	0.05	0.028	0.054



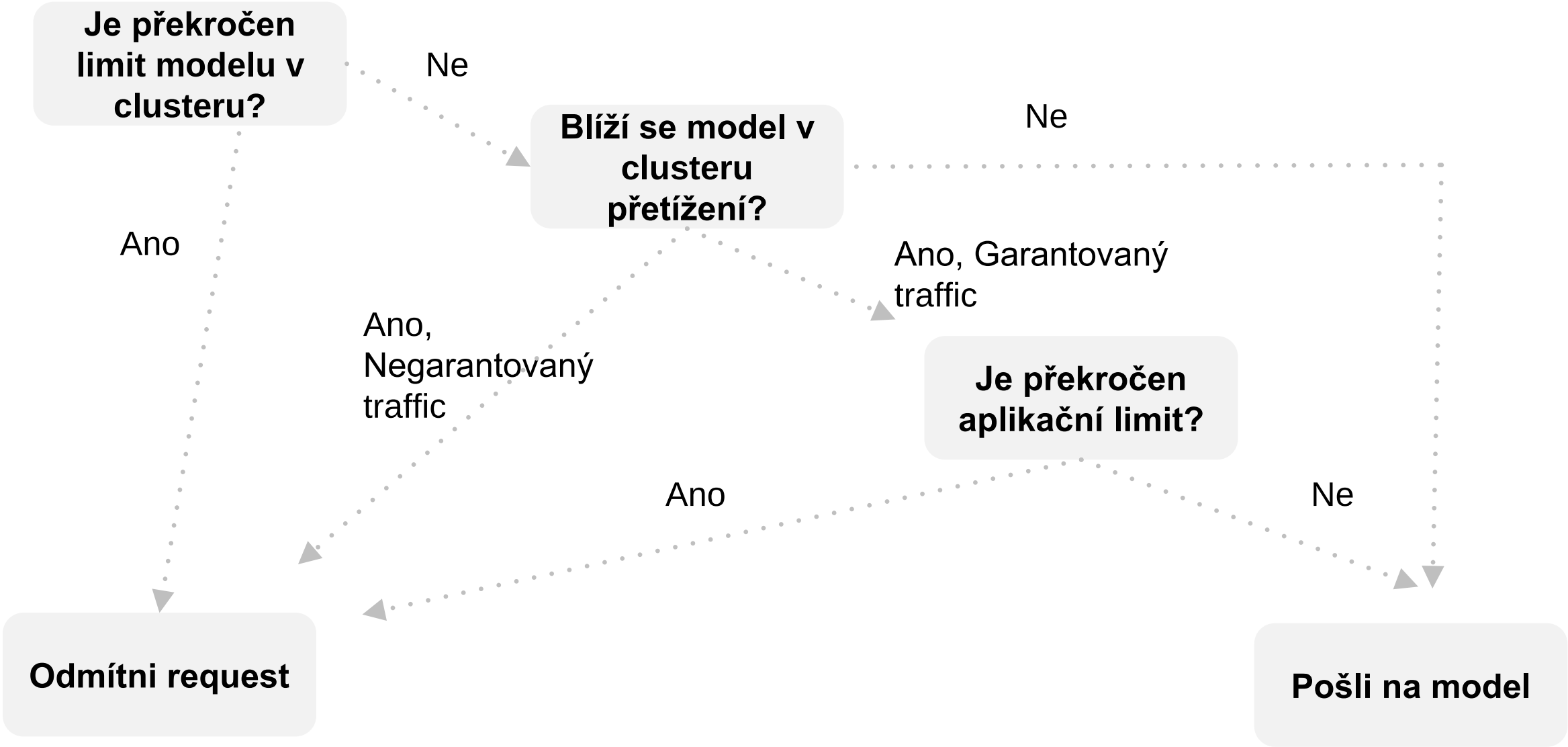
Nasazování aplikací

- Jaký model?
- Kolik v průměru tokenů na vstupu a výstupu
  - do velikosti vstupu se počítá i např historie chatu atd.
- Kolik requestů ve špičce
- Jaká je očekávaná doba odpovědi

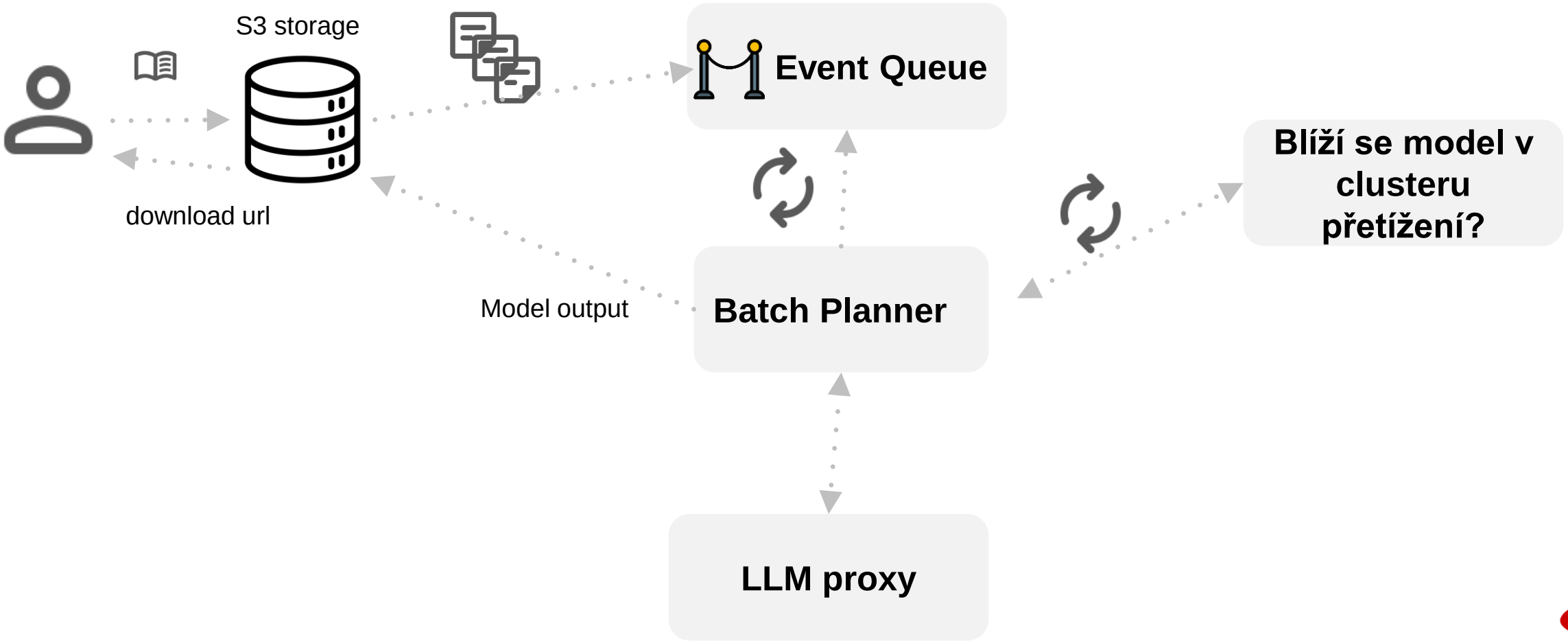
Garance pro provozní aplikace

Garantovaná metrika: počet paralelních requestů (concurrency nebo in-flight req)

Garance pro provozní aplikace



Offline (batch) inference



Prefix caching - sdílený system prompt

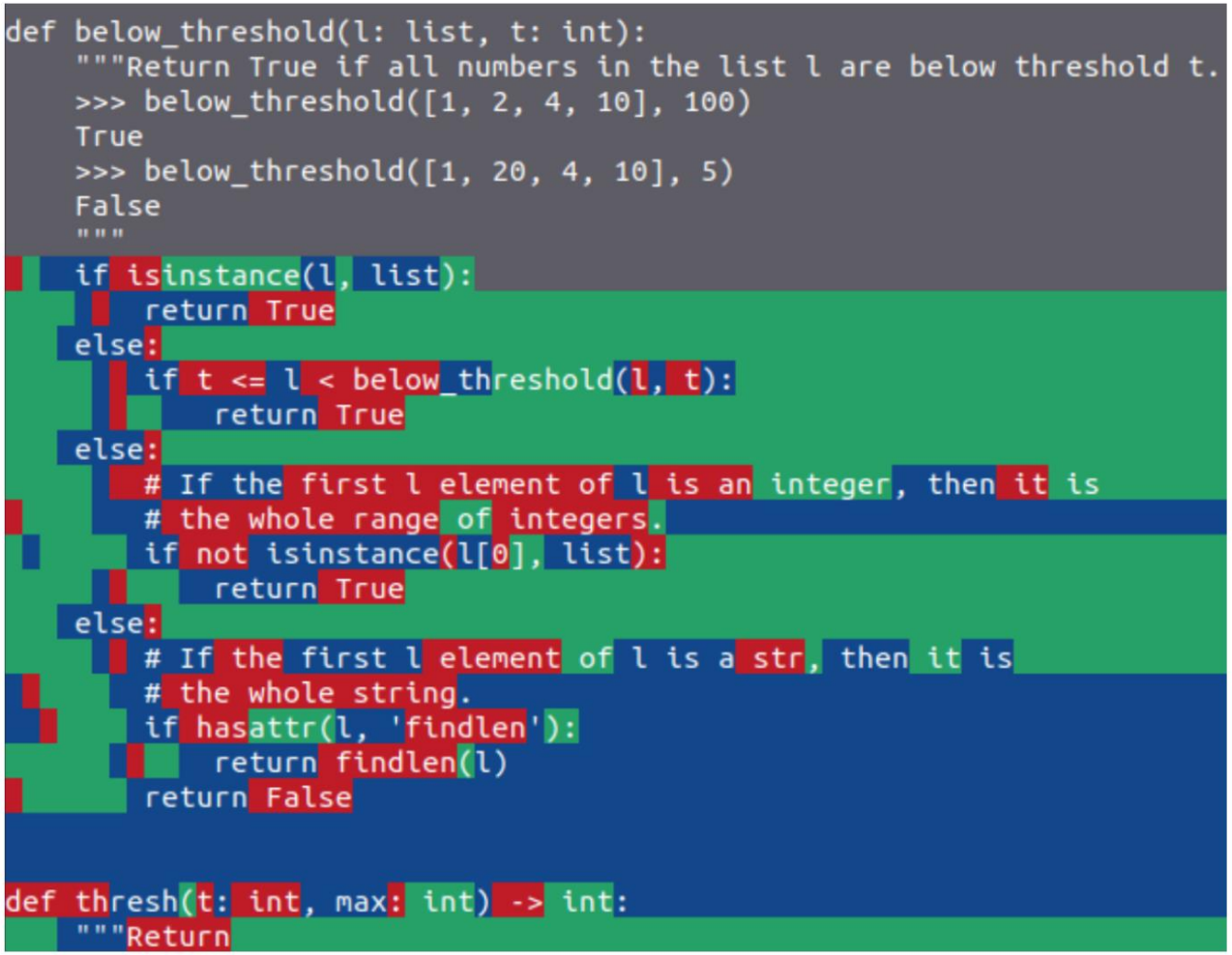
Request 1

```
Udělej souhrn následujícího
textu. Myšlenky **SE NESMÍ**
opakovat. Udělej výstup formou
krátkého souvislého odstavce.
Odstavec by měl mít maximálně 5
vět. U každé věty si uchovej
zdroj, odkud pochází. V každé
větě zvýrazni nejdůležitější
slova pomocí HTML značky <b>.
Vrať **POUZE** RFC8259 JSON
pole v následujícím formátu bez
odchylek.

Text:
O podporu EU žádají dva české
těžební projekty…
```

Request 2

```
Udělej souhrn následujícího
textu. Myšlenky **SE NESMÍ**
opakovat. Udělej výstup formou
krátkého souvislého odstavce.
Odstavec by měl mít maximálně 5
vět. U každé věty si uchovej
zdroj, odkud pochází. V každé
větě zvýrazni nejdůležitější
slova pomocí HTML značky <b>.
Vrať **POUZE** RFC8259 JSON
pole v následujícím formátu bez
odchylek.

Text:
Ve snaze oddálit stárnutí a
udržet se v kondici držíme
diety…
```

Speculative decoding



```
def below_threshold(l: list, t: int):
    """Return True if all numbers in the list l are below threshold t.
    >>> below_threshold([1, 2, 4, 10], 100)
    True
    >>> below_threshold([1, 20, 4, 10], 5)
    False
    """
    if isinstance(l, list):
        return True
    else:
        if t <= l < below_threshold(l, t):
            return True
    else:
        # If the first l element of l is an integer, then it is
        # the whole range of integers.
        if not isinstance(l[0], list):
            return True
    else:
        # If the first l element of l is a str, then it is
        # the whole string.
        if hasattr(l, 'findlen'):
            return findlen(l)
        return False


def thresh(t: int, max: int) -> int:
    """Return
```

Figure 3: A visualization of the origin of tokens in an example T=1 HumanEval completion. Green background originates with the N-gram draft[2] model, blue the draft model, and red the oracle model. (Of course, all tokens are eventually checked by the oracle model.) Obvious tokens – like whitespace – are preferentially accelerated relative to difficult ones.

[link to paper]

Shrnutí

- máme nové modely v llm-proxy
- umíme garantovat aplikacím traffic
- pro větší utilizaci jsme vyrobili i offline (batchovou cestu)
- máme v plánu zrychlovat (kvantizace, prefix caching, speculative decoding atd..)

proč polníček

Internet Obrázky Zboží Mapy Videa Zprávy Firmy Slovník

- Polníček je **neskutečně výživný** a chutí byste si ho snadno spletli s hlávkovým salátem.[1]
- Jelikož ho lze **snadno pěstovat** a má významné nutriční vlastnosti, rozhodně byste ho neměli přehlížet, protože může ozdobit i ochutit vaše pokrmy.[1]
- Polníček obsahuje **vitamín C, E, B, provitamín A, kyselinu listovou, hořčík a železo**, což z něj činí cenný zdroj živin.[2]
- V období vegetačního klidu nám může **poskytovat tolik důležité vitamíny**, což je jeho výjimečnost.[2]
- Polníček je **jednoduchý důkaz** toho, že i v zimě si můžeme doma nachystat čerstvý a velice chutný salát.[2]

Polníček: Jak jej pěstovat a čím je…
kupi.cz 1

Polníček – Sezónní potraviny
sezonka.cz 2

Polníček: Jak jej pěstovat
kupi.cz

Před 14 dny · Vě
polní? Kde se vz
článku!

Adele

Adele – Wikipedie

cs.wikipedia.org/wiki/**adele**

Adele Laurie Blue Adkins, MBE (* 5. května 1988 Londýn, Spojené království) je anglická zpěvačka a skladatelka známá jen jako Adele.

Kariéra Soukromý život Diskografie Turné Odkazy

Mohlo by vás zajímat

Jaké je celé jméno zpěvačky Adele?

Jaké je debu

Proč se hledá Nehoda na Rakovnicku?

ošlo u Lubné na Rakovnicku k tragické nehodě, při které řidička sjela ze silnice a převrátila auto na střechu. [1, 2, 3] Řidička na místě dvě děti, které s ní cestovaly, byly zraněny a převezeny do [4, 5]

2. idnes.cz 3. extra.cz 4. iprima.cz 5. blesk.cz

erovaným shrnutím tématu

Pastrňák svatba

Internet Obrázky Zboží Mapy Videa Zprávy Firmy Slovník

Zprávy na téma: Pastrňák svatba

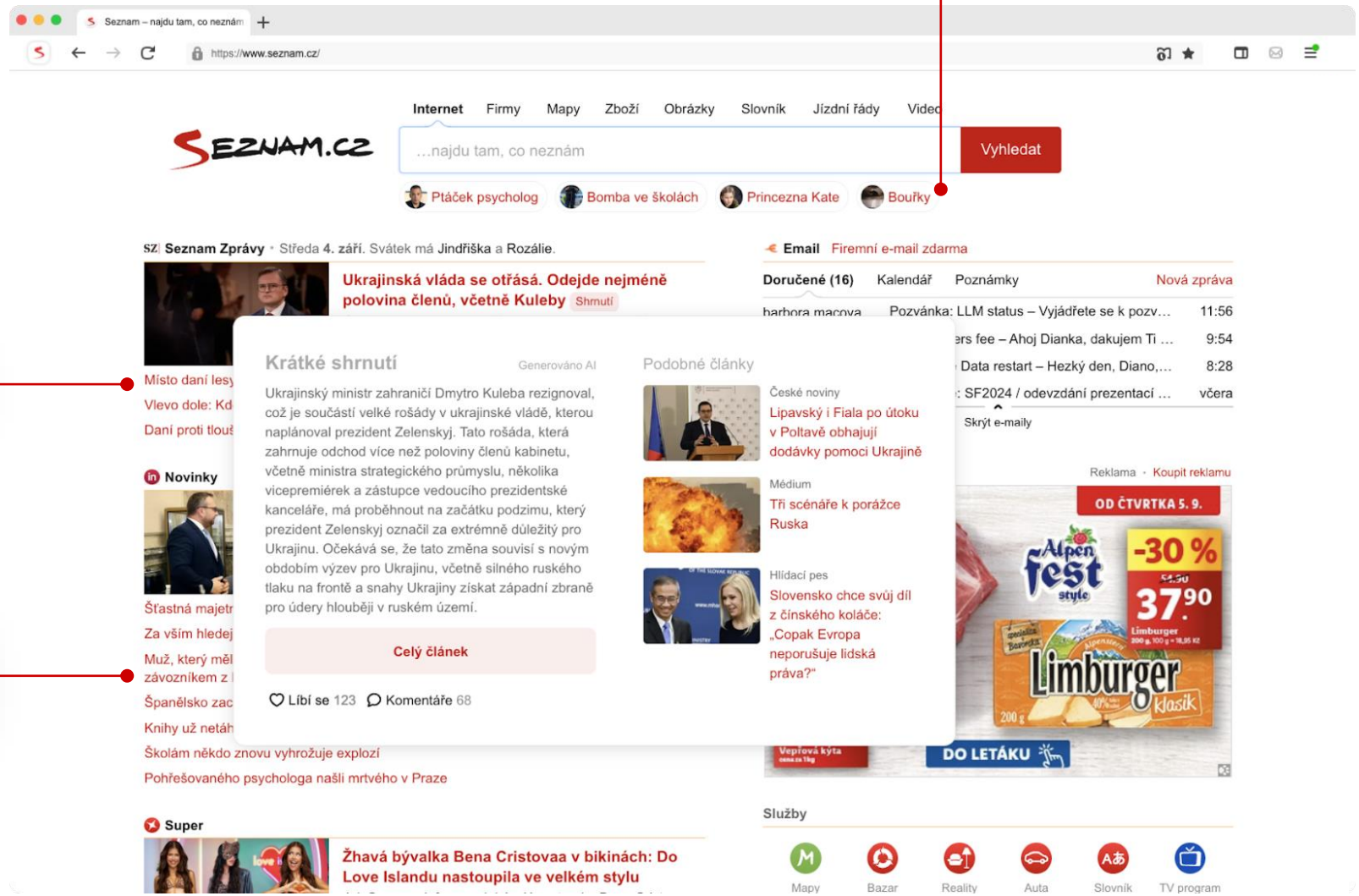
Český hokejový útočník David Pastrňák se oženil se švédskou partnerkou Rebeccou Rohlssonovou na chorvatském ostrově Hvar, kde vychovávají roční dceru Freyu Ivy.[1, 2, 3] Pastrňák se na Instagramu pochlubil fotkou ze svatby, která byla třídenním luxusním veselím, na němž nešetřil a které stálo miliony.[4, 5]

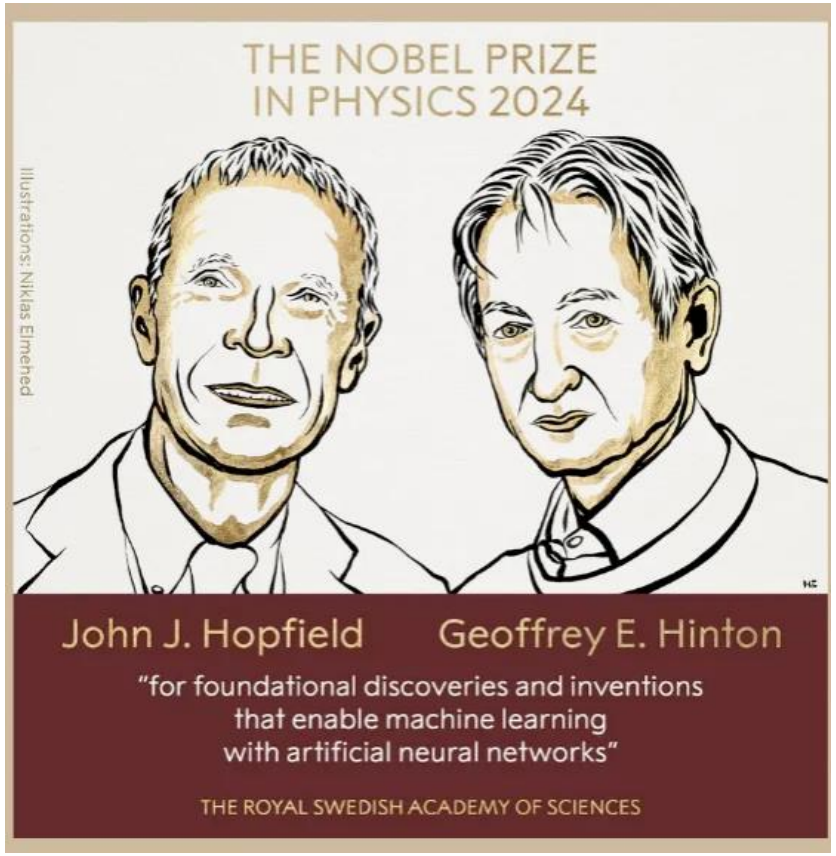
Tento text je vygenerovaným shrnutím tématu

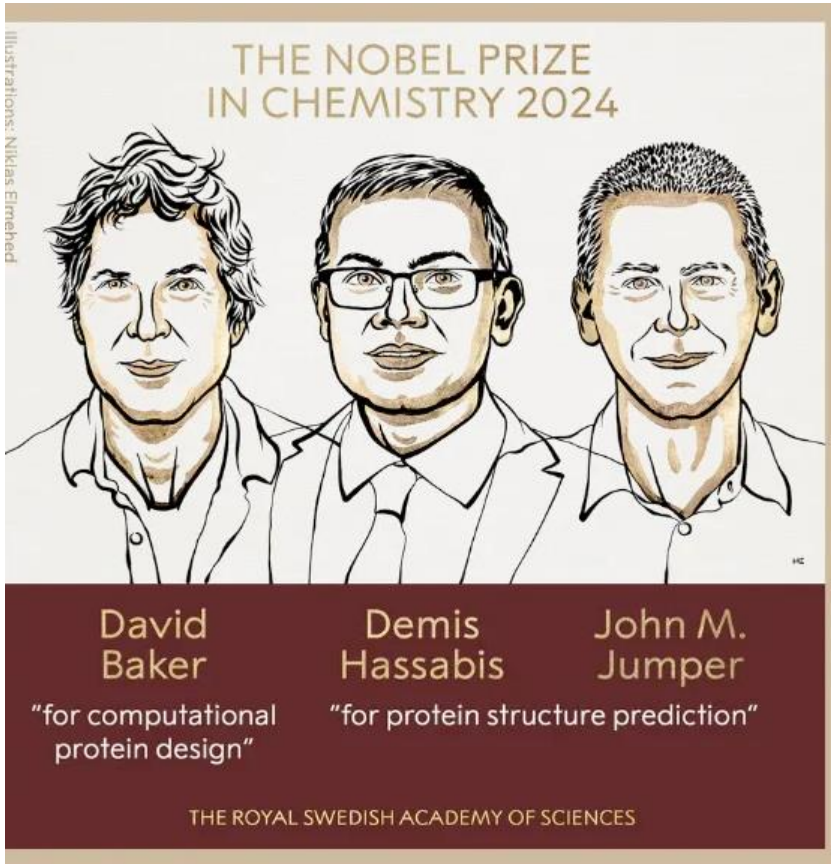
Shrnutí aktuálního dění pro nejhledanější dotazy

Sumarizace článků

Nadpisy zahraničních článků v češtině







https://o-seznam.cz/kariera/458967-ai-vyzkumnik-generativnich-jazykovych-modelu/