Qwen3.8-Flash-Next na 1× RTX PRO 6000

Výsledky benchmarku a odporúčaná konfigurácia

Report vychádza z dodaných výsledkov benchmarku. Merania ani závery neboli pri prevode do Markdownu nezávisle overované.

Všetkých 5 konfigurácií prebehlo bez chyby, OOM aj zásahu watchdogu. Odstávka produkcie trvala približne **14 min** (22:57:56 - 23:12:05 UTC). Podľa reportu produkcia opäť beží (`GetServerStatus: OK`, porty 50051 a 8080). FlashInfer sa nekompiloval: `fused_moe_120` sa načítal z cache.

Stručný prehľad

Oblasť	Výsledok
GPU	1× RTX PRO 6000 Blackwell, 96 GB
vLLM	0.31.0
RAM systému	91.9 GiB
Odporúčaná konfigurácia	**GM: CUDA graphs + MTP(3), BF16 KV**
Decode, 1 request (GM)	**159.1 tok/s**
Celkový výkon, 4 requesty (GM)	**432 tok/s**
Najvyšší celkový výkon	**474 tok/s (GMF)**
RAM samotného modelu	Približne **50 GiB**, z toho 47.7 GiB pinned PLE
Overenie kvality	Zatiaľ nevykonané; vrátane slovenčiny a vision

CUDA graphs v praxi

Pri `enforce_eager=False` vLLM 0.31.0 použije pre tento model:

• breakable CUDA graphs, `cudagraph_mode=FULL_AND_PIECEWISE`, **bez torch.compile**;

• capture sizes `[1, 2, 4, 8]`, s MTP `[1 … 32]`.

Capture trvá 6-10 s a zaberie 0.23-0.48 GiB VRAM. KV kapacita kvôli tomu klesne len o 1.6 %.

Porovnanie konfigurácií

Spoločné nastavenia: `max_num_seqs=4`, `gpu_memory_utilization=0.90`, `max_model_len=32768`.

Krátky prompt má 87 tokenov, výstup 256 tokenov (greedy). Prebehli 3 opakovania; uvedené hodnoty predstavujú medián. TTFT znamená čas do prvého tokenu. Pomlčka označuje neuvedené meranie.

Výkon

Konfigurácia	Režim	TTFT	Decode: 1 request	Spolu: n=1	Spolu: n=2	Spolu: n=4
E	Eager	0.078 s	31.2 tok/s	-	-	-
G	CUDA graphs	0.062 s	98.1 tok/s	96 tok/s	162 tok/s	251 tok/s
GM	Graphs + MTP(3)	0.078 s	**159.1 tok/s**	152 tok/s	251 tok/s	432 tok/s
GF	Graphs + FP8 KV	0.063 s	98.0 tok/s	-	-	-
GMF	Graphs + MTP(3) + FP8 KV	0.078 s	157.2 tok/s	139 tok/s	276 tok/s	**474 tok/s**

Pamäť a KV kapacita

Konfigurácia	VRAM za behu	Špička VRAM pri inite	RAM systému	KV kapacita
E	90.1 GiB	90.1 GiB	55.6 GiB	270 882 tokenov
G	85.6 GiB	88.7 GiB	55.8 GiB	266 513 tokenov
GM	86.0 GiB	90.7 GiB	55.8 GiB	117 028 tokenov
GF	85.6 GiB	90.0 GiB	55.5 GiB	**399 769 tokenov**
GMF	86.4 GiB	90.7 GiB	55.7 GiB	153 600 tokenov

Poznámky k meraniam

Konfigurácia	Poznámky
E	Baseline; init 96 s.
G	3.1× oproti eager; rozptyl ±0.1 %.
GM	Acceptance 53 %; približne 2.6 tokenu na krok; rozptyl ±6 %; init 164 s.
GF	+50 % KV kapacity; výkon bez zmeny.
GMF	Najvyšší celkový výkon; FP8 KV bez kalibrácie, škála 1.0.

Hodnoty RAM sú celkové využitie systému pri vypnutej produkcii. Pred štartom enginu bolo obsadených približne 5.5 GiB, takže samotný model si berie približne 50 GiB, z toho 47.7 GiB je pinned PLE tabuľka. Systém má 91.9 GiB.

Škálovanie kontextu: TTFT a prefill

Kontext	G: TTFT (BF16 KV)	G: prefill	GF: TTFT (FP8 KV)	GF: prefill	GMF: TTFT
1K	0.118 s	8.7K tok/s	0.106 s	9.6K tok/s	0.107 s
8K	0.684 s	12.0K tok/s	0.595 s	13.8K tok/s	0.610 s

Kontext	G: TTFT (BF16 KV)	G: prefill	GF: TTFT (FP8 KV)	GF: prefill	GMF: TTFT
32K	**2.43 s**	13.2K tok/s	**2.28 s**	14.0K tok/s	2.33 s

VRAM aj RAM sú pri 1K až 32K konštantné, pretože KV cache je predalokovaná. Všetky dlhé prompty prebehli stabilne.

Súbežnosť a PLE/UVA

• **Celkový výkon rastie so súbežnosťou:** G 96 → 162 → 251 tok/s; GMF 139 → 276 → 474 tok/s. Report uvádza rýchlosť jedného requestu pri n=4 približne 77 (G) a 126 (GMF) tok/s, čo je pri MoE bežné.

• **UVA offload podľa meraní nie je úzke hrdlo:** prenos z hostu na GPU počas inferencie je 20-140 MB/s, zlomok kapacity PCIe 5.0.

• **TTFT pri prvom novom tvare batchu** vyskočí na 0.3-0.75 s (G pri n=4, GM pri n=2). Pri ďalších behoch tohto tvaru sa už neopakoval: GMF pri n=4 mal 0.06 s. Report to pripisuje jednorazovej JIT kompilácii alebo autotune.

MTP a FP8 KV

MTP

MTP funguje; recipe používa `num_speculative_tokens=3`. Draft vrstva vo FP8 beží na SM120 cez **DEEPGEMM**. Pôvodný predpoklad VLLM_CUTLASS/TRITON sa nepotvrdil.

• Prínos: +62 % decode pri 1 requeste, +72-89 % celkového výkonu pri 4 requestoch.

• Cena: +2.67 GiB váh; KV kapacita klesne na 117K tokenov (s FP8 KV 153K).

• Hodnoty `num_speculative_tokens` 1 a 2 neboli testované.

FP8 KV

FP8 KV je natívne podporovaná a v benchmarku výkon neznižovala; prefill bol o 6-14 % rýchlejší. Kapacita stúpne o 50 % (s MTP o 31 %). Nárast je menší než 2×, lebo stav GDN vrstiev zostáva v plnej presnosti.

Neoverená kvalita: checkpoint nemá kalibrované KV škály (použije sa škála 1.0) a dopad na kvalitu nebol meraný.

Odporúčaná konfigurácia podľa benchmarku

Pre 1× RTX PRO 6000 Blackwell 96 GB, vLLM 0.31.0 a približne 92 GB RAM report odporúča **GM: CUDA graphs + MTP(3), BF16 KV cache**.

Nasledujúci blok predstavuje parametre konfigurácie, nie samostatný spustiteľný skript:

```
enforce_eager = False  # breakable CUDA graphs, bez torch.compile
speculative_config = {"method": "mtp", "num_speculative_tokens": 3}
engram_config = {"cpu_offload": True}
max_model_len = 32768
max_num_seqs = 4
max_num_batched_tokens = 4096
gpu_memory_utilization = 0.90

# Prostredie:
# PYTORCH_ALLOC_CONF=pinned_max_round_threshold_mb:1024
# MAX_JOBS=4
# HF_HUB_OFFLINE=1
# Lokálny snapshot a predkompilovaný FlashInfer fused_moe_120.
```

Dôvody odporúčania:

• **Výkon:** približne 159 tok/s pre jedného používateľa a 430 tok/s spolu pri 4 requestoch.

• **BF16 KV:** vyhne sa zatiaľ neoverenému dopadu FP8 KV na kvalitu.

• **Kapacita:** 117K tokenov KV predstavuje približne 29K tokenov na request pri štyroch súbežných requestoch. Report očakáva pri vyššom zaťažení zaradenie do fronty namiesto pádu; toto správanie nie je samo osebe garanciou pre všetky produkčné záťaže.

Ak treba kapacitu pre 4× plný 32K kontext, report navrhuje GMF, ale až po overení kvality na reálnych slovenských dokumentoch:

```
kv_cache_dtype = "fp8"
attention_config = {"indexer_kv_dtype": "fp8"}
```

Pred nasadením treba vyriešiť

Oblasť	Stav alebo potrebný krok
Kvalita výstupu	Texty v benchmarku neboli ukladané. Chýba porovnanie greedy výstupu s MTP a bez neho; nebola testovaná slovenčina ani vision.
Warmup	Po štarte zahriať batch 1/2/4, aby prvé requesty nemali TTFT spike.
GPU	Model potrebuje PRO 6000 samostatne; vedľa `qwen3.6-27b` sa na ňu nezmestí.
RAM	Natrvalo zaberie približne 50 GiB RAM.
Životný cyklus	Studený štart s MTP trvá približne 2.7 min; odporúčané je držať model načítaný a nevyhadzovať ho.
Swap	Pri každom načítaní checkpointu sa zaplnil na 8 GiB, hoci voľnej RAM bolo dosť; report uvádza tlak page cache. Na zváženie je `vm.swappiness`. Systémové nastavenia neboli menené.

`ModelManager` ani `model_catalog.json` neboli menené.

Podkladové výsledky

Report uvádza výsledky v `scratchpad/bench/`: JSON a log pre každú konfiguráciu plus CSV pamäte a PCIe. Tieto podkladové súbory nie sú súčasťou tohto Markdown dokumentu.