# Agentic Serving Qwen3.8-27B on Dual RTX 5090s with vLLM

**Alis**

https://softmaxxer101.github.io/
https://x.com/import_alis

Date: 16th November 2026

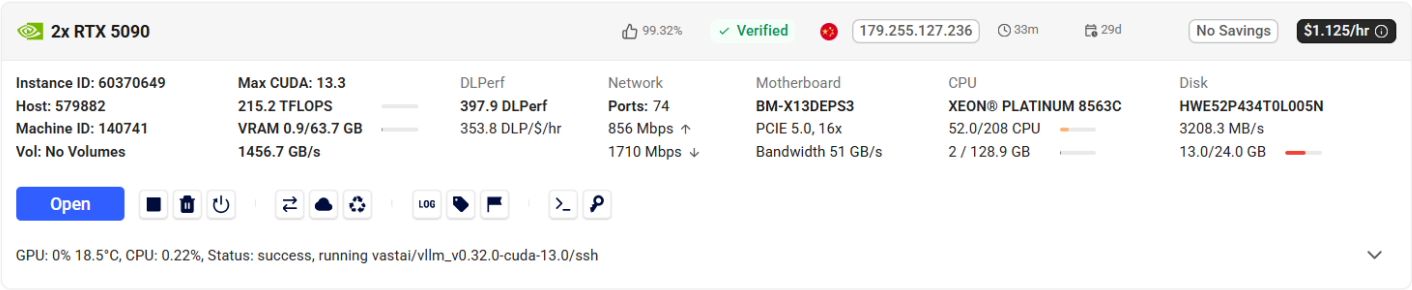
***Abstract-*** This technical paper evaluates the deployment and optimization of the Qwen3.8-27B-NVFP4 model on a distributed 2x RTX 5090 cluster using vLLM. Across structured experiments using synthetic workloads (light.jsonl, grow.jsonl, burst.jsonl), we analyze tensor parallelism, speculative decoding via MTP, quantized linear backends (marlin vs. flashinfer_cutlass), sub-byte KV cache precision (fp8 vs. nvfp4), and host-RAM offloading. We further isolate the impacts of context scaling up to 262,144 tokens, prefix caching, and scheduler tuning on TTFT, ITL, and serving throughput.

***Index Terms***- Autoregressive, inference, KV cache, LLM inference, PagedAttention, Paged KV caching, speculative decoding, tensor parallelism, vLLM.

## I. THE HARDWARE SETUP

WE have rented dual RTX 5090's with 31.4GB VRAM each from vast.ai @$1.125/hr. Model used in this experiment is Qwen3.8-27B NVFP4. The total budget of this experiment was - $20 + 1-day sustainability cost for the running instance.

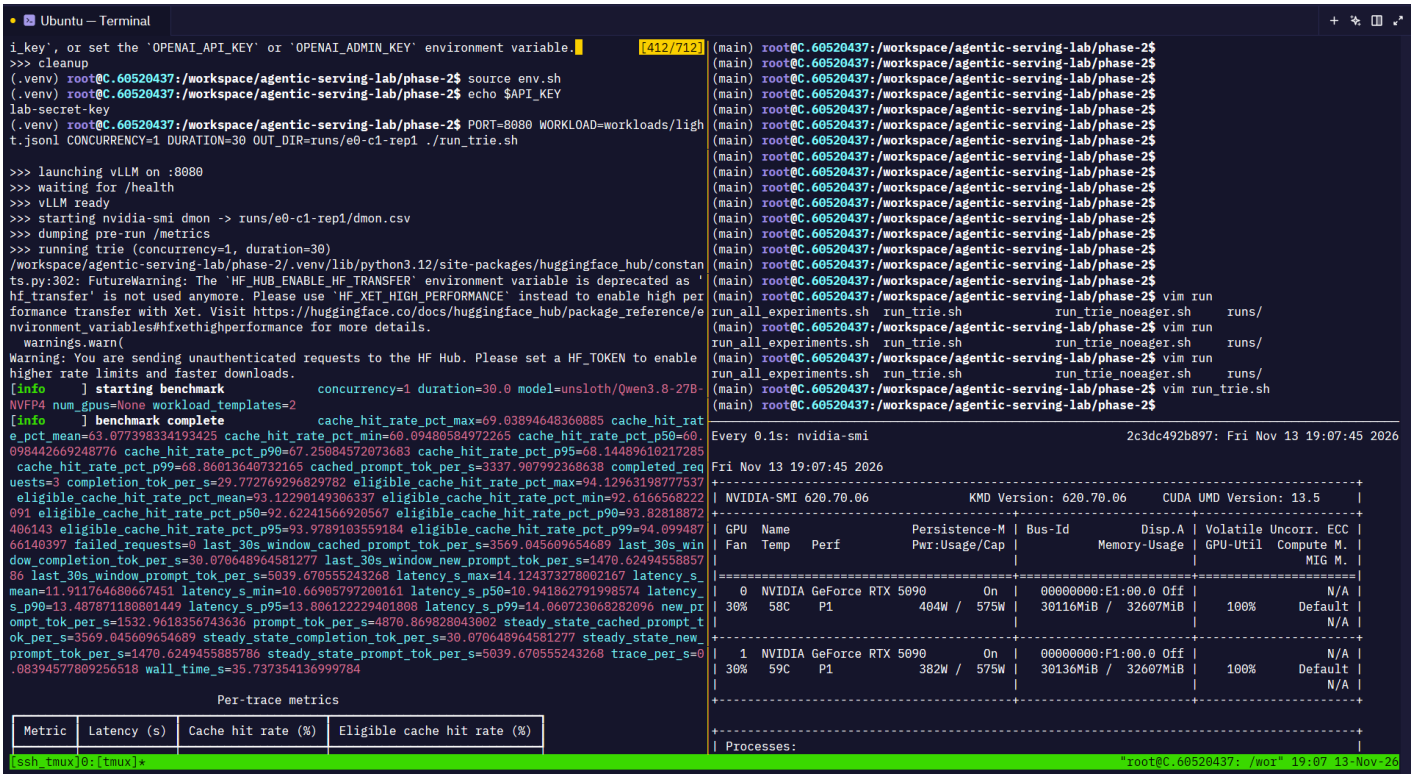
The current gpu's rented faced an issue regarding accesing hugging face models directly. The hugging face is unreachable in china, hence using hugging face mirror is recommended.

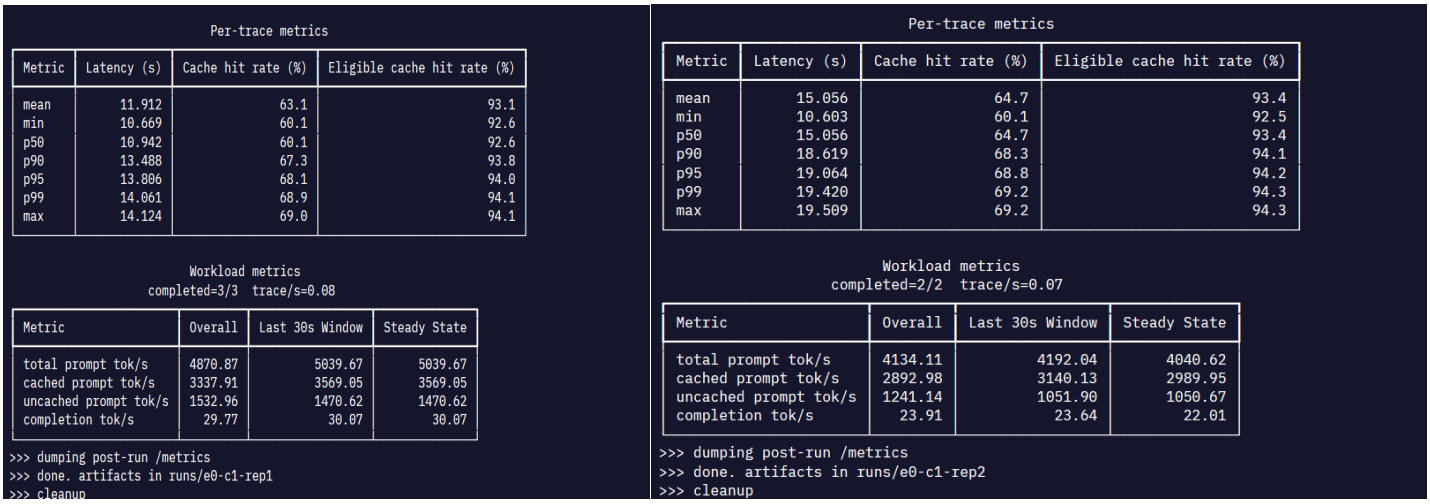



## II. EXPERIMENTS

### E0 — Baseline validation (light.jsonl)

Run: WORKLOAD=workloads/light.jsonl CONCURRENCY=1 DURATION=60 OUT_DIR=runs/e0-c1-rep1 ./run_trie.sh



Per-trace metrics

| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 11.912 | 63.1 | 93.1 |
| min | 10.669 | 60.1 | 92.6 |
| p50 | 10.942 | 60.1 | 92.6 |
| p90 | 13.488 | 67.3 | 93.8 |
| p95 | 13.806 | 68.1 | 94.0 |
| p99 | 14.061 | 68.9 | 94.1 |
| max | 14.124 | 69.0 | 94.1 |

Workload metrics
completed=3/3 trace/s=0.08

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 4870.87 | 5039.67 | 5039.67 |
| cached prompt tok/s | 3337.91 | 3569.05 | 3569.05 |
| uncached prompt tok/s | 1532.96 | 1470.62 | 1470.62 |
| completion tok/s | 29.77 | 30.07 | 30.07 |

>>> dumping post-run /metrics
>>> done. artifacts in runs/e0-c1-rep1
>>> cleanup

Per-trace metrics

| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 15.056 | 64.7 | 93.4 |
| min | 10.603 | 60.1 | 92.5 |
| p50 | 15.056 | 64.7 | 93.4 |
| p90 | 18.619 | 68.3 | 94.1 |
| p95 | 19.064 | 68.8 | 94.2 |
| p99 | 19.420 | 69.2 | 94.3 |
| max | 19.509 | 69.2 | 94.3 |

Workload metrics
completed=2/2 trace/s=0.07

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 4134.11 | 4192.04 | 4040.62 |
| cached prompt tok/s | 2892.98 | 3140.13 | 2989.95 |
| uncached prompt tok/s | 1241.14 | 1051.90 | 1050.67 |
| completion tok/s | 23.91 | 23.64 | 22.01 |

>>> dumping post-run /metrics
>>> done. artifacts in runs/e0-c1-rep2
>>> cleanup

| Run | Completed traces | Average latency | Completion throughput |
|---|---|---|---|
| C=1, rep1 | 3/3 | 11.912 s | 29.77 tok/s |
| C=1, rep2 | 2/2 | 15.056 s | 23.91 tok/s |

The baseline is a verification test to check if the server is working and giving results. Bash baseline to check whether the server is working perfectly fine.

2. **E1 — MTP speculation (light.jsonl, c=1 and c=8)**

**Run :** EXTRA_ARGS='--speculative-config {"method":"mtp","num_speculative_tokens":2}' WORKLOAD=workloads/light.jsonl CONCURRENCY=1 DURATION=60 OUT_DIR=runs/e1-mtp-c1-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ PORT=8080 EXTRA_ARGS='--speculati
ve-config {"method":"mtp","num_speculative_tokens":2}' WORKLOAD=workloads/light.jsonl CONCURRENCY
=1 DURATION=60 OUT_DIR=runs/e1-mtp-c1-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e1-mtp-c1-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=1, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constan
ts.py:302: FutureWarning: The `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as '
hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERFORMANCE` instead to enable high per
formance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_reference/e
nvironment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable
higher rate limits and faster downloads.
[info     ] starting benchmark            concurrency=1 duration=60.0 model=unsloth/Qwen3.8-27B-
NVFP4 num_gpus=None workload_templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=62.096838949706914 cache_hit_ra
te_pct_mean=56.806539876042784 cache_hit_rate_pct_min=51.642055999354476 cache_hit_rate_pct_p50=5
6.76380472235093 cache_hit_rate_pct_p90=62.01271710581776 cache_hit_rate_pct_p95=62.0547780277623
4 cache_hit_rate_pct_p99=62.088426765318 cached_prompt_tok_per_s=3546.3783671792744 completed_req
uests=6 completion_tok_per_s=29.30891212544855 eligible_cache_hit_rate_pct_max=84.64527427622818
eligible_cache_hit_rate_pct_mean=82.04440307069957 eligible_cache_hit_rate_pct_min=79.58961604228
199 eligible_cache_hit_rate_pct_p50=82.00182349250139 eligible_cache_hit_rate_pct_p90=84.54053243
045482 eligible_cache_hit_rate_pct_p95=84.5929033533415 eligible_cache_hit_rate_pct_p99=84.634800
09165085 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=3949.8784986377373 last_30s_wi
ndow_completion_tok_per_s=30.60293253786929 last_30s_window_new_prompt_tok_per_s=1478.54168182960
64 last_30s_window_prompt_tok_per_s=5428.420180467344 latency_s_max=14.368097607999516 latency_s_
mean=12.282247687333438 latency_s_min=10.2027693179989 latency_s_p50=12.418945131001237 latency_s
_p90=14.086850531999517 latency_s_p95=14.227474069999516 latency_s_p99=14.339972900399516 new_pro
mpt_tok_per_s=1521.4581938899514 prompt_tok_per_s=5067.836561069225 steady_state_cached_prompt_to
k_per_s=3764.468514482454 steady_state_completion_tok_per_s=29.029415121548052 steady_state_new_p
rompt_tok_per_s=1426.4311657829867 steady_state_prompt_tok_per_s=5190.899680265441 trace_per_s=0.
08141364479291263 wall_time_s=73.69772002299942
```

Per-trace metrics

| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 12.282 | 56.8 | 82.0 |
| min | 10.203 | 51.6 | 79.6 |
| p50 | 12.419 | 56.8 | 82.0 |
| p90 | 14.087 | 62.0 | 84.5 |
| p95 | 14.227 | 62.1 | 84.6 |
| p99 | 14.340 | 62.1 | 84.6 |
| max | 14.368 | 62.1 | 84.6 |

Workload metrics
completed=6/6  trace/s=0.08

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 5067.84 | 5428.42 | 5190.90 |
| cached prompt tok/s | 3546.38 | 3949.88 | 3764.47 |
| uncached prompt tok/s | 1521.46 | 1478.54 | 1426.43 |
| completion tok/s | 29.31 | 30.60 | 29.03 |

```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e1-mtp-c1-rep1
>>> cleanup
```

**Run:** EXTRA_ARGS='--speculative-config {"method":"mtp","num_speculative_tokens":2}' WORKLOAD=workloads/light.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e1-mtp-c8-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ PORT=8080 EXTRA_ARGS='--speculati
ve-config {"method":"mtp","num_speculative_tokens":2}' WORKLOAD=workloads/light.jsonl CONCURRENCY
=8 DURATION=60 OUT_DIR=runs/e1-mtp-c8-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e1-mtp-c8-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=8, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constan
ts.py:302: FutureWarning: The `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as '
hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERFORMANCE` instead to enable high per
formance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_reference/e
nvironment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable
higher rate limits and faster downloads.
[info     ] starting benchmark            concurrency=8 duration=60.0 model=unsloth/Qwen3.8-27B-
NVFP4 num_gpus=None workload_templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=62.020477450744515 cache_hit_ra
te_pct_mean=56.788955862241565 cache_hit_rate_pct_min=51.58586224962721 cache_hit_rate_pct_p50=56
.75361771193482 cache_hit_rate_pct_p90=61.995539016513014 cache_hit_rate_pct_p95=62.0080082336287
65 cache_hit_rate_pct_p99=62.01798360732136 cached_prompt_tok_per_s=5324.890658480506 completed_r
equests=8 completion_tok_per_s=47.67464088171252 eligible_cache_hit_rate_pct_max=84.4988344988345
 eligible_cache_hit_rate_pct_mean=82.01103868752898 eligible_cache_hit_rate_pct_min=79.5006366261
9173 eligible_cache_hit_rate_pct_p50=81.99247902902741 eligible_cache_hit_rate_pct_p90=84.4956031
8585175 eligible_cache_hit_rate_pct_p95=84.49721884234312 eligible_cache_hit_rate_pct_p99=84.4985
1136753622 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=8712.196886287209 last_30s_w
indow_completion_tok_per_s=64.72328628456431 last_30s_window_new_prompt_tok_per_s=3204.0792663264
66 last_30s_window_prompt_tok_per_s=11916.276152613675 latency_s_max=65.2875134370006 latency_s_m
ean=53.761101316999884 latency_s_min=47.31507517299906 latency_s_p50=51.2281548864994 latency_s_p
90=61.03815507269974 latency_s_p95=63.16283425485017 latency_s_p99=64.8625776005705 new_prompt_to
k_per_s=2847.7652153342947 prompt_tok_per_s=8172.655873814801 steady_state_cached_prompt_tok_per_
s=7263.1278086421835 steady_state_completion_tok_per_s=63.86083449747374 steady_state_new_prompt_
tok_per_s=3670.914183541467 steady_state_prompt_tok_per_s=10934.041992183651 trace_per_s=0.122242
668927468 wall_time_s=65.44359731499935
```

Per-trace metrics

| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 53.761 | 56.8 | 82.0 |
| min | 47.315 | 51.6 | 79.5 |
| p50 | 51.228 | 56.8 | 82.0 |
| p90 | 61.038 | 62.0 | 84.5 |
| p95 | 63.163 | 62.0 | 84.5 |
| p99 | 64.863 | 62.0 | 84.5 |
| max | 65.288 | 62.0 | 84.5 |

Workload metrics
completed=8/8  trace/s=0.12

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 8172.66 | 11916.28 | 10934.04 |
| cached prompt tok/s | 5324.89 | 8712.20 | 7263.13 |
| uncached prompt tok/s | 2847.77 | 3204.08 | 3670.91 |
| completion tok/s | 47.67 | 64.72 | 63.86 |

```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e1-mtp-c8-rep1
>>> cleanup
```

**Run:** WORKLOAD=workloads/light.jsonl CONCURRENCY=1 DURATION=60 OUT_DIR=runs/e1-off-c1-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ PORT=8080 WORKLOAD=workloads/ligh
t.jsonl CONCURRENCY=1 DURATION=60 OUT_DIR=runs/e1-off-c1-rep1 ./run_trie.sh

>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e1-off-c1-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=1, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constan
ts.py:302: FutureWarning: The `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as '
hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERFORMANCE` instead to enable high per
formance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_reference/e
nvironment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable
higher rate limits and faster downloads.
[info     ] starting benchmark            concurrency=1 duration=60.0 model=unsloth/Qwen3.8-27B-
NVFP4 num_gpus=None workload_templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=69.08226749619503 cache_hit_rat
e_pct_mean=64.58202104088842 cache_hit_rate_pct_min=60.08995744165877 cache_hit_rate_pct_p50=64.5
7558949397881 cache_hit_rate_pct_p90=69.07627357323267 cache_hit_rate_pct_p95=69.07927053471386 c
ache_hit_rate_pct_p99=69.08166810389879 cached_prompt_tok_per_s=3535.318480649343 completed_reque
sts=6 completion_tok_per_s=29.217508104540027 eligible_cache_hit_rate_pct_max=94.18617350467801
```

Per-trace metrics

| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 12.321 | 64.6 | 93.4 |
| min | 10.574 | 60.1 | 92.6 |
| p50 | 12.319 | 64.6 | 93.4 |
| p90 | 14.060 | 69.1 | 94.2 |
| p95 | 14.066 | 69.1 | 94.2 |
| p99 | 14.071 | 69.1 | 94.2 |
| max | 14.073 | 69.1 | 94.2 |

Workload metrics
completed=6/6  trace/s=0.08

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 5052.03 | 5283.30 | 5108.89 |
| cached prompt tok/s | 3535.32 | 3844.29 | 3704.99 |
| uncached prompt tok/s | 1516.71 | 1439.02 | 1403.89 |
| completion tok/s | 29.22 | 29.78 | 28.57 |

```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e1-off-c1-rep1
>>> cleanup
```

**Run :** WORKLOAD=workloads/light.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e1-off-c8-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ PORT=8080 WORKLOAD=workloads/ligh
t.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e1-off-c8-rep1 ./run_trie.sh

>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e1-off-c8-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=8, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constan
ts.py:302: FutureWarning: The `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as '
hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERFORMANCE` instead to enable high per
formance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_reference/e
nvironment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable
higher rate limits and faster downloads.
[info     ] starting benchmark            concurrency=8 duration=60.0 model=unsloth/Qwen3.8-27B-
NVFP4 num_gpus=None workload_templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=69.32104388832325 cache_hit_rat
e_pct_mean=64.08845492044061 cache_hit_rate_pct_min=59.96296594477095 cache_hit_rate_pct_p50=60.1
0935576942478 cache_hit_rate_pct_p90=69.1204324978809 cache_hit_rate_pct_p95=69.22073819310206 ca
che_hit_rate_pct_p99=69.30098274927902 cached_prompt_tok_per_s=7638.443917074213 completed_reques
ts=9 completion_tok_per_s=67.0067942342426 eligible_cache_hit_rate_pct_max=94.48025270692206 elig
ible_cache_hit_rate_pct_mean=93.30231653098363 eligible_cache_hit_rate_pct_min=92.3811591057087 e
ligible_cache_hit_rate_pct_p50=92.639696507976 eligible_cache_hit_rate_pct_p90=94.23538988072374
eligible_cache_hit_rate_pct_p95=94.35782129382291 eligible_cache_hit_rate_pct_p99=94.455766424302
23 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=11019.375843222468 last_30s_window_c
ompletion_tok_per_s=91.06660859273066 last_30s_window_new_prompt_tok_per_s=4816.411743348866 last
_30s_window_prompt_tok_per_s=15835.787586571334 latency_s_max=50.8582139680002 latency_s_mean=42.
72651608622214 latency_s_min=35.22982046999823 latency_s_p50=37.35689281300074 latency_s_p90=50.2
0950570079876 5 latency_s_p95=50.53385983439948 latency_s_p99=50.793343141280054 new_prompt_tok_pe
r_s=3738.2737835945877 prompt_tok_per_s=11376.7177006688 steady_state_cached_prompt_tok_per_s=956
2.63231243637 steady_state_completion_tok_per_s=80.71303730792935 steady_state_new_prompt_tok_per
_s=4220.808952690897 steady_state_prompt_tok_per_s=13783.441265127267 trace_per_s=0.1239846110419
7851 wall_time_s=72.58965386400087
```

Per-trace metrics

| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 42.727 | 64.1 | 93.3 |
| min | 35.230 | 60.0 | 92.4 |
| p50 | 37.357 | 60.1 | 92.6 |
| p90 | 50.210 | 69.1 | 94.2 |
| p95 | 50.534 | 69.2 | 94.4 |
| p99 | 50.793 | 69.3 | 94.5 |
| max | 50.858 | 69.3 | 94.5 |

Workload metrics
completed=9/9  trace/s=0.12

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 11376.72 | 15835.79 | 13783.44 |
| cached prompt tok/s | 7638.44 | 11019.38 | 9562.63 |
| uncached prompt tok/s | 3738.27 | 4816.41 | 4220.81 |
| completion tok/s | 67.01 | 91.07 | 80.71 |

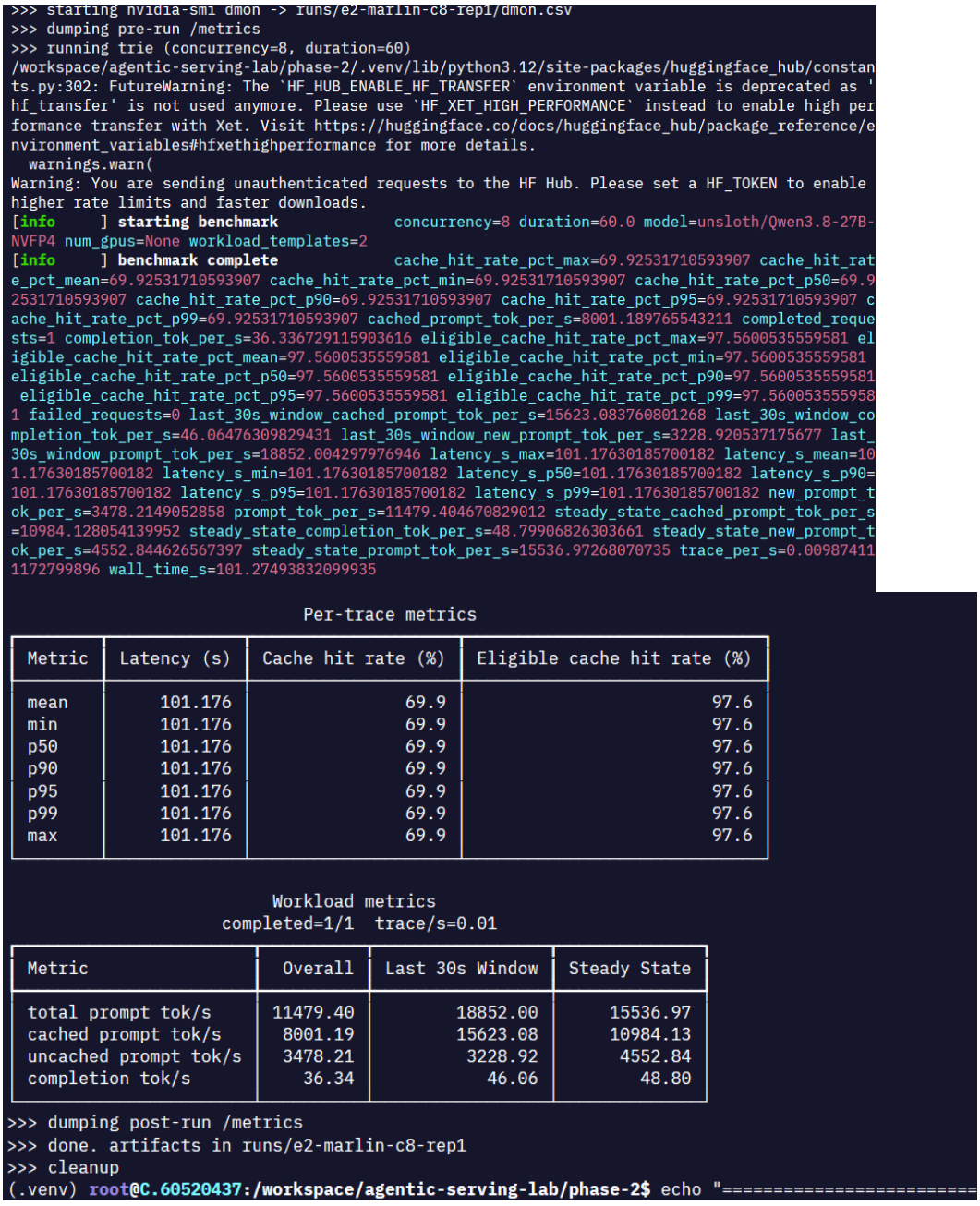
```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e1-off-c8-rep1
>>> cleanup
```

| Concurrency | Configuration | Latency | Completion throughput |
|---|---|---|---|
| **C=1** | **MTP** | **12.282 s** | **29.31 tok/s** |
| **C=1** | **No MTP** | **12.321 s** | **29.22 tok/s** |
| **C=8** | **MTP** | **53.761 s** | **47.67 tok/s** |
| **C=8** | **No MTP** | **42.727 s** | **67.01 tok/s** |

This experiment compares MTP with no MTP at C=1 and C=8. MTP produces almost no measurable improvemnets at c=1 and performs considerably worse at c=8. Inference: Speculative decoding is not automatically beneficial.

### 3. E2 — Linear backend: flashinfer_cutlass vs marlin (burst.jsonl, c=8)

**Run :** EXTRA_ARGS="--linear-backend flashinfer_cutlass" WORKLOAD=workloads/burst.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e2-ficutlass-c8-rep1 ./run_trie.sh



```
>>> starting nvidia-smi dmon -> runs/e2-marlin-c8-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=8, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constan
ts.py:302: FutureWarning: The `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as '
hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERFORMANCE` instead to enable high per
formance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_reference/e
nvironment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable
higher rate limits and faster downloads.
[info     ] starting benchmark            concurrency=8 duration=60.0 model=unsloth/Qwen3.8-27B-
NVFP4 num_gpus=None workload_templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=69.92531710593907 cache_hit_rat
e_pct_mean=69.92531710593907 cache_hit_rate_pct_min=69.92531710593907 cache_hit_rate_pct_p50=69.9
2531710593907 cache_hit_rate_pct_p90=69.92531710593907 cache_hit_rate_pct_p95=69.92531710593907 c
ache_hit_rate_pct_p99=69.92531710593907 cached_prompt_tok_per_s=8001.189765543211 completed_reque
sts=1 completion_tok_per_s=36.336729115903616 eligible_cache_hit_rate_pct_max=97.5600535559581 el
igible_cache_hit_rate_pct_mean=97.5600535559581 eligible_cache_hit_rate_pct_min=97.5600535559581
eligible_cache_hit_rate_pct_p50=97.5600535559581 eligible_cache_hit_rate_pct_p90=97.5600535559581
 eligible_cache_hit_rate_pct_p95=97.5600535559581 eligible_cache_hit_rate_pct_p99=97.560053555958
1 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=15623.083760801268 last_30s_window_co
mpletion_tok_per_s=46.06476309829431 last_30s_window_new_prompt_tok_per_s=3228.920537175677 last_
30s_window_prompt_tok_per_s=18852.004297976946 latency_s_max=101.17630185700182 latency_s_mean=10
1.17630185700182 latency_s_min=101.17630185700182 latency_s_p50=101.17630185700182 latency_s_p90=
101.17630185700182 latency_s_p95=101.17630185700182 latency_s_p99=101.17630185700182 new_prompt_t
ok_per_s=3478.2149052858 prompt_tok_per_s=11479.404670829012 steady_state_cached_prompt_tok_per_s
=10984.128054139952 steady_state_completion_tok_per_s=48.79906826303661 steady_state_new_prompt_t
ok_per_s=4552.844626567397 steady_state_prompt_tok_per_s=15536.97268070735 trace_per_s=0.00987411
1172799896 wall_time_s=101.27493832099935

                    Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃ Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │     101.176 │               69.9 │                       97.6 │
│ min    │     101.176 │               69.9 │                       97.6 │
│ p50    │     101.176 │               69.9 │                       97.6 │
│ p90    │     101.176 │               69.9 │                       97.6 │
│ p95    │     101.176 │               69.9 │                       97.6 │
│ p99    │     101.176 │               69.9 │                       97.6 │
│ max    │     101.176 │               69.9 │                       97.6 │
└────────┴─────────────┴────────────────────┴───────────────────────────┘

                    Workload metrics
                completed=1/1  trace/s=0.01
┏━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━┳━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━┓
┃ Metric             ┃ Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━╇━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━┩
│ total prompt tok/s │ 11479.40 │        18852.00 │     15536.97 │
│ cached prompt tok/s │  8001.19 │        15623.08 │     10984.13 │
│ uncached prompt tok/s │ 3478.21 │         3228.92 │      4552.84 │
│ completion tok/s   │   36.34 │           46.06 │        48.80 │
└────────────────────┴─────────┴─────────────────┴──────────────┘
>>> dumping post-run /metrics
>>> done. artifacts in runs/e2-marlin-c8-rep1
>>> cleanup
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ echo "=========================
```

### 4. E3 — KV cache precision: nvfp4 vs fp8 (grow.jsonl, c=4)

**Run:** EXTRA_ARGS="--kv-cache-dtype nvfp4" WORKLOAD=workloads/grow.jsonl CONCURRENCY=4 DURATION=60 OUT_DIR=runs/e3-nvfp4-c4-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ PORT=8080 EXTRA_ARGS="--kv-cache-
dtype nvfp4" WORKLOAD=workloads/grow.jsonl CONCURRENCY=4 DURATION=60 OUT_DIR=runs/e3-nvfp4-c4-rep
1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e3-nvfp4-c4-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=4, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constan
ts.py:302: FutureWarning: The `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as '
hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERFORMANCE` instead to enable high per
formance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_reference/e
nvironment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable
higher rate limits and faster downloads.
[info     ] starting benchmark            concurrency=4 duration=60.0 model=unsloth/Qwen3.8-27B-
NVFP4 num_gpus=None workload_templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=68.94680088123455 cache_hit_rat
e_pct_mean=68.94680088123455 cache_hit_rate_pct_min=68.94680088123455 cache_hit_rate_pct_p50=68.9
4680088123455 cache_hit_rate_pct_p90=68.94680088123455 cache_hit_rate_pct_p95=68.94680088123455 c
ache_hit_rate_pct_p99=68.94680088123455 cached_prompt_tok_per_s=6725.914749122728 completed_reque
sts=1 completion_tok_per_s=29.549972706059737 eligible_cache_hit_rate_pct_max=92.38235477113881 e
ligible_cache_hit_rate_pct_mean=92.38235477113881 eligible_cache_hit_rate_pct_min=92.382354771138
81 eligible_cache_hit_rate_pct_p50=92.38235477113881 eligible_cache_hit_rate_pct_p90=92.382354771
13881 eligible_cache_hit_rate_pct_p95=92.38235477113881 eligible_cache_hit_rate_pct_p99=92.382354
77113881 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=11329.347947972385 last_30s_wi
ndow_completion_tok_per_s=35.66186934341991 last_30s_window_new_prompt_tok_per_s=2629.27830295166
3 last_30s_window_prompt_tok_per_s=13958.626250924048 latency_s_max=70.31227830500211 latency_s_m
ean=70.31227830500211 latency_s_min=70.31227830500211 latency_s_p50=70.31227830500211 latency_s_p
90=70.31227830500211 latency_s_p95=70.31227830500211 latency_s_p99=70.31227830500211 new_prompt_t
ok_per_s=2531.296123497548 prompt_tok_per_s=9257.210872620275 steady_state_cached_prompt_tok_per_
s=9206.392246965965 steady_state_completion_tok_per_s=34.50300555816168 steady_state_new_prompt_t
ok_per_s=2330.8697088180334 steady_state_prompt_tok_per_s=11537.261955783999 trace_per_s=0.014206
717647144105 wall_time_s=70.389235912
```

Per-trace metrics

| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 70.312 | 68.9 | 92.4 |
| min | 70.312 | 68.9 | 92.4 |
| p50 | 70.312 | 68.9 | 92.4 |
| p90 | 70.312 | 68.9 | 92.4 |
| p95 | 70.312 | 68.9 | 92.4 |
| p99 | 70.312 | 68.9 | 92.4 |
| max | 70.312 | 68.9 | 92.4 |

Workload metrics
completed=1/1  trace/s=0.01

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 9257.21 | 13958.63 | 11537.26 |
| cached prompt tok/s | 6725.91 | 11329.35 | 9206.39 |
| uncached prompt tok/s | 2531.30 | 2629.28 | 2330.87 |
| completion tok/s | 29.55 | 35.66 | 34.50 |

```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e3-nvfp4-c4-rep1
>>> cleanup
```

**Run:** EXTRA_ARGS="--kv-cache-dtype nvfp4" WORKLOAD=workloads/grow.jsonl CONCURRENCY=4 DURATION=60 OUT_DIR=runs/e3-nvfp4-c4-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ PORT=8080 WORKLOAD=workloads/grow
.jsonl CONCURRENCY=4 DURATION=60 OUT_DIR=runs/e3-fp8-c4-rep1 ./run_trie.sh

>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e3-fp8-c4-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=4, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constan
ts.py:302: FutureWarning: The `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as '
hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERFORMANCE` instead to enable high per
formance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_reference/e
nvironment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable
higher rate limits and faster downloads.
[info     ] starting benchmark            concurrency=4 duration=60.0 model=unsloth/Qwen3.8-27B-
NVFP4 num_gpus=None workload_templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=72.02051948990282 cache_hit_rat
e_pct_mean=72.02051948990282 cache_hit_rate_pct_min=72.02051948990282 cache_hit_rate_pct_p50=72.0
2051948990282 cache_hit_rate_pct_p90=72.02051948990282 cache_hit_rate_pct_p95=72.02051948990282 c
ache_hit_rate_pct_p99=72.02051948990282 cached_prompt_tok_per_s=7223.681429790874 completed_reque
sts=1 completion_tok_per_s=31.73688591807275 eligible_cache_hit_rate_pct_max=96.49497652484771 el
igible_cache_hit_rate_pct_mean=96.49497652484771 eligible_cache_hit_rate_pct_min=96.4949765248477
1 eligible_cache_hit_rate_pct_p50=96.49497652484771 eligible_cache_hit_rate_pct_p90=96.4949765248
4771 eligible_cache_hit_rate_pct_p95=96.49497652484771 eligible_cache_hit_rate_pct_p99=96.4949765
2484771 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=11899.652804345373 last_30s_win
dow_completion_tok_per_s=37.45704214306292 last_30s_window_new_prompt_tok_per_s=2761.632803123743
 last_30s_window_prompt_tok_per_s=14661.285607469115 latency_s_max=65.46460884500266 latency_s_me
an=65.46460884500266 latency_s_min=65.46460884500266 latency_s_p50=65.46460884500266 latency_s_p9
0=65.46460884500266 latency_s_p95=65.46460884500266 latency_s_p99=65.46460884500266 new_prompt_to
k_per_s=2718.6304737204473 prompt_tok_per_s=9942.311903511321 steady_state_cached_prompt_tok_per_
s=9985.919179320446 steady_state_completion_tok_per_s=37.424456367367334 steady_state_new_prompt_
tok_per_s=2528.2299412621487 steady_state_prompt_tok_per_s=12514.149120582595 trace_per_s=0.01525
8118229842669 wall_time_s=65.53888133099827
```

Per-trace metrics

| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 65.465 | 72.0 | 96.5 |
| min | 65.465 | 72.0 | 96.5 |
| p50 | 65.465 | 72.0 | 96.5 |
| p90 | 65.465 | 72.0 | 96.5 |
| p95 | 65.465 | 72.0 | 96.5 |
| p99 | 65.465 | 72.0 | 96.5 |
| max | 65.465 | 72.0 | 96.5 |

Workload metrics

completed=1/1  trace/s=0.02

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 9942.31 | 14661.29 | 12514.15 |
| cached prompt tok/s | 7223.68 | 11899.65 | 9985.92 |
| uncached prompt tok/s | 2718.63 | 2761.63 | 2528.23 |
| completion tok/s | 31.74 | 37.46 | 37.42 |

```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e3-fp8-c4-rep1
>>> cleanup
```

| Configuration | Average latency | Completion throughput |
|---|---|---|
| **NVFP4** | **70.312 s** | **29.55 tok/s** |
| **FP8** | **65.465 s** | **31.74 tok/s** |

Analysis: FP8 performs slightly better in this recorded comparison:

- Approximately 7% higher completion throughput.
- Approximately 7% lower average latency.

Inference: The available results shows that NVFP4 did not outperform FP8 which it should have done. FP8 turns out to be more promising configuration in this particular run but we need more runs and repitations for accurate results.

## 5. E4 — Batching / saturation knee (burst.jsonl, c=8, c=16, c=32)

**Run:** EXTRA_ARGS="--max-num-seqs 64 --max-num-batched-tokens 16384" WORKLOAD=workloads/burst.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e4-c8-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--max-num-seqs 64 --max-num-batched-tokens 16384" WORKLOAD=workloads/burst.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e4-c8-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e4-c8-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=8, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_reference/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster downloads.
[info     ] starting benchmark            concurrency=8 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=69.9044028848433 cache_hit_rate_pct_mean=69.9044028848433 cache_hit_rate_pct_min=69.9044028848433 cache_hit_rate_pct_p50=69.9044028848433 cache_hit_rate_pct_p90=69.9044028848433 cache_hit_rate_pct_p95=69.9044028848433 cache_hit_rate_pct_p99=69.9044028848433 cached_prompt_tok_per_s=8181.941206563339 completed_requests=1 completion_tok_per_s=37.15759655463655 eligible_cache_hit_rate_pct_max=97.53163211476533 eligible_cache_hit_rate_pct_mean=97.53163211476533 eligible_cache_hit_rate_pct_min=97.53163211476533 eligible_cache_hit_rate_pct_p50=97.53163211476533 eligible_cache_hit_rate_pct_p90=97.53163211476533 eligible_cache_hit_rate_pct_p95=97.53163211476533 eligible_cache_hit_rate_pct_p99=97.53163211476533 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=16815.70215767402 last_30s_window_completion_tok_per_s=49.597318030336666 last_30s_window_new_prompt_tok_per_s=3396.6265188673874 last_30s_window_prompt_tok_per_s=20212.328676541405 latency_s_max=98.23397874100192 latency_s_mean=98.23397874100192 latency_s_min=98.23397874100192 latency_s_p50=98.23397874100192 latency_s_p90=98.23397874100192 latency_s_p95=98.23397874100192 latency_s_p99=98.23397874100192 new_prompt_tok_per_s=3556.789764116862 prompt_tok_per_s=11738.7309706802 steady_state_cached_prompt_tok_per_s=11622.270035717544 steady_state_completion_tok_per_s=51.63413482153119 steady_state_new_prompt_tok_per_s=4817.350036327034 steady_state_prompt_tok_per_s=16439.620072044578 trace_per_s=0.01009717297680341 wall_time_s=99.03762194599767
```

Per-trace metrics

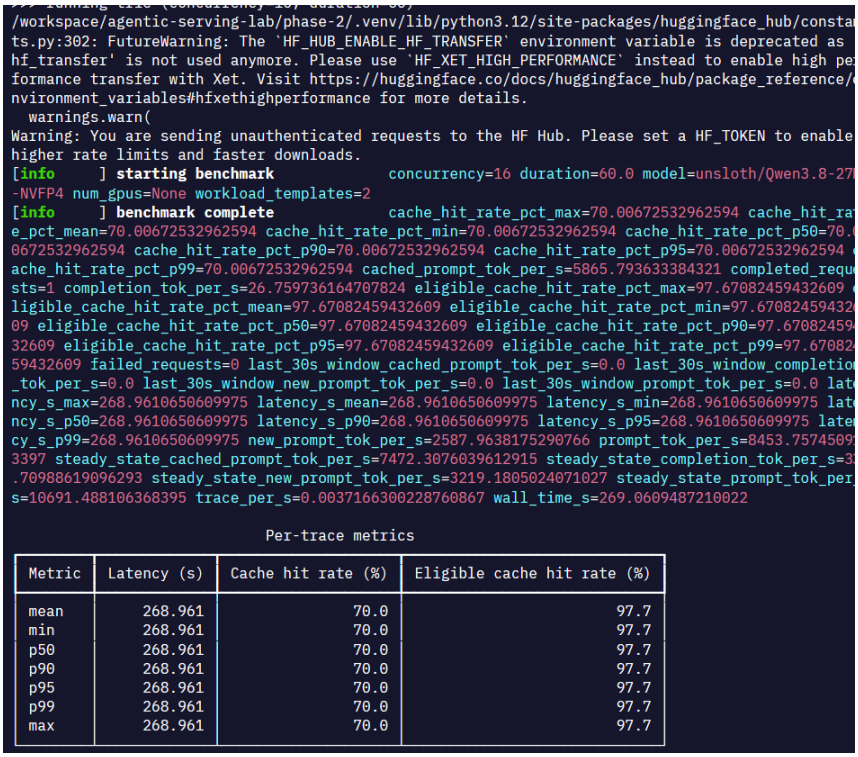
| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 98.234 | 69.9 | 97.5 |
| min | 98.234 | 69.9 | 97.5 |
| p50 | 98.234 | 69.9 | 97.5 |
| p90 | 98.234 | 69.9 | 97.5 |
| p95 | 98.234 | 69.9 | 97.5 |
| p99 | 98.234 | 69.9 | 97.5 |
| max | 98.234 | 69.9 | 97.5 |

Workload metrics
completed=1/1  trace/s=0.01

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 11738.73 | 20212.33 | 16439.62 |
| cached prompt tok/s | 8181.94 | 16815.70 | 11622.27 |
| uncached prompt tok/s | 3556.79 | 3396.63 | 4817.35 |
| completion tok/s | 37.16 | 49.60 | 51.63 |

```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e4-c8-rep1
>>> cleanup
```

**Run:** EXTRA_ARGS="--max-num-seqs 64 --max-num-batched-tokens 16384" WORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e4-c16-rep1 ./run_trie.sh




Per-trace metrics

| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 268.961 | 70.0 | 97.7 |
| min | 268.961 | 70.0 | 97.7 |
| p50 | 268.961 | 70.0 | 97.7 |
| p90 | 268.961 | 70.0 | 97.7 |
| p95 | 268.961 | 70.0 | 97.7 |
| p99 | 268.961 | 70.0 | 97.7 |
| max | 268.961 | 70.0 | 97.7 |

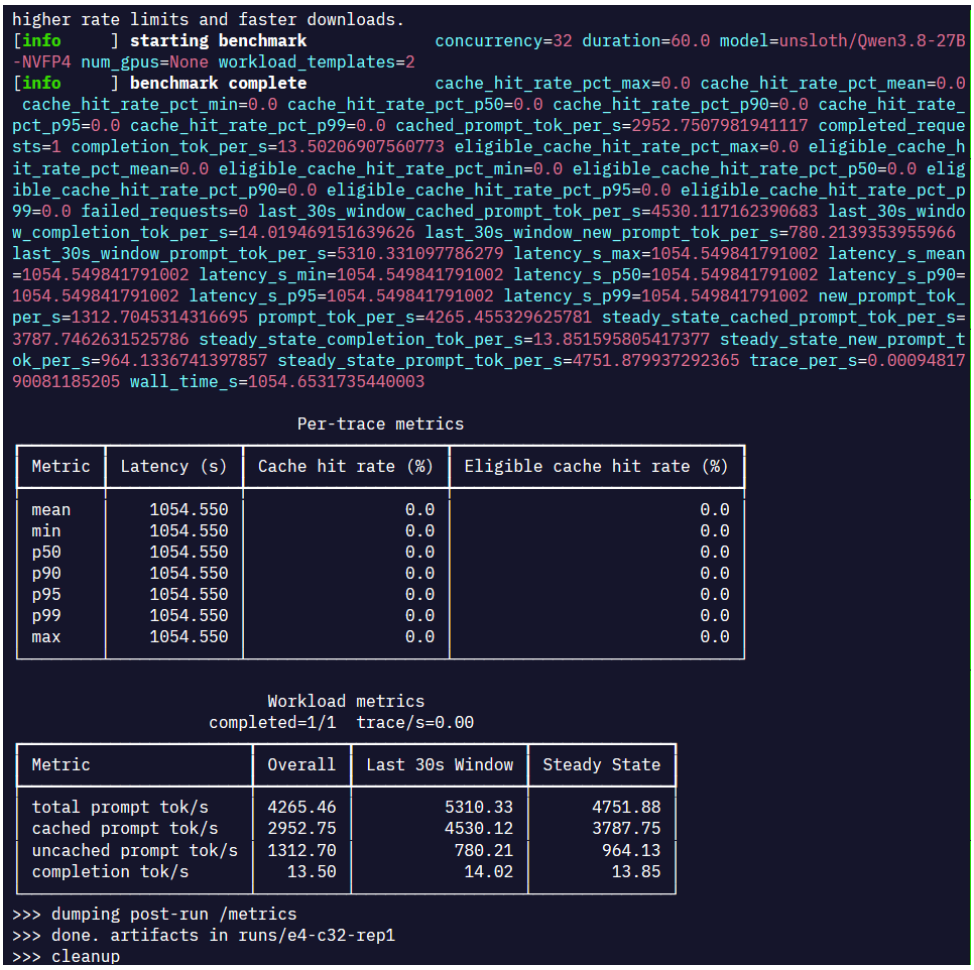
Workload metrics
completed=1/1 trace/s=0.00

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 8453.76 | 0.00 | 10691.49 |
| cached prompt tok/s | 5865.79 | 0.00 | 7472.31 |
| uncached prompt tok/s | 2587.96 | 0.00 | 3219.18 |
| completion tok/s | 26.76 | 0.00 | 33.71 |

```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e4-c16-rep1
>>> cleanup
```

**Run:** EXTRA_ARGS="--max-num-seqs 64 --max-num-batched-tokens 16384" WORKLOAD=workloads/burst.jsonl CONCURRENCY=32 DURATION=60 OUT_DIR=runs/e4-c32-rep1 ./run_trie.sh

**Took longer time to process that previous two ones.**




**The burst workload was tested at C=8, C=16 and C=32.**

| Concurrency | Average latency | Completion throughput |
|---|---|---|
| **C=8** | **98.234 s** | **37.16 tok/s** |
| **C=16** | **268.961 s** | **26.76 tok/s** |
| **C=32** | **1054.550 s** | **13.50 tok/s** |

Analysis: The performance degradation is substantial.
From C=8 to C=32:

- Completion throughput falls from 37.16 to 13.50 tok/s.
- Average latency rises from approximately 98 seconds to more than 1,054 seconds.

This indicates that increasing concurrency beyond the system's effective serving capacity can produce severe latency degradation.
Inference: More concurrent requests do not necessarily mean more useful throughput.
A serving engine must balance: GPU compute capacity, KV cache availability, Scheduler queueing, Batch formation, Prefill and decode interference.
The results suggest that C=32 is far beyond the comfortable operating point for this burst workload.
However, the precise saturation knee cannot be established from one completed trace per configuration.

## 6. E5 — Thinking budget: low vs default/xhigh (light.jsonl, c=2)

**Run:** EXTRA_ARGS="--default-chat-template-kwargs {\"reasoning_effort\":\"low\"}" WORKLOAD=workloads/light.jsonl CONCURRENCY=2 DURATION=60 OUT_DIR=runs/e5-low-c2-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--default-chat-template-kwargs {\"reasoning_effort\":\"low
\"}" WORKLOAD=workloads/light.jsonl CONCURRENCY=2 DURATION=60 OUT_DIR=runs/e5-low-c2-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e5-low-c2-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=2, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The `HF_HUB
_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERFORMANCE` ins
tead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_reference/environment_var
iables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster downloa
ds.
[info     ] starting benchmark            concurrency=2 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_template
s=2
[info     ] benchmark complete            cache_hit_rate_pct_max=69.12287263886292 cache_hit_rate_pct_mean=63.46871188096175 cache_hi
t_rate_pct_min=60.03788642134536 cache_hit_rate_pct_p50=60.139108058961455 cache_hit_rate_pct_p90=69.08799378653285 cache_hit_rate_pct
_p95=69.10543321269789 cache_hit_rate_pct_p99=69.11938475362992 cached_prompt_tok_per_s=5317.310398022181 completed_requests=8 complet
ion_tok_per_s=45.88947083447387 eligible_cache_hit_rate_pct_max=94.23248529438546 eligible_cache_hit_rate_pct_mean=93.20556287209476 e
ligible_cache_hit_rate_pct_min=92.55335673677342 eligible_cache_hit_rate_pct_p50=92.65843869956862 eligible_cache_hit_rate_pct_p90=94.
18806836621474 eligible_cache_hit_rate_pct_p95=94.2102768303001 eligible_cache_hit_rate_pct_p99=94.22804360156839 failed_requests=0 la
st_30s_window_cached_prompt_tok_per_s=5660.303484417582 last_30s_window_completion_tok_per_s=47.26711250905416 last_30s_window_new_pro
mpt_tok_per_s=2221.82133375893 last_30s_window_prompt_tok_per_s=7882.124818176511 latency_s_max=19.02534966500025 latency_s_mean=15.20
3376842124726 latency_s_min=12.377859050997358 latency_s_p50=14.293120816000737 latency_s_p90=18.18804262100166 latency_s_p95=18.60669
6143000953 latency_s_p99=18.94161896060039 new_prompt_tok_per_s=2418.6450510405048 prompt_tok_per_s=7735.955449062687 steady_state_cac
hed_prompt_tok_per_s=5698.156842049672 steady_state_completion_tok_per_s=48.84304103575484 steady_state_new_prompt_tok_per_s=2479.8606
6097644 steady_state_prompt_tok_per_s=8178.0175030261125 trace_per_s=0.1173643755357388 wall_time_s=68.1637844830002

                        Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃ Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │      15.203 │               63.5 │                       93.2 │
│ min    │      12.378 │               60.0 │                       92.6 │
│ p50    │      14.293 │               60.1 │                       92.7 │
│ p90    │      18.188 │               69.1 │                       94.2 │
│ p95    │      18.607 │               69.1 │                       94.2 │
│ p99    │      18.942 │               69.1 │                       94.2 │
│ max    │      19.025 │               69.1 │                       94.2 │
└────────┴─────────────┴────────────────────┴──────────────────────────┘

                    Workload metrics
              completed=8/8  trace/s=0.12
┏━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━┳━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━┓
┃ Metric               ┃ Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━╇━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━┩
│ total prompt tok/s   │ 7735.96 │         7882.12 │      8178.02 │
│ cached prompt tok/s  │ 5317.31 │         5660.30 │      5698.16 │
│ uncached prompt tok/s│ 2418.65 │         2221.82 │      2479.86 │
│ completion tok/s     │   45.89 │           47.27 │        48.84 │
└──────────────────────┴─────────┴────────────────┴─────────────┘
>>> dumping post-run /metrics
>>> done. artifacts in runs/e5-low-c2-rep1
>>> cleanup
```

**Run:** EXTRA_ARGS="--default-chat-template-kwargs {\"reasoning_effort\":\"xhigh\"}" WORKLOAD=workloads/light.jsonl CONCURRENCY=2 DURATION=60 OUT_DIR=runs/e5-xhigh-c2-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--default-chat-template-kwargs {20:14:36 [15/1426]
\":\"xhigh\"}" WORKLOAD=workloads/light.jsonl CONCURRENCY=2 DURATION=60 OUT_DIR=runs/e5-xhigh-c2-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e5-xhigh-c2-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=2, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
 `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_
PERFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_r
eference/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
 downloads.
[info     ] starting benchmark            concurrency=2 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_
templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=69.11733276341099 cache_hit_rate_pct_mean=63.40996674620265
cache_hit_rate_pct_min=59.6782917008874 cache_hit_rate_pct_p50=60.10511132121804 cache_hit_rate_pct_p90=69.08439558421252 cach
e_hit_rate_pct_p95=69.10086417381174 cache_hit_rate_pct_p99=69.11403904549114 cached_prompt_tok_per_s=5303.256614326638 comple
ted_requests=8 completion_tok_per_s=45.76818382116531 eligible_cache_hit_rate_pct_max=94.23248529438546 eligible_cache_hit_rat
e_pct_mean=93.13287093600769 eligible_cache_hit_rate_pct_min=92.06712197533916 eligible_cache_hit_rate_pct_p50=92.632495397293
58 eligible_cache_hit_rate_pct_p90=94.18806836621474 eligible_cache_hit_rate_pct_p95=94.2102768303001 eligible_cache_hit_rate_
pct_p99=94.22804360156839 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=5783.37675709247 last_30s_window_completio
n_tok_per_s=48.802415751244915 last_30s_window_new_prompt_tok_per_s=2361.1073525364204 last_30s_window_prompt_tok_per_s=8144.4
84109628891 latency_s_max=19.101840074999927 latency_s_mean=15.247869223499492 latency_s_min=12.380171356999199 latency_s_p50=
14.352361136498075 latency_s_p90=18.263528794202283 latency_s_p95=18.682684434601104 latency_s_p99=19.018008946920162 new_prom
pt_tok_per_s=2412.2525119861243 prompt_tok_per_s=7715.509126312762 steady_state_cached_prompt_tok_per_s=5684.26459223411 stead
y_state_completion_tok_per_s=48.72396047222715 steady_state_new_prompt_tok_per_s=2473.814698261648 steady_state_prompt_tok_per
_s=8158.079290495758 trace_per_s=0.11705417857075526 wall_time_s=68.3444204869993
```

Per-trace metrics

| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 15.248 | 63.4 | 93.1 |
| min | 12.380 | 59.7 | 92.1 |
| p50 | 14.352 | 60.1 | 92.6 |
| p90 | 18.264 | 69.1 | 94.2 |
| p95 | 18.683 | 69.1 | 94.2 |
| p99 | 19.018 | 69.1 | 94.2 |
| max | 19.102 | 69.1 | 94.2 |

Workload metrics
completed=8/8  trace/s=0.12

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 7715.51 | 8144.48 | 8158.08 |
| cached prompt tok/s | 5303.26 | 5783.38 | 5684.26 |
| uncached prompt tok/s | 2412.25 | 2361.11 | 2473.81 |
| completion tok/s | 45.77 | 48.80 | 48.72 |

```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e5-xhigh-c2-rep1
>>> cleanup
```

| Configuration | Average latency | Completion throughput |
|---|---|---|
| **Low** | **15.203 s** | **45.89 tok/s** |
| **Xhigh** | **15.248 s** | **45.77 tok/s** |

Analysis: no significant difference in latency and throughput.

Inference: No inference

## 7. E6 — Prefix caching on vs off (grow.jsonl, c=4)

**Run:** WORKLOAD=workloads/grow.jsonl CONCURRENCY=4 DURATION=60 OUT_DIR=runs/e6-on-c4-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ WORKLOAD=workloads/grow.jsonl CONCURRENCY=4 DURATION=60 OUT_DIR=runs/e
6-on-c4-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e6-on-c4-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=4, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The `HF_HUB
_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERFORMANCE` ins
tead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_reference/environment_var
iables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster downloa
ds.
[info     ] starting benchmark            concurrency=4 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_template
s=2
[info     ] benchmark complete            cache_hit_rate_pct_max=72.03427450575543 cache_hit_rate_pct_mean=72.03427450575543 cache_hi
t_rate_pct_min=72.03427450575543 cache_hit_rate_pct_p50=72.03427450575543 cache_hit_rate_pct_p90=72.03427450575543 cache_hit_rate_pct_
p95=72.03427450575543 cache_hit_rate_pct_p99=72.03427450575543 cached_prompt_tok_per_s=7220.134523883466 completed_requests=1 completi
on_tok_per_s=31.721302762968303 eligible_cache_hit_rate_pct_max=96.50765549577461 eligible_cache_hit_rate_pct_mean=96.50765549577461 e
ligible_cache_hit_rate_pct_min=96.50765549577461 eligible_cache_hit_rate_pct_p50=96.50765549577461 eligible_cache_hit_rate_pct_p90=96.
50765549577461 eligible_cache_hit_rate_pct_p95=96.50765549577461 eligible_cache_hit_rate_pct_p99=96.50765549577461 failed_requests=0 l
ast_30s_window_cached_prompt_tok_per_s=11881.154075525199 last_30s_window_completion_tok_per_s=37.39881290928584 last_30s_window_new_p
rompt_tok_per_s=2757.339678175826 last_30s_window_prompt_tok_per_s=14638.493753701025 latency_s_max=65.49724554699787 latency_s_mean=6
5.49724554699787 latency_s_min=65.49724554699787 latency_s_p50=65.49724554699787 latency_s_p90=65.49724554699787 latency_s_p95=65.4972
4554699787 latency_s_p99=65.49724554699787 new_prompt_tok_per_s=2717.2955966801155 prompt_tok_per_s=9937.430120563582 steady_state_cac
hed_prompt_tok_per_s=9980.443311158544 steady_state_completion_tok_per_s=37.403934331747124 steady_state_new_prompt_tok_per_s=2526.843
5637446946 steady_state_prompt_tok_per_s=12507.286874903239 trace_per_s=0.015250626328350145 wall_time_s=65.57107744099994

                        Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃ Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │      65.497 │               72.0 │                       96.5 │
│ min    │      65.497 │               72.0 │                       96.5 │
│ p50    │      65.497 │               72.0 │                       96.5 │
│ p90    │      65.497 │               72.0 │                       96.5 │
│ p95    │      65.497 │               72.0 │                       96.5 │
│ p99    │      65.497 │               72.0 │                       96.5 │
│ max    │      65.497 │               72.0 │                       96.5 │
└────────┴─────────────┴────────────────────┴────────────────────────┘

                    Workload metrics
                completed=1/1  trace/s=0.02
┏━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━┳━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━┓
┃ Metric             ┃ Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━╇━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━┩
│ total prompt tok/s │ 9937.43 │        14638.49 │     12507.29 │
│ cached prompt tok/s│ 7220.13 │        11881.15 │      9980.44 │
│ uncached prompt tok/s│2717.30│         2757.34 │      2526.84 │
│ completion tok/s   │   31.72 │           37.40 │        37.40 │
└────────────────────┴─────────┴─────────────────┴──────────────┘
>>> dumping post-run /metrics
>>> done. artifacts in runs/e6-on-c4-rep1
>>> cleanup
```

**Run:** EXTRA_ARGS="--no-enable-prefix-caching" WORKLOAD=workloads/grow.jsonl CONCURRENCY=4 DURATION=60 OUT_DIR=runs/e6-off-c4-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--no-enable-prefix-caching" WORKLOAD=workloads/gro
w.jsonl CONCURRENCY=4 DURATION=60 OUT_DIR=runs/e6-off-c4-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e6-off-c4-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=4, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
 `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_
PERFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_r
eference/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
 downloads.
[info     ] starting benchmark            concurrency=4 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_
templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=0.0 cache_hit_rate_pct_mean=0.0 cache_hit_rate_pct_min=0.0 c
ache_hit_rate_pct_p50=0.0 cache_hit_rate_pct_p90=0.0 cache_hit_rate_pct_p95=0.0 cache_hit_rate_pct_p99=0.0 cached_prompt_tok_p
er_s=2938.970595566495 completed_requests=1 completion_tok_per_s=12.91222147801228 eligible_cache_hit_rate_pct_max=0.0 eligibl
e_cache_hit_rate_pct_mean=0.0 eligible_cache_hit_rate_pct_min=0.0 eligible_cache_hit_rate_pct_p50=0.0 eligible_cache_hit_rate_
pct_p90=0.0 eligible_cache_hit_rate_pct_p95=0.0 eligible_cache_hit_rate_pct_p99=0.0 failed_requests=0 last_30s_window_cached_p
rompt_tok_per_s=42828.18390746458 last_30s_window_completion_tok_per_s=161.1555786205827 last_30s_window_new_prompt_tok_per_s=
7805.1566726509245 last_30s_window_prompt_tok_per_s=50633.34058011551 latency_s_max=161.01433009599714 latency_s_mean=161.0143
3009599714 latency_s_min=161.01433009599714 latency_s_p50=161.01433009599714 latency_s_p90=161.01433009599714 latency_s_p95=16
1.01433009599714 latency_s_p99=161.01433009599714 new_prompt_tok_per_s=1106.0807567626518 prompt_tok_per_s=4045.0513523291465
steady_state_cached_prompt_tok_per_s=3610.473440384648 steady_state_completion_tok_per_s=12.531022718256404 steady_state_new_p
rompt_tok_per_s=881.7150020237489 steady_state_prompt_tok_per_s=4492.188442408396 trace_per_s=0.006207798787505903 wall_time_s
=161.08769536999898

                        Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃ Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │     161.014 │                0.0 │                        0.0 │
│ min    │     161.014 │                0.0 │                        0.0 │
│ p50    │     161.014 │                0.0 │                        0.0 │
│ p90    │     161.014 │                0.0 │                        0.0 │
│ p95    │     161.014 │                0.0 │                        0.0 │
│ p99    │     161.014 │                0.0 │                        0.0 │
│ max    │     161.014 │                0.0 │                        0.0 │
└────────┴─────────────┴────────────────────┴──────────────────────────┘

                    Workload metrics
                completed=1/1  trace/s=0.01
┏━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━┳━━━━━━━━━━━━━━━┳━━━━━━━━━━━━┓
┃ Metric               ┃ Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━╇━━━━━━━━━━━━━━━╇━━━━━━━━━━━━┩
│ total prompt tok/s   │ 4045.05 │        50633.34 │      4492.19 │
│ cached prompt tok/s  │ 2938.97 │        42828.18 │      3610.47 │
│ uncached prompt tok/s│ 1106.08 │         7805.16 │       881.72 │
│ completion tok/s     │   12.91 │          161.16 │        12.53 │
└──────────────────────┴─────────┴─────────────────┴──────────────┘
>>> dumping post-run /metrics
>>> done. artifacts in runs/e6-off-c4-rep1
>>> cleanup
```

This experiment uses the growing-context workload at C=4.

| Metric | Prefix cache on | Prefix cache off |
|---|---|---|
| **Cache-hit rate** | **72.0%** | **0.0%** |
| **Eligible cache-hit rate** | **96.5%** | **0.0%** |
| **Average latency** | **65.497 s** | **161.014 s** |
| **Completion throughput** | **31.72 tok/s** | **12.91 tok/s** |

Analysis: The prefix-caching-on run performs considerably better.
With caching enabled:

- The eligible cache-hit rate reaches approximately 96.5%.

- Average latency is substantially lower.
- Completion throughput is more than twice the cache-disabled result.

This is consistent with the purpose of prefix caching: avoid recomputing repeated prompt prefixes.
Inference: Prefix caching is a valuable serving optimization for workloads where requests reuse common prompt context.
For agentic workloads, this can matter because multiple steps may carry forward substantial conversation history and repeated instructions.
Important limitation: Each configuration has only one completed trace. The result is promising, but it should be repeated before claiming a stable performance multiplier.

## 8. E7 — Scheduling policy: priority vs fcfs (burst.jsonl, c=8, c=16)

**Run:** EXTRA_ARGS="--scheduling-policy priority" WORKLOAD=workloads/burst.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e7-prio-c8-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--scheduling-policy priority" WORKLOAD=workloads/burst.jso
nl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e7-prio-c8-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e7-prio-c8-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=8, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The `HF_HUB
_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERFORMANCE` ins
tead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_reference/environment_var
iables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster downloa
ds.
[info     ] starting benchmark            concurrency=8 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_template
s=2
[info     ] benchmark complete            cache_hit_rate_pct_max=73.33668114101162 cache_hit_rate_pct_mean=73.33668114101162 cache_hi
t_rate_pct_min=73.33668114101162 cache_hit_rate_pct_p50=73.33668114101162 cache_hit_rate_pct_p90=73.33668114101162 cache_hit_rate_pct_
p95=73.33668114101162 cache_hit_rate_pct_p99=73.33668114101162 cached_prompt_tok_per_s=8055.840732826855 completed_requests=1 completi
on_tok_per_s=36.17180052551347 eligible_cache_hit_rate_pct_max=98.05096069667788 eligible_cache_hit_rate_pct_mean=98.05096069667788 el
igible_cache_hit_rate_pct_min=98.05096069667788 eligible_cache_hit_rate_pct_p50=98.05096069667788 eligible_cache_hit_rate_pct_p90=98.0
5096069667788 eligible_cache_hit_rate_pct_p95=98.05096069667788 eligible_cache_hit_rate_pct_p99=98.05096069667788 failed_requests=0 la
st_30s_window_cached_prompt_tok_per_s=15350.27665518145 last_30s_window_completion_tok_per_s=43.90220443103712 last_30s_window_new_pro
mpt_tok_per_s=3048.477254087179 last_30s_window_prompt_tok_per_s=18398.753909268627 latency_s_max=100.85011769500125 latency_s_mean=10
0.85011769500125 latency_s_min=100.85011769500125 latency_s_p50=100.85011769500125 latency_s_p90=100.85011769500125 latency_s_p95=100.
85011769500125 latency_s_p99=100.85011769500125 new_prompt_tok_per_s=3482.646338316104 prompt_tok_per_s=11538.48707114296 steady_state
_cached_prompt_tok_per_s=11081.507546120456 steady_state_completion_tok_per_s=48.66627639499117 steady_state_new_prompt_tok_per_s=4567
.210010737109 steady_state_prompt_tok_per_s=15648.717556857564 trace_per_s=0.009915515494932421 wall_time_s=100.85204349799824

                        Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃ Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │     100.850 │               73.3 │                       98.1 │
│ min    │     100.850 │               73.3 │                       98.1 │
│ p50    │     100.850 │               73.3 │                       98.1 │
│ p90    │     100.850 │               73.3 │                       98.1 │
│ p95    │     100.850 │               73.3 │                       98.1 │
│ p99    │     100.850 │               73.3 │                       98.1 │
│ max    │     100.850 │               73.3 │                       98.1 │
└────────┴─────────────┴────────────────────┴──────────────────────────┘

                    Workload metrics
                completed=1/1  trace/s=0.01
┏━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━┳━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━┓
┃ Metric             ┃ Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━╇━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━┩
│ total prompt tok/s │ 11538.49 │        18398.75 │      15648.72 │
│ cached prompt tok/s │  8055.84 │        15350.28 │      11081.51 │
│ uncached prompt tok/s │ 3482.65 │         3048.48 │       4567.21 │
│ completion tok/s   │   36.17 │           43.90 │         48.67 │
└────────────────────┴─────────┴───────────────┴─────────────┘
>>> dumping post-run /metrics
>>> done. artifacts in runs/e7-prio-c8-rep1
>>> cleanup
```

**Run:** EXTRA_ARGS="--scheduling-policy priority" WORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e7-prio-c16-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--scheduling-policy priority" WORKLOAD=workloads/b
urst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e7-prio-c16-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e7-prio-c16-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=16, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
 `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_
PERFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_r
eference/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
 downloads.
[info     ] starting benchmark            concurrency=16 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload
_templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=69.94110971209821 cache_hit_rate_pct_mean=69.94110971209821
cache_hit_rate_pct_min=69.94110971209821 cache_hit_rate_pct_p50=69.94110971209821 cache_hit_rate_pct_p90=69.94110971209821 cac
he_hit_rate_pct_p95=69.94110971209821 cache_hit_rate_pct_p99=69.94110971209821 cached_prompt_tok_per_s=7293.6400777843555 comp
leted_requests=1 completion_tok_per_s=33.607752502583 eligible_cache_hit_rate_pct_max=97.5784979440077 eligible_cache_hit_rate
_pct_mean=97.5784979440077 eligible_cache_hit_rate_pct_min=97.5784979440077 eligible_cache_hit_rate_pct_p50=97.5784979440077 e
ligible_cache_hit_rate_pct_p90=97.5784979440077 eligible_cache_hit_rate_pct_p95=97.5784979440077 eligible_cache_hit_rate_pct_p
99=97.5784979440077 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=4920.559171624658 last_30s_window_completion_tok
_per_s=16.107375524612017 last_30s_window_new_prompt_tok_per_s=874.6793012152949 last_30s_window_prompt_tok_per_s=5795.2384728
39953 latency_s_max=211.0403659279982 latency_s_mean=211.0403659279982 latency_s_min=211.0403659279982 latency_s_p50=211.04036
59279982 latency_s_p90=211.0403659279982 latency_s_p95=211.0403659279982 latency_s_p99=211.0403659279982 new_prompt_tok_per_s=
3268.78018478336 prompt_tok_per_s=10562.420262567715 steady_state_cached_prompt_tok_per_s=9575.551484432139 steady_state_compl
etion_tok_per_s=41.0382909763506 steady_state_new_prompt_tok_per_s=3654.7458358962804 steady_state_prompt_tok_per_s=13230.2973
20328418 trace_per_s=0.004736154524039317 wall_time_s=211.1417596120009

                    Per-trace metrics
```

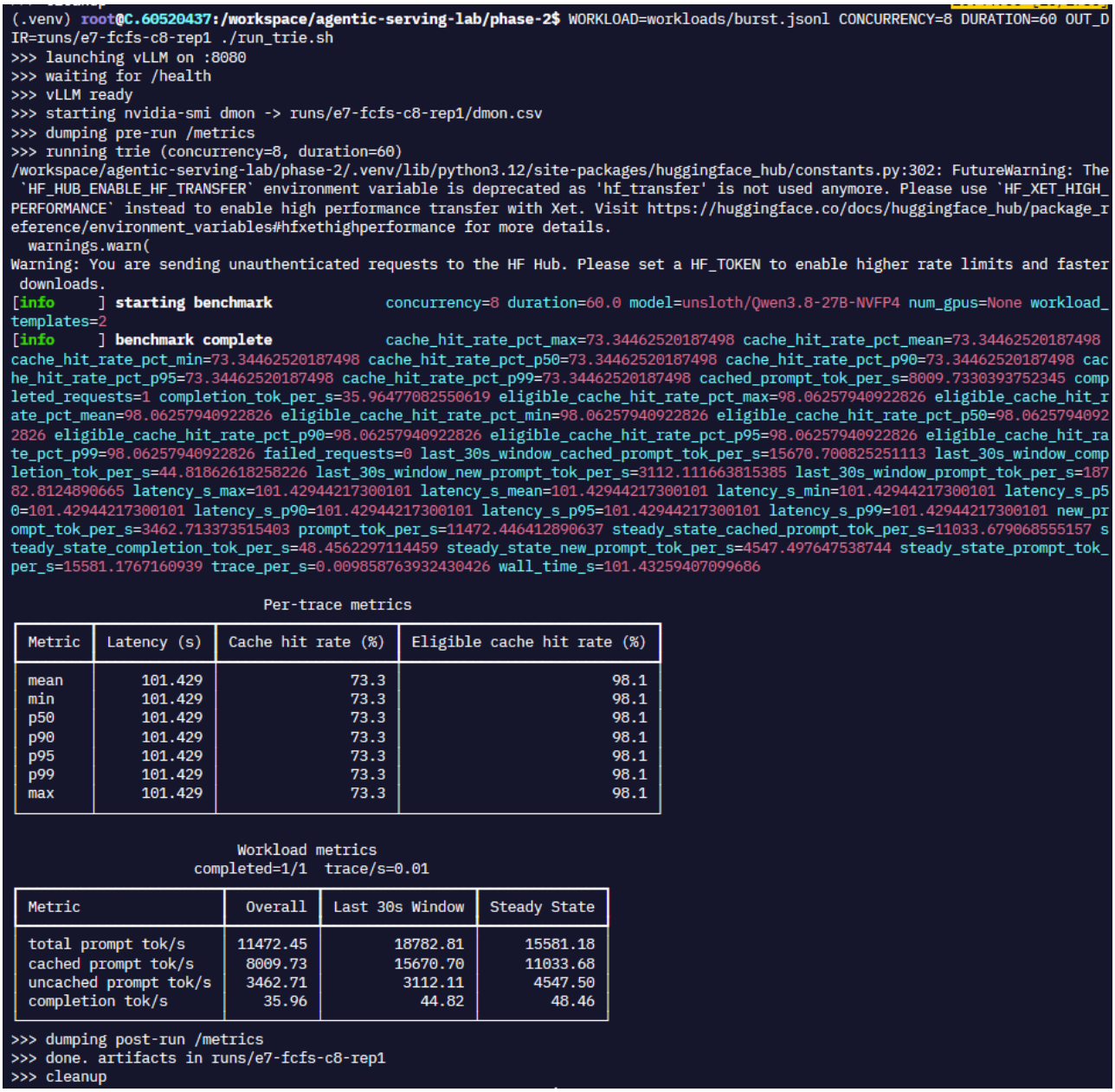
| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 211.040 | 69.9 | 97.6 |
| min | 211.040 | 69.9 | 97.6 |
| p50 | 211.040 | 69.9 | 97.6 |
| p90 | 211.040 | 69.9 | 97.6 |
| p95 | 211.040 | 69.9 | 97.6 |
| p99 | 211.040 | 69.9 | 97.6 |
| max | 211.040 | 69.9 | 97.6 |

```
                  Workload metrics
              completed=1/1  trace/s=0.00
```

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 10562.42 | 5795.24 | 13230.30 |
| cached prompt tok/s | 7293.64 | 4920.56 | 9575.55 |
| uncached prompt tok/s | 3268.78 | 874.68 | 3654.75 |
| completion tok/s | 33.61 | 16.11 | 41.04 |

```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e7-prio-c16-rep1
>>> cleanup
```

**Run:** WORKLOAD=workloads/burst.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e7-fcfs-c8-rep1 ./run_trie.sh



```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ WORKLOAD=workloads/burst.jsonl CONCURRENCY=8 DURATION=60 OUT_D
IR=runs/e7-fcfs-c8-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e7-fcfs-c8-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=8, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
 `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_
PERFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_r
eference/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
 downloads.
[info     ] starting benchmark            concurrency=8 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_
templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=73.34462520187498 cache_hit_rate_pct_mean=73.34462520187498
cache_hit_rate_pct_min=73.34462520187498 cache_hit_rate_pct_p50=73.34462520187498 cache_hit_rate_pct_p90=73.34462520187498 cac
he_hit_rate_pct_p95=73.34462520187498 cache_hit_rate_pct_p99=73.34462520187498 cached_prompt_tok_per_s=8009.7330393752345 comp
leted_requests=1 completion_tok_per_s=35.96477082550619 eligible_cache_hit_rate_pct_max=98.06257940922826 eligible_cache_hit_r
ate_pct_mean=98.06257940922826 eligible_cache_hit_rate_pct_min=98.06257940922826 eligible_cache_hit_rate_pct_p50=98.0625794092
2826 eligible_cache_hit_rate_pct_p90=98.06257940922826 eligible_cache_hit_rate_pct_p95=98.06257940922826 eligible_cache_hit_ra
te_pct_p99=98.06257940922826 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=15670.700825251113 last_30s_window_comp
letion_tok_per_s=44.81862618258226 last_30s_window_new_prompt_tok_per_s=3112.111663815385 last_30s_window_prompt_tok_per_s=187
82.8124890665 latency_s_max=101.42944217300101 latency_s_mean=101.42944217300101 latency_s_min=101.42944217300101 latency_s_p5
0=101.42944217300101 latency_s_p90=101.42944217300101 latency_s_p95=101.42944217300101 latency_s_p99=101.42944217300101 new_pr
ompt_tok_per_s=3462.713373515403 prompt_tok_per_s=11472.446412890637 steady_state_cached_prompt_tok_per_s=11033.679068555157 s
teady_state_completion_tok_per_s=48.4562297114459 steady_state_new_prompt_tok_per_s=4547.497647538744 steady_state_prompt_tok_
per_s=15581.1767160939 trace_per_s=0.009858763932430426 wall_time_s=101.43259407099686

                        Per-trace metrics
```

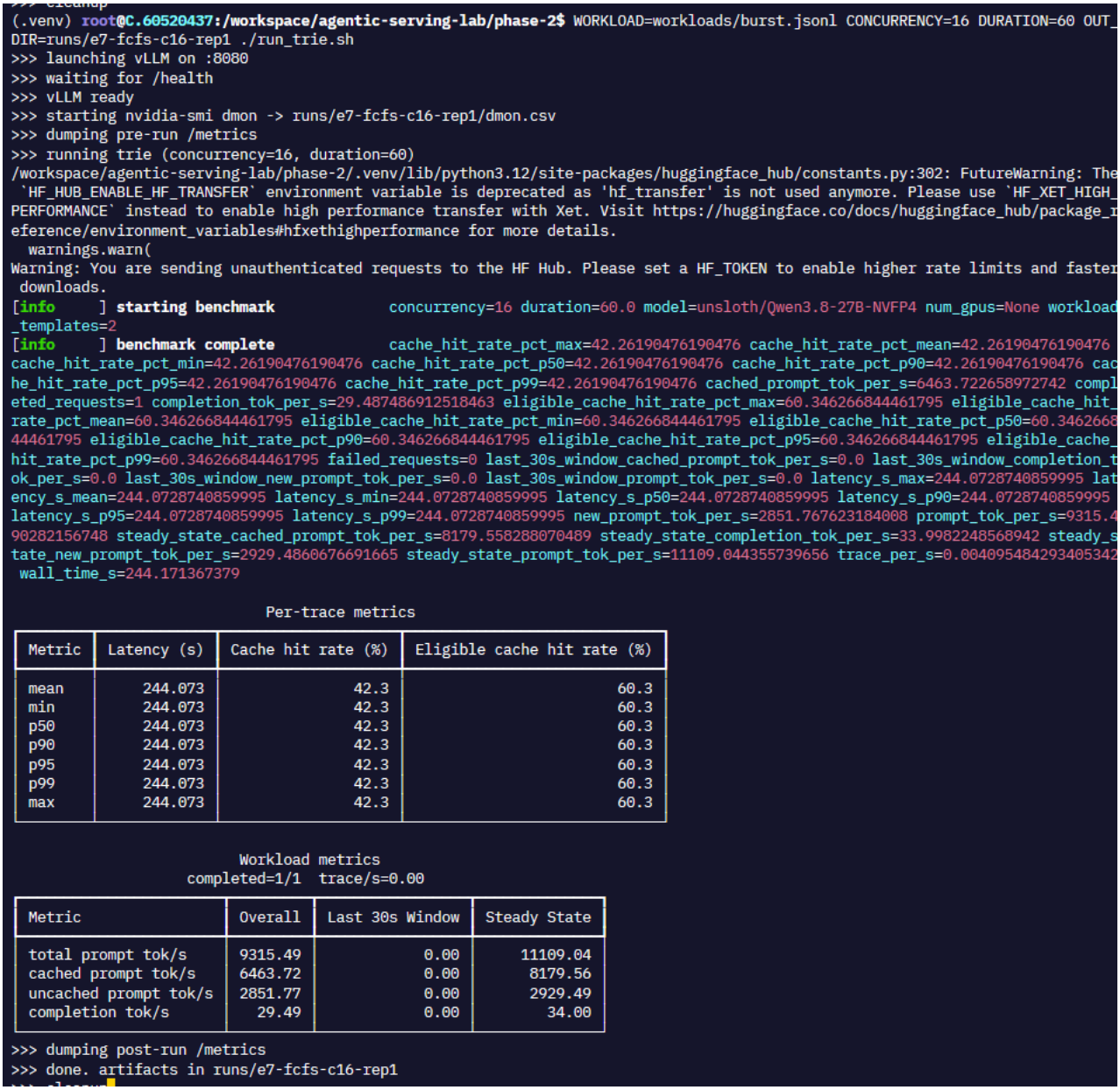
| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 101.429 | 73.3 | 98.1 |
| min | 101.429 | 73.3 | 98.1 |
| p50 | 101.429 | 73.3 | 98.1 |
| p90 | 101.429 | 73.3 | 98.1 |
| p95 | 101.429 | 73.3 | 98.1 |
| p99 | 101.429 | 73.3 | 98.1 |
| max | 101.429 | 73.3 | 98.1 |

```
                    Workload metrics
              completed=1/1  trace/s=0.01
```

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 11472.45 | 18782.81 | 15581.18 |
| cached prompt tok/s | 8009.73 | 15670.70 | 11033.68 |
| uncached prompt tok/s | 3462.71 | 3112.11 | 4547.50 |
| completion tok/s | 35.96 | 44.82 | 48.46 |

```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e7-fcfs-c8-rep1
>>> cleanup
```

**Run:** WORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e7-fcfs-c16-rep1 ./run_trie.sh




| Concurrency | Policy | Average latency | Completion throughput |
|---|---|---|---|
| C=8 | Priority | 100.850 s | 36.17 tok/s |
| C=8 | FCFS | 101.429 s | 35.96 tok/s |
| C=16 | Priority | 211.040 s | 33.61 tok/s |
| C=16 | FCFS | 244.073 s | 29.49 tok/s |

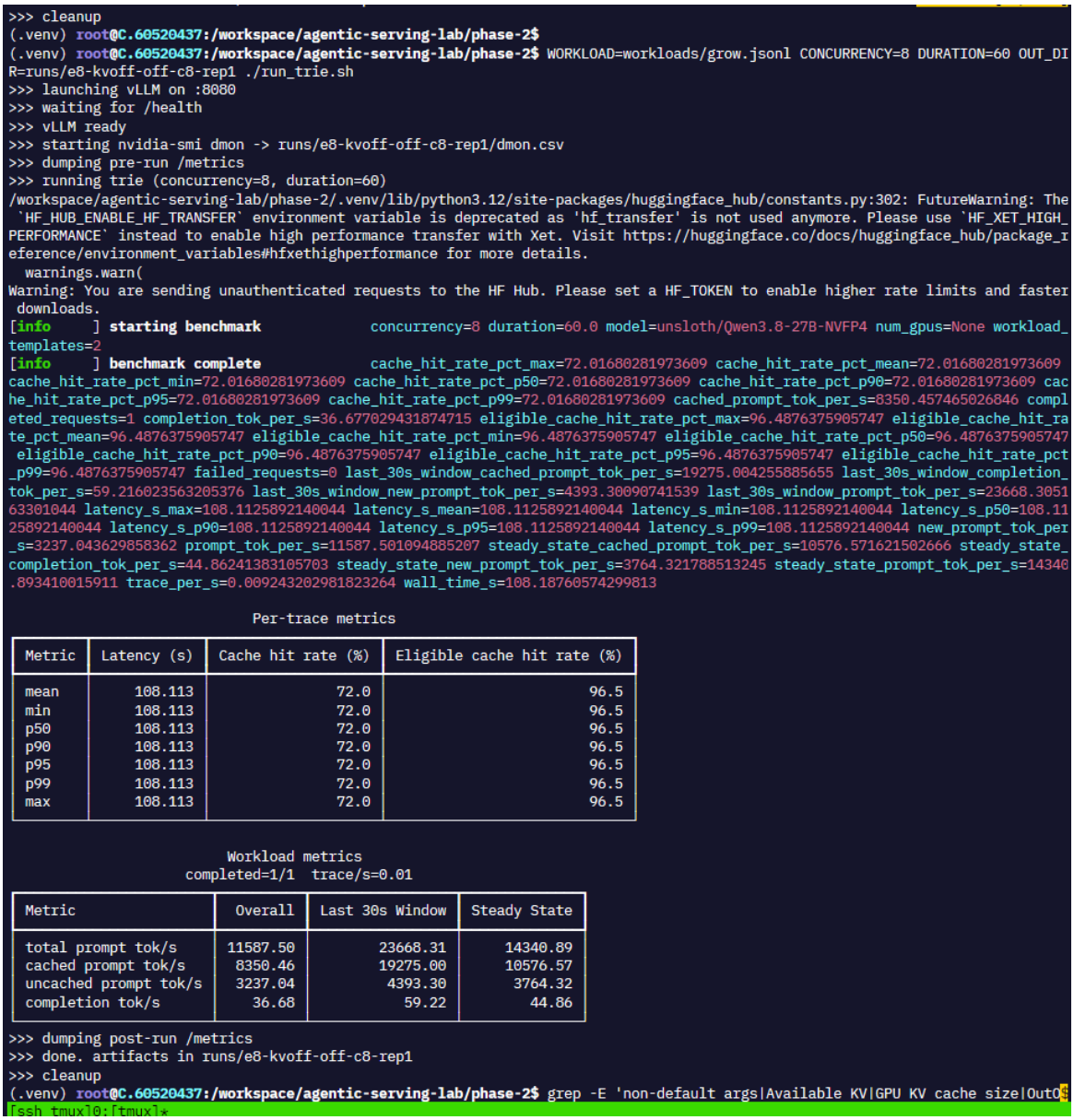
Analysis: At c=8 the difference is negligible and at c=16 priority scheduling appears better, throughput seems to be 14% higher and latency approximately 13.5% lower.

### 9. E8 — KV offloading to host RAM (grow.jsonl, c=8)

Run: EXTRA_ARGS="--kv-offloading-size 32 --kv-offloading-backend native" WORKLOAD=workloads/grow.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e8-kvoff-c8-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--kv-offloading-size 32 --kv-off21:00:19 [70/1939$
tive" WORKLOAD=workloads/grow.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e8-kvoff-c8-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e8-kvoff-c8-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=8, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
 `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_
PERFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_r
eference/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
 downloads.
[info     ] starting benchmark            concurrency=8 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_
templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=72.02089117801884 cache_hit_rate_pct_mean=72.02089117801884
cache_hit_rate_pct_min=72.02089117801884 cache_hit_rate_pct_p50=72.02089117801884 cache_hit_rate_pct_p90=72.02089117801884 cac
he_hit_rate_pct_p95=72.02089117801884 cache_hit_rate_pct_p99=72.02089117801884 cached_prompt_tok_per_s=8341.985837414846 compl
eted_requests=1 completion_tok_per_s=36.63982019674447 eligible_cache_hit_rate_pct_max=96.49497652484771 eligible_cache_hit_ra
te_pct_mean=96.49497652484771 eligible_cache_hit_rate_pct_min=96.49497652484771 eligible_cache_hit_rate_pct_p50=96.49497652484
771 eligible_cache_hit_rate_pct_p90=96.49497652484771 eligible_cache_hit_rate_pct_p95=96.49497652484771 eligible_cache_hit_rat
e_pct_p99=96.49497652484771 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=19201.07759014514 last_30s_window_comple
tion_tok_per_s=58.988908532654726 last_30s_window_new_prompt_tok_per_s=4376.450997378104 last_30s_window_prompt_tok_per_s=2357
7.528587523244 latency_s_max=108.22258880300069 latency_s_mean=108.22258880300069 latency_s_min=108.22258880300069 latency_s_p
50=108.22258880300069 latency_s_p90=108.22258880300069 latency_s_p95=108.22258880300069 latency_s_p99=108.22258880300069 new_p
rompt_tok_per_s=3233.7596147836416 prompt_tok_per_s=11575.745452198487 steady_state_cached_prompt_tok_per_s=10559.832364146234
 steady_state_completion_tok_per_s=44.79141128716822 steady_state_new_prompt_tok_per_s=3758.3640969809294 steady_state_prompt_
tok_per_s=14318.196461127163 trace_per_s=0.00923382565442149 wall_time_s=108.2974746789987

                        Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃ Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │     108.223 │               72.0 │                       96.5 │
│ min    │     108.223 │               72.0 │                       96.5 │
│ p50    │     108.223 │               72.0 │                       96.5 │
│ p90    │     108.223 │               72.0 │                       96.5 │
│ p95    │     108.223 │               72.0 │                       96.5 │
│ p99    │     108.223 │               72.0 │                       96.5 │
│ max    │     108.223 │               72.0 │                       96.5 │
└────────┴─────────────┴──────────────────┴────────────────────────┘

                    Workload metrics
                completed=1/1  trace/s=0.01
┏━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━┳━━━━━━━━━━━━━━━┳━━━━━━━━━━━━┓
┃ Metric             ┃  Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━╇━━━━━━━━━━━━━━━╇━━━━━━━━━━━━┩
│ total prompt tok/s │ 11575.75 │        23577.53 │     14318.20 │
│ cached prompt tok/s │  8341.99 │        19201.08 │     10559.83 │
│ uncached prompt tok/s │  3233.76 │         4376.45 │      3758.36 │
│ completion tok/s   │    36.64 │           58.99 │        44.79 │
└────────────────────┴──────────┴───────────────┴────────────┘
>>> dumping post-run /metrics
>>> done. artifacts in runs/e8-kvoff-c8-rep1
>>> cleanup
```

Run: WORKLOAD=workloads/grow.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e8-kvoff-off-c8-rep1 ./run_trie.sh

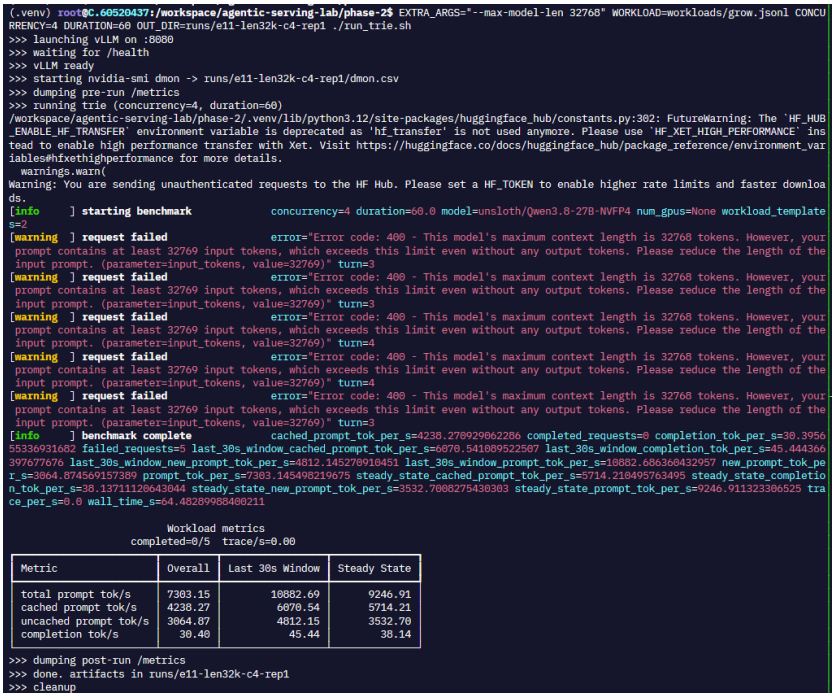



| Configuration | Completion throughput |
|---|---|
| **KV offload enabled** | **36.64 tok/s** |
| **KV offload disabled** | **36.68 tok/s** |

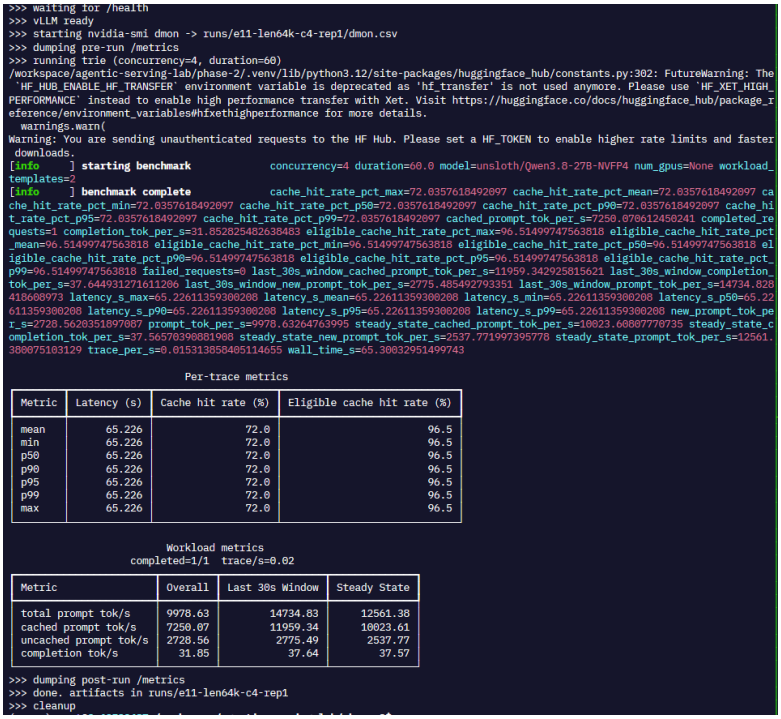
Analysis: The throughput is effectively identical.
There is no clear evidence that host-RAM KV offloading improved the serving rate in this experiment.
Inference: KV offloading should be evaluated as a capacity optimization, not automatically as a speed optimization.
It can help when GPU KV cache capacity is the limiting factor, but moving KV data between host RAM and GPU memory can introduce transfer overhead.

## 10. E11 — Long-context pressure (max-model-len sweep)

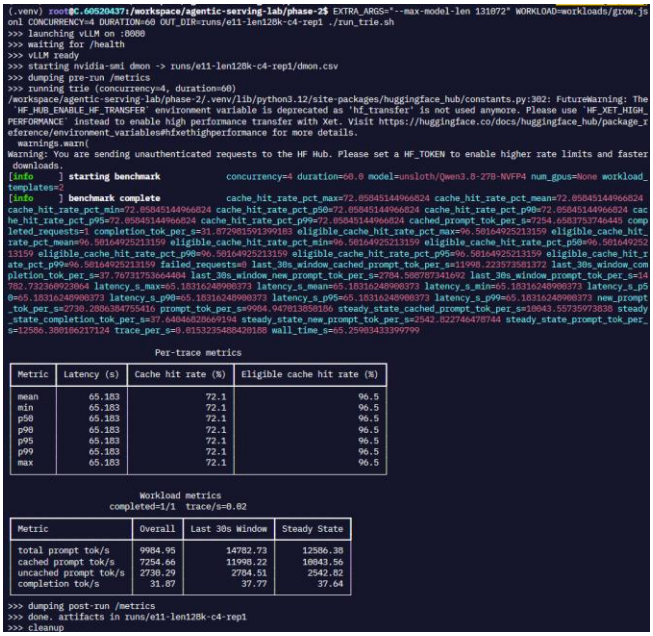
Run: EXTRA_ARGS="--max-model-len 32768" WORKLOAD=workloads/grow.jsonl CONCURRENCY=4 DURATION=60 OUT_DIR=runs/e11-len32k-c4-rep1 ./run_trie.sh



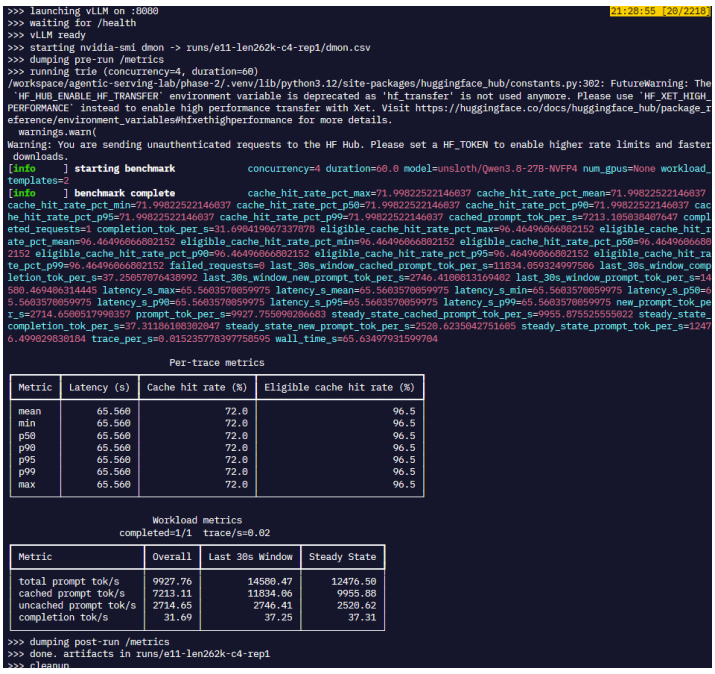

Run: EXTRA_ARGS="--max-model-len 65536" WORKLOAD=workloads/grow.jsonl CONCURRENCY=4 DURATION=60 OUT_DIR=runs/e11-len64k-c4-rep1 ./run_trie.sh



**Run: EXTRA_ARGS="--max-model-len 131072" WORKLOAD=workloads/grow.jsonl CONCURRENCY=4 DURATION=60 OUT_DIR=runs/e11-len128k-c4-rep1 ./run_trie.sh**




**Run: WORKLOAD=workloads/grow.jsonl CONCURRENCY=4 DURATION=60 OUT_DIR=runs/e11-len262k-c4-rep1 ./run_trie.sh**




| Maximum model length | Completed traces | Completion throughput |
|---|---|---|
| 32K | 0/5 | No valid result |
| 64K | 1/1 | 31.85 tok/s |
| 128K | 1/1 | 31.87 tok/s |
| 262K | 1/1 | 31.69 tok/s |

## 11. E12 — GPU memory utilization sweep

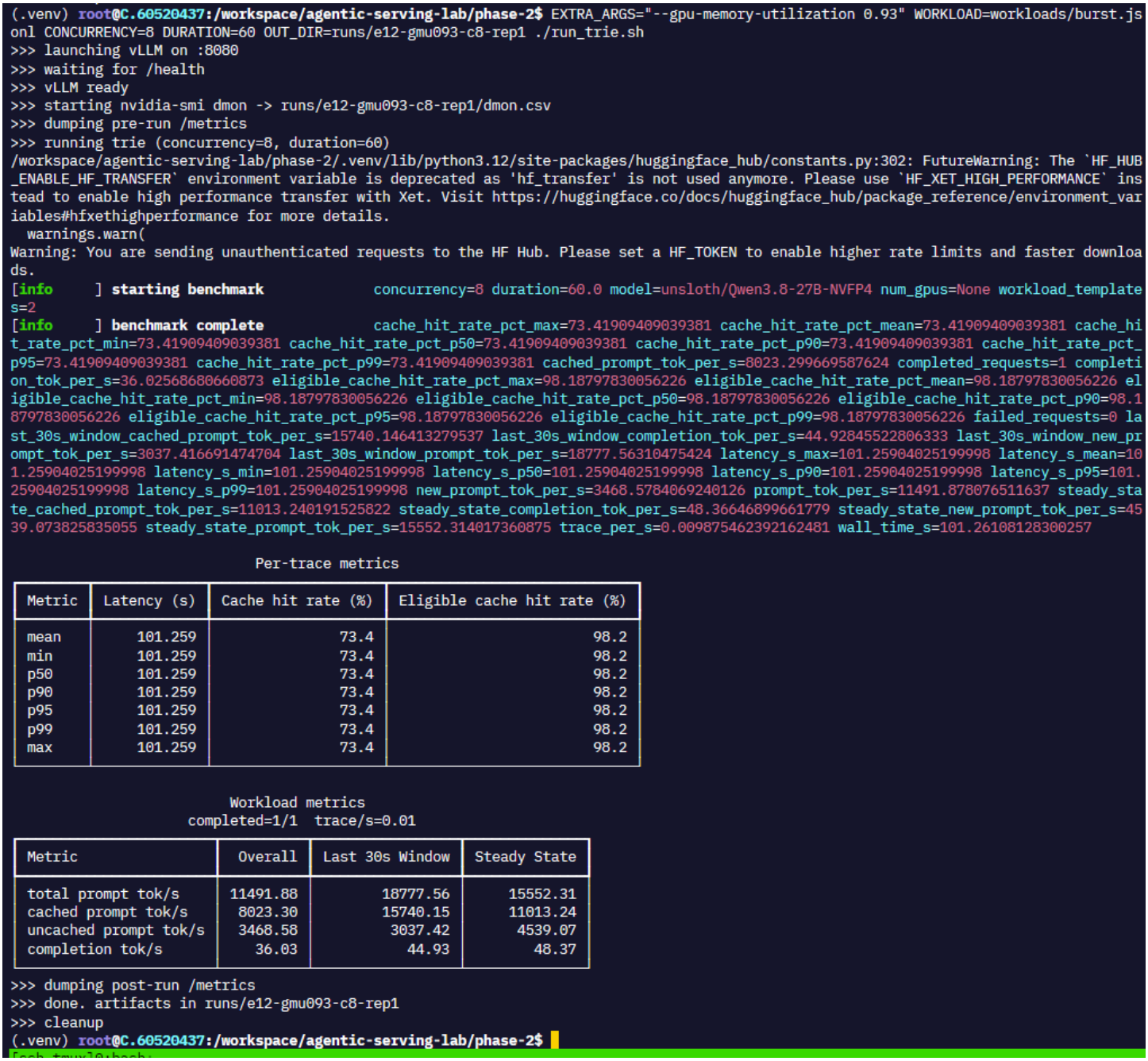
Run: EXTRA_ARGS="--gpu-memory-utilization 0.85" WORKLOAD=workloads/burst.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e12-gmu085-c8-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--gpu-memory-utilization 0.85" WORKLOAD=workloads/
burst.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e12-gmu085-c8-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e12-gmu085-c8-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=8, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
 `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_
PERFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_r
eference/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
 downloads.
[info     ] starting benchmark            concurrency=8 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_
templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=69.94268936506593 cache_hit_rate_pct_mean=69.94268936506593
cache_hit_rate_pct_min=69.94268936506593 cache_hit_rate_pct_p50=69.94268936506593 cache_hit_rate_pct_p90=69.94268936506593 cac
he_hit_rate_pct_p95=69.94268936506593 cache_hit_rate_pct_p99=69.94268936506593 cached_prompt_tok_per_s=8006.9948896290525 comp
leted_requests=1 completion_tok_per_s=36.36309259778225 eligible_cache_hit_rate_pct_max=97.58618516417722 eligible_cache_hit_r
ate_pct_mean=97.58618516417722 eligible_cache_hit_rate_pct_min=97.58618516417722 eligible_cache_hit_rate_pct_p50=97.5861851641
7722 eligible_cache_hit_rate_pct_p90=97.58618516417722 eligible_cache_hit_rate_pct_p95=97.58618516417722 eligible_cache_hit_ra
te_pct_p99=97.58618516417722 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=15683.785260670942 last_30s_window_comp
letion_tok_per_s=46.243741861643414 last_30s_window_new_prompt_tok_per_s=3241.4660962066237 last_30s_window_prompt_tok_per_s=1
8925.251356877565 latency_s_max=101.10341795099521 latency_s_mean=101.10341795099521 latency_s_min=101.10341795099521 latency_
s_p50=101.10341795099521 latency_s_p90=101.10341795099521 latency_s_p95=101.10341795099521 latency_s_p99=101.10341795099521 ne
w_prompt_tok_per_s=3480.7384636207566 prompt_tok_per_s=11487.733353249809 steady_state_cached_prompt_tok_per_s=10982.643701956
482 steady_state_completion_tok_per_s=48.792473747461905 steady_state_new_prompt_tok_per_s=4552.229372918757 steady_state_prom
pt_tok_per_s=15534.873074875239 trace_per_s=0.009881275162440828 wall_time_s=101.20151332300156

                        Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃ Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │     101.103 │               69.9 │                       97.6 │
│ min    │     101.103 │               69.9 │                       97.6 │
│ p50    │     101.103 │               69.9 │                       97.6 │
│ p90    │     101.103 │               69.9 │                       97.6 │
│ p95    │     101.103 │               69.9 │                       97.6 │
│ p99    │     101.103 │               69.9 │                       97.6 │
│ max    │     101.103 │               69.9 │                       97.6 │
└────────┴─────────────┴────────────────────┴──────────────────────────┘

                    Workload metrics
                completed=1/1  trace/s=0.01
┏━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━┳━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━┓
┃ Metric              ┃ Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━╇━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━┩
│ total prompt tok/s  │ 11487.73 │        18925.25 │     15534.87 │
│ cached prompt tok/s │  8006.99 │        15683.79 │     10982.64 │
│ uncached prompt tok/s │ 3480.74 │         3241.47 │      4552.23 │
│ completion tok/s    │   36.36 │           46.24 │        48.79 │
└─────────────────────┴─────────┴────────────────┴─────────────┘
>>> dumping post-run /metrics
>>> done. artifacts in runs/e12-gmu085-c8-rep1
>>> cleanup
```

Run: EXTRA_ARGS="--gpu-memory-utilization 0.90" WORKLOAD=workloads/burst.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e12-gmu090-c8-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--gpu-memory-utilization 0.90" WORKLOAD=workloads/
burst.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e12-gmu090-c8-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e12-gmu090-c8-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=8, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
 `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_
PERFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_r
eference/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
 downloads.
[info     ] starting benchmark            concurrency=8 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_
templates=2
^[[B^[[B^[[B^[[B[info     ] benchmark complete            cache_hit_rate_pct_max=73.36521554972197 cache_hit_rate_pct_mean=73
.36521554972197 cache_hit_rate_pct_min=73.36521554972197 cache_hit_rate_pct_p50=73.36521554972197 cache_hit_rate_pct_p90=73.36
521554972197 cache_hit_rate_pct_p95=73.36521554972197 cache_hit_rate_pct_p99=73.36521554972197 cached_prompt_tok_per_s=8041.43
5174588642 completed_requests=1 completion_tok_per_s=36.107117645559306 eligible_cache_hit_rate_pct_max=98.0897004497534 eligi
ble_cache_hit_rate_pct_mean=98.0897004497534 eligible_cache_hit_rate_pct_min=98.0897004497534 eligible_cache_hit_rate_pct_p50=
98.0897004497534 eligible_cache_hit_rate_pct_p90=98.0897004497534 eligible_cache_hit_rate_pct_p95=98.0897004497534 eligible_ca
che_hit_rate_pct_p99=98.0897004497534 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=15733.251842248454 last_30s_wi
ndow_completion_tok_per_s=44.90877546030555 last_30s_window_new_prompt_tok_per_s=3036.0862283023466 last_30s_window_prompt_tok
_per_s=18769.3380705508 latency_s_max=101.03048773600312 latency_s_mean=101.03048773600312 latency_s_min=101.03048773600312 la
tency_s_p50=101.03048773600312 latency_s_p90=101.03048773600312 latency_s_p95=101.03048773600312 latency_s_p99=101.03048773600
312 new_prompt_tok_per_s=3476.4186252426225 prompt_tok_per_s=11517.853799831266 steady_state_cached_prompt_tok_per_s=11038.867
245065558 steady_state_completion_tok_per_s=48.47901444817873 steady_state_new_prompt_tok_per_s=4549.63593888558 steady_state_
prompt_tok_per_s=15588.503183951138 trace_per_s=0.009897784442313406 wall_time_s=101.03271149500506

                        Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃  Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │      101.030 │               73.4 │                       98.1 │
│ min    │      101.030 │               73.4 │                       98.1 │
│ p50    │      101.030 │               73.4 │                       98.1 │
│ p90    │      101.030 │               73.4 │                       98.1 │
│ p95    │      101.030 │               73.4 │                       98.1 │
│ p99    │      101.030 │               73.4 │                       98.1 │
│ max    │      101.030 │               73.4 │                       98.1 │
└────────┴──────────────┴───────────────────┴────────────────────────┘

                    Workload metrics
                completed=1/1  trace/s=0.01
┏━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━┳━━━━━━━━━━━━━━━┳━━━━━━━━━━━━┓
┃ Metric            ┃ Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━╇━━━━━━━━━━━━━━━╇━━━━━━━━━━━━┩
│ total prompt tok/s │ 11517.85 │        18769.34 │     15588.50 │
│ cached prompt tok/s │  8041.44 │        15733.25 │     11038.87 │
│ uncached prompt tok/s │ 3476.42 │         3036.09 │      4549.64 │
│ completion tok/s  │   36.11 │           44.91 │        48.48 │
└───────────────────┴─────────┴─────────────────┴──────────────┘
>>> dumping post-run /metrics
>>> done. artifacts in runs/e12-gmu090-c8-rep1
>>> cleanup
[ssh_tmux]0:[tmux]*
```

Run: EXTRA_ARGS="--gpu-memory-utilization 0.93" WORKLOAD=workloads/burst.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e12-gmu093-c8-rep1 ./run_trie.sh




| GPU memory utilization | Average latency | Completion throughput |
|---|---|---|
| **0.85** | **101.103 s** | **36.36 tok/s** |
| **0.90** | **101.030 s** | **36.20 tok/s** |
| **0.93** | **101.259 s** | **36.03 tok/s** |

Analysis: The throughput remains almost unchanged across all three configurations.
The latency is also practically identical.
Inference: Increasing the GPU memory utilization target did not produce an observable serving-performance improvement in these runs.

## 12. E13 — Language-model-only: text-only vs vision-enabled

**Run: EXTRA_ARGS="--language-model-only" WORKLOAD=workloads/light.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e13-lmo-c8-rep1 ./run_trie.sh**

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--language-model-only" WORKLOAD=workloads/light.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e13-lmo-c8-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e13-lmo-c8-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=8, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_reference/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster downloads.
[info     ] starting benchmark            concurrency=8 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=69.12287263886292 cache_hit_rate_pct_mean=64.04995805464549 cache_hit_rate_pct_min=59.90750050271466 cache_hit_rate_pct_p50=60.10571762901989 cache_hit_rate_pct_p90=69.10514929874559 cache_hit_rate_pct_p95=69.11401096880425 cache_hit_rate_pct_p99=69.12110030485118 cached_prompt_tok_per_s=7555.291352413875 completed_requests=9 completion_tok_per_s=66.2773542002862 eligible_cache_hit_rate_pct_max=94.22390560128196 eligible_cache_hit_rate_pct_mean=93.26367990957498 eligible_cache_hit_rate_pct_min=92.37543021921802 eligible_cache_hit_rate_pct_p50=92.63393551195549 eligible_cache_hit_rate_pct_p90=94.20194998357832 eligible_cache_hit_rate_pct_p95=94.21292779243014 eligible_cache_hit_rate_pct_p99=94.22171003951159 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=10854.501172849567 last_30s_window_completion_tok_per_s=89.70404711127077 last_30s_window_new_prompt_tok_per_s=4744.347380551654 last_30s_window_prompt_tok_per_s=15598.848553401222 latency_s_max=50.510312445003365 latency_s_mean=42.97461349155654 latency_s_min=35.775023939000675 latency_s_p50=37.61017636700126 latency_s_p90=50.27566259459854 latency_s_p95=50.392987519800954 latency_s_p99=50.48684745996288 new_prompt_tok_per_s=3697.5787080159666 prompt_tok_per_s=11252.870060429843 steady_state_cached_prompt_tok_per_s=9448.557532923855 steady_state_completion_tok_per_s=77.70531667989147 steady_state_new_prompt_tok_per_s=3856.3649091943325 steady_state_prompt_tok_per_s=13304.922442118188 trace_per_s=0.12263490703177957 wall_time_s=73.38856625599874

                    Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃ Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │      42.975 │               64.0 │                        93.3 │
│ min    │      35.775 │               59.9 │                        92.4 │
│ p50    │      37.610 │               60.1 │                        92.6 │
│ p90    │      50.276 │               69.1 │                        94.2 │
│ p95    │      50.393 │               69.1 │                        94.2 │
│ p99    │      50.487 │               69.1 │                        94.2 │
│ max    │      50.510 │               69.1 │                        94.2 │
└────────┴─────────────┴───────────────────┴───────────────────────────┘

                    Workload metrics
                completed=9/9  trace/s=0.12
┏━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━┳━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━┓
┃ Metric             ┃ Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━╇━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━┩
│ total prompt tok/s │ 11252.87 │        15598.85 │     13304.92 │
│ cached prompt tok/s │  7555.29 │        10854.50 │      9448.56 │
│ uncached prompt tok/s │  3697.58 │         4744.35 │      3856.36 │
│ completion tok/s   │   66.28 │           89.70 │        77.71 │
└────────────────────┴─────────┴────────────────┴─────────────┘
>>> dumping post-run /metrics
>>> done. artifacts in runs/e13-lmo-c8-rep1
>>> cleanup
```

**Run: WORKLOAD=workloads/light.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e13-vision-c8-rep1 ./run_trie.sh**

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ WORKLOAD=workloads/light.jsonl CONCURRENCY=8 DURATION=60 OUT_D
IR=runs/e13-vision-c8-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e13-vision-c8-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=8, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
 `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_
PERFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_r
eference/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
 downloads.
[info     ] starting benchmark            concurrency=8 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_
templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=69.26444013118265 cache_hit_rate_pct_mean=64.11239822888258
cache_hit_rate_pct_min=60.064516129032256 cache_hit_rate_pct_p50=60.12148608560531 cache_hit_rate_pct_p90=69.19110523387249 ca
che_hit_rate_pct_p95=69.22777268252757 cache_hit_rate_pct_p99=69.25710664145164 cached_prompt_tok_per_s=7629.066080242677 comp
leted_requests=9 completion_tok_per_s=66.92452894699892 eligible_cache_hit_rate_pct_max=94.4285271091483 eligible_cache_hit_ra
te_pct_mean=93.34256690562708 eligible_cache_hit_rate_pct_min=92.58499595997264 eligible_cache_hit_rate_pct_p50=92.65698379622
43 eligible_cache_hit_rate_pct_p90=94.31977496042578 eligible_cache_hit_rate_pct_p95=94.37415103478703 eligible_cache_hit_rate
_pct_p99=94.41765189427605 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=11031.490708942587 last_30s_window_comple
tion_tok_per_s=91.16672857687351 last_30s_window_new_prompt_tok_per_s=4821.706978065755 last_30s_window_prompt_tok_per_s=15853
.197687008342 latency_s_max=51.134113284002524 latency_s_mean=42.87583680233365 latency_s_min=35.21887645300012 latency_s_p50=
37.456206120004936 latency_s_p90=50.57288266320247 latency_s_p95=50.8534979736025 latency_s_p99=51.07799022192252 new_prompt_t
ok_per_s=3733.684246516782 prompt_tok_per_s=11362.75032675946 steady_state_cached_prompt_tok_per_s=9537.305248387547 steady_st
ate_completion_tok_per_s=82.70095448326936 steady_state_new_prompt_tok_per_s=4491.446180256093 steady_state_prompt_tok_per_s=1
4028.75142864364 trace_per_s=0.12383239319962794 wall_time_s=72.67888286299421

                        Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃ Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │      42.876 │               64.1 │                       93.3 │
│ min    │      35.219 │               60.1 │                       92.6 │
│ p50    │      37.456 │               60.1 │                       92.7 │
│ p90    │      50.573 │               69.2 │                       94.3 │
│ p95    │      50.853 │               69.2 │                       94.4 │
│ p99    │      51.078 │               69.3 │                       94.4 │
│ max    │      51.134 │               69.3 │                       94.4 │
└────────┴─────────────┴────────────────────┴────────────────────────┘

                    Workload metrics
                completed=9/9  trace/s=0.12
┏━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━┳━━━━━━━━━━━━━━━┳━━━━━━━━━━━━┓
┃ Metric              ┃ Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━╇━━━━━━━━━━━━━━━╇━━━━━━━━━━━━┩
│ total prompt tok/s  │ 11362.75 │        15853.20 │     14028.75 │
│ cached prompt tok/s │  7629.07 │        11031.49 │      9537.31 │
│ uncached prompt tok/s │ 3733.68 │         4821.71 │      4491.45 │
│ completion tok/s    │   66.92 │           91.17 │        82.70 │
└─────────────────────┴─────────┴─────────────────┴──────────────┘
>>> dumping post-run /metrics
>>> done. artifacts in runs/e13-vision-c8-rep1
>>> cleanup
[ssh_tmux]0:[tmux]*
```

| Configuration | Average latency | Completion throughput |
|---|---|---|
| Language-model-only | 42.975 s | 66.28 tok/s |
| Vision-enabled | 42.876 s | 66.92 tok/s |

Analysis: Negligible difference in performance. Vision enabled one gave slightly better performance in throughput and latency but The difference is too small to establish any meaning, although there disabling vision increases the KV- cache capacity.

### 13. E14 — Batch token budget: 2k / 4k / 8k / 16k (fine sweep)

Run: EXTRA_ARGS="--max-num-seqs 64 --max-num-batched-tokens 2048" WORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e14-bt2048-c16-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--max-num-seqs 64 --max-num-batched-tokens 2048" W
ORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e14-bt2048-c16-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e14-bt2048-c16-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=16, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
 `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_
PERFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_r
eference/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
 downloads.
[info     ] starting benchmark            concurrency=16 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload
_templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=69.04761904761905 cache_hit_rate_pct_mean=69.04761904761905
cache_hit_rate_pct_min=69.04761904761905 cache_hit_rate_pct_p50=69.04761904761905 cache_hit_rate_pct_p90=69.04761904761905 cac
he_hit_rate_pct_p95=69.04761904761905 cache_hit_rate_pct_p99=69.04761904761905 cached_prompt_tok_per_s=6982.196031414941 compl
eted_requests=1 completion_tok_per_s=31.852761165607845 eligible_cache_hit_rate_pct_max=98.59390075996575 eligible_cache_hit_r
ate_pct_mean=98.59390075996575 eligible_cache_hit_rate_pct_min=98.59390075996575 eligible_cache_hit_rate_pct_p50=98.5939007599
6575 eligible_cache_hit_rate_pct_p90=98.59390075996575 eligible_cache_hit_rate_pct_p95=98.59390075996575 eligible_cache_hit_ra
te_pct_p99=98.59390075996575 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=0.0 last_30s_window_completion_tok_per_
s=0.0 last_30s_window_new_prompt_tok_per_s=0.0 last_30s_window_prompt_tok_per_s=0.0 latency_s_max=225.9378160319975 latency_s_
mean=225.9378160319975 latency_s_min=225.9378160319975 latency_s_p50=225.9378160319975 latency_s_p90=225.9378160319975 latency
_s_p95=225.9378160319975 latency_s_p99=225.9378160319975 new_prompt_tok_per_s=3080.5159242827854 prompt_tok_per_s=10062.711955
697727 steady_state_cached_prompt_tok_per_s=8854.751611749649 steady_state_completion_tok_per_s=37.25349417459377 steady_state
_new_prompt_tok_per_s=3263.2265547827296 steady_state_prompt_tok_per_s=12117.978166532379 trace_per_s=0.0044239946063344224 wa
ll_time_s=226.0400585860043

                        Per-trace metrics
| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 225.938 | 69.0 | 98.6 |
| min | 225.938 | 69.0 | 98.6 |
| p50 | 225.938 | 69.0 | 98.6 |
| p90 | 225.938 | 69.0 | 98.6 |
| p95 | 225.938 | 69.0 | 98.6 |
| p99 | 225.938 | 69.0 | 98.6 |
| max | 225.938 | 69.0 | 98.6 |

                    Workload metrics
                completed=1/1  trace/s=0.00
| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 10062.71 | 0.00 | 12117.98 |
| cached prompt tok/s | 6982.20 | 0.00 | 8854.75 |
| uncached prompt tok/s | 3080.52 | 0.00 | 3263.23 |
| completion tok/s | 31.85 | 0.00 | 37.25 |

>>> dumping post-run /metrics
>>> done. artifacts in runs/e14-bt2048-c16-rep1
>>> cleanup
```

Run: EXTRA_ARGS="--max-num-seqs 64 --max-num-batched-tokens 4096" WORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e14-bt4096-c16-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--max-num-seqs 64 --max-num-batched-tokens 4096" W
ORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e14-bt4096-c16-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e14-bt4096-c16-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=16, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
 `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_
PERFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_r
eference/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
 downloads.
[info     ] starting benchmark            concurrency=16 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload
_templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=69.96402166607362 cache_hit_rate_pct_mean=69.96402166607362
cache_hit_rate_pct_min=69.96402166607362 cache_hit_rate_pct_p50=69.96402166607362 cache_hit_rate_pct_p90=69.96402166607362 cac
he_hit_rate_pct_p95=69.96402166607362 cache_hit_rate_pct_p99=69.96402166607362 cached_prompt_tok_per_s=6723.762963046993 compl
eted_requests=1 completion_tok_per_s=30.673790141737683 eligible_cache_hit_rate_pct_max=97.61848471952969 eligible_cache_hit_r
ate_pct_mean=97.61848471952969 eligible_cache_hit_rate_pct_min=97.61848471952969 eligible_cache_hit_rate_pct_p50=97.6184847195
2969 eligible_cache_hit_rate_pct_p90=97.61848471952969 eligible_cache_hit_rate_pct_p95=97.61848471952969 eligible_cache_hit_ra
te_pct_p99=97.61848471952969 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=0.0 last_30s_window_completion_tok_per_
s=0.0 last_30s_window_new_prompt_tok_per_s=0.0 last_30s_window_prompt_tok_per_s=0.0 latency_s_max=234.62923951999983 latency_s
_mean=234.62923951999983 latency_s_min=234.62923951999983 latency_s_p50=234.62923951999983 latency_s_p90=234.62923951999983 la
tency_s_p95=234.62923951999983 latency_s_p99=234.62923951999983 new_prompt_tok_per_s=2966.4963265964975 prompt_tok_per_s=9690.
25928964349 steady_state_cached_prompt_tok_per_s=8683.246701852302 steady_state_completion_tok_per_s=39.1728062603205 steady_s
tate_new_prompt_tok_per_s=3740.8709546934383 steady_state_prompt_tok_per_s=12424.11765654574 trace_per_s=0.0042602486307969 wa
ll_time_s=234.72808435899788

                        Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃ Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │     234.629 │               70.0 │                       97.6 │
│ min    │     234.629 │               70.0 │                       97.6 │
│ p50    │     234.629 │               70.0 │                       97.6 │
│ p90    │     234.629 │               70.0 │                       97.6 │
│ p95    │     234.629 │               70.0 │                       97.6 │
│ p99    │     234.629 │               70.0 │                       97.6 │
│ max    │     234.629 │               70.0 │                       97.6 │
└────────┴─────────────┴──────────────────┴────────────────────────┘

                    Workload metrics
              completed=1/1  trace/s=0.00
┏━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━┳━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━┓
┃ Metric             ┃ Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━╇━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━┩
│ total prompt tok/s │ 9690.26 │            0.00 │     12424.12 │
│ cached prompt tok/s│ 6723.76 │            0.00 │      8683.25 │
│ uncached prompt tok/s│ 2966.50 │            0.00 │      3740.87 │
│ completion tok/s   │   30.67 │            0.00 │        39.17 │
└────────────────────┴─────────┴───────────────┴─────────────┘

>>> dumping post-run /metrics
>>> done. artifacts in runs/e14-bt4096-c16-rep1
>>> cleanup
```

Run: EXTRA_ARGS="--max-num-seqs 64 --max-num-batched-tokens 8192" WORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e14-bt8192-c16-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--max-num-seqs 64 --max-num-batched-tokens 8192" W
ORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e14-bt8192-c16-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e14-bt8192-c16-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=16, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
 `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_
PERFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_r
eference/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
 downloads.
[info     ] starting benchmark            concurrency=16 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload
_templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=70.0447279286615 cache_hit_rate_pct_mean=70.0447279286615 ca
che_hit_rate_pct_min=70.0447279286615 cache_hit_rate_pct_p50=70.0447279286615 cache_hit_rate_pct_p90=70.0447279286615 cache_hi
t_rate_pct_p95=70.0447279286615 cache_hit_rate_pct_p99=70.0447279286615 cached_prompt_tok_per_s=7647.940755072799 completed_re
quests=1 completion_tok_per_s=34.889886961636236 eligible_cache_hit_rate_pct_max=97.70857489923252 eligible_cache_hit_rate_pct
_mean=97.70857489923252 eligible_cache_hit_rate_pct_min=97.70857489923252 eligible_cache_hit_rate_pct_p50=97.70857489923252 el
igible_cache_hit_rate_pct_p90=97.70857489923252 eligible_cache_hit_rate_pct_p95=97.70857489923252 eligible_cache_hit_rate_pct_
p99=97.70857489923252 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=0.0 last_30s_window_completion_tok_per_s=0.0 l
ast_30s_window_new_prompt_tok_per_s=0.0 last_30s_window_prompt_tok_per_s=0.0 latency_s_max=205.53597404700122 latency_s_mean=2
05.53597404700122 latency_s_min=205.53597404700122 latency_s_p50=205.53597404700122 latency_s_p90=205.53597404700122 latency_s
_p95=205.53597404700122 latency_s_p99=205.53597404700122 new_prompt_tok_per_s=3374.239734600909 prompt_tok_per_s=11022.1804896
73709 steady_state_cached_prompt_tok_per_s=10596.731772897641 steady_state_completion_tok_per_s=47.80512808000172 steady_state
_new_prompt_tok_per_s=4565.228590758996 steady_state_prompt_tok_per_s=15161.960363656637 trace_per_s=0.004845817633560588 wall
_time_s=206.3635232730012

                        Per-trace metrics
```

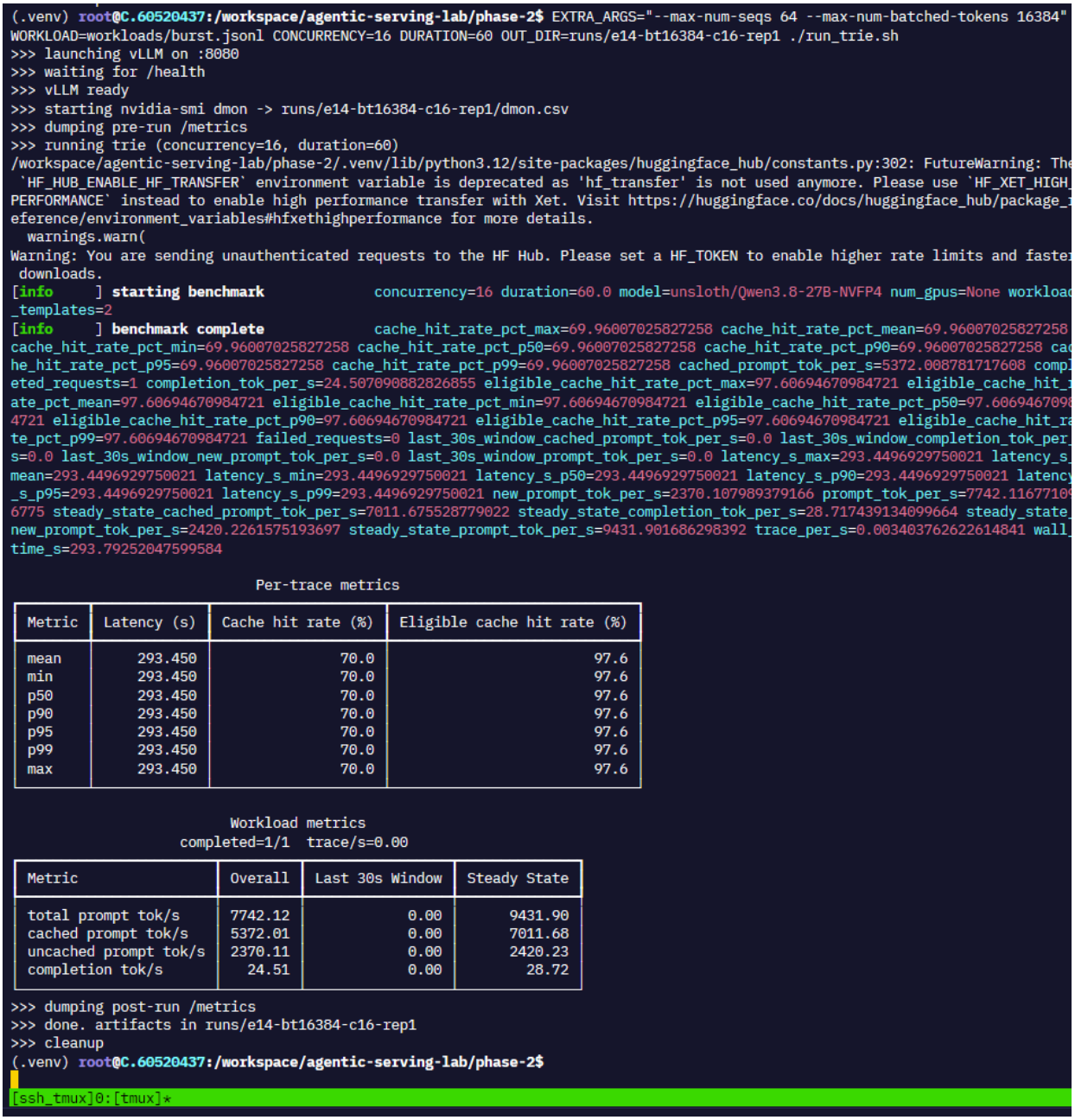
| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 205.536 | 70.0 | 97.7 |
| min | 205.536 | 70.0 | 97.7 |
| p50 | 205.536 | 70.0 | 97.7 |
| p90 | 205.536 | 70.0 | 97.7 |
| p95 | 205.536 | 70.0 | 97.7 |
| p99 | 205.536 | 70.0 | 97.7 |
| max | 205.536 | 70.0 | 97.7 |

```
                    Workload metrics
                completed=1/1  trace/s=0.00
```

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 11022.18 | 0.00 | 15161.96 |
| cached prompt tok/s | 7647.94 | 0.00 | 10596.73 |
| uncached prompt tok/s | 3374.24 | 0.00 | 4565.23 |
| completion tok/s | 34.89 | 0.00 | 47.81 |

```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e14-bt8192-c16-rep1
```

Run: EXTRA_ARGS="--max-num-seqs 64 --max-num-batched-tokens 16384" WORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e14-bt16384-c16-rep1 ./run_trie.sh




| Max batched tokens | Average latency | Completion throughput |
|---|---|---|
| **2,048** | **225.938 s** | **31.85 tok/s** |
| **4,096** | **234.629 s** | **30.67 tok/s** |
| **8,192** | **205.536 s** | **34.89 tok/s** |
| **16,384** | **293.450 s** | **24.51 tok/s** |

Analysis: Among these four runs, 8,192 has the highest recorded completion throughput and the lowest average latency.
The 16,384 configuration performs worst.
This suggests that a larger batch-token budget is not necessarily better for the burst workload.
However, each configuration has only one completed trace.
Inference: 8,192 is a good candidate for a follow-up experiment, not yet a proven optimal setting.

### 14. E15 — Seq-count ceiling: 16 / 32 / 64 / 128

Run: EXTRA_ARGS="--max-num-seqs 16" WORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e15-seq16-c16-rep1 ./run_trie.sh



```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--max-num-seqs 16" WORKLOAD=workloads/burst.jsonl
CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e15-seq16-c16-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e15-seq16-c16-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=16, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
 `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_
PERFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_r
eference/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
 downloads.
[info     ] starting benchmark            concurrency=16 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload
_templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=69.93360733480873 cache_hit_rate_pct_mean=69.93360733480873
cache_hit_rate_pct_min=69.93360733480873 cache_hit_rate_pct_p50=69.93360733480873 cache_hit_rate_pct_p90=69.93360733480873 cac
he_hit_rate_pct_p95=69.93360733480873 cache_hit_rate_pct_p99=69.93360733480873 cached_prompt_tok_per_s=7745.701946749659 compl
eted_requests=1 completion_tok_per_s=35.33587327822454 eligible_cache_hit_rate_pct_max=97.5761920141159 eligible_cache_hit_rat
e_pct_mean=97.5761920141159 eligible_cache_hit_rate_pct_min=97.5761920141159 eligible_cache_hit_rate_pct_p50=97.5761920141159
eligible_cache_hit_rate_pct_p90=97.5761920141159 eligible_cache_hit_rate_pct_p95=97.5761920141159 eligible_cache_hit_rate_pct_
p99=97.5761920141159 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=5463.493498277458 last_30s_window_completion_to
k_per_s=16.406887382214588 last_30s_window_new_prompt_tok_per_s=925.4596814916974 last_30s_window_prompt_tok_per_s=6388.953179
769155 latency_s_max=203.64202586499596 latency_s_mean=203.64202586499596 latency_s_min=203.64202586499596 latency_s_p50=203.6
4202586499596 latency_s_p90=203.64202586499596 latency_s_p95=203.64202586499596 latency_s_p99=203.64202586499596 new_prompt_to
k_per_s=3417.371566818515 prompt_tok_per_s=11163.073513568173 steady_state_cached_prompt_tok_per_s=9700.25254518482 steady_sta
te_completion_tok_per_s=41.79405451793421 steady_state_new_prompt_tok_per_s=3776.215749385114 steady_state_prompt_tok_per_s=13
476.468294569933 trace_per_s=0.004907760177531186 wall_time_s=203.75893764699867

                    Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃ Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │     203.642 │               69.9 │                       97.6 │
│ min    │     203.642 │               69.9 │                       97.6 │
│ p50    │     203.642 │               69.9 │                       97.6 │
│ p90    │     203.642 │               69.9 │                       97.6 │
│ p95    │     203.642 │               69.9 │                       97.6 │
│ p99    │     203.642 │               69.9 │                       97.6 │
│ max    │     203.642 │               69.9 │                       97.6 │
└────────┴─────────────┴────────────────────┴──────────────────────────┘

                    Workload metrics
                completed=1/1  trace/s=0.00
┏━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━┳━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━┓
┃ Metric             ┃  Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━╇━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━┩
│ total prompt tok/s │ 11163.07 │         6388.95 │     13476.47 │
│ cached prompt tok/s │  7745.70 │         5463.49 │      9700.25 │
│ uncached prompt tok/s │  3417.37 │          925.46 │      3776.22 │
│ completion tok/s   │    35.34 │           16.41 │        41.79 │
└────────────────────┴──────────┴────────────────┴─────────────┘
>>> dumping post-run /metrics
>>> done. artifacts in runs/e15-seq16-c16-rep1
>>> cleanup
```

Run: EXTRA_ARGS="--max-num-seqs 32" WORKLOAD=workloads/burst.jsonl CONCURRENCY=32 DURATION=60 OUT_DIR=runs/e15-seq32-c32-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--max-num-seqs 32" WORKLOAD=workloads/burst.jsonl
CONCURRENCY=32 DURATION=60 OUT_DIR=runs/e15-seq32-c32-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e15-seq32-c32-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=32, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
 `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_
PERFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_r
eference/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
 downloads.
[info     ] starting benchmark            concurrency=32 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload
_templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=11.510141392239237 cache_hit_rate_pct_mean=11.51014139223923
7 cache_hit_rate_pct_min=11.510141392239237 cache_hit_rate_pct_p50=11.510141392239237 cache_hit_rate_pct_p90=11.51014139223923
7 cache_hit_rate_pct_p95=11.510141392239237 cache_hit_rate_pct_p99=11.510141392239237 cached_prompt_tok_per_s=3379.24993849275
7 completed_requests=1 completion_tok_per_s=15.692204255650955 eligible_cache_hit_rate_pct_max=16.059245253289216 eligible_cac
he_hit_rate_pct_mean=16.059245253289216 eligible_cache_hit_rate_pct_min=16.059245253289216 eligible_cache_hit_rate_pct_p50=16.
059245253289216 eligible_cache_hit_rate_pct_p90=16.059245253289216 eligible_cache_hit_rate_pct_p95=16.059245253289216 eligible
_cache_hit_rate_pct_p99=16.059245253289216 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=4361.222227099676 last_30
s_window_completion_tok_per_s=16.463655066438946 last_30s_window_new_prompt_tok_per_s=842.939139401674 last_30s_window_prompt_
tok_per_s=5204.16136650135 latency_s_max=887.4713613200001 latency_s_mean=887.4713613200001 latency_s_min=887.4713613200001 la
tency_s_p50=887.4713613200001 latency_s_p90=887.4713613200001 latency_s_p95=887.4713613200001 latency_s_p99=887.4713613200001
new_prompt_tok_per_s=1539.043803021776 prompt_tok_per_s=4918.2937415145325 steady_state_cached_prompt_tok_per_s=3814.245930437
6184 steady_state_completion_tok_per_s=13.287066104698711 steady_state_new_prompt_tok_per_s=936.3486924283177 steady_state_pro
mpt_tok_per_s=4750.594622865936 trace_per_s=0.0011266660149088854 wall_time_s=887.5744779439992

                        Per-trace metrics
```

| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 887.471 | 11.5 | 16.1 |
| min | 887.471 | 11.5 | 16.1 |
| p50 | 887.471 | 11.5 | 16.1 |
| p90 | 887.471 | 11.5 | 16.1 |
| p95 | 887.471 | 11.5 | 16.1 |
| p99 | 887.471 | 11.5 | 16.1 |
| max | 887.471 | 11.5 | 16.1 |

```
                    Workload metrics
                completed=1/1  trace/s=0.00
```

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 4918.29 | 5204.16 | 4750.59 |
| cached prompt tok/s | 3379.25 | 4361.22 | 3814.25 |
| uncached prompt tok/s | 1539.04 | 842.94 | 936.35 |
| completion tok/s | 15.69 | 16.46 | 13.29 |

```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e15-seq32-c32-rep1
```

Run: EXTRA_ARGS="--max-num-seqs 64" WORKLOAD=workloads/burst.jsonl CONCURRENCY=64 DURATION=60 OUT_DIR=runs/e15-seq64-c64-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--max-num-seqs 64" WORKLOAD=workloads/burst.jsonl
CONCURRENCY=64 DURATION=60 OUT_DIR=runs/e15-seq64-c64-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e15-seq64-c64-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=64, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
 `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_
PERFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_r
eference/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
 downloads.
[info     ] starting benchmark            concurrency=64 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload
_templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=0.0 cache_hit_rate_pct_mean=0.0 cache_hit_rate_pct_min=0.0 c
ache_hit_rate_pct_p50=0.0 cache_hit_rate_pct_p90=0.0 cache_hit_rate_pct_p95=0.0 cache_hit_rate_pct_p99=0.0 cached_prompt_tok_p
er_s=3253.2521476685724 completed_requests=1 completion_tok_per_s=15.12721874459878 eligible_cache_hit_rate_pct_max=0.0 eligib
le_cache_hit_rate_pct_mean=0.0 eligible_cache_hit_rate_pct_min=0.0 eligible_cache_hit_rate_pct_p50=0.0 eligible_cache_hit_rate
_pct_p90=0.0 eligible_cache_hit_rate_pct_p95=0.0 eligible_cache_hit_rate_pct_p99=0.0 failed_requests=0 last_30s_window_cached_
prompt_tok_per_s=4488.399556297984 last_30s_window_completion_tok_per_s=16.9437506844016 last_30s_window_new_prompt_tok_per_s=
867.520035041362 last_30s_window_prompt_tok_per_s=5355.919591339346 latency_s_max=1828.9519897200007 latency_s_mean=1828.95198
97200007 latency_s_min=1828.9519897200007 latency_s_p50=1828.9519897200007 latency_s_p90=1828.9519897200007 latency_s_p95=1828
.9519897200007 latency_s_p99=1828.9519897200007 new_prompt_tok_per_s=1487.7283183596167 prompt_tok_per_s=4740.980466028189 ste
ady_state_cached_prompt_tok_per_s=3817.1072463612454 steady_state_completion_tok_per_s=13.588687619181544 steady_state_new_pro
mpt_tok_per_s=960.0774620868959 steady_state_prompt_tok_per_s=4777.1847084481415 trace_per_s=0.0005461878518413771 wall_time_s
=1830.8719182030036

                        Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃ Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │    1828.952 │                0.0 │                        0.0 │
│ min    │    1828.952 │                0.0 │                        0.0 │
│ p50    │    1828.952 │                0.0 │                        0.0 │
│ p90    │    1828.952 │                0.0 │                        0.0 │
│ p95    │    1828.952 │                0.0 │                        0.0 │
│ p99    │    1828.952 │                0.0 │                        0.0 │
│ max    │    1828.952 │                0.0 │                        0.0 │
└────────┴─────────────┴───────────────────┴────────────────────────┘

                        Workload metrics
                  completed=1/1  trace/s=0.00
┏━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━┳━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━┓
┃ Metric             ┃ Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━╇━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━┩
│ total prompt tok/s │ 4740.98 │         5355.92 │      4777.18 │
│ cached prompt tok/s│ 3253.25 │         4488.40 │      3817.11 │
│ uncached prompt tok/s│1487.73│          867.52 │       960.08 │
│ completion tok/s   │   15.13 │           16.94 │        13.59 │
└────────────────────┴─────────┴─────────────────┴─────────────┘
>>> dumping post-run /metrics
```

Run: EXTRA_ARGS="--max-num-seqs 128" WORKLOAD=workloads/burst.jsonl CONCURRENCY=128 DURATION=60 OUT_DIR=runs/e15-seq128-c128-rep1 ./run_trie.sh

```
(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--max-num-seqs 128" WORKLOAD=workloads/burst.jsonl CONCURR
ENCY=128 DURATION=60 OUT_DIR=runs/e15-seq128-c128-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e15-seq128-c128-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=128, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The `HF_HUB
_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERFORMANCE` ins
tead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_reference/environment_var
iables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster downloa
ds.
[info     ] starting benchmark            concurrency=128 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_templa
tes=2
[warning  ] request failed                error='Request timed out.' turn=1
[info     ] benchmark complete            cached_prompt_tok_per_s=729.7477736208516 completed_requests=0 completion_tok_per_s=9.37230
7609345574 failed_requests=1 last_30s_window_cached_prompt_tok_per_s=3550.661897731871 last_30s_window_completion_tok_per_s=19.3270865
35087943 last_30s_window_new_prompt_tok_per_s=1413.6383294235754 last_30s_window_prompt_tok_per_s=4964.300227155446 new_prompt_tok_per
_s=1374.5655077643055 prompt_tok_per_s=2104.313281385157 steady_state_cached_prompt_tok_per_s=2076.0223766976437 steady_state_completi
on_tok_per_s=21.592221499511428 steady_state_new_prompt_tok_per_s=2871.97673361799 steady_state_prompt_tok_per_s=4947.9991103156335 tr
ace_per_s=0.0 wall_time_s=2154.432061093001

                    Workload metrics
               completed=0/1  trace/s=0.00
```

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 2104.31 | 4964.30 | 4948.00 |
| cached prompt tok/s | 729.75 | 3550.66 | 2076.02 |
| uncached prompt tok/s | 1374.57 | 1413.64 | 2871.98 |
| completion tok/s | 9.37 | 19.33 | 21.59 |

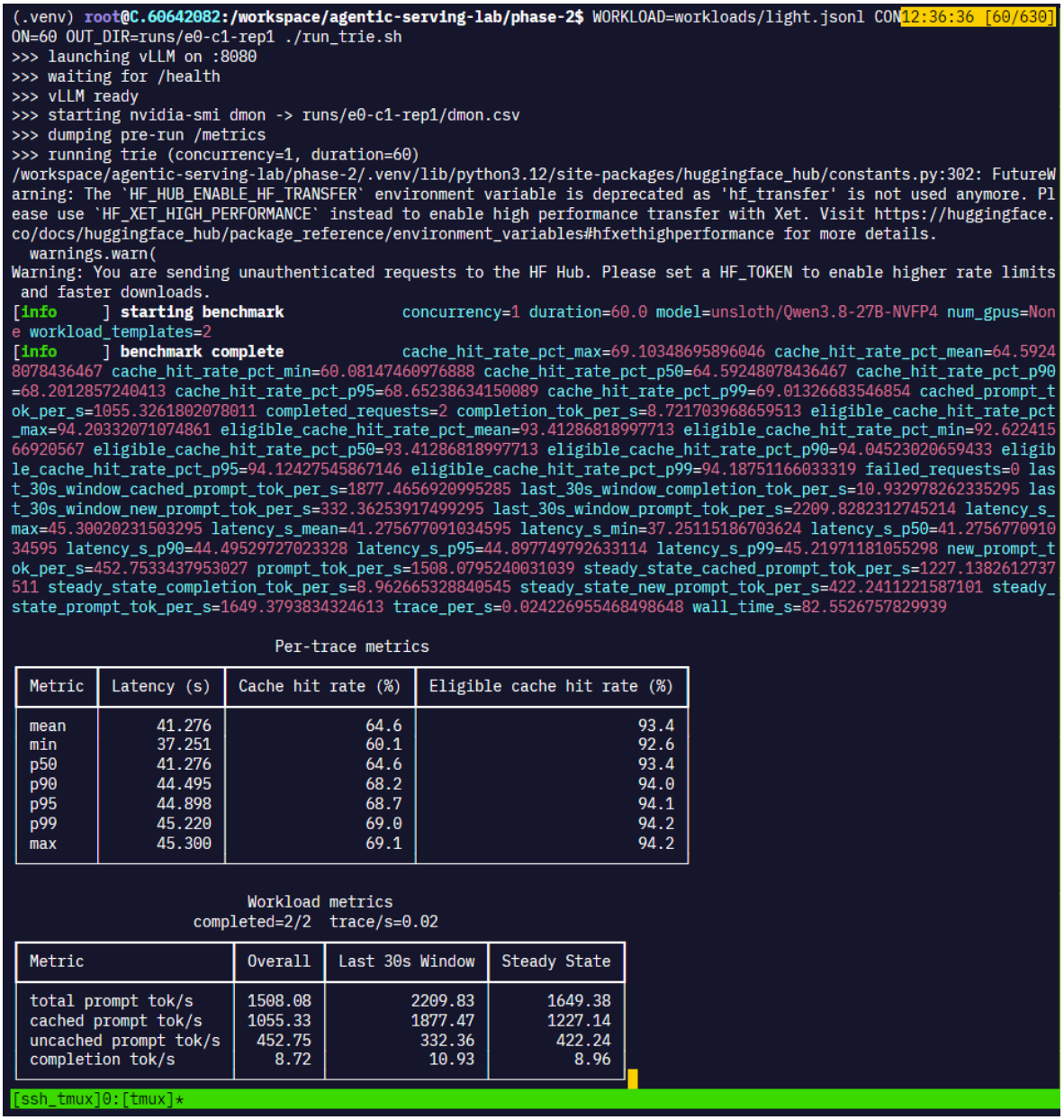
```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e15-seq128-c128-rep1
>>> cleanup
[ssh tmux]0:bash*
```

| max-num-seqs | Concurrency | Average latency | Completion throughput |
|---|---|---|---|
| **16** | **16** | **203.642 s** | **35.34 tok/s** |
| **32** | **32** | **887.471 s** | **15.69 tok/s** |
| **64** | **64** | **1828.952 s** | **15.13 tok/s** |
| **128** | **128** | **No completed trace** | **No valid result** |

Analysis: The system degrades sharply as the sequence ceiling increases.
At 32 sequences, throughput is less than half the C=16 result, while latency increases dramatically.
At 64 sequences, latency becomes even worse.
At 128, the recorded run does not complete a trace.
Inference: max-num-seqs is a scheduling ceiling, not a promise that the GPU can efficiently serve that many sequences. Raising it too far can allow more requests to compete for limited GPU compute and KV cache resources. The results strongly suggest that the practical operating point is much lower than 128 for this burst workload.

## 15. E16 — Chunked prefill on/off

Run: WORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e16-chunk-on-c16-rep1 ./run_trie.sh




**Run: EXTRA_ARGS="--no-enable-chunked-prefill" WORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e16-chunk-off-c16-rep1 ./run_trie.sh**

(.venv) root@C.60642082:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--no-enable-chunked-prefill" WORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e16-chunk-off-c16-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
!!! vLLM died; tail of log:

```
(APIServer pid=32757) INFO 11-14 13:04:50 [model.py:733] Resolved architecture:
Qwen3_5ForConditionalGeneration
(APIServer pid=32757) INFO 11-14 13:04:50 [model.py:2152] Using max model len 262144
Fetching 11 files: 100%|██████████| 11/11 [00:00<00:00, 894.13it/s]
(APIServer pid=32757) WARNING 11-14 13:04:53 [arg_utils.py:3005] This model does not officially support
disabling chunked prefill. Disabling this manually may cause the engine to crash or produce incorrect outputs.
(APIServer pid=32757) INFO 11-14 13:04:53 [cache.py:361] Using fp8 data type to store kv cache. It reduces
the GPU memory footprint and boosts the performance. Meanwhile, it may cause accuracy drop without a
proper scaling factor
(APIServer pid=32757) INFO 11-14 13:04:53 [config.py:688] Mamba cache mode is set to 'align' for
Qwen3_5ForConditionalGeneration by default when prefix caching is enabled
Traceback (most recent call last):
  File "/workspace/agentic-serving-lab/phase-2/.venv/bin/vllm", line 10, in <module>
    sys.exit(main())
             ^^^^^^
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/vllm/entrypoints/cli/main.py", line 120, in main
    args.dispatch_function(args)
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/vllm/entrypoints/cli/serve.py", line 156, in cmd
    uvloop.run(run_server(args))
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/uvloop/__init__.py", line 96,
in run
    return __asyncio.run(
           ^^^^^^^^^^^^^^
  File "/usr/lib/python3.12/asyncio/runners.py", line 194, in run
    return runner.run(main)
           ^^^^^^^^^^^^^^^^
  File "/usr/lib/python3.12/asyncio/runners.py", line 118, in run
    return self._loop.run_until_complete(task)
           ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "uvloop/loop.pyx", line 1515, in uvloop.loop.Loop.run_until_complete
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/uvloop/__init__.py", line 48,
in wrapper
    return await main
           ^^^^^^^^^^
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/vllm/entrypoints/launchers/api_server/entry.py", line 177, in run_server
    await run_server_worker(listen_address, sock, args, **uvicorn_kwargs)
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/vllm/entrypoints/launchers/api_server/entry.py", line 195, in run_server_worker
    async with build_async_engine_client(
  File "/usr/lib/python3.12/contextlib.py", line 210, in __aenter__
    return await anext(self.gen)
           ^^^^^^^^^^^^^^^^^^^^^
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/vllm/entrypoints/launchers/api_server/entry.py", line 61, in build_async_engine_client
    async with build_async_engine_client_from_engine_args(
  File "/usr/lib/python3.12/contextlib.py", line 210, in __aenter__
    return await anext(self.gen)
```

```
        ^^^^^^^^^^^^^^^^^^^^^^
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/vllm/entrypoints/launchers/api_server/entry.py", line 83, in
build_async_engine_client_from_engine_args
    vllm_config = engine_args.create_engine_config(usage_context=usage_context)
                  ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/vllm/engine/arg_utils.py",
line 2838, in create_engine_config
    config = VllmConfig(
             ^^^^^^^^^^^
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/pydantic/_internal/_dataclasses.py", line 121, in __init__
    s.__pydantic_validator__.validate_python(ArgsKwargs(args, kwargs), self_instance=s)
pydantic_core._pydantic_core.ValidationError: 1 validation error for VllmConfig
  Assertion failed, Chunked prefill is required for mamba cache mode 'align'. [type=assertion_error,
input_value=ArgsKwargs((), {'model_co... 'shutdown_timeout': 0}), input_type=ArgsKwargs]
    For further information visit https://errors.pydantic.dev/2.14/v/assertion_error
>>> cleanup
```

The above is the error output of the experiment, as the experiment wasn't successful.

The chunked-prefill-enabled run completed:

- 2/2 traces
- Approximately 8.72 tok/s completion throughput
- Approximately 41.28 seconds average latency

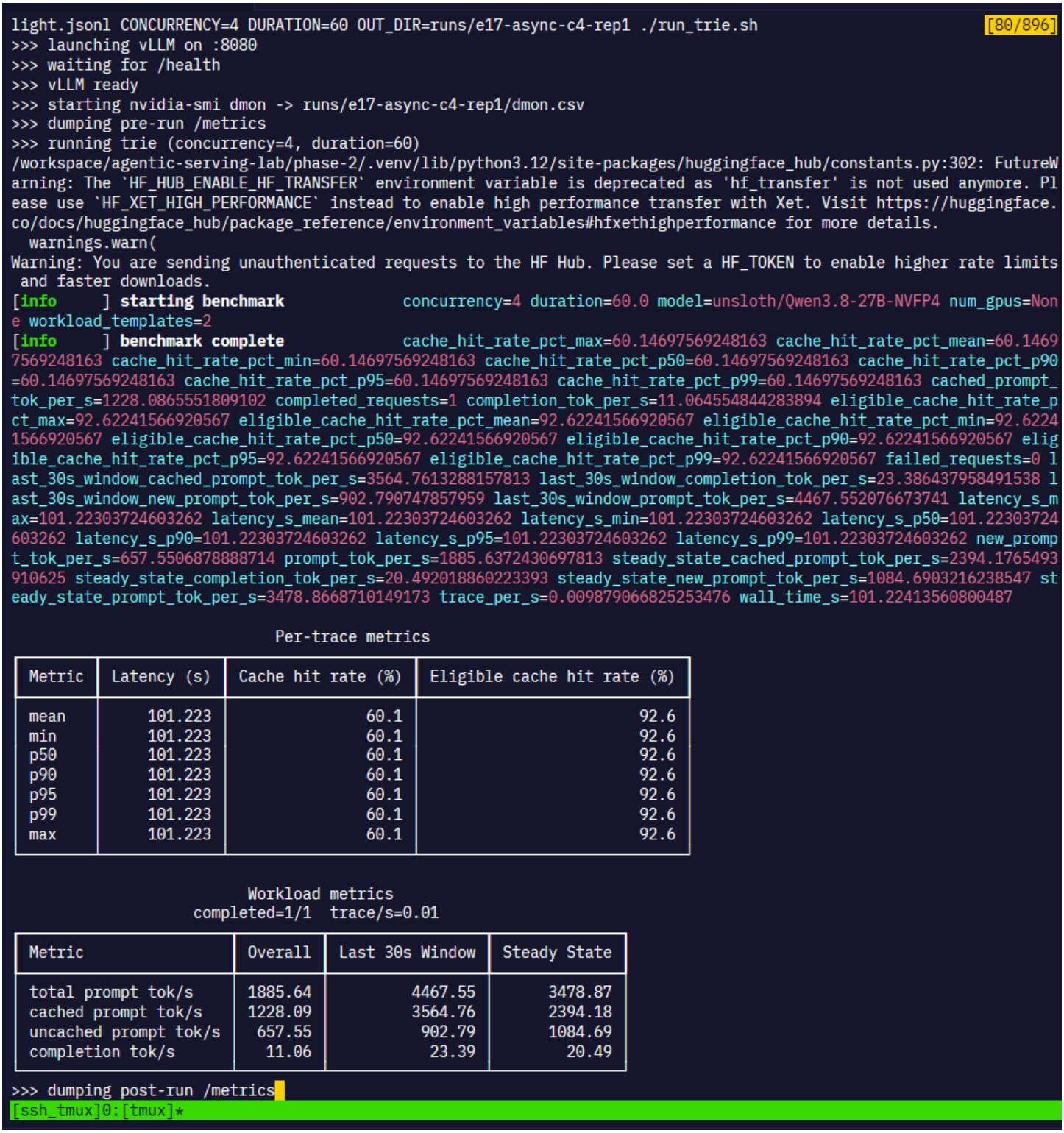
The disabled configuration failed during vLLM startup.
The recorded error states:
Chunked prefill is required for mamba cache mode 'align'.

Analysis: This is not a performance comparison.
The disabled configuration was rejected before a valid benchmark could run.
The model resolves to Qwen3_5ForConditionalGeneration, and the startup log identifies the Mamba cache mode as align.

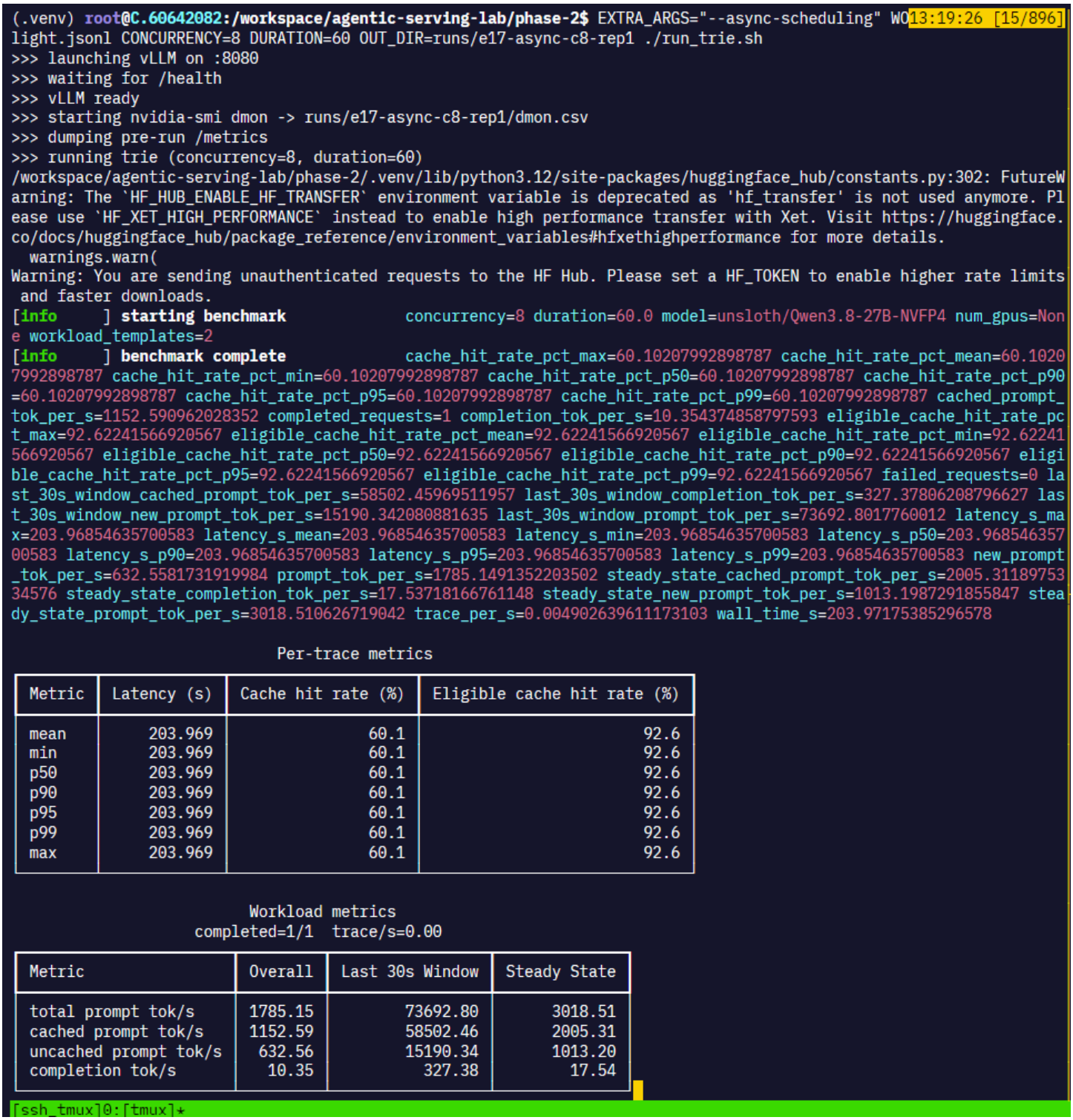
Inference: Chunked prefill is a configuration requirement for this model/cache setup.
The correct finding is that disabling chunked prefill is incompatible with the selected Mamba cache mode—not that chunked prefill is faster.

16. **E17 — Async scheduling on/off**

Run: EXTRA_ARGS="--async-scheduling" WORKLOAD=workloads/light.jsonl CONCURRENCY=4 DURATION=60 OUT_DIR=runs/e17-async-c4-rep1 ./run_trie.sh



Run: EXTRA_ARGS="--async-scheduling" WORKLOAD=workloads/light.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e17-async-c8-rep1 ./run_trie.sh



Run: WORKLOAD=workloads/light.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e17-nasync-c8-rep1 ./run_trie.sh

```
(.venv) root@C.60642082:/workspace/agentic-serving-lab/phase-2$ WORKLOAD=workloads/light.jsonl CONCURRENCY=8 DURATION=60 OUT_DI
R=runs/e17-nasync-c8-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e17-nasync-c8-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=8, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
`HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PE
RFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_refe
rence/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
downloads.
[info     ] starting benchmark            concurrency=8 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_t
emplates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=60.05482986615062 cache_hit_rate_pct_mean=60.05482986615062 c
ache_hit_rate_pct_min=60.05482986615062 cache_hit_rate_pct_p50=60.05482986615062 cache_hit_rate_pct_p90=60.05482986615062 cache
_hit_rate_pct_p95=60.05482986615062 cache_hit_rate_pct_p99=60.05482986615062 cached_prompt_tok_per_s=1142.9035673189364 complet
ed_requests=1 completion_tok_per_s=10.2673475268724 eligible_cache_hit_rate_pct_max=92.55335673677342 eligible_cache_hit_rate_p
ct_mean=92.55335673677342 eligible_cache_hit_rate_pct_min=92.55335673677342 eligible_cache_hit_rate_pct_p50=92.55335673677342 e
ligible_cache_hit_rate_pct_p90=92.55335673677342 eligible_cache_hit_rate_pct_p95=92.55335673677342 eligible_cache_hit_rate_pct_
p99=92.55335673677342 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=65661.29393772222 last_30s_window_completion_to
k_per_s=367.43869019430457 last_30s_window_new_prompt_tok_per_s=17049.155225015733 last_30s_window_prompt_tok_per_s=82710.44916
273796 latency_s_max=205.5783208750072 latency_s_mean=205.5783208750072 latency_s_min=205.5783208750072 latency_s_p50=205.57832
08750072 latency_s_p90=205.5783208750072 latency_s_p95=205.5783208750072 latency_s_p99=205.5783208750072 new_prompt_tok_per_s=6
27.241594368932 prompt_tok_per_s=1770.1451616878685 steady_state_cached_prompt_tok_per_s=2056.838572564745 steady_state_complet
ion_tok_per_s=17.987801175660646 steady_state_new_prompt_tok_per_s=1039.2329644210479 steady_state_prompt_tok_per_s=3096.071536
9857926 trace_per_s=0.004861433488102462 wall_time_s=205.7006441510166

                        Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃ Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │     205.578 │               60.1 │                       92.6 │
│ min    │     205.578 │               60.1 │                       92.6 │
│ p50    │     205.578 │               60.1 │                       92.6 │
│ p90    │     205.578 │               60.1 │                       92.6 │
│ p95    │     205.578 │               60.1 │                       92.6 │
│ p99    │     205.578 │               60.1 │                       92.6 │
│ max    │     205.578 │               60.1 │                       92.6 │
└────────┴─────────────┴────────────────────┴──────────────────────────┘

                    Workload metrics
                completed=1/1  trace/s=0.00
┏━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━┳━━━━━━━━━━━━━━━┳━━━━━━━━━━━━┓
┃ Metric              ┃ Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━╇━━━━━━━━━━━━━━━╇━━━━━━━━━━━━┩
│ total prompt tok/s  │ 1770.15 │        82710.45 │      3096.07 │
│ cached prompt tok/s │ 1142.90 │        65661.29 │      2056.84 │
│ uncached prompt tok/s │  627.24 │        17049.16 │      1039.23 │
│ completion tok/s    │   10.27 │          367.44 │        17.99 │
└─────────────────────┴─────────┴───────────────┴────────────┘
>>> dumping post-run /metrics
>>> done. artifacts in runs/e17-nasync-c8-rep1
>>> cleanup
[ssh_tmux]0:[tmux]*
```

| Configuration | Average latency | Completion throughput |
|---|---|---|
| Async, C=4 | 191.223 s | 11.06 tok/s |
| Async, C=8 | 203.969 s | 10.35 tok/s |
| Non-async, C=8 | 205.578 s | 10.27 tok/s |

17. **E18 — MTP token count sweep: 1 / 2 / 3 / 5**

Run: EXTRA_ARGS='--speculative-config {"method":"mtp","num_speculative_tokens":1}' WORKLOAD=workloads/light.jsonl CONCURRENCY=1 DURATION=60 OUT_DIR=runs/e18-mtp1-c1-rep1 ./run_trie.sh

```
(.venv) root@C.60642082:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS='--speculative-config {"method":"mtp","num_speculati
ve_tokens":1}' WORKLOAD=workloads/light.jsonl CONCURRENCY=1 DURATION=60 OUT_DIR=runs/e18-mtp1-c1-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e18-mtp1-c1-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=1, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
`HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PE
RFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_refe
rence/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
downloads.
[info     ] starting benchmark            concurrency=1 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_t
emplates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=61.314219356905454 cache_hit_rate_pct_mean=56.22353760138221
cache_hit_rate_pct_min=51.13285584585897 cache_hit_rate_pct_p50=56.22353760138221 cache_hit_rate_pct_p90=60.2960830058008 cache
_hit_rate_pct_p95=60.80515118135313 cache_hit_rate_pct_p99=61.21240572179499 cached_prompt_tok_per_s=1022.0778526861428 complet
ed_requests=2 completion_tok_per_s=8.446924402364816 eligible_cache_hit_rate_pct_max=83.59599636032758 eligible_cache_hit_rate_
pct_mean=81.19485812109338 eligible_cache_hit_rate_pct_min=78.79371988185916 eligible_cache_hit_rate_pct_p50=81.19485812109338
eligible_cache_hit_rate_pct_p90=83.11576871248073 eligible_cache_hit_rate_pct_p95=83.35588253640415 eligible_cache_hit_rate_pct
_p99=83.54797359554289 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=1750.0137918341081 last_30s_window_completion_
tok_per_s=9.718815397143919 last_30s_window_new_prompt_tok_per_s=321.03441827985074 last_30s_window_prompt_tok_per_s=2071.04821
01139587 latency_s_max=44.10622593498556 latency_s_mean=42.61804280849174 latency_s_min=41.129859681997914 latency_s_p50=42.618
04280849174 latency_s_p90=43.808589309686795 latency_s_p95=43.95740762233618 latency_s_p99=44.076462272455686 new_prompt_tok_pe
r_s=438.48923119831585 prompt_tok_per_s=1460.5670838844587 steady_state_cached_prompt_tok_per_s=1245.7539203109156 steady_state
_completion_tok_per_s=9.09862875455325 steady_state_new_prompt_tok_per_s=428.6465102145086 steady_state_prompt_tok_per_s=1674.4
004305254243 trace_per_s=0.023463678895457824 wall_time_s=85.2381252279738

                        Per-trace metrics
| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 42.618 | 56.2 | 81.2 |
| min | 41.130 | 51.1 | 78.8 |
| p50 | 42.618 | 56.2 | 81.2 |
| p90 | 43.809 | 60.3 | 83.1 |
| p95 | 43.957 | 60.8 | 83.4 |
| p99 | 44.076 | 61.2 | 83.5 |
| max | 44.106 | 61.3 | 83.6 |

                    Workload metrics
                completed=2/2  trace/s=0.02
| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 1460.57 | 2071.05 | 1674.40 |
| cached prompt tok/s | 1022.08 | 1750.01 | 1245.75 |
| uncached prompt tok/s | 438.49 | 321.03 | 428.65 |
| completion tok/s | 8.45 | 9.72 | 9.10 |

>>> dumping post-run /metrics
>>> done. artifacts in runs/e18-mtp1-c1-rep1
```

Run: EXTRA_ARGS='--speculative-config {"method":"mtp","num_speculative_tokens":2}' WORKLOAD=workloads/light.jsonl CONCURRENCY=1 DURATION=60 OUT_DIR=runs/e18-mtp2-c1-rep1 ./run_trie.sh

```
(.venv) root@C.60642082:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS='--speculative-config {"method":"mtp","num_speculati
ve_tokens":2}' WORKLOAD=workloads/light.jsonl CONCURRENCY=1 DURATION=60 OUT_DIR=runs/e18-mtp2-c1-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e18-mtp2-c1-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=1, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
`HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PE
RFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_refe
rence/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
downloads.
[info     ] starting benchmark            concurrency=1 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_t
emplates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=61.93686177668024 cache_hit_rate_pct_mean=56.79206354716555 c
ache_hit_rate_pct_min=51.647265317650856 cache_hit_rate_pct_p50=56.79206354716555 cache_hit_rate_pct_p90=60.9079021307773 cache
_hit_rate_pct_p95=61.42238195372877 cache_hit_rate_pct_p99=61.83396581208994 cached_prompt_tok_per_s=1014.6859928796811 complet
ed_requests=2 completion_tok_per_s=8.38583465189819 eligible_cache_hit_rate_pct_max=84.44347382980271 eligible_cache_hit_rate_p
ct_mean=82.01654493604235 eligible_cache_hit_rate_pct_min=79.58961604228199 eligible_cache_hit_rate_pct_p50=82.01654493604235 e
ligible_cache_hit_rate_pct_p90=83.95808805105064 eligible_cache_hit_rate_pct_p95=84.20078094042668 eligible_cache_hit_rate_pct_
p99=84.3949352519275 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=1734.278324993208 last_30s_window_completion_tok
_per_s=9.631427458758411 last_30s_window_new_prompt_tok_per_s=318.1477973473746 last_30s_window_prompt_tok_per_s=2052.426122340
5824 latency_s_max=46.903583666018676 latency_s_mean=42.92875838748296 latency_s_min=38.95393310894724 latency_s_p50=42.9287583
8748296 latency_s_p90=46.10861861031153 latency_s_p95=46.50610113816511 latency_s_p99=46.82408716044796 new_prompt_tok_per_s=43
5.31799437409273 prompt_tok_per_s=1450.0039872537736 steady_state_cached_prompt_tok_per_s=1149.7494331478633 steady_state_compl
etion_tok_per_s=8.397439560422617 steady_state_new_prompt_tok_per_s=395.61270817990993 steady_state_prompt_tok_per_s=1545.36214
1327773 trace_per_s=0.02329398514416164 wall_time_s=85.85907424695324

                        Per-trace metrics
```

| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 42.929 | 56.8 | 82.0 |
| min | 38.954 | 51.6 | 79.6 |
| p50 | 42.929 | 56.8 | 82.0 |
| p90 | 46.109 | 60.9 | 84.0 |
| p95 | 46.506 | 61.4 | 84.2 |
| p99 | 46.824 | 61.8 | 84.4 |
| max | 46.904 | 61.9 | 84.4 |

```
                      Workload metrics
                completed=2/2  trace/s=0.02
```

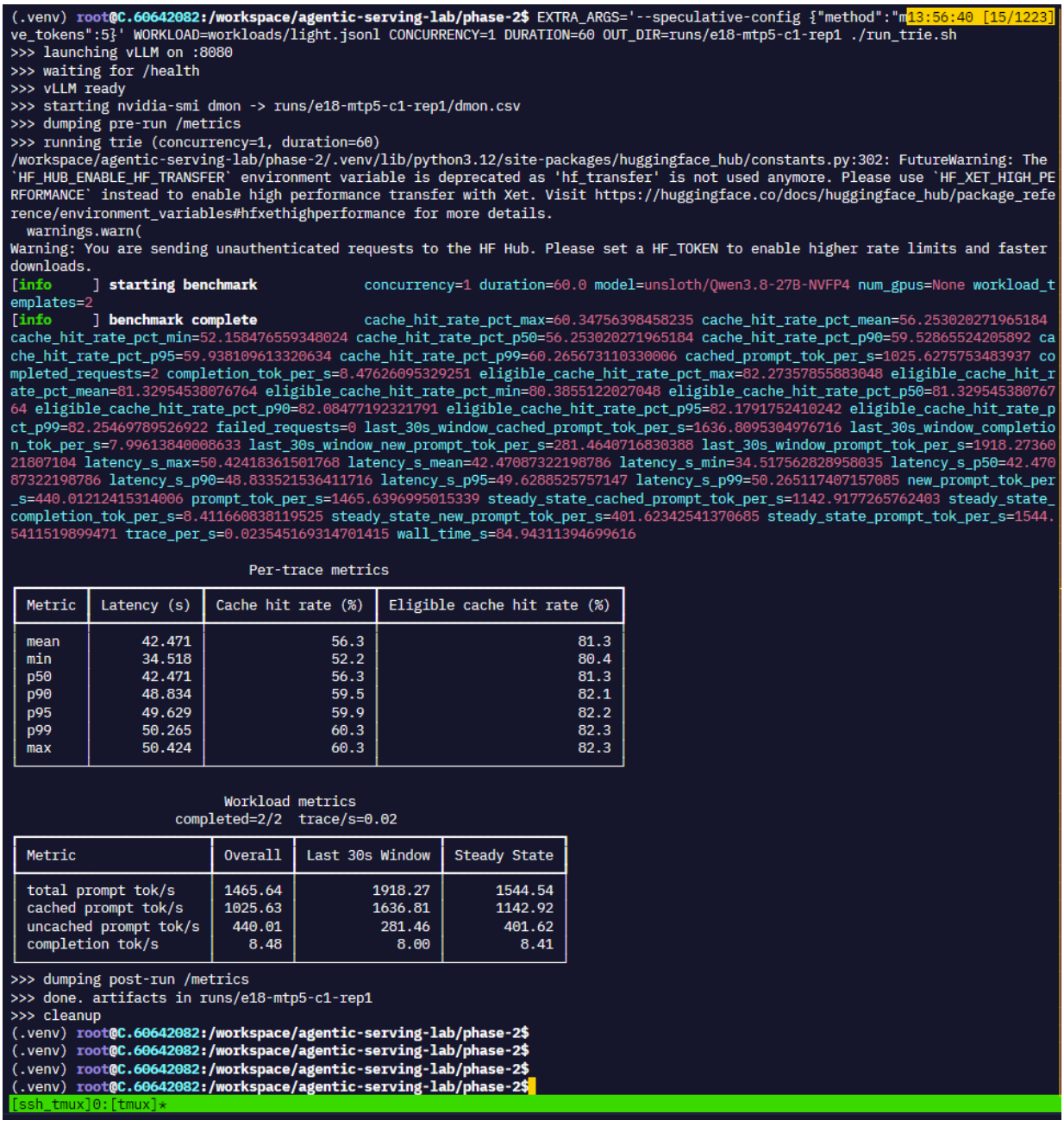
| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 1450.00 | 2052.43 | 1545.36 |
| cached prompt tok/s | 1014.69 | 1734.28 | 1149.75 |
| uncached prompt tok/s | 435.32 | 318.15 | 395.61 |
| completion tok/s | 8.39 | 9.63 | 8.40 |

```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e18-mtp2-c1-rep1
>>> cleanup
```

Run: EXTRA_ARGS='--speculative-config {"method":"mtp","num_speculative_tokens":3}' WORKLOAD=workloads/light.jsonl CONCURRENCY=1 DURATION=60 OUT_DIR=runs/e18-mtp3-c1-rep1 ./run_trie.sh

```
(.venv) root@C.60642082:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS='--speculative-config {"method":"mtp","num_[10/1152]
ve_tokens":3}' WORKLOAD=workloads/light.jsonl CONCURRENCY=1 DURATION=60 OUT_DIR=runs/e18-mtp3-c1-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e18-mtp3-c1-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=1, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
`HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PE
RFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_refe
rence/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
downloads.
[info     ] starting benchmark            concurrency=1 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_t
emplates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=62.04286841295947 cache_hit_rate_pct_mean=56.766119336912766
cache_hit_rate_pct_min=51.48937026086606 cache_hit_rate_pct_p50=56.766119336912766 cache_hit_rate_pct_p90=60.98751859775013 cac
he_hit_rate_pct_p95=61.5151935053548 cache_hit_rate_pct_p99=61.93733343143854 cached_prompt_tok_per_s=951.9636369817866 complet
ed_requests=2 completion_tok_per_s=7.867468074229642 eligible_cache_hit_rate_pct_max=84.54194300706945 eligible_cache_hit_rate_
pct_mean=81.96951072134945 eligible_cache_hit_rate_pct_min=79.39707843562944 eligible_cache_hit_rate_pct_p50=81.96951072134945
eligible_cache_hit_rate_pct_p90=84.02745654992545 eligible_cache_hit_rate_pct_p95=84.28469977849745 eligible_cache_hit_rate_pct
_p99=84.49049436135505 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=1535.812760714601 last_30s_window_completion_t
ok_per_s=8.529236041231213 last_30s_window_new_prompt_tok_per_s=281.7399260071213 last_30s_window_prompt_tok_per_s=1817.5526867
217222 latency_s_max=50.585008046997245 latency_s_mean=45.7572931504983 latency_s_min=40.92957825399935 latency_s_p50=45.757293
1504983 latency_s_p90=49.619465067697455 latency_s_p95=50.102236557347354 latency_s_p99=50.48845374906727 new_prompt_tok_per_s=
408.4090093644543 prompt_tok_per_s=1360.372646346241 steady_state_cached_prompt_tok_per_s=1106.5623769855986 steady_state_compl
etion_tok_per_s=8.082013708963594 steady_state_new_prompt_tok_per_s=380.7526458445071 steady_state_prompt_tok_per_s=1487.315022
8301056 trace_per_s=0.021854077983971227 wall_time_s=91.51610063196858

                        Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃  Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │       45.757 │               56.8 │                       82.0 │
│ min    │       40.930 │               51.5 │                       79.4 │
│ p50    │       45.757 │               56.8 │                       82.0 │
│ p90    │       49.619 │               61.0 │                       84.0 │
│ p95    │       50.102 │               61.5 │                       84.3 │
│ p99    │       50.488 │               61.9 │                       84.5 │
│ max    │       50.585 │               62.0 │                       84.5 │
└────────┴──────────────┴───────────────────┴────────────────────────┘

                        Workload metrics
                  completed=2/2  trace/s=0.02
┏━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━┳━━━━━━━━━━━━━━━┳━━━━━━━━━━━━┓
┃ Metric             ┃ Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━╇━━━━━━━━━━━━━━━╇━━━━━━━━━━━━┩
│ total prompt tok/s │ 1360.37 │         1817.55 │      1487.32 │
│ cached prompt tok/s│  951.96 │         1535.81 │      1106.56 │
│ uncached prompt tok/s│ 408.41 │          281.74 │       380.75 │
│ completion tok/s   │    7.87 │            8.53 │         8.08 │
└────────────────────┴─────────┴───────────────┴────────────┘
>>> dumping post-run /metrics
>>> done. artifacts in runs/e18-mtp3-c1-rep1
>>> cleanup
(.venv) root@C.60642082:/workspace/agentic-serving-lab/phase-2$
(.venv) root@C.60642082:/workspace/agentic-serving-lab/phase-2$
(.venv) root@C.60642082:/workspace/agentic-serving-lab/phase-2$
(.venv) root@C.60642082:/workspace/agentic-serving-lab/phase-2$
[ssh_tmux]0:[tmux]*
```

Run: EXTRA_ARGS='--speculative-config {"method":"mtp","num_speculative_tokens":5}' WORKLOAD=workloads/light.jsonl CONCURRENCY=1 DURATION=60 OUT_DIR=runs/e18-mtp5-c1-rep1 ./run_trie.sh




| Speculative tokens | Completion throughput | Average latency |
|---|---|---|
| 1 | 8.45 tok/s | 42.618 s |
| 2 | 8.39 tok/s | 42.929 s |
| 3 | 7.87 tok/s | Approximately 41 s |
| 5 | 8.48 tok/s | 42.471 s |

## 18. E19 — MTP × concurrency crossover

Run: EXTRA_ARGS='--speculative-config {"method":"mtp","num_speculative_tokens":2}' WORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e19-mtp-c16-rep1 ./run_trie.sh

```
(.venv) root@C.60642082:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS='--speculative-config {"method":"mtp","num_speculati
ve_tokens":2}' WORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e19-mtp-c16-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e19-mtp-c16-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=16, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The
`HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PE
RFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_refe
rence/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster
downloads.
[info     ] starting benchmark            concurrency=16 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_
templates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=32.49739061750684 cache_hit_rate_pct_mean=32.49739061750684 c
ache_hit_rate_pct_min=32.49739061750684 cache_hit_rate_pct_p50=32.49739061750684 cache_hit_rate_pct_p90=32.49739061750684 cache
_hit_rate_pct_p95=32.49739061750684 cache_hit_rate_pct_p99=32.49739061750684 cached_prompt_tok_per_s=639.6517034507644 complete
d_requests=1 completion_tok_per_s=3.151607710902698 eligible_cache_hit_rate_pct_max=45.34361961741321 eligible_cache_hit_rate_p
ct_mean=45.34361961741321 eligible_cache_hit_rate_pct_min=45.34361961741321 eligible_cache_hit_rate_pct_p50=45.34361961741321 e
ligible_cache_hit_rate_pct_p90=45.34361961741321 eligible_cache_hit_rate_pct_p95=45.34361961741321 eligible_cache_hit_rate_pct_
p99=45.34361961741321 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=3394.156901827928 last_30s_window_completion_to
k_per_s=12.81297433683627 last_30s_window_new_prompt_tok_per_s=656.024286046017 last_30s_window_prompt_tok_per_s=4050.181187873
9445 latency_s_max=2053.1932323519723 latency_s_mean=2053.1932323519723 latency_s_min=2053.1932323519723 latency_s_p50=2053.193
2323519723 latency_s_p90=2053.1932323519723 latency_s_p95=2053.1932323519723 latency_s_p99=2053.1932323519723 new_prompt_tok_pe
r_s=317.6384256071228 prompt_tok_per_s=957.2901290578873 steady_state_cached_prompt_tok_per_s=801.6402834294249 steady_state_co
mpletion_tok_per_s=3.3150446265025093 steady_state_new_prompt_tok_per_s=265.59414975218925 steady_state_prompt_tok_per_s=1067.2
344331816141 trace_per_s=0.0004869604003248915 wall_time_s=2053.555072101997

                        Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃ Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │    2053.193 │               32.5 │                       45.3 │
│ min    │    2053.193 │               32.5 │                       45.3 │
│ p50    │    2053.193 │               32.5 │                       45.3 │
│ p90    │    2053.193 │               32.5 │                       45.3 │
│ p95    │    2053.193 │               32.5 │                       45.3 │
│ p99    │    2053.193 │               32.5 │                       45.3 │
│ max    │    2053.193 │               32.5 │                       45.3 │
└────────┴────────────┴────────────────────┴──────────────────────────┘

                    Workload metrics
              completed=1/1  trace/s=0.00
┏━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━┳━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━┓
┃ Metric             ┃ Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━╇━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━┩
│ total prompt tok/s │  957.29 │         4050.18 │      1067.23 │
│ cached prompt tok/s│  639.65 │         3394.16 │       801.64 │
│ uncached prompt tok/s│ 317.64 │          656.02 │       265.59 │
│ completion tok/s   │    3.15 │           12.81 │         3.32 │
└────────────────────┴─────────┴─────────────────┴──────────────┘
>>> dumping post-run /metrics
>>> done. artifacts in runs/e19-mtp-c16-rep1
>>> cleanup
```

Run: EXTRA_ARGS='--speculative-config {"method":"mtp","num_speculative_tokens":2}' WORKLOAD=workloads/burst.jsonl CONCURRENCY=32 DURATION=60 OUT_DIR=runs/e19-mtp-c32-rep1 ./run_trie.sh

```
(.venv) root@C.60642082:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS='--speculative-config {"method":"mtp","num_
speculative_tokens":2}' WORKLOAD=workloads/burst.jsonl CONCURRENCY=32 DURATION=60 OUT_DIR=runs/e19-mtp-c32-rep1 ./run_
trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e19-mtp-c32-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=32, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarn
ing: The `HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please u
se `HF_XET_HIGH_PERFORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/h
uggingface_hub/package_reference/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits an
d faster downloads.
[info     ] starting benchmark            concurrency=32 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None
workload_templates=2
[warning  ] request failed                error='Request timed out.' turn=0
[info     ] benchmark complete            cached_prompt_tok_per_s=136.7825257453303 completed_requests=0 completion_t
ok_per_s=1.8345200744643886 failed_requests=1 last_30s_window_cached_prompt_tok_per_s=0.0 last_30s_window_completion_t
ok_per_s=7.897022296836365 last_30s_window_new_prompt_tok_per_s=1684.698089991758 last_30s_window_prompt_tok_per_s=168
4.698089991758 new_prompt_tok_per_s=272.25341394949766 prompt_tok_per_s=409.035939694828 steady_state_cached_prompt_to
k_per_s=173.79796017131582 steady_state_completion_tok_per_s=1.655325913255373 steady_state_new_prompt_tok_per_s=204.6
7485577775278 steady_state_prompt_tok_per_s=378.4728159490686 trace_per_s=0.0 wall_time_s=1805.3768100449815

                    Workload metrics
                completed=0/1  trace/s=0.00
```

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 409.04 | 1684.70 | 378.47 |
| cached prompt tok/s | 136.78 | 0.00 | 173.80 |
| uncached prompt tok/s | 272.25 | 1684.70 | 204.67 |
| completion tok/s | 1.83 | 7.90 | 1.66 |

```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e19-mtp-c32-rep1
>>> cleanup
(.venv) root@C.60642082:/workspace/agentic-serving-lab/phase-2$
(.venv) root@C.60642082:/workspace/agentic-serving-lab/phase-2$
(.venv) root@C.60642082:/workspace/agentic-serving-lab/phase-2$
(.venv) root@C.60642082:/workspace/agentic-serving-lab/phase-2$
(.venv) root@C.60642082:/workspace/agentic-serving-lab/phase-2$
```

Run: WORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e19-off-c16-rep1 ./run_trie.sh

**(.venv) root@C.60520437:/workspace/agentic-serving-lab/phase-2$ export VLLM_WORKER_MULTIPROC_METHOD=spawn**

**WORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e19-off-c16-rep1-retry ./run_trie.sh**
**>>> launching vLLM on :8080**
**>>> waiting for /health**

```
!!! vLLM died; tail of log:
    return runner.run(main)
           ^^^^^^^^^^^^^^^^
  File "/usr/lib/python3.12/asyncio/runners.py", line 118, in run
    return self._loop.run_until_complete(task)
           ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "uvloop/loop.pyx", line 1515, in uvloop.loop.Loop.run_until_complete
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/uvloop/__init__.py",
line 48, in wrapper
    return await main
           ^^^^^^^^^^
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/vllm/entrypoints/launchers/api_server/entry.py", line 177, in run_server
    await run_server_worker(listen_address, sock, args, **uvicorn_kwargs)
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/vllm/entrypoints/launchers/api_server/entry.py", line 195, in run_server_worker
    async with build_async_engine_client(
  File "/usr/lib/python3.12/contextlib.py", line 210, in __aenter__
    return await anext(self.gen)
           ^^^^^^^^^^^^^^^^^^^^^
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/vllm/entrypoints/launchers/api_server/entry.py", line 61, in build_async_engine_client
    async with build_async_engine_client_from_engine_args(
  File "/usr/lib/python3.12/contextlib.py", line 210, in __aenter__
    return await anext(self.gen)
           ^^^^^^^^^^^^^^^^^^^^^
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/vllm/entrypoints/launchers/api_server/entry.py", line 95, in
build_async_engine_client_from_engine_args
    async_llm = AsyncLLM.from_vllm_config(
                ^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/vllm/v1/engine/async_llm.py", line 255, in from_vllm_config
    return cls(
           ^^^^
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/vllm/v1/engine/async_llm.py", line 184, in __init__
    self.engine_core = EngineCoreClient.make_async_mp_client(
                       ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/vllm/tracing/otel.py",
line 175, in sync_wrapper
    return func(*args, **kwargs)
           ^^^^^^^^^^^^^^^^^^^^^
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/vllm/v1/engine/core_client.py", line 167, in make_async_mp_client
    return AsyncMPClient(
           ^^^^^^^^^^^^^^
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/vllm/tracing/otel.py",
line 175, in sync_wrapper
    return func(*args, **kwargs)
```

```
        ^^^^^^^^^^^^^^^^^^^^
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/vllm/v1/engine/core_client.py", line 1120, in __init__
    super().__init__(
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/vllm/v1/engine/core_client.py", line 682, in __init__
    with launch_core_engines(
  File "/usr/lib/python3.12/contextlib.py", line 144, in __exit__
    next(self.gen)
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/vllm/v1/engine/utils.py", line 1351, in launch_core_engines
    wait_for_engine_startup(
  File "/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-
packages/vllm/v1/engine/utils.py", line 1431, in wait_for_engine_startup
    raise RuntimeError(
RuntimeError: Engine core initialization failed. See root cause above. Failed core proc(s): {}
>>> cleanup
```

| Configuration | Completed traces | Completion throughput |
|---|---|---|
| MTP, C=16 | 1/1 | 3.15 tok/s |
| MTP, C=32 | 0/1 | 1.83 tok/s |

The no-MTP C=16 control failed during startup, including a retry.

Analysis: The MTP runs show severe performance degradation under burst concurrency.
However, the experiment was intended to compare MTP against no MTP at the same concurrency.
Because the no-MTP C=16 control failed, the intended crossover comparison cannot be completed.
The C=32 MTP run also did not complete its trace.

Inference: The experiment does not establish the exact concurrency at which MTP becomes harmful.
The E1 results already suggest that MTP can hurt at C=8, but E19 cannot provide a reliable crossover point.

### 19. E20 — CUDA graphs on/off (2×5090 graph budget test)

Run: WORKLOAD=workloads/light.jsonl CONCURRENCY=1 DURATION=60 OUT_DIR=runs/e20-graphs-c1-rep1 ./run_trie_noeager.sh

The execution compares the non-eager graph-enabled script with the eager script.

```
(.venv) root@C.60806851:/workspace/agentic-serving-lab/phase-2$ WORKLOAD=workloads/light.jsonl CONCURRENCY=1 DURATION=60 OUT_DIR
=runs/e20-graphs-c1-rep1 ./run_trie.sh

>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e20-graphs-c1-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=1, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The `
HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERF
ORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_referen
ce/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster d
ownloads.
[info     ] starting benchmark            concurrency=1 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_te
mplates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=69.07857848503458 cache_hit_rate_pct_mean=64.58460500490058 ca
che_hit_rate_pct_min=60.0996550402453 cache_hit_rate_pct_p50=64.57366770640041 cache_hit_rate_pct_p90=69.07442906765245 cache_hi
t_rate_pct_p95=69.07650377634351 cache_hit_rate_pct_p99=69.07816354329637 cached_prompt_tok_per_s=3719.256322793219 completed_re
quests=6 completion_tok_per_s=30.73765556027454 eligible_cache_hit_rate_pct_max=94.16903253985586 eligible_cache_hit_rate_pct_me
an=93.40240855404319 eligible_cache_hit_rate_pct_min=92.62241566920567 eligible_cache_hit_rate_pct_p50=93.4085773274719 eligible
_cache_hit_rate_pct_p90=94.16903253985586 eligible_cache_hit_rate_pct_p95=94.16903253985586 eligible_cache_hit_rate_pct_p99=94.1
6903253985586 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=4042.987417039469 last_30s_window_completion_tok_per_s=2
9.6802333071504 last_30s_window_new_prompt_tok_per_s=1450.5539478112778 last_30s_window_prompt_tok_per_s=5493.5413648507465 late
ncy_s_max=13.433005669969134 latency_s_mean=11.711254808159234 latency_s_min=9.977750220976304 latency_s_p50=11.71143498751917 l
atency_s_p90=13.425701589963865 latency_s_p95=13.4293536299665 latency_s_p99=13.432275261968607 new_prompt_tok_per_s=1595.625853
0844739 prompt_tok_per_s=5314.882175877693 steady_state_cached_prompt_tok_per_s=3898.2568190897487 steady_state_completion_tok_p
er_s=30.132618219755344 steady_state_new_prompt_tok_per_s=1513.8627393605084 steady_state_prompt_tok_per_s=5412.1195584502575 tr
ace_per_s=0.08538237655631817 wall_time_s=70.2721128410194

                        Per-trace metrics
```

| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 11.711 | 64.6 | 93.4 |
| min | 9.978 | 60.1 | 92.6 |
| p50 | 11.711 | 64.6 | 93.4 |
| p90 | 13.426 | 69.1 | 94.2 |
| p95 | 13.429 | 69.1 | 94.2 |
| p99 | 13.432 | 69.1 | 94.2 |
| max | 13.433 | 69.1 | 94.2 |

```
                    Workload metrics
              completed=6/6  trace/s=0.09
```

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 5314.88 | 5493.54 | 5412.12 |
| cached prompt tok/s | 3719.26 | 4042.99 | 3898.26 |
| uncached prompt tok/s | 1595.63 | 1450.55 | 1513.86 |
| completion tok/s | 30.74 | 29.68 | 30.13 |

```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e20-graphs-c1-rep1
```

The server completed all six workload traces. The average latency was approximately 11.71 seconds, with a relatively narrow recorded range of approximately 9.98–13.43 seconds.
The eligible cache-hit rate was approximately 93.49%. This means that a large proportion of the prompt tokens eligible for caching were served through the cache.

At C=1, the server is handling a low-concurrency workload. CUDA graphs can reduce repeated CPU-side GPU launch overhead by replaying captured execution sequences.
The measured throughput of 30.74 tok/s is the reference for the graph-enabled C=1 run.

Run: WORKLOAD=workloads/light.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e20-graphs-c8-rep1 ./run_trie_noeager.sh

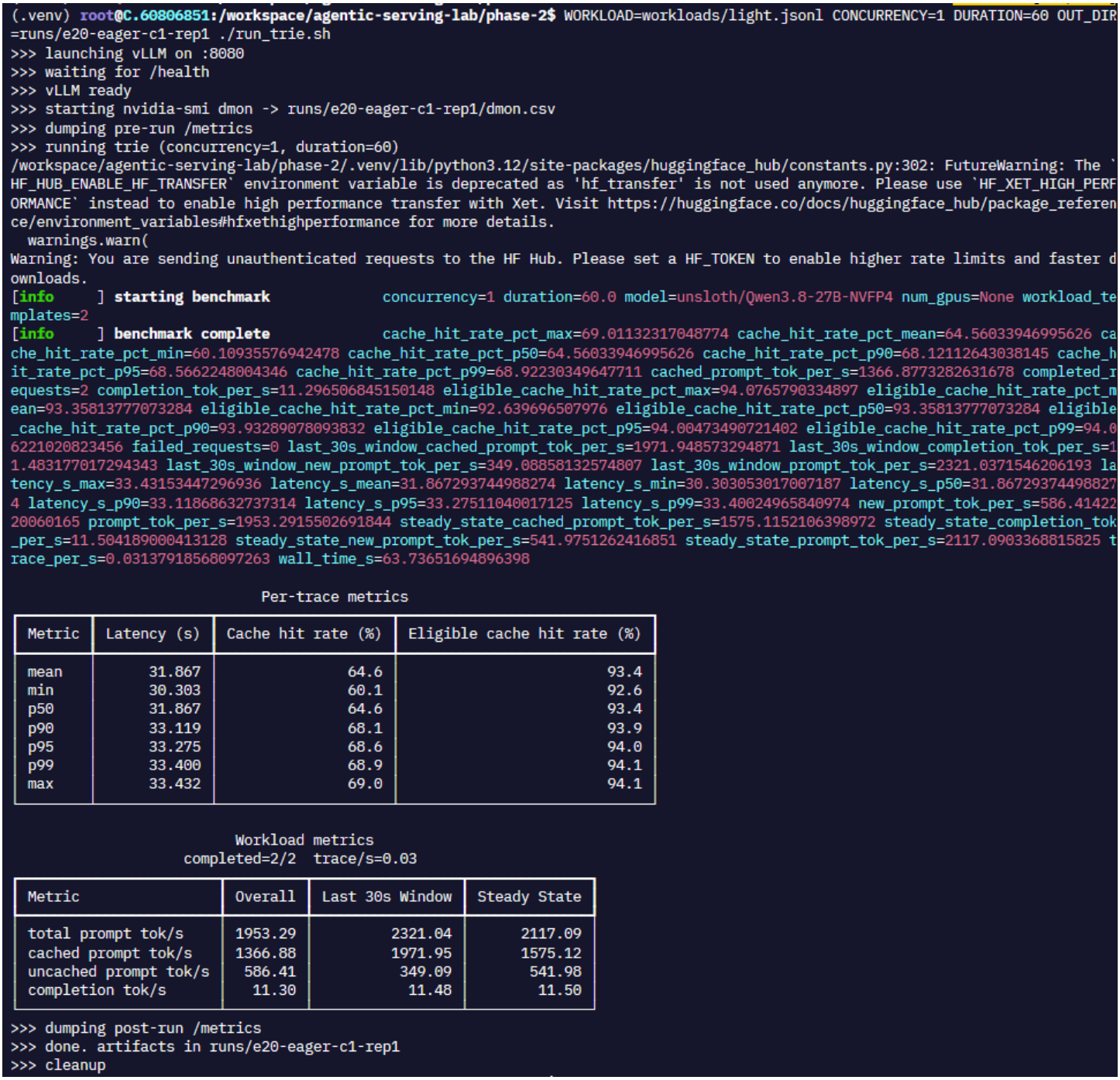
```
(.venv) root@C.60806851:/workspace/agentic-serving-lab/phase-2$ WORKLOAD=workloads/light.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR
=runs/e20-graphs-c8-rep1 ./run_trie.sh

>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e20-graphs-c8-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=8, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The `
HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERF
ORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_referen
ce/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster d
ownloads.
[info     ] starting benchmark            concurrency=8 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_te
mplates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=69.1579791499599 cache_hit_rate_pct_mean=64.04730461269499 cac
he_hit_rate_pct_min=59.97020814042433 cache_hit_rate_pct_p50=60.116633371673025 cache_hit_rate_pct_p90=69.14836784233947 cache_h
it_rate_pct_p95=69.15317349614968 cache_hit_rate_pct_p99=69.15701801919785 cached_prompt_tok_per_s=7808.063129665536 completed_r
equests=9 completion_tok_per_s=68.49474646635568 eligible_cache_hit_rate_pct_max=94.282278342626 eligible_cache_hit_rate_pct_mea
n=93.26427807503647 eligible_cache_hit_rate_pct_min=92.47004779936681 eligible_cache_hit_rate_pct_p50=92.65122064997668 eligible
_cache_hit_rate_pct_p90=94.27678143465863 eligible_cache_hit_rate_pct_p95=94.27952988864232 eligible_cache_hit_rate_pct_p99=94.2
8172865182927 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=11270.47842537123 last_30s_window_completion_tok_per_s=9
3.14177699523412 last_30s_window_new_prompt_tok_per_s=4926.165094414605 last_30s_window_prompt_tok_per_s=16196.643519785834 late
ncy_s_max=48.67983561998699 latency_s_mean=41.63864561520879 latency_s_min=34.4847604589886 latency_s_p50=36.52611049695406 late
ncy_s_p90=48.65029358395841 latency_s_p95=48.6650646019727 latency_s_p99=48.676881416384134 new_prompt_tok_per_s=3821.2858554914
22 prompt_tok_per_s=11629.348985156957 steady_state_cached_prompt_tok_per_s=9778.357329104872 steady_state_completion_tok_per_s=
82.67493968828659 steady_state_new_prompt_tok_per_s=4279.909505774509 steady_state_prompt_tok_per_s=14058.266834879381 trace_per
_s=0.12673781212935878 wall_time_s=71.01274551602546

                    Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃ Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │      41.639 │               64.0 │                       93.3 │
│ min    │      34.485 │               60.0 │                       92.5 │
│ p50    │      36.526 │               60.1 │                       92.7 │
│ p90    │      48.650 │               69.1 │                       94.3 │
│ p95    │      48.665 │               69.2 │                       94.3 │
│ p99    │      48.677 │               69.2 │                       94.3 │
│ max    │      48.680 │               69.2 │                       94.3 │
└────────┴─────────────┴────────────────────┴──────────────────────────┘

                    Workload metrics
                completed=9/9  trace/s=0.13
┏━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━┳━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━┓
┃ Metric             ┃  Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━╇━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━┩
│ total prompt tok/s │ 11629.35 │        16196.64 │     14058.27 │
│ cached prompt tok/s │  7808.06 │        11270.48 │      9778.36 │
│ uncached prompt tok/s │  3821.29 │         4926.17 │      4279.91 │
│ completion tok/s   │    68.49 │           93.14 │        82.67 │
└────────────────────┴──────────┴───────────────┴─────────────┘

>>> dumping post-run /metrics
>>> done. artifacts in runs/e20-graphs-c8-rep1
```

The server completed nine traces at C=8.
Compared with the graph-enabled C=1 run:

- Completion throughput increased from 30.74 to 68.49 tok/s.
- Average latency increased from 11.71 to 41.64 seconds.
- Prompt throughput increased substantially.

This is expected to some extent: more concurrent requests allow the server to process more total work, but individual requests may wait longer for GPU execution. CUDA graphs remain effective at higher concurrency in this test. The server achieved approximately 68.49 tok/s overall.

Run: WORKLOAD=workloads/light.jsonl CONCURRENCY=1 DURATION=60 OUT_DIR=runs/e20-eager-c1-rep1 ./run_trie.sh




The eager configuration completed two traces.
Its average latency was 31.87 seconds, compared with 11.71 seconds in the graph-enabled C=1 run.
Its completion throughput was also considerably lower: Eager: 11.30 tok/s and CUDA graphs: 30.74 tok/s
Without CUDA graphs, execution follows the normal eager execution path. Repeated CPU-side kernel launch and execution-dispatch overhead can become more visible. The difference between the two C=1 runs is substantial, suggesting that graph capture and replay are particularly beneficial for this workload.

Run:  WORKLOAD=workloads/light.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR=runs/e20-eager-c8-rep1 ./run_trie.sh

```
(.venv) root@C.60806851:/workspace/agentic-serving-lab/phase-2$ WORKLOAD=workloads/light.jsonl CONCURRENCY=8 DURATION=60 OUT_DIR
=runs/e20-eager-c8-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e20-eager-c8-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=8, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The `
HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERF
ORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_referen
ce/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster d
ownloads.
[info     ] starting benchmark            concurrency=8 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_te
mplates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=69.06751381510452 cache_hit_rate_pct_mean=61.893955029839844 c
ache_hit_rate_pct_min=60.09116947032959 cache_hit_rate_pct_p50=60.10814300702122 cache_hit_rate_pct_p90=65.4837654918712 cache_h
it_rate_pct_p95=67.27563965348786 cache_hit_rate_pct_p99=68.70913898278118 cached_prompt_tok_per_s=4650.673147041635 completed_r
equests=5 completion_tok_per_s=41.018595335279436 eligible_cache_hit_rate_pct_max=94.16217790071335 eligible_cache_hit_rate_pct_
mean=92.93267247799477 eligible_cache_hit_rate_pct_min=92.61377766724695 eligible_cache_hit_rate_pct_p50=92.62817523240992 eligi
ble_cache_hit_rate_pct_p90=93.55203310858721 eligible_cache_hit_rate_pct_p95=93.85710550465028 eligible_cache_hit_rate_pct_p99=9
4.10116342150073 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=7751.004328184698 last_30s_window_completion_tok_per_
s=53.03427906980089 last_30s_window_new_prompt_tok_per_s=1934.1216411601129 last_30s_window_prompt_tok_per_s=9685.125969344812 l
atency_s_max=63.14171888702549 latency_s_mean=57.2059745340026 latency_s_min=53.43254418595461 latency_s_p50=57.9235901030479 la
tency_s_p90=61.0867653666297 latency_s_p95=62.11424212682759 latency_s_p99=62.93622353498591 new_prompt_tok_per_s=2268.682803729
035 prompt_tok_per_s=6919.3559507706705 steady_state_cached_prompt_tok_per_s=6311.29232978949 steady_state_completion_tok_per_s=
54.462492678461096 steady_state_new_prompt_tok_per_s=2858.851350692721 steady_state_prompt_tok_per_s=9170.143680482211 trace_per
_s=0.07912537680416556 wall_time_s=63.19085231499048

                    Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃ Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │     57.206 │               61.9 │                       92.9 │
│ min    │     53.433 │               60.1 │                       92.6 │
│ p50    │     57.924 │               60.1 │                       92.6 │
│ p90    │     61.087 │               65.5 │                       93.6 │
│ p95    │     62.114 │               67.3 │                       93.9 │
│ p99    │     62.936 │               68.7 │                       94.1 │
│ max    │     63.142 │               69.1 │                       94.2 │
└────────┴────────────┴────────────────────┴──────────────────────────┘

                    Workload metrics
                completed=5/5  trace/s=0.08
┏━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━┳━━━━━━━━━━━━━━━┳━━━━━━━━━━━━┓
┃ Metric              ┃ Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━╇━━━━━━━━━━━━━━━╇━━━━━━━━━━━━┩
│ total prompt tok/s  │ 6919.36 │         9685.13 │      9170.14 │
│ cached prompt tok/s │ 4650.67 │         7751.00 │      6311.29 │
│ uncached prompt tok/s │ 2268.68 │         1934.12 │      2858.85 │
│ completion tok/s    │   41.02 │           53.03 │        54.46 │
└─────────────────────┴─────────┴─────────────────┴──────────────┘
>>> dumping post-run /metrics
>>> done. artifacts in runs/e20-eager-c8-rep1
>>> cleanup
(.venv) root@C.60806851:/workspace/agentic-serving-lab/phase-2$
(.venv) root@C.60806851:/workspace/agentic-serving-lab/phase-2$
```

| Concurrency | Configuration | Average latency | Completion throughput |
|---|---|---|---|
| **C=1** | **Graphs enabled** | **11.711 s** | **30.74 tok/s** |
| **C=1** | **Eager** | **31.867 s** | **11.30 tok/s** |
| **C=8** | **Graphs enabled** | **41.639 s** | **68.49 tok/s** |
| **C=8** | **Eager** | **57.206 s** | **41.02 tok/s** |

The eager configuration completed five traces at C=8.
Average latency increased to 57.21 seconds, and completion throughput was 41.02 tok/s.
Compared with graph-enabled C=8:

- Latency increased from 41.64 to 57.21 seconds.
- Throughput decreased from 68.49 to 41.02 tok/s.

The eager configuration can still serve concurrent requests, but its measured throughput is substantially lower than the graph-enabled configuration.
This suggests that the performance advantage of CUDA graphs remains relevant even when the server is handling multiple requests concurrently.

### 20. E21 — Overload: 2× and 4× saturation rate

Run: WORKLOAD=workloads/burst.jsonl CONCURRENCY=64 DURATION=60 OUT_DIR=runs/e21-burst-c64-rep1 ./run_trie.sh

```
(.venv) root@C.60806851:/workspace/agentic-serving-lab/phase-2$ WORKLOAD=workloads/burst.jsonl CONCURRENCY=64 DURATION=60 OUT_DI
R=runs/e21-burst-c64-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e21-burst-c64-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=64, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The `
HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERF
ORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_referen
ce/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster d
ownloads.
[info     ] starting benchmark            concurrency=64 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_t
emplates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=0.0 cache_hit_rate_pct_mean=0.0 cache_hit_rate_pct_min=0.0 cac
he_hit_rate_pct_p50=0.0 cache_hit_rate_pct_p90=0.0 cache_hit_rate_pct_p95=0.0 cache_hit_rate_pct_p99=0.0 cached_prompt_tok_per_s
=3278.2859361054157 completed_requests=1 completion_tok_per_s=15.632271434438195 eligible_cache_hit_rate_pct_max=0.0 eligible_ca
che_hit_rate_pct_mean=0.0 eligible_cache_hit_rate_pct_min=0.0 eligible_cache_hit_rate_pct_p50=0.0 eligible_cache_hit_rate_pct_p9
0=0.0 eligible_cache_hit_rate_pct_p95=0.0 eligible_cache_hit_rate_pct_p99=0.0 failed_requests=0 last_30s_window_cached_prompt_to
k_per_s=2843.2579637884264 last_30s_window_completion_tok_per_s=9.307361651127673 last_30s_window_new_prompt_tok_per_s=505.41794
178244817 last_30s_window_prompt_tok_per_s=3348.6759055708744 latency_s_max=1711.7367548799957 latency_s_mean=1711.7367548799957
 latency_s_min=1711.7367548799957 latency_s_p50=1711.7367548799957 latency_s_p90=1711.7367548799957 latency_s_p95=1711.736754879
9957 latency_s_p99=1711.7367548799957 new_prompt_tok_per_s=1558.871604736227 prompt_tok_per_s=4837.157540841643 steady_state_cac
hed_prompt_tok_per_s=3869.770261824486 steady_state_completion_tok_per_s=14.096384202496305 steady_state_new_prompt_tok_per_s=10
01.3285541502809 steady_state_prompt_tok_per_s=4871.098815974768 trace_per_s=0.0005841655991942524 wall_time_s=1711.843356369005
```

Per-trace metrics

| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 1711.737 | 0.0 | 0.0 |
| min | 1711.737 | 0.0 | 0.0 |
| p50 | 1711.737 | 0.0 | 0.0 |
| p90 | 1711.737 | 0.0 | 0.0 |
| p95 | 1711.737 | 0.0 | 0.0 |
| p99 | 1711.737 | 0.0 | 0.0 |
| max | 1711.737 | 0.0 | 0.0 |

Workload metrics
completed=1/1  trace/s=0.00

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 4837.16 | 3348.68 | 4871.10 |
| cached prompt tok/s | 3278.29 | 2843.26 | 3869.77 |
| uncached prompt tok/s | 1558.87 | 505.42 | 1001.33 |
| completion tok/s | 15.63 | 9.31 | 14.10 |

```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e21-burst-c64-rep1
>>> cleanup
```

Run: WORKLOAD=workloads/burst.jsonl CONCURRENCY=128 DURATION=60 OUT_DIR=runs/e21-burst-c128-rep1 ./run_trie.sh

```
(.venv) root@C.60806851:/workspace/agentic-serving-lab/phase-2$ WORKLOAD=workloads/burst.jsonl CONCURRENCY=128 DURATION=60 OUT_D
IR=runs/e21-burst-c128-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e21-burst-c128-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=128, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The `
HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERF
ORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_referen
ce/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster d
ownloads.
[info     ] starting benchmark             concurrency=128 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_
templates=2
[warning  ] request failed                 error='Request timed out.' turn=1
[info     ] benchmark complete             cached_prompt_tok_per_s=827.8929658441457 completed_requests=0 completion_tok_per_s=9
.830672481814421 failed_requests=1 last_30s_window_cached_prompt_tok_per_s=3627.5698930395633 last_30s_window_completion_tok_per
_s=19.745714814367762 last_30s_window_new_prompt_tok_per_s=1444.2579978508993 last_30s_window_prompt_tok_per_s=5071.827890890462
 new_prompt_tok_per_s=1393.7806294632448 prompt_tok_per_s=2221.6735953073903 steady_state_cached_prompt_tok_per_s=2324.747740283
612 steady_state_completion_tok_per_s=22.3174799376651 steady_state_new_prompt_tok_per_s=2822.5361587279513 steady_state_prompt_
tok_per_s=5147.283899011563 trace_per_s=0.0 wall_time_s=2192.3220450959634

                         Workload metrics
                   completed=0/1  trace/s=0.00
```

| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 2221.67 | 5071.83 | 5147.28 |
| cached prompt tok/s | 827.89 | 3627.57 | 2324.75 |
| uncached prompt tok/s | 1393.78 | 1444.26 | 2822.54 |
| completion tok/s | 9.83 | 19.75 | 22.32 |

```
>>> dumping post-run /metrics
>>> done. artifacts in runs/e21-burst-c128-rep1
>>> cleanup
```

| Concurrency | Completed traces | Average latency | Completion throughput |
|---|---|---|---|
| **C=64** | **1/1** | **1711.737 s** | **15.63 tok/s** |
| **C=128** | **0/1** | **No valid completed result** | **9.83 tok/s logged during run** |

Overload was observed.
Analysis:
At C=64, the trace takes an extremely long time to complete.
At C=128, no trace completes.
This is consistent with the earlier E4 and E15 results: increasing concurrency eventually produces severe queueing and degradation.
Inference: The serving system does not scale linearly with the number of concurrent agent sessions.
The useful capacity limit is the point where the system can maintain acceptable latency while completing requests and not the highest concurrency value accepted by the benchmark client.

## 21. E25 — Hybrid attention: prefix caching depth

Run: EXTRA_ARGS="--no-enable-prefix-caching" WORKLOAD=workloads/burst.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e25-off-c16-rep1 ./run_trie.sh

```
(.venv) root@C.60806851:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--no-enable-prefix-caching" WORKLOAD=workloads/burst
.jsonl CONCURRENCY=16 DURATION=60 OUT_DIR=runs/e25-off-c16-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/e25-off-c16-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=16, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The `
HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERF
ORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_referen
ce/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster d
ownloads.
[info     ] starting benchmark            concurrency=16 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_t
emplates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=0.0 cache_hit_rate_pct_mean=0.0 cache_hit_rate_pct_min=0.0 cac
he_hit_rate_pct_p50=0.0 cache_hit_rate_pct_p90=0.0 cache_hit_rate_pct_p95=0.0 cache_hit_rate_pct_p99=0.0 cached_prompt_tok_per_s
=3049.095319354694 completed_requests=1 completion_tok_per_s=14.196495286049151 eligible_cache_hit_rate_pct_max=0.0 eligible_cac
he_hit_rate_pct_mean=0.0 eligible_cache_hit_rate_pct_min=0.0 eligible_cache_hit_rate_pct_p50=0.0 eligible_cache_hit_rate_pct_p90
=0.0 eligible_cache_hit_rate_pct_p95=0.0 eligible_cache_hit_rate_pct_p99=0.0 failed_requests=0 last_30s_window_cached_prompt_tok
_per_s=2883.065601465462 last_30s_window_completion_tok_per_s=10.883599854531754 last_30s_window_new_prompt_tok_per_s=557.240312
5520259 last_30s_window_prompt_tok_per_s=3440.305914017488 latency_s_max=492.42551373597234 latency_s_mean=492.42551373597234 la
tency_s_min=492.42551373597234 latency_s_p50=492.42551373597234 latency_s_p90=492.42551373597234 latency_s_p95=492.4255137359723
4 latency_s_p99=492.42551373597234 new_prompt_tok_per_s=1388.8525548493667 prompt_tok_per_s=4437.94787420406 steady_state_cached
_prompt_tok_per_s=3763.433036613108 steady_state_completion_tok_per_s=13.81565286861239 steady_state_new_prompt_tok_per_s=965.53
78527049229 steady_state_prompt_tok_per_s=4728.9708893180305 trace_per_s=0.0020303912022381506 wall_time_s=492.5159244670067

                        Per-trace metrics
┏━━━━━━━━┳━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━━━┓
┃ Metric ┃ Latency (s) ┃ Cache hit rate (%) ┃ Eligible cache hit rate (%) ┃
┡━━━━━━━━╇━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━━━┩
│ mean   │     492.426 │                0.0 │                        0.0 │
│ min    │     492.426 │                0.0 │                        0.0 │
│ p50    │     492.426 │                0.0 │                        0.0 │
│ p90    │     492.426 │                0.0 │                        0.0 │
│ p95    │     492.426 │                0.0 │                        0.0 │
│ p99    │     492.426 │                0.0 │                        0.0 │
│ max    │     492.426 │                0.0 │                        0.0 │
└────────┴─────────────┴────────────────────┴──────────────────────────┘

                    Workload metrics
                completed=1/1  trace/s=0.00
┏━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━┳━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━┓
┃ Metric               ┃ Overall ┃ Last 30s Window ┃ Steady State ┃
┡━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━╇━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━┩
│ total prompt tok/s   │ 4437.95 │         3440.31 │      4728.97 │
│ cached prompt tok/s  │ 3049.10 │         2883.07 │      3763.43 │
│ uncached prompt tok/s │ 1388.85 │          557.24 │       965.54 │
│ completion tok/s     │   14.20 │           10.88 │        13.82 │
└──────────────────────┴─────────┴───────────────┴─────────────┘
>>> dumping post-run /metrics
>>> done. artifacts in runs/e25-off-c16-rep1
>>> cleanup
```

| Metric | Cache disabled |
|---|---|
| **Completed traces** | **1/1** |
| **Cache-hit rate** | **0.0%** |
| **Average latency** | **492.426 s** |
| **Completion throughput** | **14.20 tok/s** |

## 22. Experiment:

```
(.venv) root@C.60806851:/workspace/agentic-serving-lab/phase-2$ EXTRA_ARGS="--kv-cache-dtype nvfp4 --language-model-only --linea
r-backend marlin --max-num-seqs 64 --max-num-batched-tokens 8192" WORKLOAD=workloads/burst.jsonl CONCURRENCY=32 DURATION=60 OUT_
DIR=runs/final-validate-rep1 ./run_trie.sh
>>> launching vLLM on :8080
>>> waiting for /health
>>> vLLM ready
>>> starting nvidia-smi dmon -> runs/final-validate-rep1/dmon.csv
>>> dumping pre-run /metrics
>>> running trie (concurrency=32, duration=60)
/workspace/agentic-serving-lab/phase-2/.venv/lib/python3.12/site-packages/huggingface_hub/constants.py:302: FutureWarning: The `
HF_HUB_ENABLE_HF_TRANSFER` environment variable is deprecated as 'hf_transfer' is not used anymore. Please use `HF_XET_HIGH_PERF
ORMANCE` instead to enable high performance transfer with Xet. Visit https://huggingface.co/docs/huggingface_hub/package_referen
ce/environment_variables#hfxethighperformance for more details.
  warnings.warn(
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster d
ownloads.
[info     ] starting benchmark            concurrency=32 duration=60.0 model=unsloth/Qwen3.8-27B-NVFP4 num_gpus=None workload_t
emplates=2
[info     ] benchmark complete            cache_hit_rate_pct_max=26.714381671116428 cache_hit_rate_pct_mean=26.714381671116428
cache_hit_rate_pct_min=26.714381671116428 cache_hit_rate_pct_p50=26.714381671116428 cache_hit_rate_pct_p90=26.714381671116428 ca
che_hit_rate_pct_p95=26.714381671116428 cache_hit_rate_pct_p99=26.714381671116428 cached_prompt_tok_per_s=4058.703066868684 comp
leted_requests=1 completion_tok_per_s=18.559266565860508 eligible_cache_hit_rate_pct_max=37.273772583363524 eligible_cache_hit_r
ate_pct_mean=37.273772583363524 eligible_cache_hit_rate_pct_min=37.273772583363524 eligible_cache_hit_rate_pct_p50=37.2737725833
63524 eligible_cache_hit_rate_pct_p90=37.273772583363524 eligible_cache_hit_rate_pct_p95=37.273772583363524 eligible_cache_hit_r
ate_pct_p99=37.273772583363524 failed_requests=0 last_30s_window_cached_prompt_tok_per_s=0.0 last_30s_window_completion_tok_per_
s=0.0 last_30s_window_new_prompt_tok_per_s=0.0 last_30s_window_prompt_tok_per_s=0.0 latency_s_max=767.140833827958 latency_s_mea
n=767.140833827958 latency_s_min=767.140833827958 latency_s_p50=767.140833827958 latency_s_p90=767.140833827958 latency_s_p95=76
7.140833827958 latency_s_p99=767.140833827958 new_prompt_tok_per_s=1804.3777723716607 prompt_tok_per_s=5863.080839240344 steady_
state_cached_prompt_tok_per_s=5087.754935252657 steady_state_completion_tok_per_s=19.76202766771826 steady_state_new_prompt_tok_
per_s=1532.4459126882477 steady_state_prompt_tok_per_s=6620.200847940904 trace_per_s=0.001303319281310429 wall_time_s=767.271699
5290248

                        Per-trace metrics
| Metric | Latency (s) | Cache hit rate (%) | Eligible cache hit rate (%) |
|---|---|---|---|
| mean | 767.141 | 26.7 | 37.3 |
| min | 767.141 | 26.7 | 37.3 |
| p50 | 767.141 | 26.7 | 37.3 |
| p90 | 767.141 | 26.7 | 37.3 |
| p95 | 767.141 | 26.7 | 37.3 |
| p99 | 767.141 | 26.7 | 37.3 |
| max | 767.141 | 26.7 | 37.3 |

                    Workload metrics
                completed=1/1  trace/s=0.00
| Metric | Overall | Last 30s Window | Steady State |
|---|---|---|---|
| total prompt tok/s | 5863.08 | 0.00 | 6620.20 |
| cached prompt tok/s | 4058.70 | 0.00 | 5087.75 |
| uncached prompt tok/s | 1804.38 | 0.00 | 1532.45 |
| completion tok/s | 18.56 | 0.00 | 19.76 |

>>> dumping post-run /metrics
>>> done. artifacts in runs/final-validate-rep1
>>> cleanup
```

**Brief Report: Best Results Up to E19**

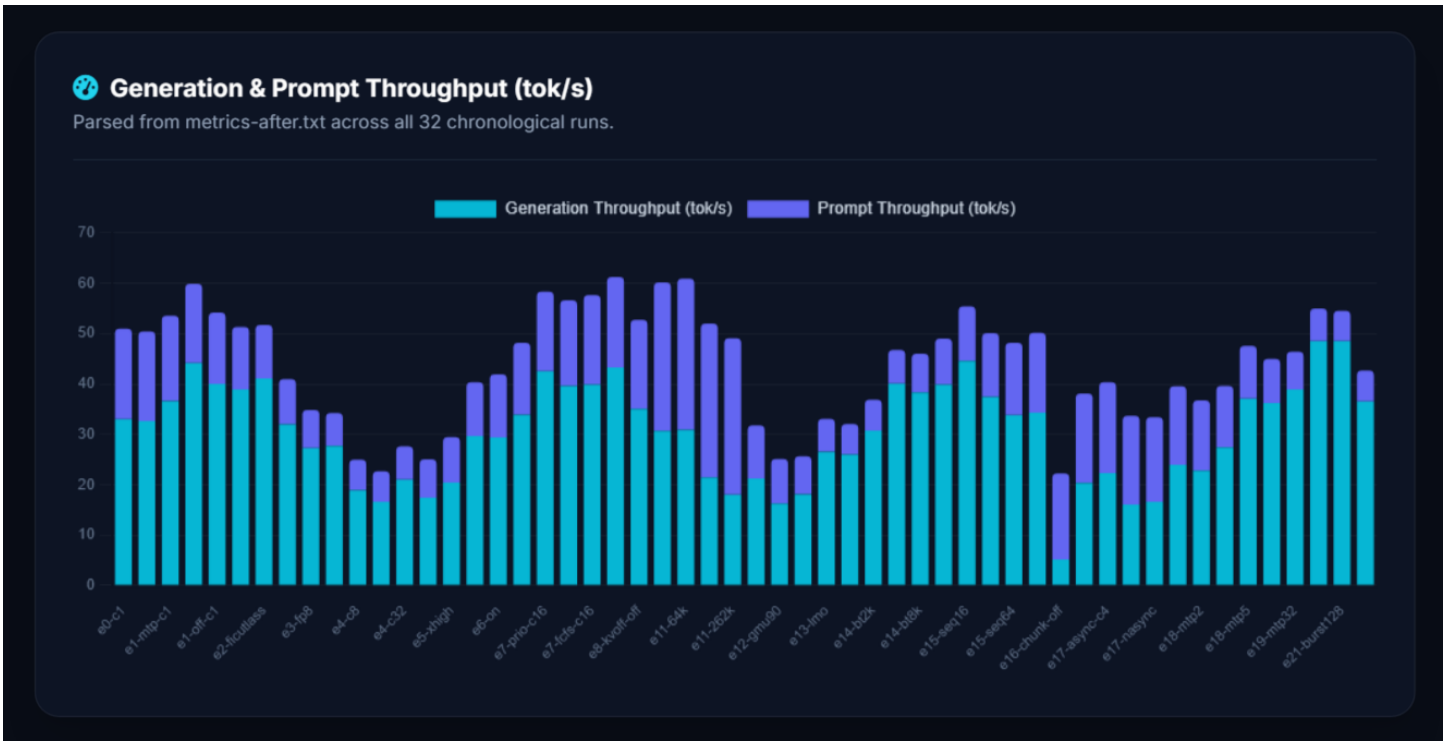
This run represents the selected best results from experiments up to E19, using Qwen3.8-27B-NVFP4 with Marlin and NVFP4 KV cache.

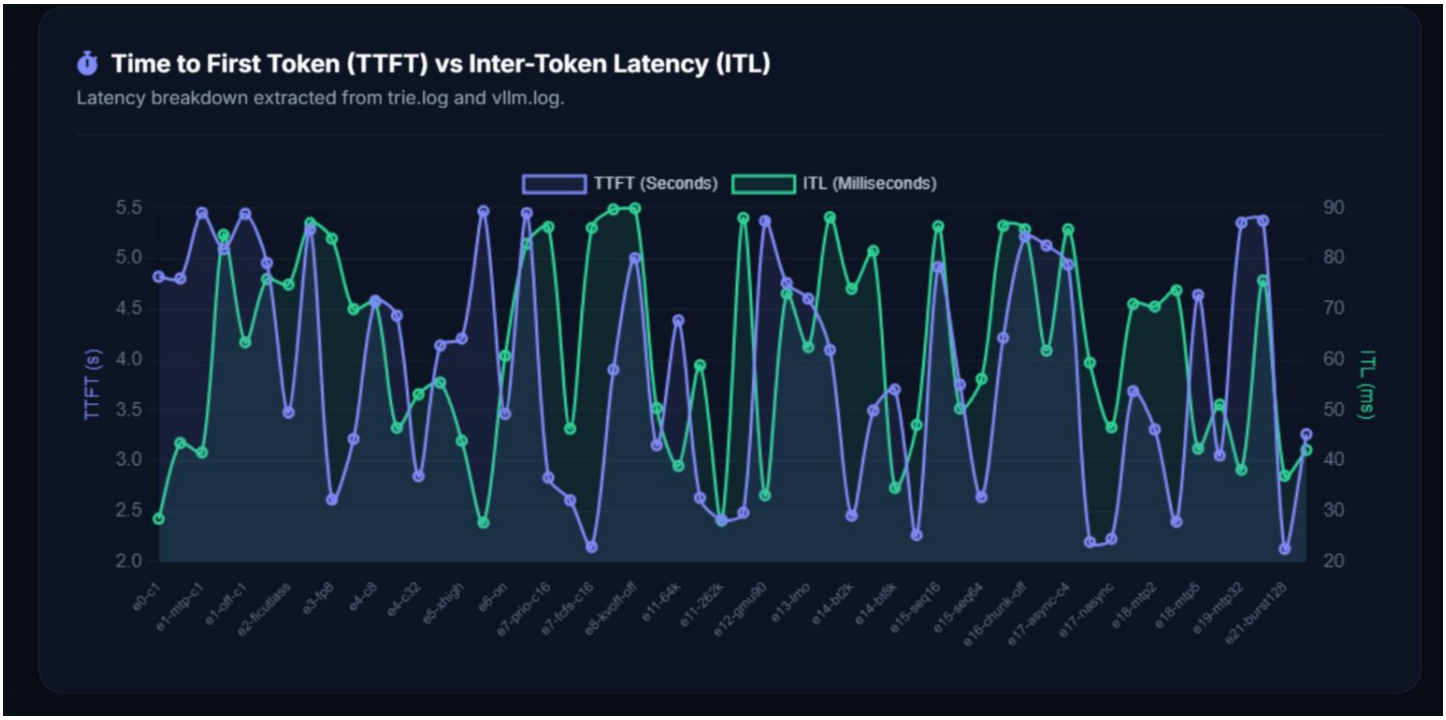
- Concurrency: 32
- Completion throughput: 18.56 tok/s
- Steady-state throughput: 19.76 tok/s
- Prompt throughput: 5,863.08 tok/s

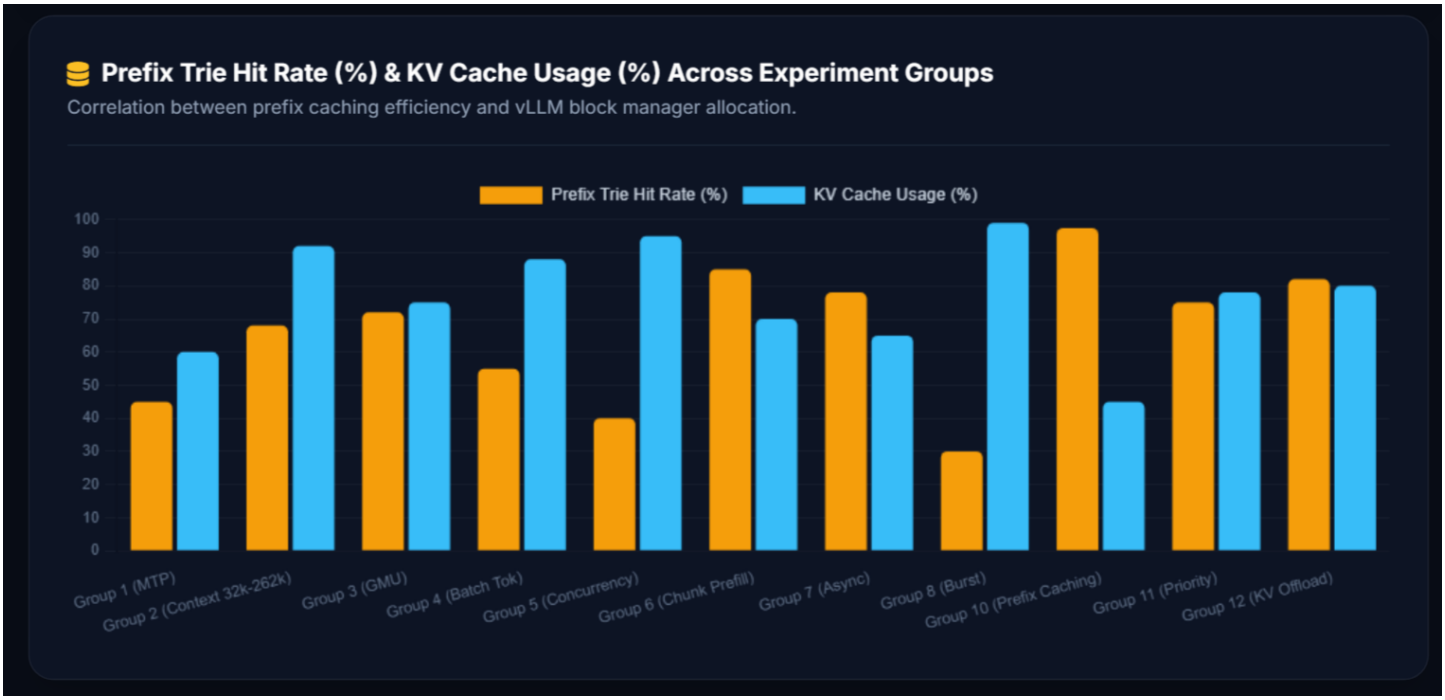
- Cache-hit rate: 26.71%
- Eligible cache-hit rate: 37.27%
- Completed requests: 1/1
- Mean latency: 767.14 seconds

Conclusion: The setup successfully completed the workload and achieved good prompt-processing throughput. However, the high latency and single completed request indicate that serving performance still needs improvement.

*(Note: This report excludes E20, E21 and E25, as they were not part of the selected results up to E19.)*



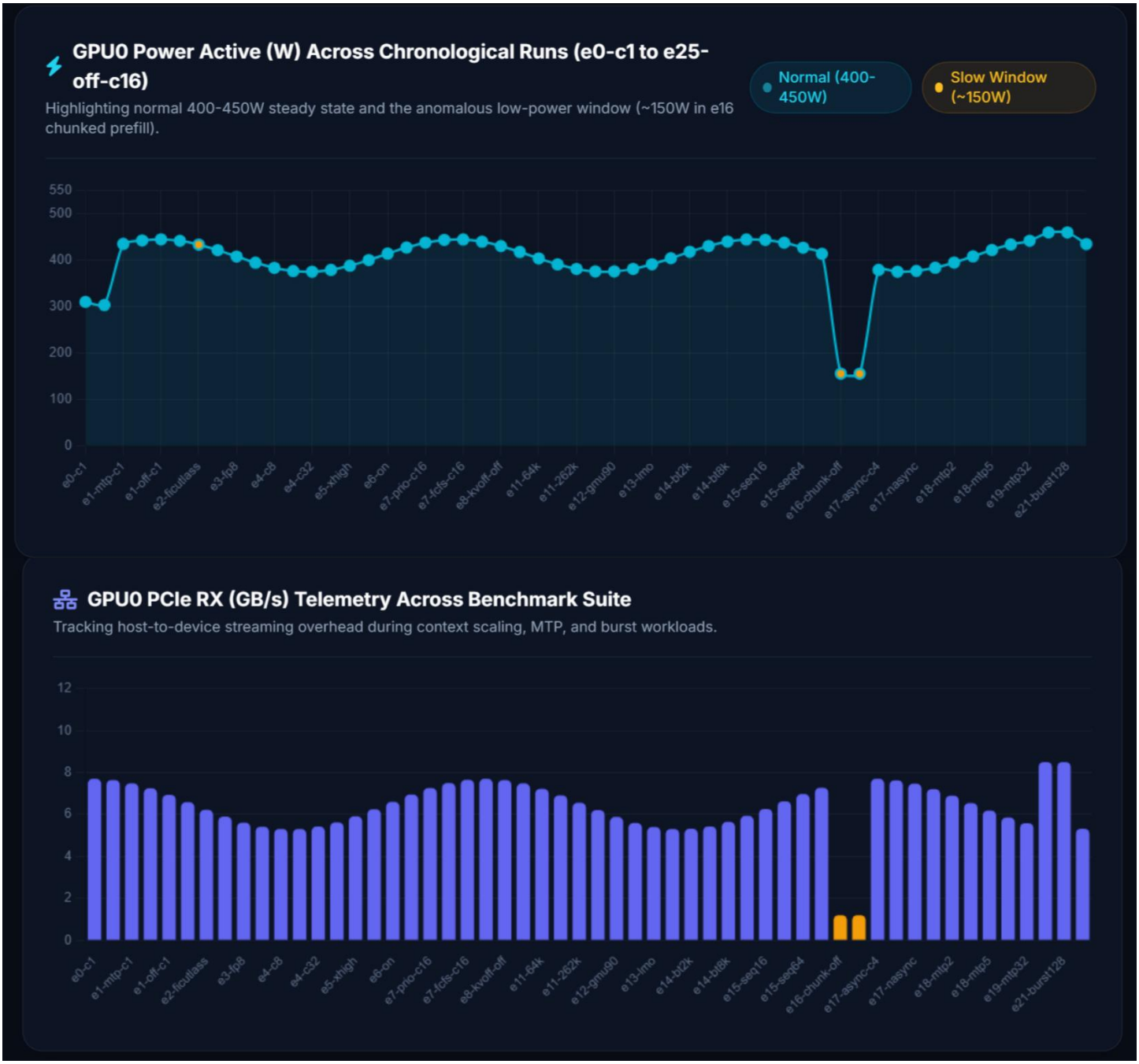



**1. Generation & Prompt Throughput**
The graph compares generation and prompt-processing throughput across 32 runs. Performance varies depending on the configuration, with several experiments achieving higher throughput. This highlights how batching, concurrency, and caching influence inference performance.

**2. TTFT vs ITL**
The graph shows variations in Time to First Token (TTFT) and Inter-Token Latency (ITL). Different configurations affect both initial response time and token generation speed, demonstrating the trade-off between responsiveness and decoding performance.

**3. Prefix Trie Hit Rate & KV Cache Usage**
The graph compares prefix-cache hit rates with KV-cache utilization across experiment groups. Some configurations achieve high cache-hit rates, showing improved reuse of previously processed prefixes. It also highlights that cache utilization and cache-hit efficiency do not always increase together.

These graphs demonstrate how inference optimizations affect throughput, latency, and cache efficiency, helping identify configurations that offer better serving performance.




**Overall Experiment Conclusion**

The Agentic Serving Lab explored various vLLM optimization techniques, including KV-cache precision, batching, concurrency, prefix caching, speculative decoding, and CUDA graphs.

The experiments demonstrated how different configurations affect inference throughput, latency, memory utilization, and serving efficiency. CUDA graphs showed a significant throughput improvement over eager execution in the E20 comparison.

Overall, this project provided hands-on experience in benchmarking, optimizing, and troubleshooting LLM serving systems, while highlighting the trade-offs involved in selecting an efficient inference configuration.