HW4

Tguo67

Question(s):

Read the paper **Taming Throughput-Latency Tradeoff in LLM Inference with Sarathi-Serve**, summarize the paper, specifically, including the points below:

- What are the motivations/challenges of this work?
- How does the design of this paper address the challenges?
- How does the paper evaluate its design (experiment settings, workloads, metrics)?
- How does the evaluation prove its claims?

Related link:

Paper of Sarathi-Serve: https://www.usenix.org/system/files/osdi24-agrawal.pdfLinks to an external site.

Answers:

What are the motivations/challenges of this work?

Prefill is long/compute-bound; decode is short/memory-bound. Existing schedulers pick one: prioritize prefills (good throughput, but multi-second generation stalls and bad tail TBT) or prioritize decodes (good TBT, poor throughput). Mixed prefill/decode also creates PP bubbles.

How does the design of this paper address the challenges?

(1) **Chunked-prefills**: split prefills into near-equal compute chunks, bounding per-iteration latency;

(2) **Stall-free batching**: always pack all ongoing decodes first, then admit limited prefill chunks/new requests under a per-iteration token budget chosen to meet a P99-TBT SLO;

(3) More uniform compute reduces PP bubbles. Implemented atop vLLM/PagedAttention.

How does the paper evaluate its design (experiment settings, workloads, metrics)?

- Models & HW: Mistral-7B (1×A100-80GB), Yi-34B (2×A100 TP-2), LLaMA2-70B (8×A40 TP-4 + PP-2), Falcon-180B (8×A100 across 2 nodes, TP-4 + PP-2).
- Workloads: Poisson arrivals using two real-ish traces
  - openchat_sharegpt4: multi-turn chats; shorter prompts (median ~1.7k) with higher variance.
  - arxiv_summarization: long documents; much longer prompts (median ~7.1k).
- Baselines: vLLM and Orca (rep. SoTA schedulers).
- Metrics:
  - Latency: TTFT (median) and TBT (P99).
  - Capacity/Throughput: max sustainable QPS under strict vs relaxed P99-TBT SLOs (SLOs set to 5× and 25× the isolated decode step time respectively).
- SLO-driven token budgets: e.g., 512 for strict, 2048 for relaxed (with an exception for 70B to control bubbles).

How does the evaluation prove its claims?

Capacity gains up to 2.6× (Mistral-7B), 3.7× (Yi-34B), and ~4.3–5.6× with PP on Falcon-180B; ~3.5× higher capacity under strict TBT SLOs vs vLLM by eliminating stalls. Ablations: chunking+hybrid together gives best TTFT & TBT; chunking overhead modest (≤~25% at 512-token chunks, near-zero at 2048). Also shows TP across nodes hurts latency; Sarathi-Serve makes PP viable on commodity networks.