Not All Prefills Are Equal:
PPD Disaggregation for Multi-turn LLM Serving

A dynamic routing system that cuts Turn 2+ time-to-first-token by ~68% on real multi-turn workloads.

Zongze Li Jingyu Liu Zhen Xu Yineng Zhang Tahseen Rabbani Ce Zhang

University of Chicago · Independent Researcher

zongzel@uchicago.edu

Multi-turn conversations are now the dominant LLM serving workload.

3.1

turns per conversation (average)

Observed average on WildChat & ShareGPT.

Chatbots, agentic systems, and tool-use loops generate sessions with many follow-up turns rather than isolated queries.

Each follow-up turn arrives carrying the entire prior conversation as its prefix — and current serving systems were not designed for that.

Standard prefill–decode disaggregation pays two hidden costs in multi-turn serving.

1. WASTED COMPUTE

Every turn re-prefills the full history.

KV transfer is strictly P→D; the response cached on D is invisible to P, so the prefill node recomputes the entire conversation history every turn.

Recomputation is up to 99% of multi-turn prefill cost.

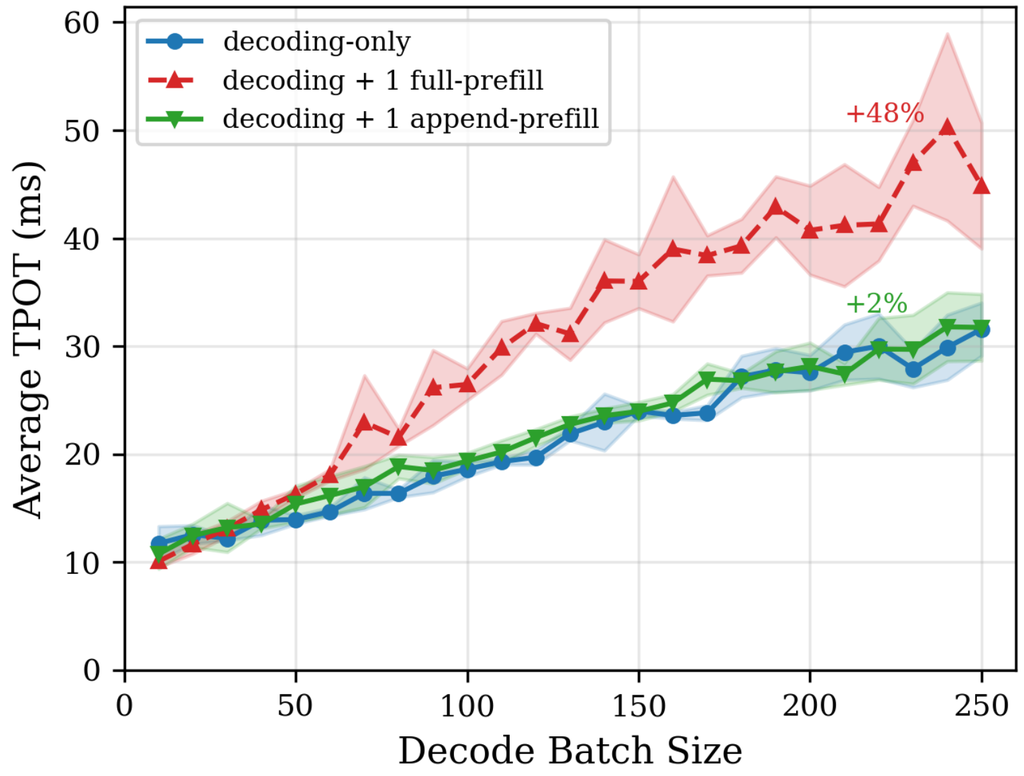
2. WASTED BANDWIDTH

Repeated KV transfers saturate the network.

Each turn ships a freshly recomputed KV cache from P to D. For Llama-3.1-8B with a 2K context, that is ~256 MB per transfer.

Under load, the P→D link queues up and the service degrades.

Append-prefill causes an order of magnitude less decode interference than full prefill.



Llama-3.1-8B on a single H100. Decode TPOT when co-located with one 1,024-token prefill.

+48%

Full prefill

new prompt, no cached context

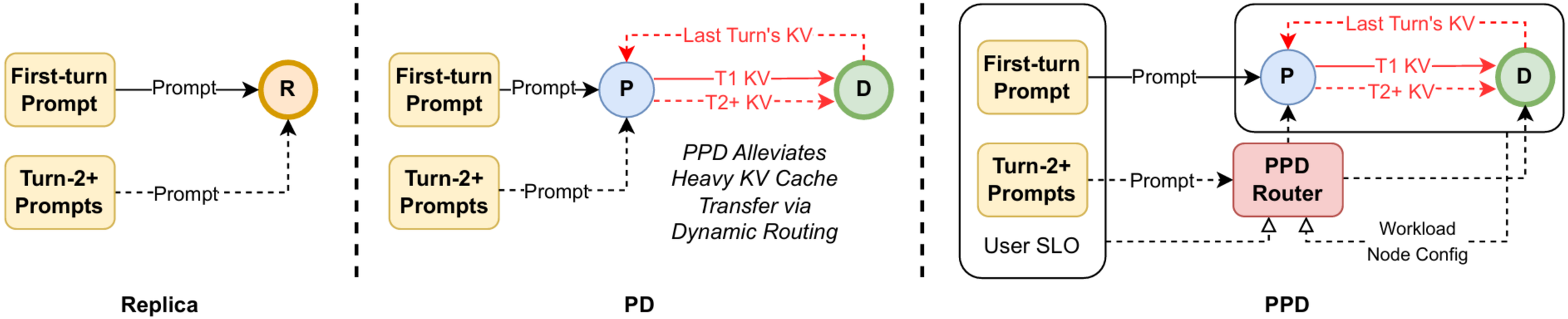
+2%

Append-prefill

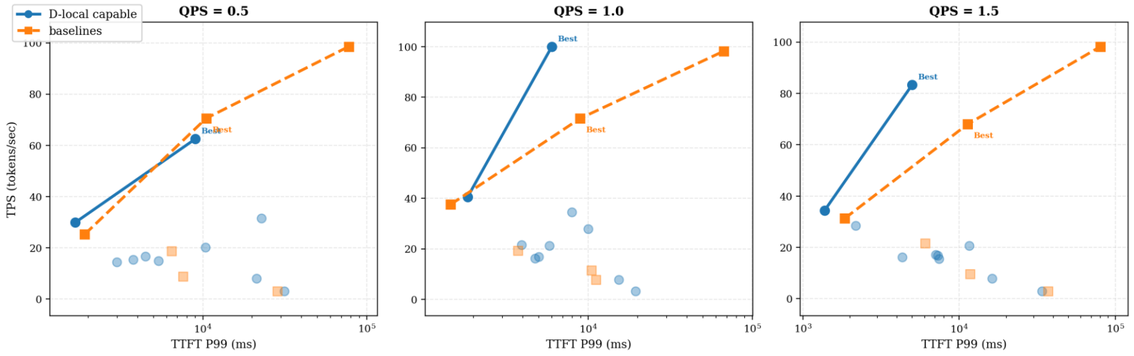
new tokens on cached KV

An order-of-magnitude gap. Decode nodes can safely absorb append-prefill — the core assumption that motivates physical PD isolation simply does not apply to Turn 2+.

Routing Turn 2+ append-prefill locally on decode nodes eliminates both costs.



No single static routing fraction wins across all service-level objectives.



P99 TTFT vs. throughput across three QPS levels. Blue: D-local-capable; orange: PD baselines.

In 92.2% of (workload, QPS) cells, the configuration that minimises Turn 2+ TTFT is not the one that minimises TPOT.

Static **x = 0** (classic PD) wins TPOT — but never wins Turn 2+ TTFT.

Static **x = 1** (always local) wins throughput — but pays TPOT under prefill-heavy traffic.

The routing choice must be made per request, not fixed system-wide.

PPD turns the routing choice into a per-request decision with one operator-facing knob.

For each Turn 2+ request, score local processing (x = 1) against the PD path (x = 0) using two SLO weights chosen offline.

Classic PD is recovered as the special case x ≡ 0; all-local is recovered as x ≡ 1.

PER-REQUEST SCORE

$$S(\psi) = w_{ttft} \cdot \Delta_{ttft} - w_{tpot} \cdot \Delta_{tpot}$$

Route locally when S > 0; otherwise forward to P.

ALGORITHM

Offline table, online lookup.

Benchmark TTFT and TPOT at x = 0 and x = 1 across a coarse workload grid.
Store the sign of the score per grid cell.
Online: snap each Turn 2+ request to its cell and return the precomputed route.
Turn 1 always uses the standard PD path (no cached KV yet).

< 1 ms decision overhead. Drop-in on vLLM disaggregated serving.

PPD cuts Turn 2+ TTFT by ~68% and restores service stability on real workloads.

−68 % Turn 2+ TTFT

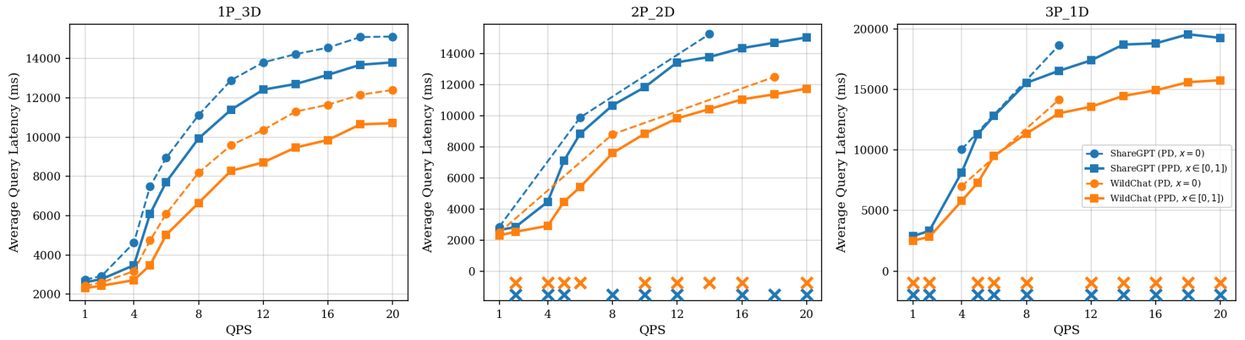
Average reduction across ShareGPT and WildChat.

100 % success rate

At every QPS level tested — where two of three PD baselines time out.

~3× less KV traffic

At the observed 3.1 turns per conversation.



Average query latency vs. QPS. Dashed = PD (x = 0); solid = PPD.
× markers mark success rate < 95%.

Conclusions: where you compute append-prefill matters as much as how you compute it.

Append-prefill is not full prefill. Order-of-magnitude less decode interference makes Turn 2+ on decode nodes architecturally viable.

No static routing wins everywhere. Workload, SLO, and the P:D ratio all shift the optimum; the decision belongs at the per-request level.

PPD is a single, monotonic operator knob. Classic PD and all-local routing are recovered as the extreme weight settings; ~68% Turn 2+ TTFT reduction and 100% success rate in real-world evaluation.

Drop-in. Built on vLLM's existing disaggregated serving — no protocol changes; composes with distributed KV cache layers.

CONTACT & MATERIALS

Zongze Li · zongzel@uchicago.edu

Scan for the paper and code release.