# Online Linear Programming for Multi-Objective Routing in LLM Serving

Zixi Chen[1], Yinyu Ye[2] , Zijie Zhou[2]

[1]School of Mathematical Sciences, Peking University, [2]Department of Industrial Engineering and Decision Analytics

In large language model serving, we introduce a multi-objective optimization framework that formulates routing as an online linear programming with interpretable decision rewards. We apply an efficient bid-price control policy based on the online linear programming that admits requests when their SLO-weighted benefit exceeds their shadow prices. To meet millisecond decision requirements, we develop a warm-started, projected first-order updates that track the evolving dual shadow prices online with predictable runtime.

## Problem

In large language model serving

- Heterogeneous SLOs
- Batch / KV-cache constraints
- Heuristics lack trade-off control

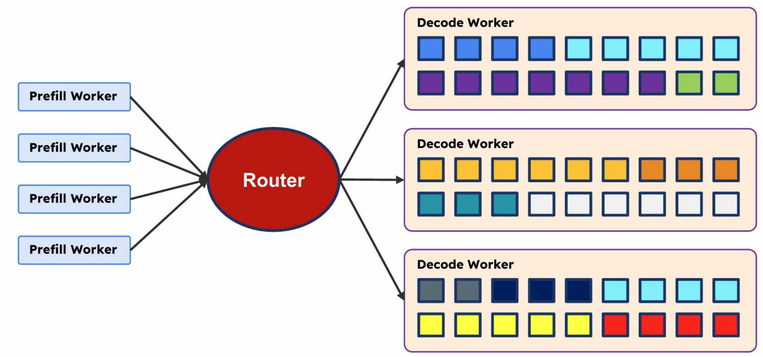



Figure. Illustration of Prefill/Decode (P/D) disaggregation with continuous batching.

## Method

- Incoming request
  ↓
- Compare SLO reward
  ↓
- Shadow price cost
  ↓
- Accept / delay + choose GPU

The bid-price margin

$$\Delta_j(g) := r_{j,g,k} - a_{(j,g,k)}^{\top} p$$

Accept if SLO value exceeds resource opportunity cost.

The corresponding online linear programming formulation at time $k$ is:

$$\max_{x} \quad \sum_{(j,g,k)\in\mathcal{A}_k} r_{j,g,k} x_{j,g,k}$$

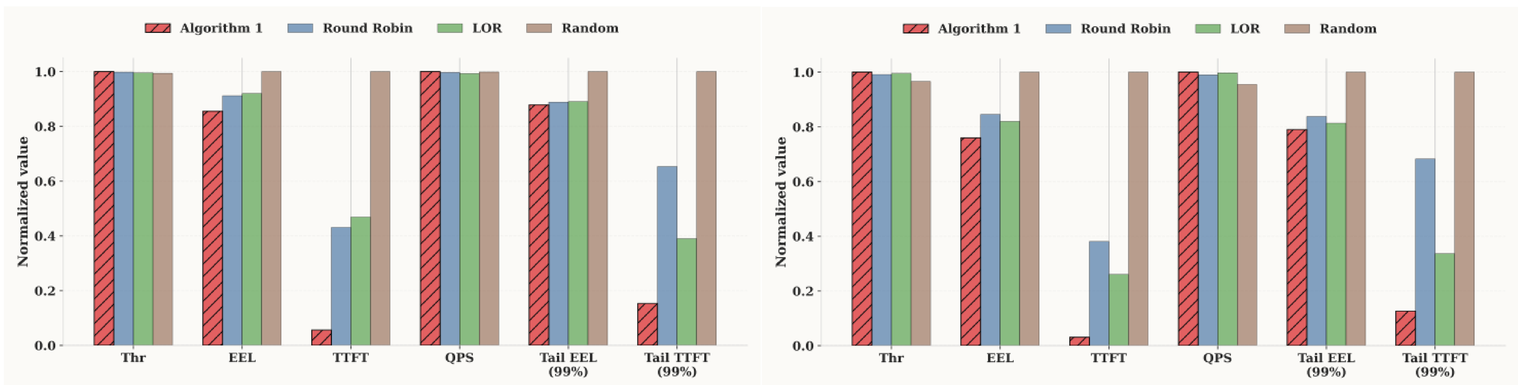
$$\text{s.t.} \quad \sum_{(j,g,k)\in\mathcal{A}_k} a_{i,(j,g,k)} x_{j,g,k} \le b_i, \quad \forall i \in \mathcal{I}, \qquad (4)$$

$$\sum_{g\in[G]} x_{j,g,k} \le 1, \quad \forall j \in \mathcal{J}_k^{(wait)}. \qquad (5)$$

*Table 1.* Relative improvement (%) over Round-Robin with noisy decode-length prediction $\hat{o}_j \sim \text{Unif}(0.8o_j, 1.2o_j)$.

| Method | Avg EEL↓ | P95 EEL↓ | P99 EEL↓ | Thr.↑ | SLO↓ |
|---|---|---|---|---|---|
| LOR | 0.67 | 4.01 | 6.19 | 0.53 | 0.98 |
| Power-of-2 | 1.30 | 3.84 | 6.06 | 0.73 | 1.84 |
| **Ours** | **45.75** | **42.49** | **25.85** | **0.90** | **11.10** |

**Takeaway:** Online LP accelerates LLM routing using shadow-price control, improving performance under resource constraints.




$(\alpha, \beta, \gamma, \sigma, \zeta_1, \zeta_2) = \left(\frac{1}{135}, \frac{1}{400}, \frac{1}{400}, 1,1,1\right)$ $(\alpha, \beta, \gamma, \sigma, \zeta_1, \zeta_2) = \left(\frac{1}{27}, \frac{1}{2048}, \frac{1}{2048}, 5,1,1\right)$

Figure. Real-data comparison between routing policies under two representative multi-objective settings.

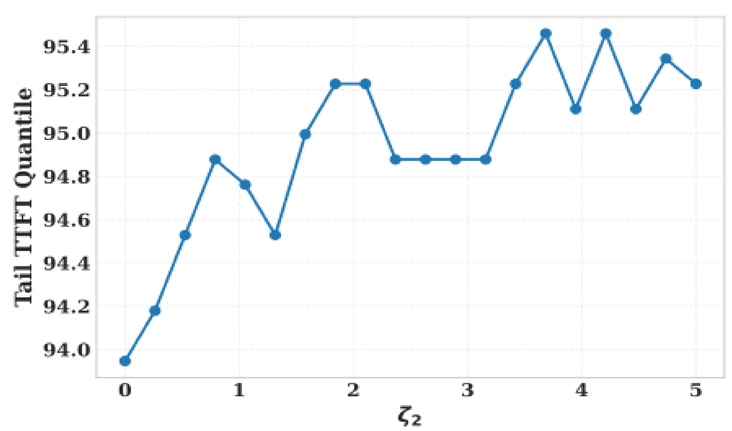



Figure. TTFT satisfaction rate as a function of the tail weight $\zeta_2$.