ICML 2026

POSITION PAPER

*Seoul · July 6–11, 2026*

# LLM Serving Needs Mathematical Optimization & Algorithmic Foundations — Not Just Heuristics

**Zijie Zhou** · Industrial Engineering & Decision Analytics, HKUST · *jerryzhou@ust.hk*

Topic

**LLM Inference Serving**

**Algorithmic Frontier**

*Optimization · Queueing*
*Online Algorithms · OR*

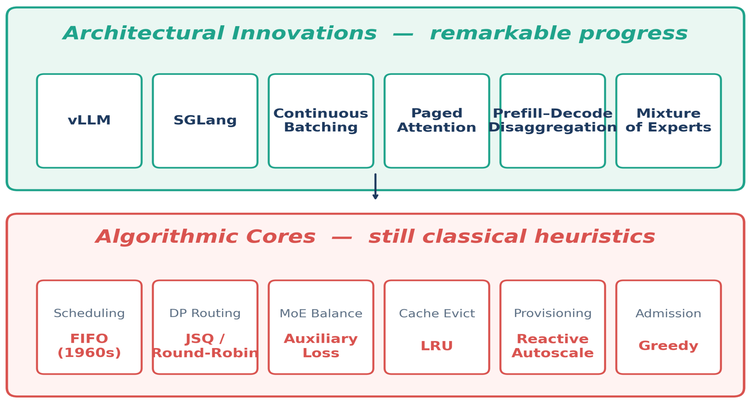
## THE GAP

### Our Position

*Architecturally, LLM serving has changed enormously.*
***Algorithmically, it still runs on classical heuristics.***

The field must develop **mathematical models and algorithms with provable guarantees** — tailored to LLM-specific structure: *dynamically growing KV memory, prefill–decode asymmetry, unknown output lengths, continuous batching.*

### Today's Serving Stack




### Four Decision Problems · hiding LLM structure

**MoE Expert Routing**

*Imbalanced experts → straggler GPUs*

Today: **auxiliary-loss balancing**

**Data-Parallel Routing**

*Barrier sync → idle GPUs*

Today: **round-robin · JSQ**

**Scheduling & Capacity**

*Unknown output, growing memory*

Today: **FIFO + reactive autoscale**

**Cache Eviction**

*Heterogeneous sizes & re-encode costs*

Today: **LRU**

## WHY THIS MATTERS

### Four Benefits Heuristics Cannot Provide

**Robustness**

Worst-case guarantees hold under workload drift, viral spikes, and adversarial inputs.

**Capacity Planning**

Closed-form stability conditions specify the minimum cluster size before deployment.

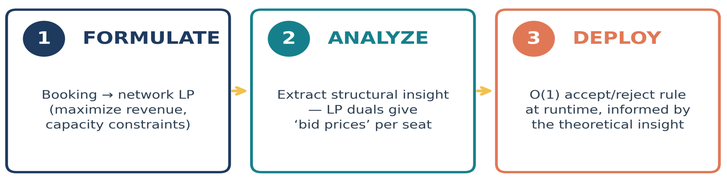
**Engineering Blueprint**

LP duals & competitive bounds reveal which constraints bind and which objectives matter.

**Calibration**

Optimality lower bounds tell engineers when to stop tuning — and where 50% gains still hide.

### A Historical Precedent · Airline Revenue Mgmt.




***American Airlines, 1980s — $1.4 B incremental revenue · Edelman Award 1991***

| Airline Revenue Mgmt. | LLM Serving |
|---|---|
| Seats on flights | GPU memory & compute |
| Booking requests | Inference requests |
| Irrevocable seat use | Sticky KV assignment |
| Unknown future demand | Unknown output length |
| Bid-prices (LP duals) | Optimization-derived thresholds & policies |

***Lesson:*** *theory's role is not as a runtime solver — but as the* ***analytical vehicle*** *that reveals the structure of good algorithms.*

## THE PATH FORWARD

### Principled Algorithms Already Working

**1 LP for MoE balancing**

*DeepSeek LPLB*

Per-batch token redistribution as an LP — ≈ 100 µs per solve, well within decode budget.

**2 Online IP for DP routing**

*Chen et al. 2026*

Adversarial guarantee: imbalance reduced by $\Omega(\sqrt{(B \log G)})$ vs. round-robin.

**3 Queueing for capacity**

*Nie et al. 2025*

Closed-form stability: $\lambda < \mu$ with $\mu = M / (\bar{b} \cdot \mathbb{E}[g(s,o)])$ — provision before deployment.

**4 Cost-aware caching**

*Zhu et al. 2023*

Least-Expected-Cost eviction matches optimal regret — up to 50× cost reduction.

### Open Frontiers

**Scheduling under prediction uncertainty**
*robustness · consistency · adversarial predictions*

**Multi-objective Pareto frontiers**
*TTFT · TPOT · throughput · fairness · energy*

**Theory of disaggregation**
*when does PD-disaggregation beat co-location?*

**Algorithmics of agentic inference**
*branching · sub-requests · variable-length pauses*

### Call to Action

**OR & algorithms:** engage deeply with systems — what binds in practice?
**Systems:** the OR toolkit reveals structure your heuristics can inherit.

***The intersection is a research frontier.***
***Let's meet there.***

**Take-away:** *LLM serving has earned the right to its own algorithmic foundations. Optimization is not just elegant theory — it is the next 10× in efficiency, robustness, and predictability.*

jerryzhou@ust.hk