

# Multi-Client Chunk Processing & Fairness

6-node hierarchical cluster, chunked responses, and concurrent clients

## Objective

Evaluate how concurrent clients and dataset size affect latency, throughput, and fairness in our 6-node Mini-3 topology using chunked Strategy responses.

## System & Experiment Setup

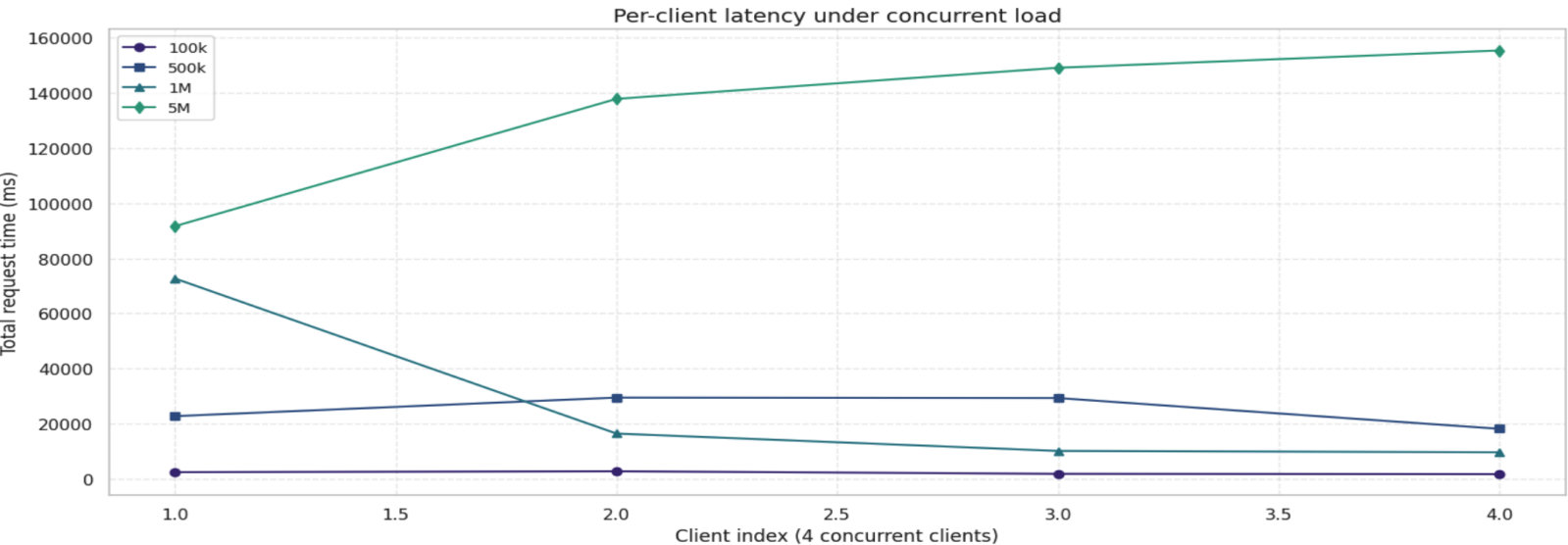
- 6 nodes: A (gateway + global leader), B/E (team leaders), C/D/F (workers).
- Chunked Strategy B (polling): clients pull chunks from a per-session buffer at A.
- Single-client baseline runs from **1K** → **10M rows** to confirm stable scaling.
- Multi-client scenarios:
  - 4 concurrent clients** on the *same* dataset: 100k, 500k, 1M, 5M rows.
  - 5 concurrent clients** on *different* datasets: 100k, 200k, 500k, 1M, 5M rows.

## Key Findings

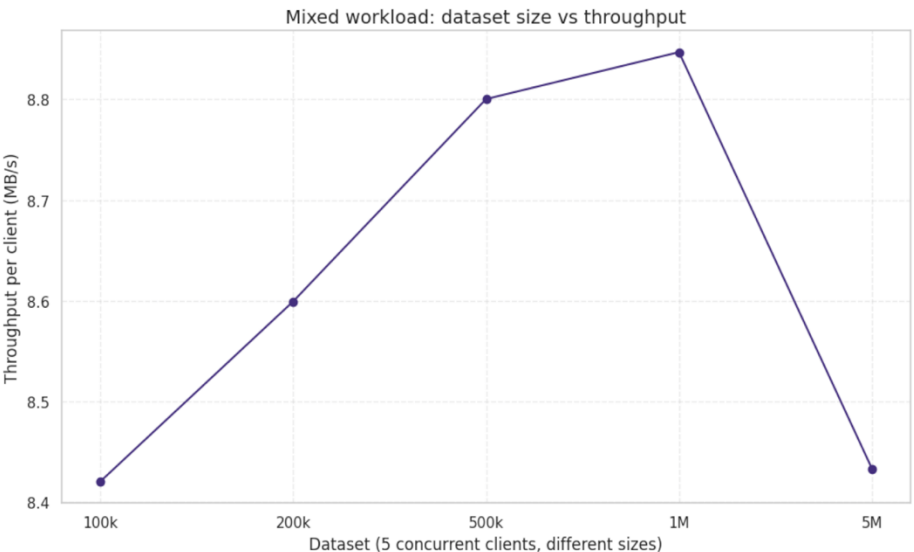
- Fairness collapse under concurrency:** For 1M rows, per-client latency ranges from ~9.7 s to ~72.6 s; for 5M rows, all four clients exceed 90 s and climb up to ~130 s.
- Cold vs warm clients:** The first client often pays for dataset load and index build; later clients reuse cached state and finish much faster.
- Multi-tenant interference:** In the mixed 5-client run, 100k–1M jobs sustain ~8–9 MB/s, while the 5M job drops to ~2–3 MB/s. Small jobs “steal” capacity from the large one.
- Implication:** Our current scheduling and deadline policy favors short queries and cached datasets but can starve big jobs and break fairness.

## Summary

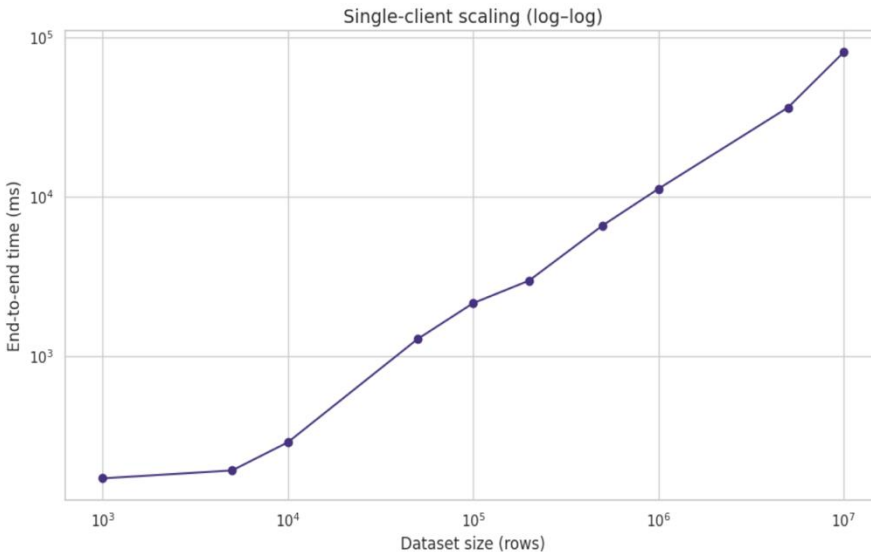
A 6-node Mini-3 hierarchy (A → B/E → C/D/F) to support multi-client concurrent chunked queries. Single-client performance scales nearly linearly from 1k → 10M rows. Under concurrency, fairness collapses: the first client often pays 70–130s load cost while later clients finish in 7–15s due to cached dataset state. Large jobs suffer partial results when deadlines fire; in mixed workloads, the 5M job is heavily starved by smaller 100k–1M jobs.



Per-client latency for 4 concurrent clients on 100k/500k/1M/5M datasets. Latency varies widely across clients for the same dataset, illustrating fairness collapse and cold vs warm client effects.



Mixed work load (5 concurrent clients, datasets 100k–5M). Smaller jobs keep high throughput (~8–9 MB/s), while the 5M job is partially served and limited to ~2–3 MB/s, showing multi-tenant interference.



Single Client Latency for different dataset sizes.

**Tested on:** 2 Windows/WSL + Ethernet | 8GB RAM (on PC-1) & 4GB RAM (on PC-2) having Cross-machine RTT: ~3-4 ms

**Tools for measurements:** chrono, gRPC timestamps, htop, Wireshark