Technical Report: Inference Latency Incident, June 2026

Executive Summary

On June 14, the production inference cluster experienced a sustained latency regression, with p99 response times rising from 180ms to 640ms over a six-hour window. Root cause was traced to a misconfigured batching window in the new sparse-attention serving path. This report summarizes the incident timeline, technical analysis, and recommended follow-up actions.

Timeline

- 09:14 — Latency alerts trigger on the primary inference cluster.
- 09:30 — On-call engineer confirms p99 latency exceeds SLO threshold.
- 10:05 — Batching configuration identified as likely cause.
- 11:40 — Rollback to previous batching parameters deployed.
- 12:15 — Latency returns to baseline; incident resolved.

Analysis

The new batching window computed a per-request wait time of

$$w = \max\left(0,\, w_{\max} - \frac{n_{\text{queued}}}{r_{\text{arrival}}}\right),$$

which under-estimated arrival rate r_{arrival} during a traffic spike, causing the scheduler to hold batches open far longer than intended and compounding queueing delay across the cluster.

Recommendations

1. Add an upper bound on batch wait time independent of estimated arrival rate.
2. Add a dedicated alert on batching queue depth, not just end-to-end latency.
3. Require a load test against production-scale traffic patterns before any batching-logic change ships.