RAG Service Batch-Queue 8-0.02 Performance Test Report

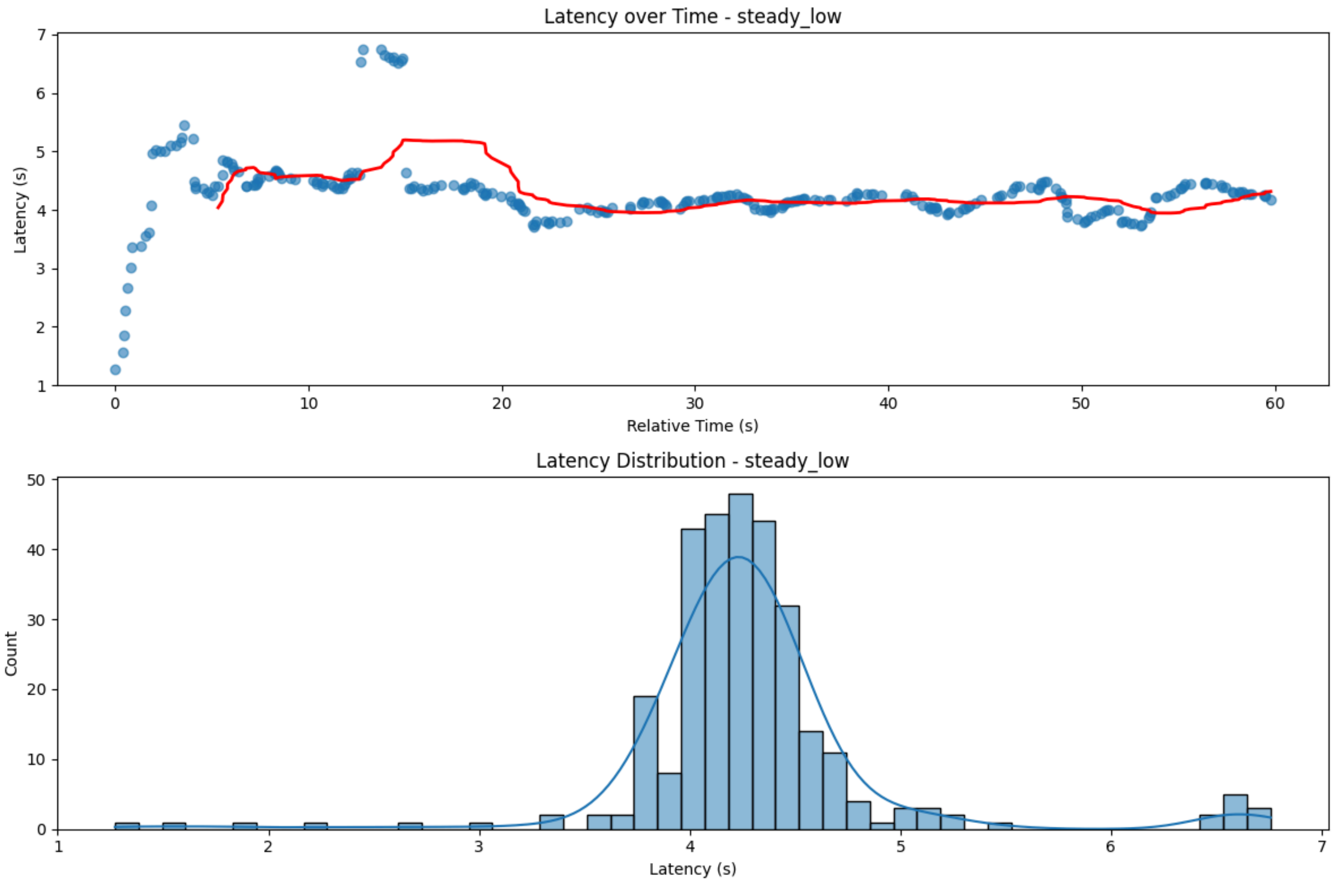
Test Date: 2025-04-02 20:27:51

Summary

Workload Pattern	Total Requests	Success Rate	Avg Latency (s)	P95 Latency (s)	Throughput (req/s)
steady_low	300	100.00%	4.2835	5.0978	5.0190
steady_high	1200	100.00%	13.3925	17.5228	20.0156
burst	1169	100.00%	13.3250	14.8331	13.0120
diurnal	3709	100.00%	14.1646	15.1038	32.4763

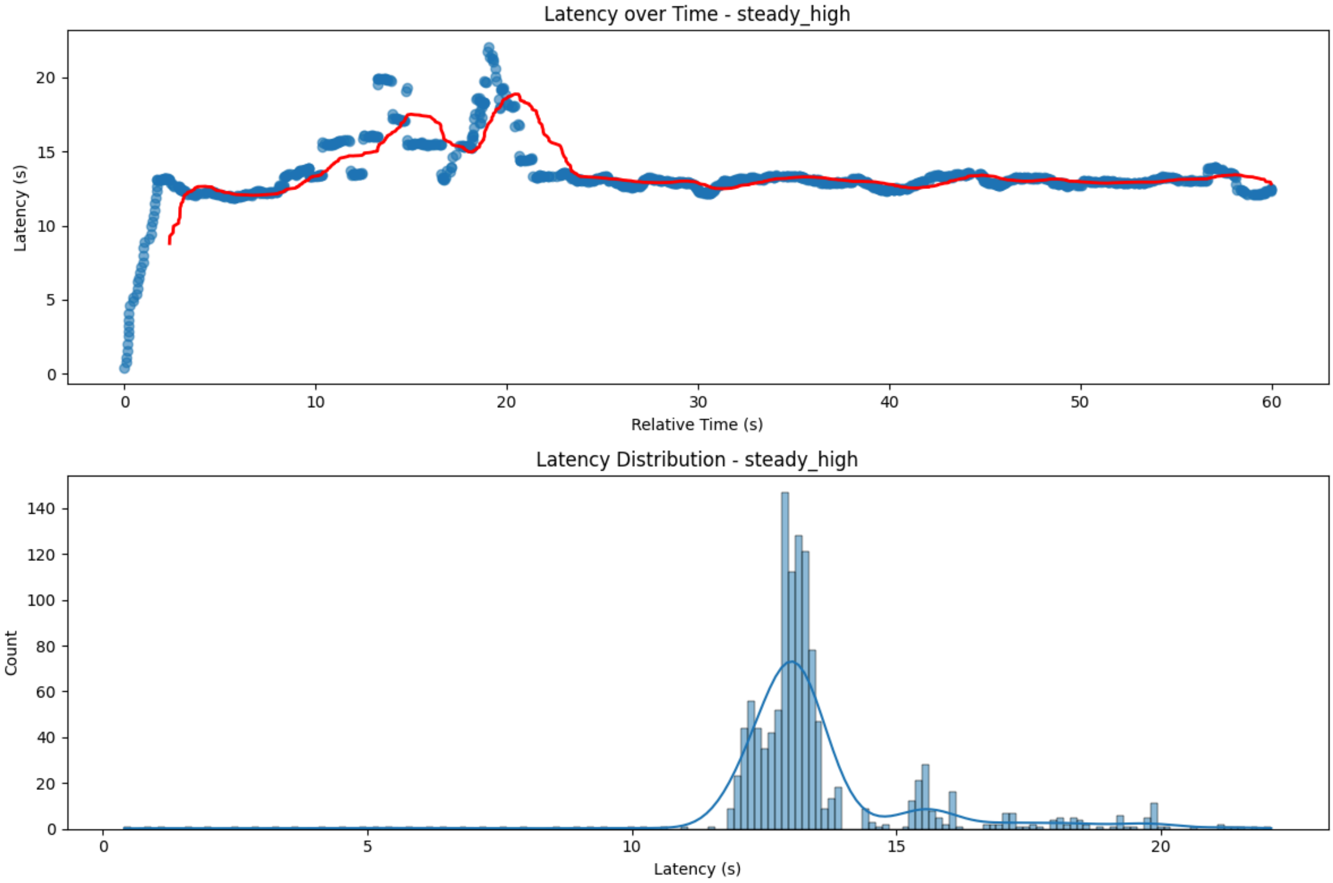
Detailed Results by Pattern

steady_low Pattern



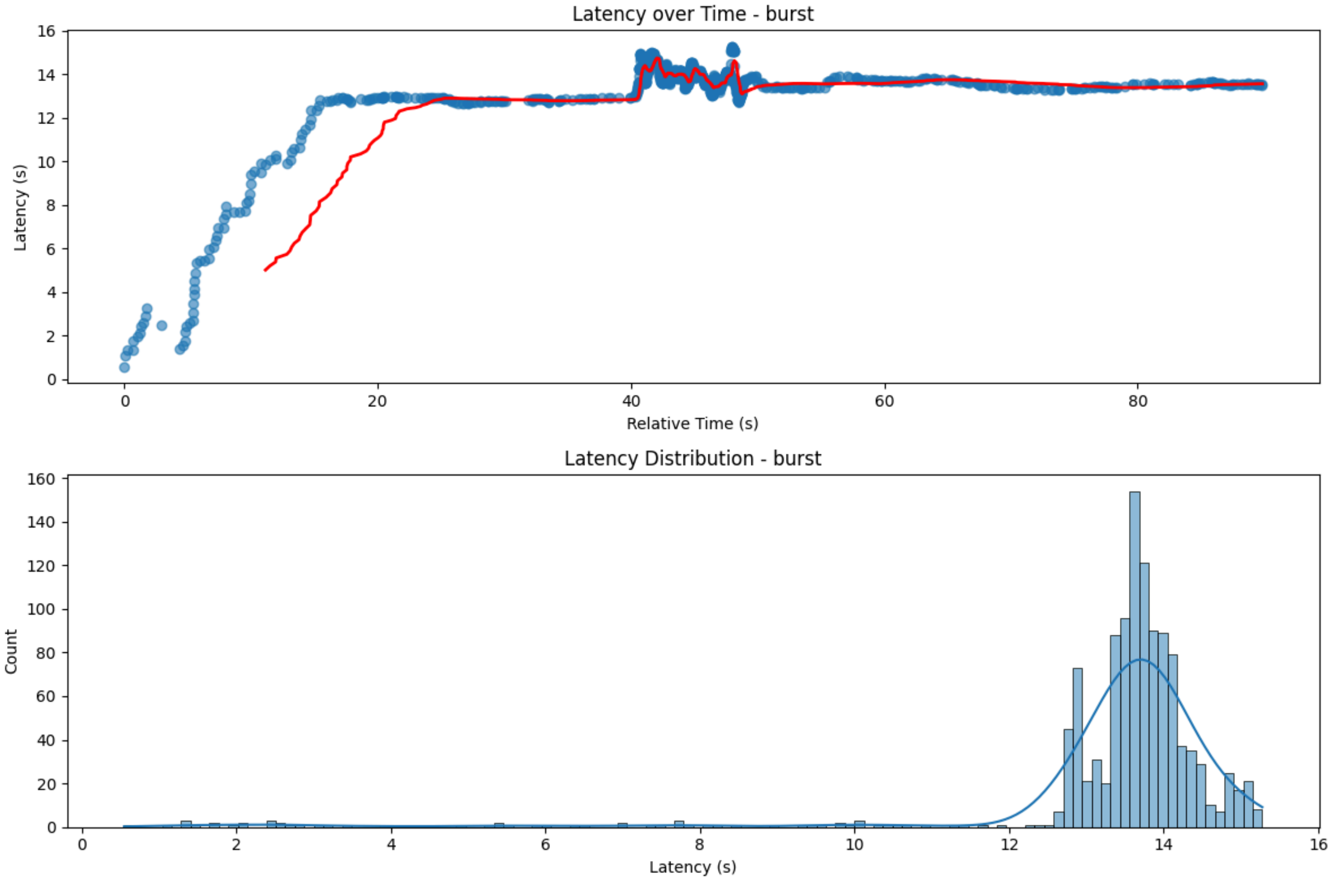
Latency distribution for steady_low workload

steady_high Pattern



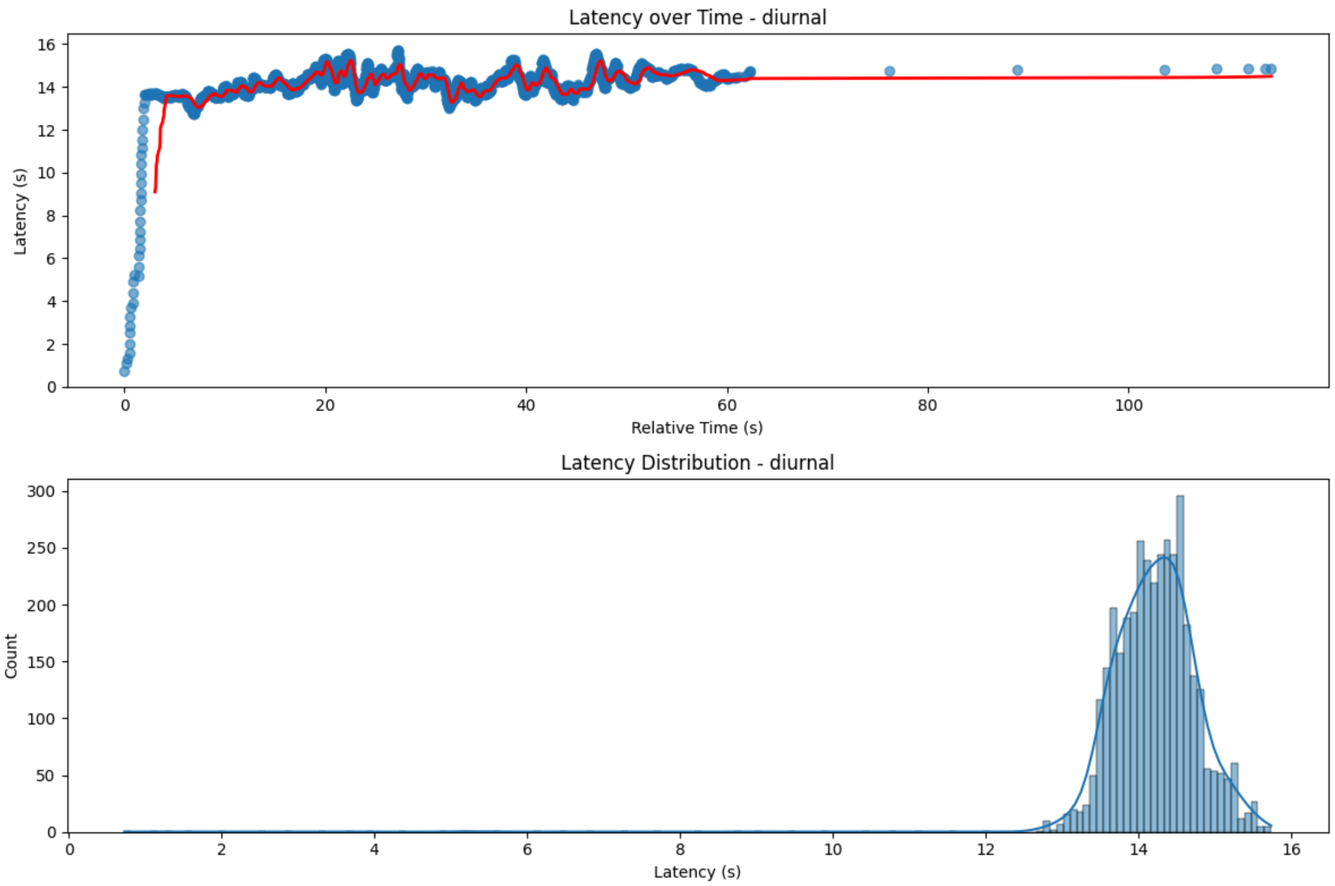
Latency distribution for steady_high workload

burst Pattern



Latency distribution for burst workload

diurnal Pattern



Latency distribution for diurnal workload

Conclusion

This report shows the performance of the Batch-Queue 8-0.02 RAG implementation across different workload patterns. These results can be used as a baseline for comparing optimized implementations.