RAG Service Base Performance Test Report

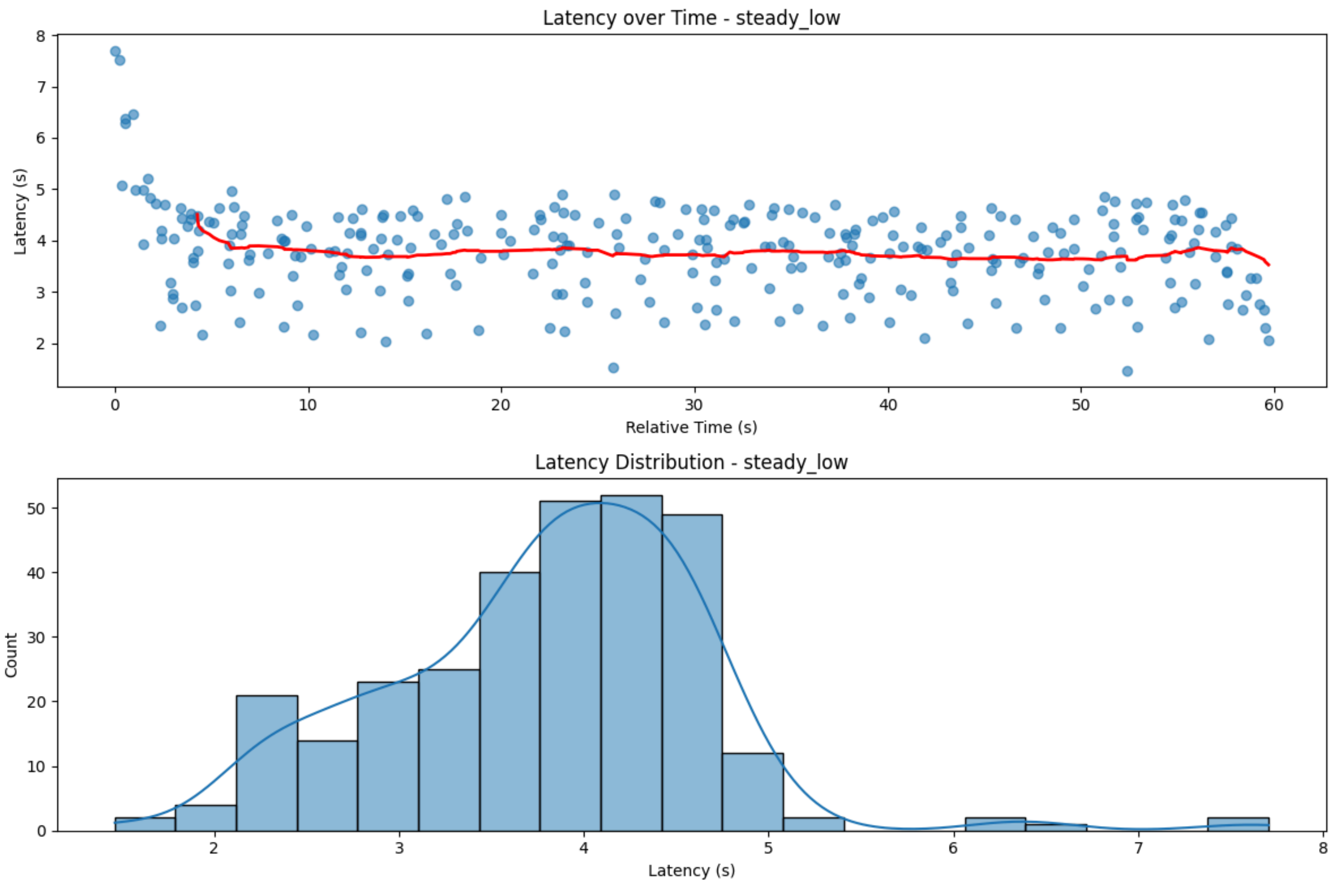
Test Date: 2025-03-25 13:36:59

Summary

Workload Pattern	Total Requests	Success Rate	Avg Latency (s)	P95 Latency (s)	Throughput (req/s)
steady_low	300	100.00%	3.7955	4.8020	5.0244
steady_high	1200	95.50%	12.8674	19.6902	19.1172
burst	1141	92.29%	12.9364	17.7489	11.8518
diurnal	3592	96.71%	13.4358	22.2544	48.1656

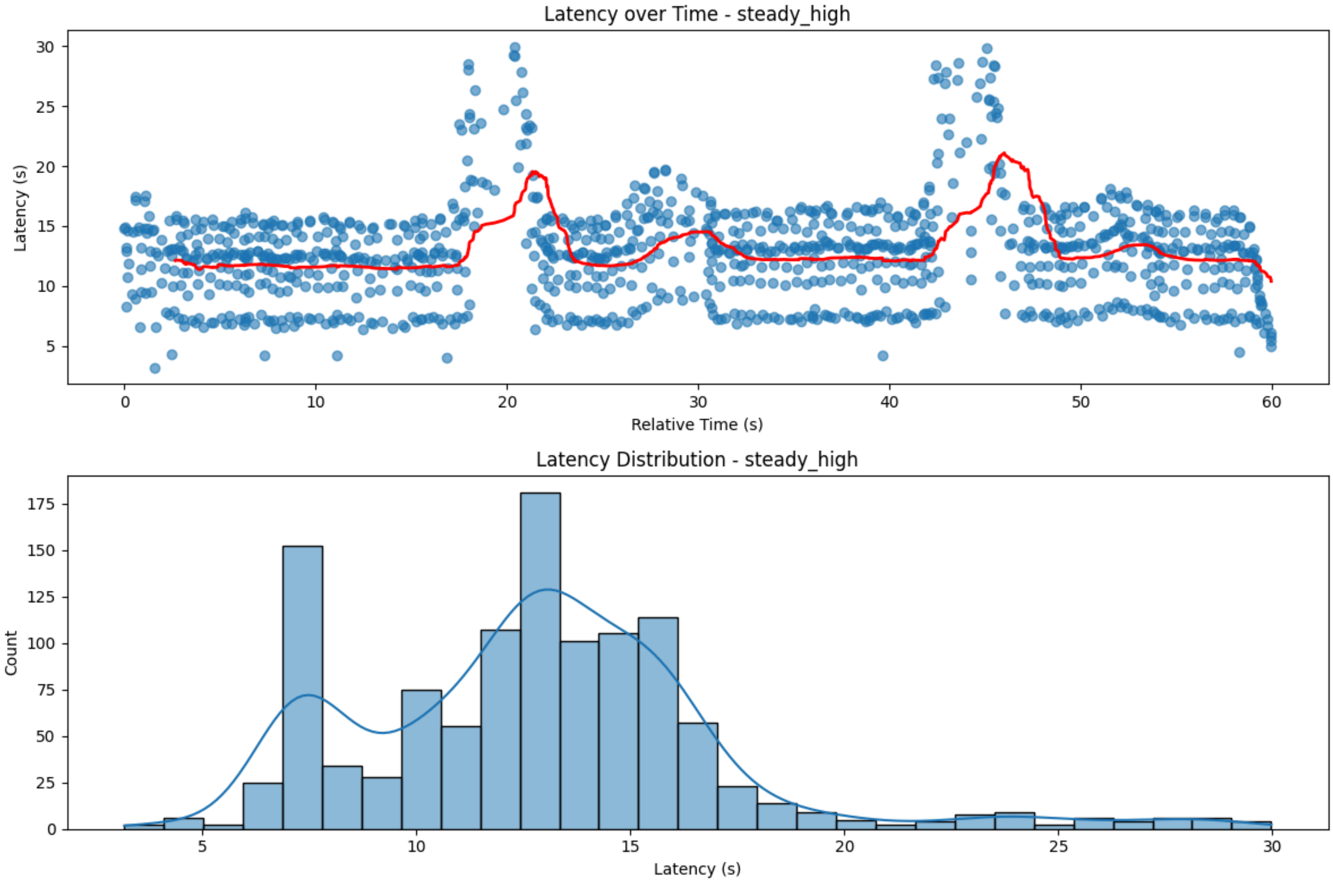
Detailed Results by Pattern

steady_low Pattern



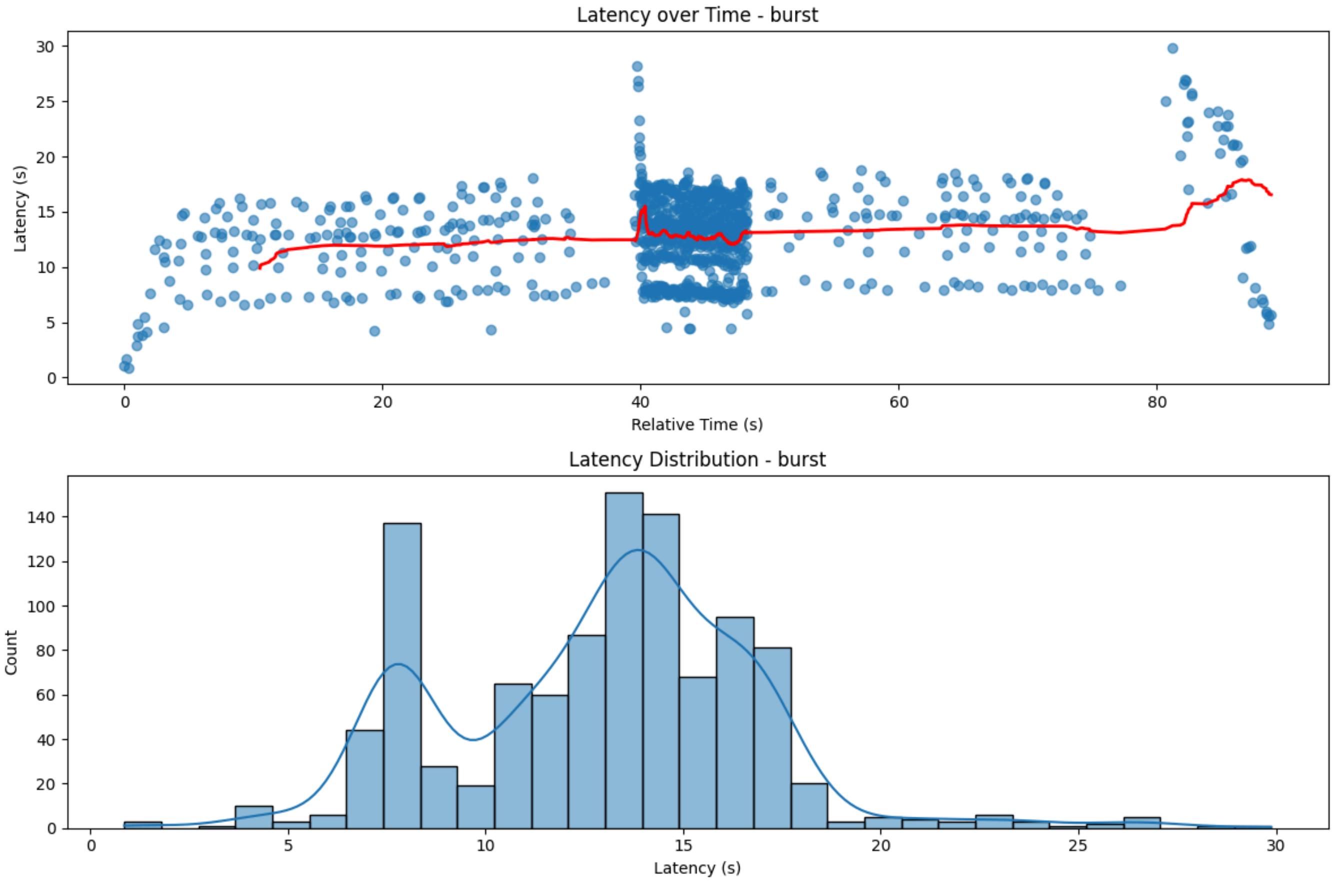
Latency distribution for steady_low workload

steady_high Pattern



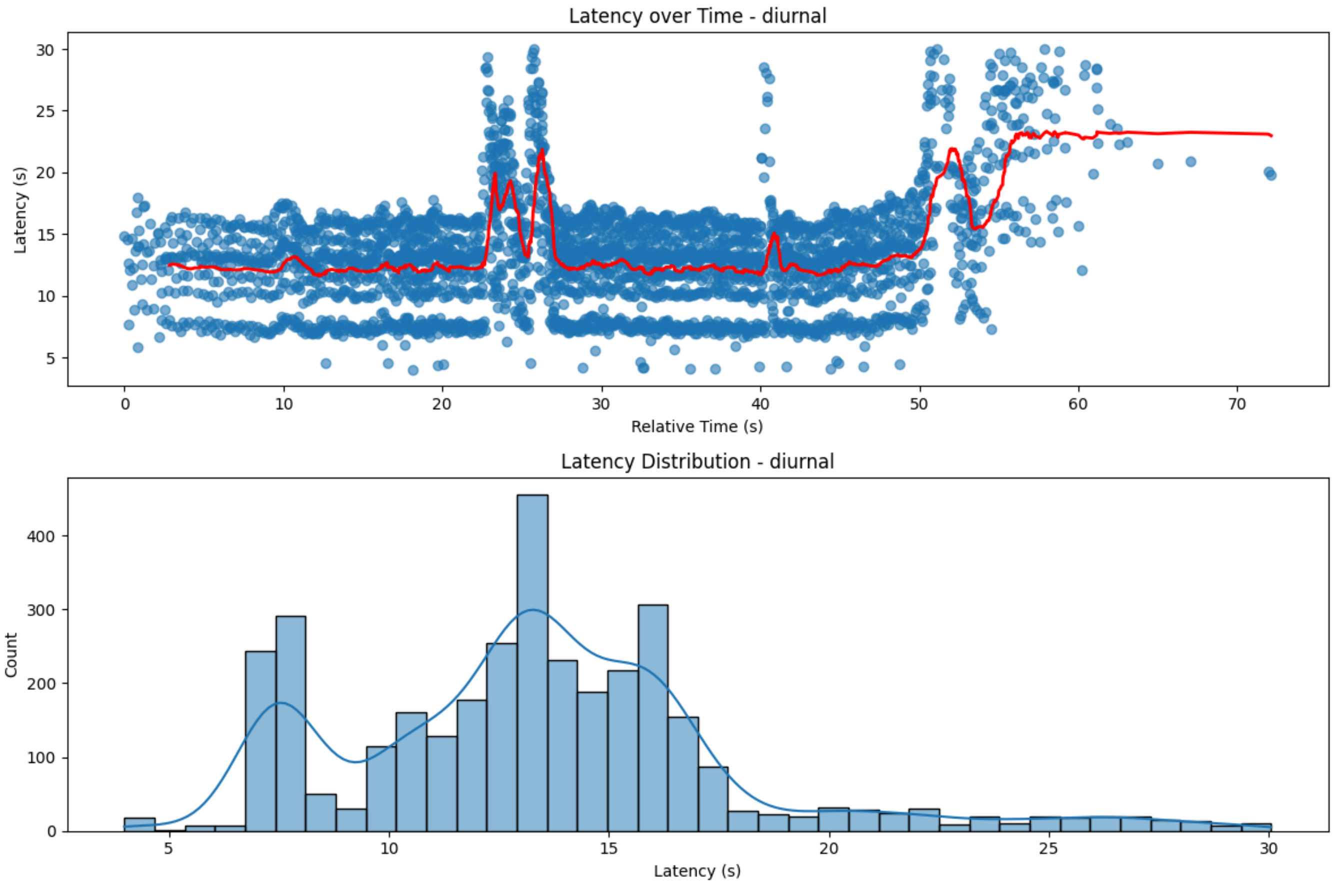
Latency distribution for steady_high workload

burst Pattern



Latency distribution for burst workload

diurnal Pattern



Latency distribution for diurnal workload

Conclusion

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