RAG Service Hugging Face 16-0.1 Performance Test Report

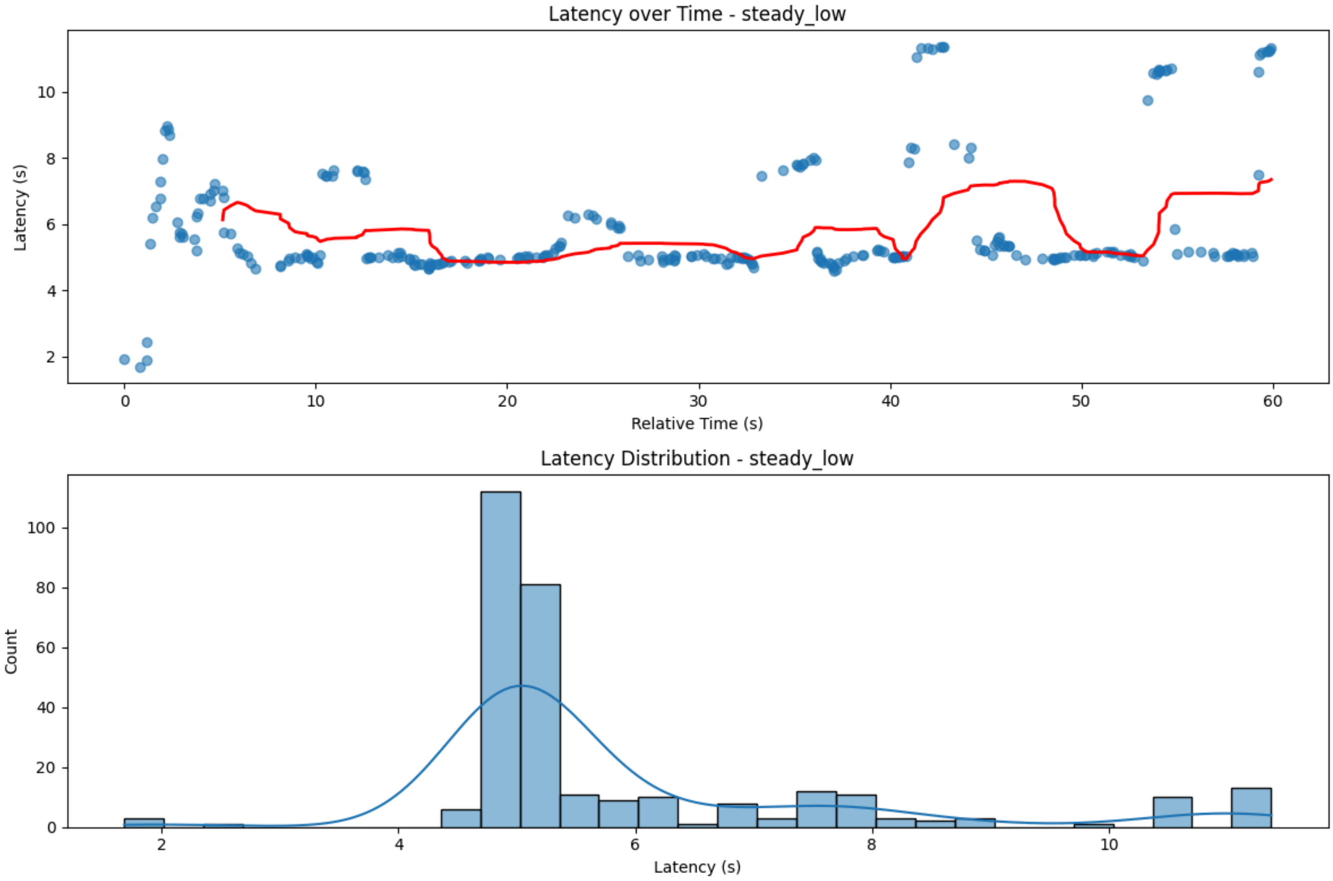
Test Date: 2025-04-03 12:59:54

Summary

Workload Pattern	Total Requests	Success Rate	Avg Latency (s)	P95 Latency (s)	Throughput (req/s)
steady_low	300	100.00%	5.8973	10.6728	5.0072
steady_high	1200	94.33%	17.5349	25.6807	18.8754
burst	1170	93.85%	15.7899	21.2934	12.3943
diurnal	3573	95.66%	16.7743	25.8764	32.1355

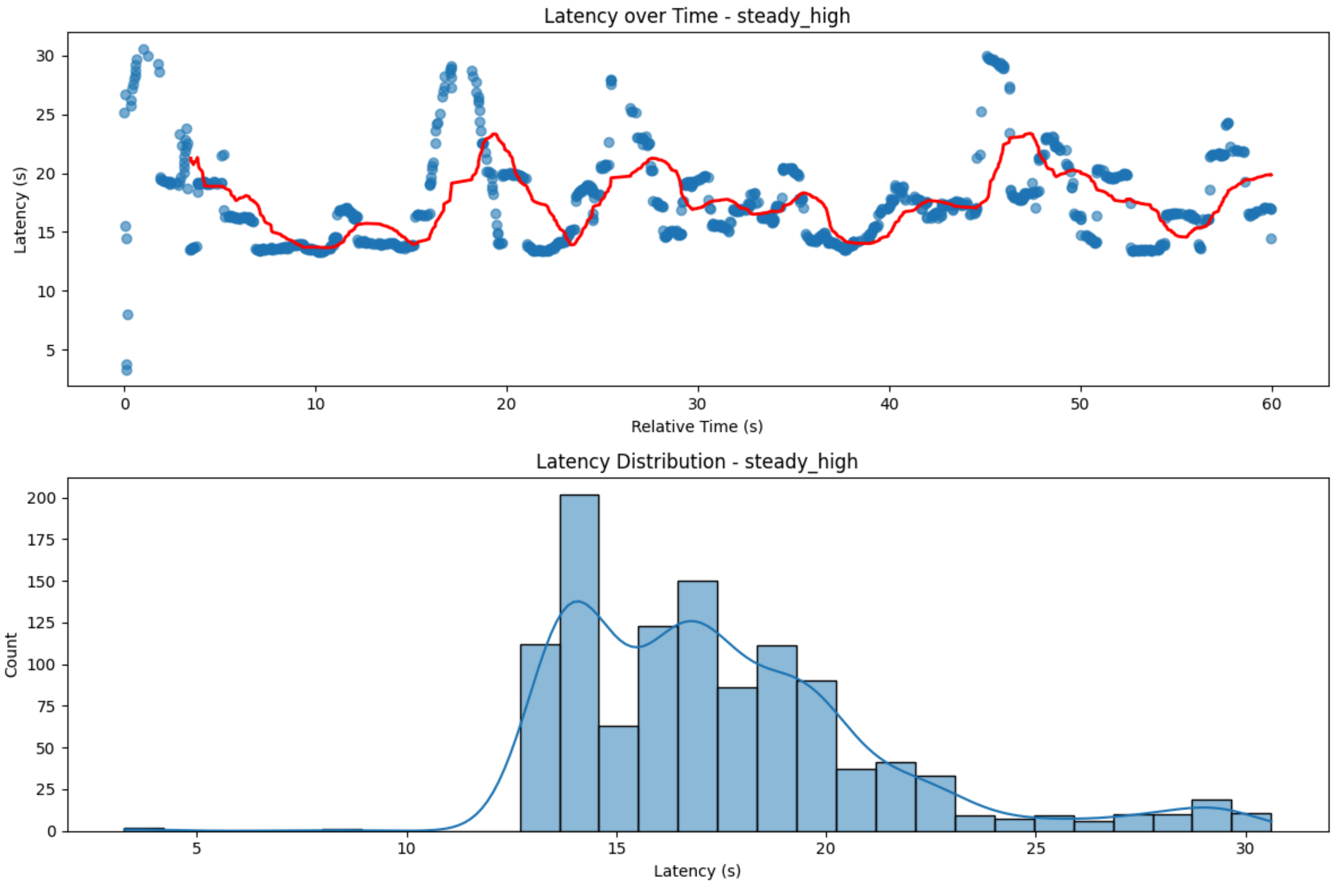
Detailed Results by Pattern

steady_low Pattern



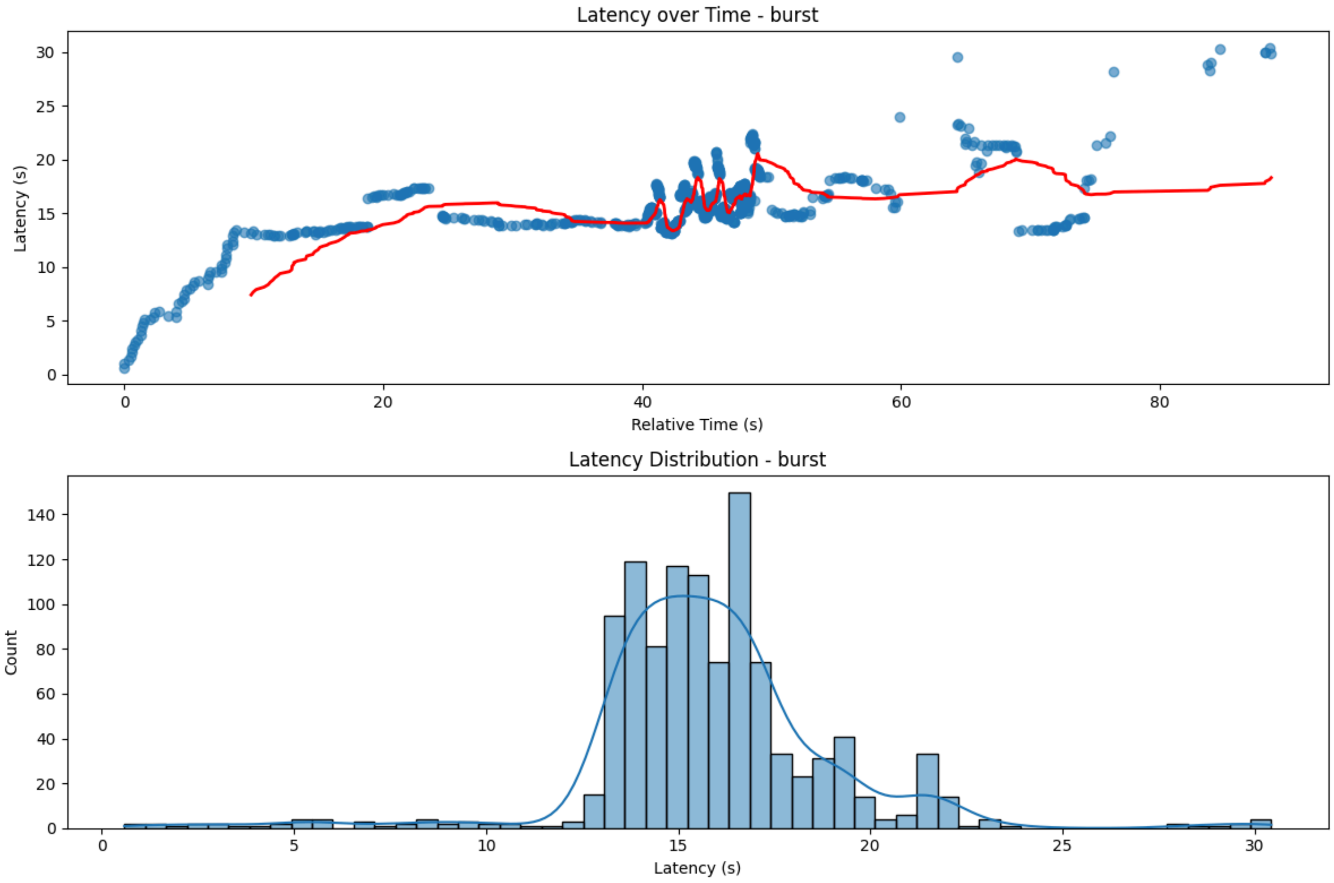
Latency distribution for steady_low workload

steady_high Pattern



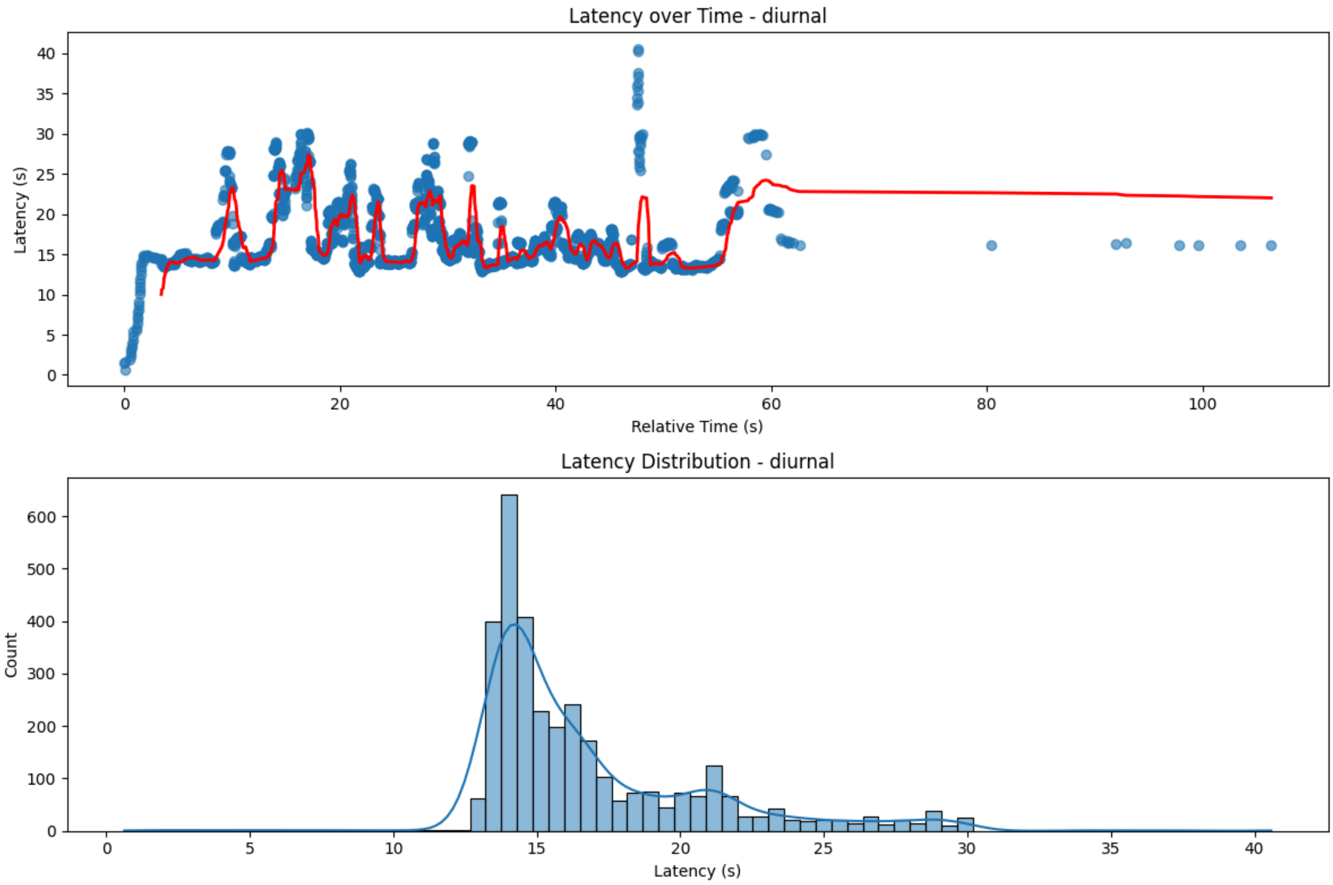
Latency distribution for steady_high workload

burst Pattern



Latency distribution for burst workload

diurnal Pattern



Latency distribution for diurnal workload

Conclusion

This report shows the performance of the Hugging Face 16-0.1 RAG implementation across different workload patterns. These results can be used as a baseline for comparing optimized implementations.