A Measurement Report on Queueing Delay

Jordan Blake

August 29th, 2026

Chapter 1 Introduction

This report studies how burst size changes the waiting time in a single-server queue. The purpose is not to claim a universal performance law, but to show a complete path from a question to a reproducible measurement and a bounded conclusion.

Chapter 2 Experimental design

Requests arrive in batches at a fixed interval. The service time is sampled from a documented distribution, and the queue is reset before each run. For a stable system with arrival rate λ and service rate μ, the traffic intensity is

$$\rho = \frac{\lambda}{\mu}, \qquad 0 \le \rho < 1. \tag{2.1}$$

Runs with $\rho \ge 1$ are reported as saturated rather than treated as normal steady-state observations.

Chapter 3 Results

Table 3.1 Illustrative queue measurements.

Batch size	Requests	Median wait (ms)	95th percentile (ms)
1	600	4.2	8.7
4	600	9.5	18.3
8	600	19.1	41.6

The tail grows faster than the median as the workload becomes burstier. Before comparing implementations, repeat each row with the same seed, warm-up rule, hardware, and clock source.

Chapter 4 Discussion

The measurements support a workload-specific association between burst size and queueing delay. They do not isolate scheduler overhead, garbage collection, or network variance. Record those factors explicitly in a future run and include confidence intervals around both central and tail estimates.

Chapter 5 Conclusion

For the tested service rate, smaller batches keep the tail shorter. The result is useful as a diagnostic baseline; capacity planning should use production traces and a saturation test.

Appendix A Run checklist

1. Record the software revision and machine configuration.
2. Declare warm-up, sampling, and outlier handling before looking at results.
3. Archive raw observations with the generated report.