SSD Performance Profiling

ECSE 4320

Ben Herman

Content

Experiment Setup

Timing Measurement:

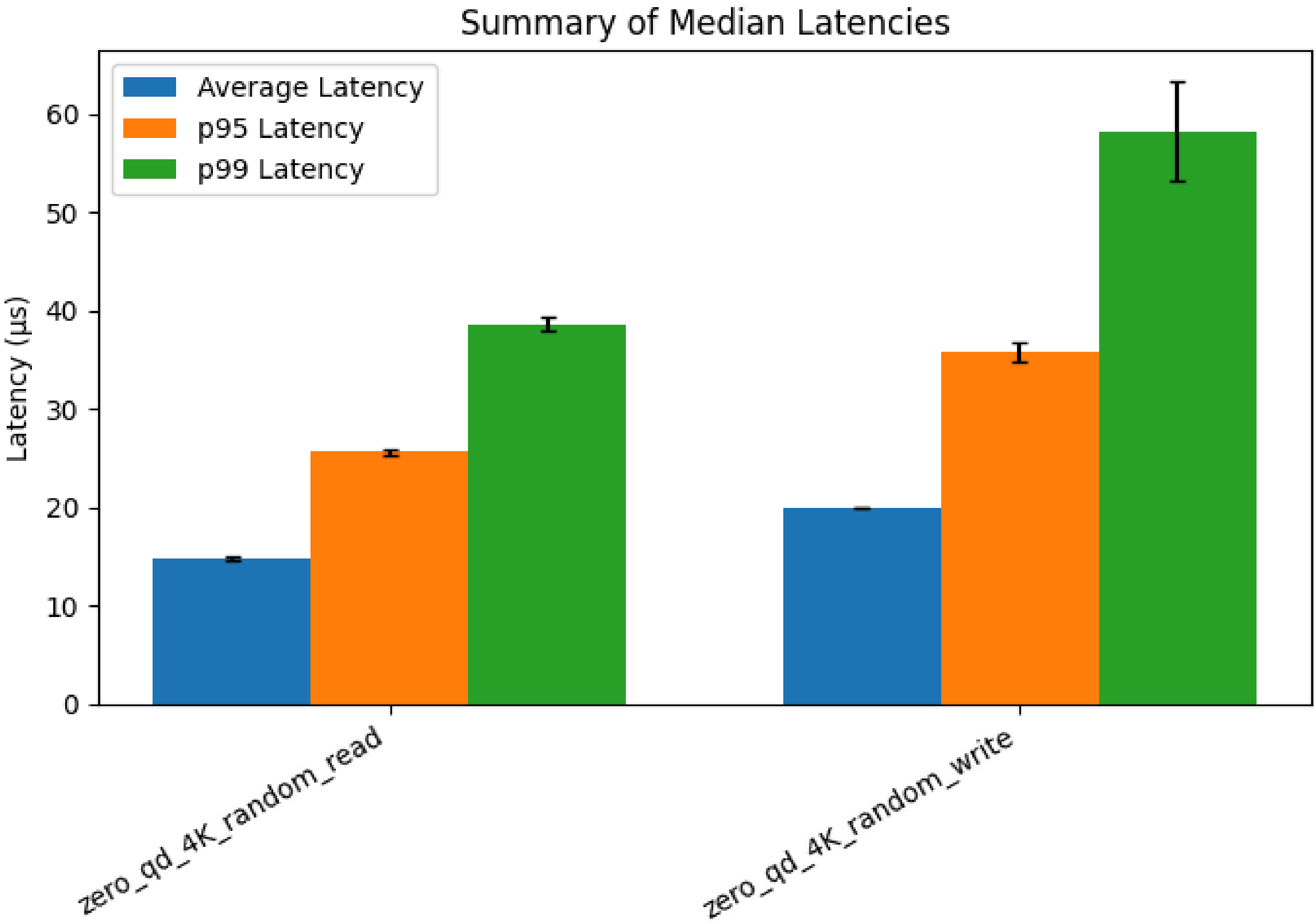
- Execution time is measured using `mach_absolute_time()`.

Conditions:

- Model: M2 Mac
- OS: Sequoia 15.6
- Powersource: Wall outlet
- Ram: 16 GB

Zero-Queue Baselines

Zero-queue latency for 4 KiB random read and write.

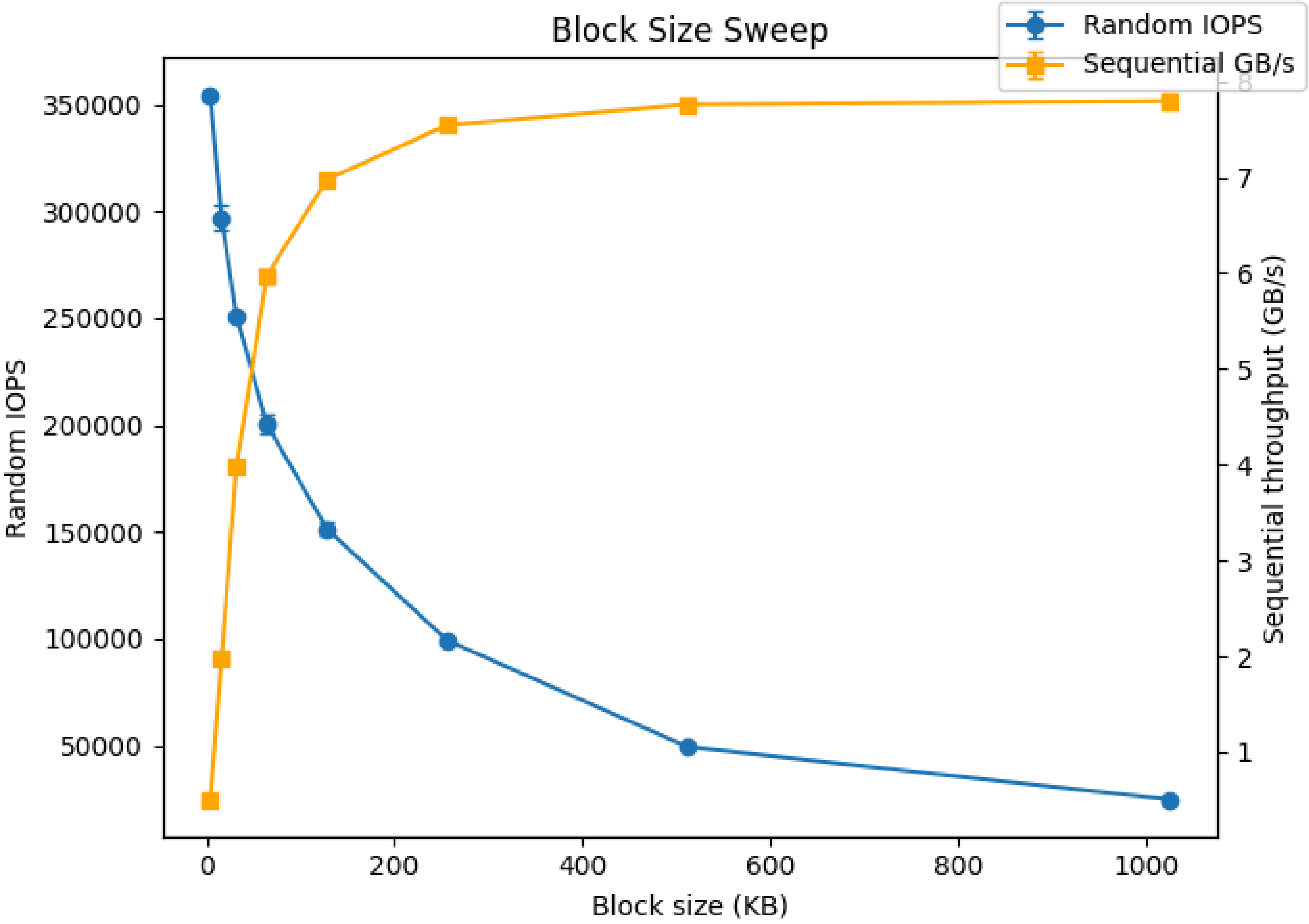


Block-Size Sweep

Impact of block size on random IOPS and sequential throughput.

We see in the graph:

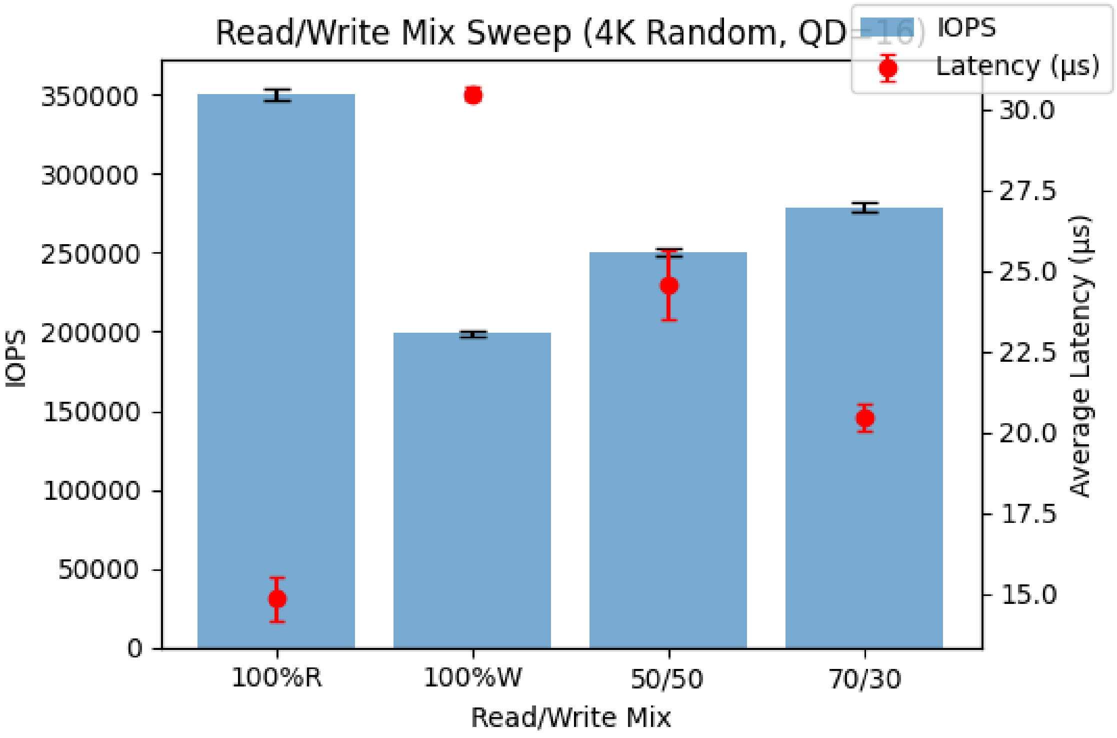
- Small blocks less than 192 KB are throughput limited by IOPS
- Large blocks more than 192 KB are throughput limited by the PCIe because its saturated with too many requests.



Read/Write Mix Sweep

Effect of varying read/write ratio at fixed block size (4 KiB random).

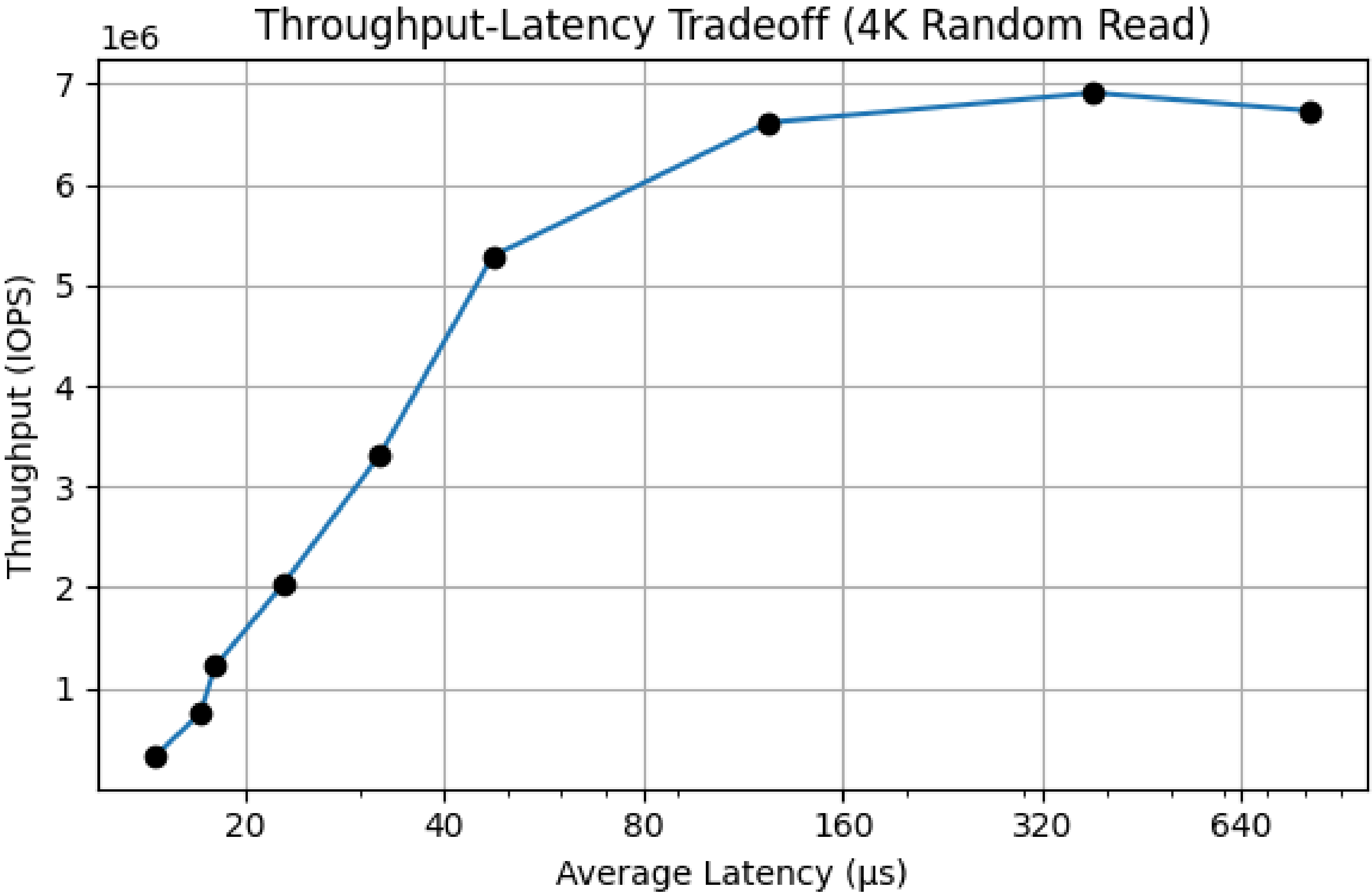
- 100% read yields highest IOPS and lowest latency
- Increasing write fraction increases latency
- We see the inbetween ratios are inbetween these extremes accordingly



Queue-Depth Sweep

Throughput-latency trade-off curve for 4 KiB random reads.

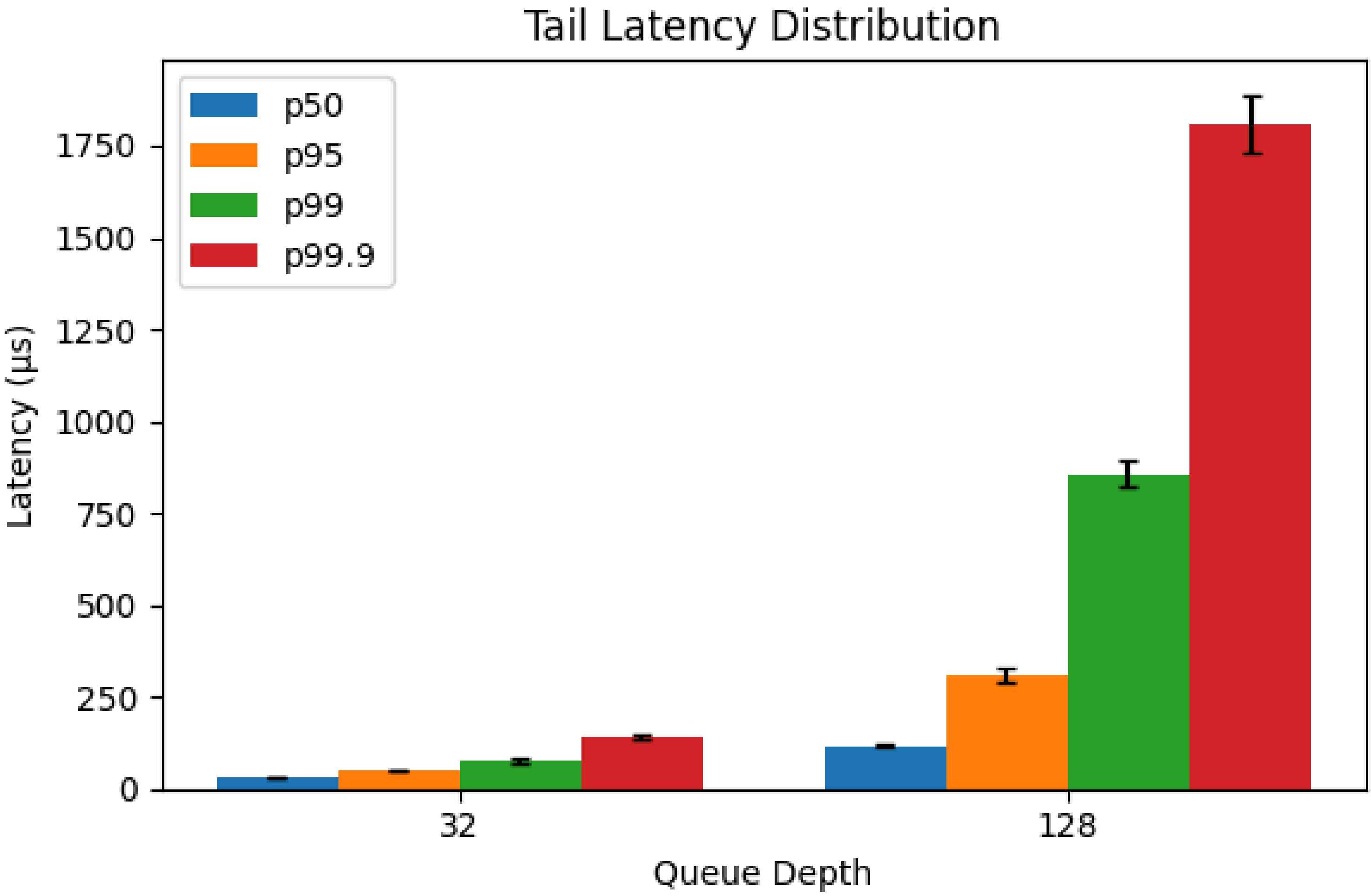
- Throughput rises with QD until saturation (QD 32-64)
- Latency grows sharply past the knee
- Little's Law relation visible: Throughput ≈ Concurrency / Latency



Tail Latency

Tail latency distribution (p50/p95/p99/p99.9) at different QDs.

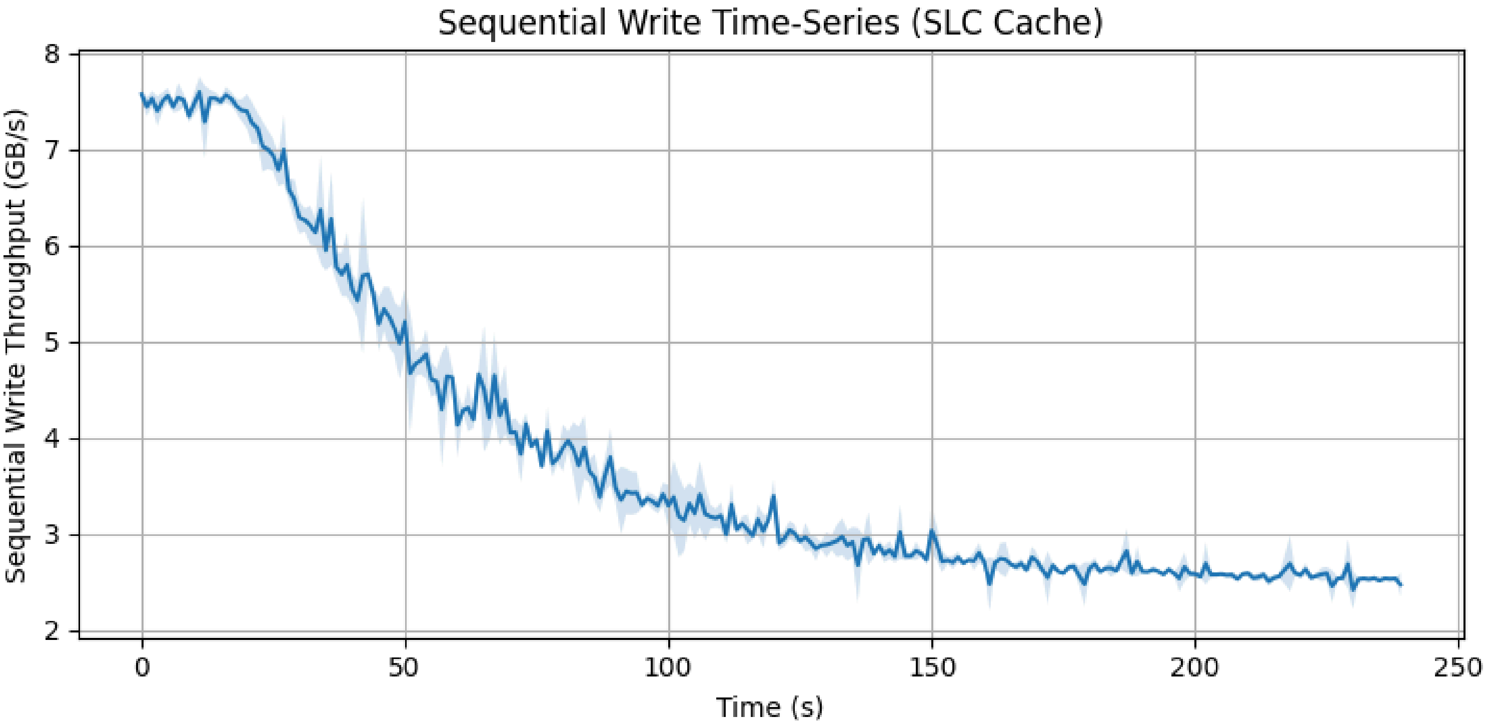
- p99.9 latency spikes significantly at high queue depth
- Important for SLA-sensitive workloads
- Highlights worst-case latency scenarios beyond average



Sequential Write Time-Series

Sequential write throughput over 240s, simulating SLC cache behavior.

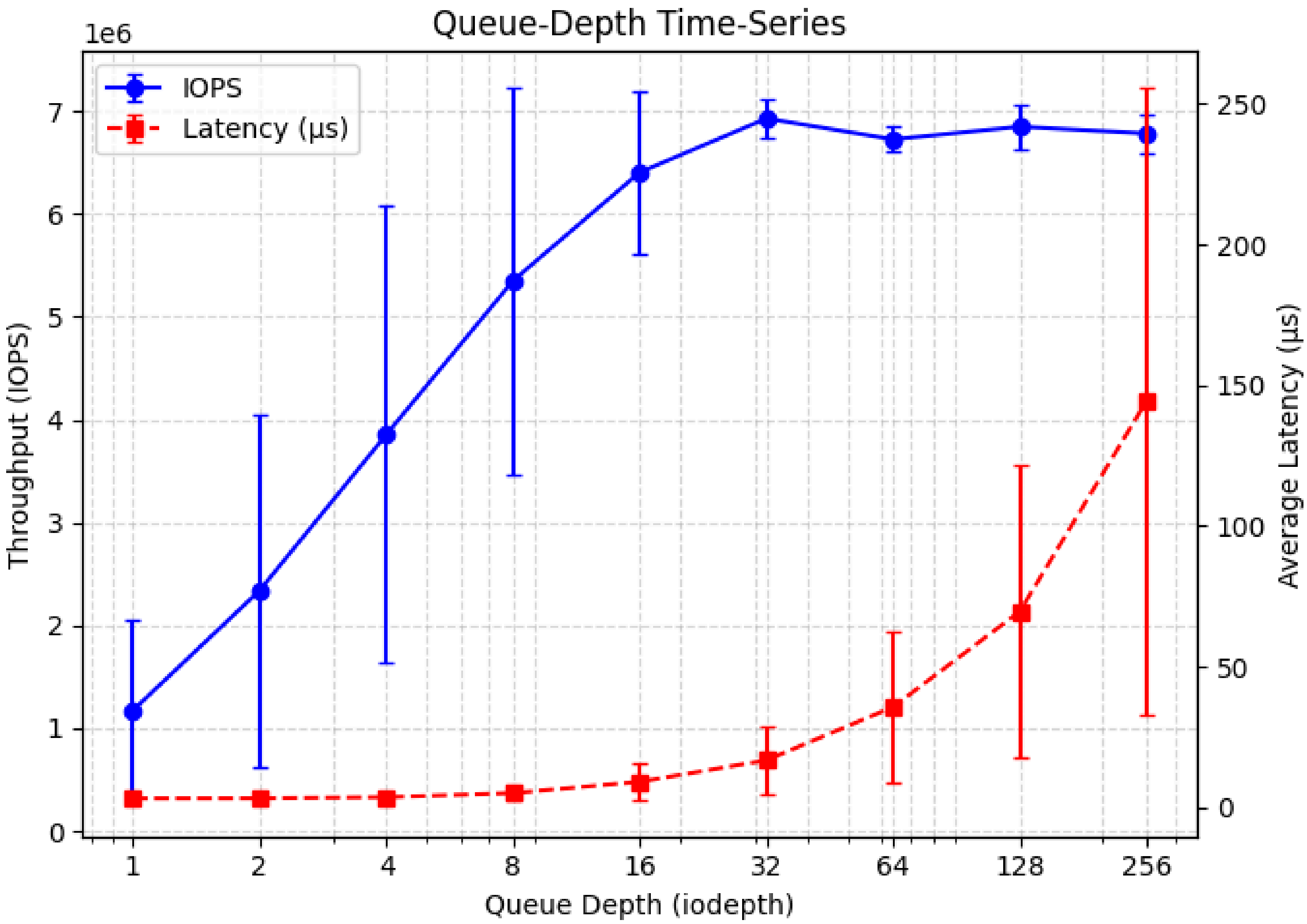
- Burst period: 7.5 GB/s for 15-21s (SLC cache)
- Steady-state decay to 2.5 GB/s
- Micro-bursts introduce variability in latency and throughput



Queue-Depth Time-Series

IOPS and latency vs iodepth (1-256) time series.

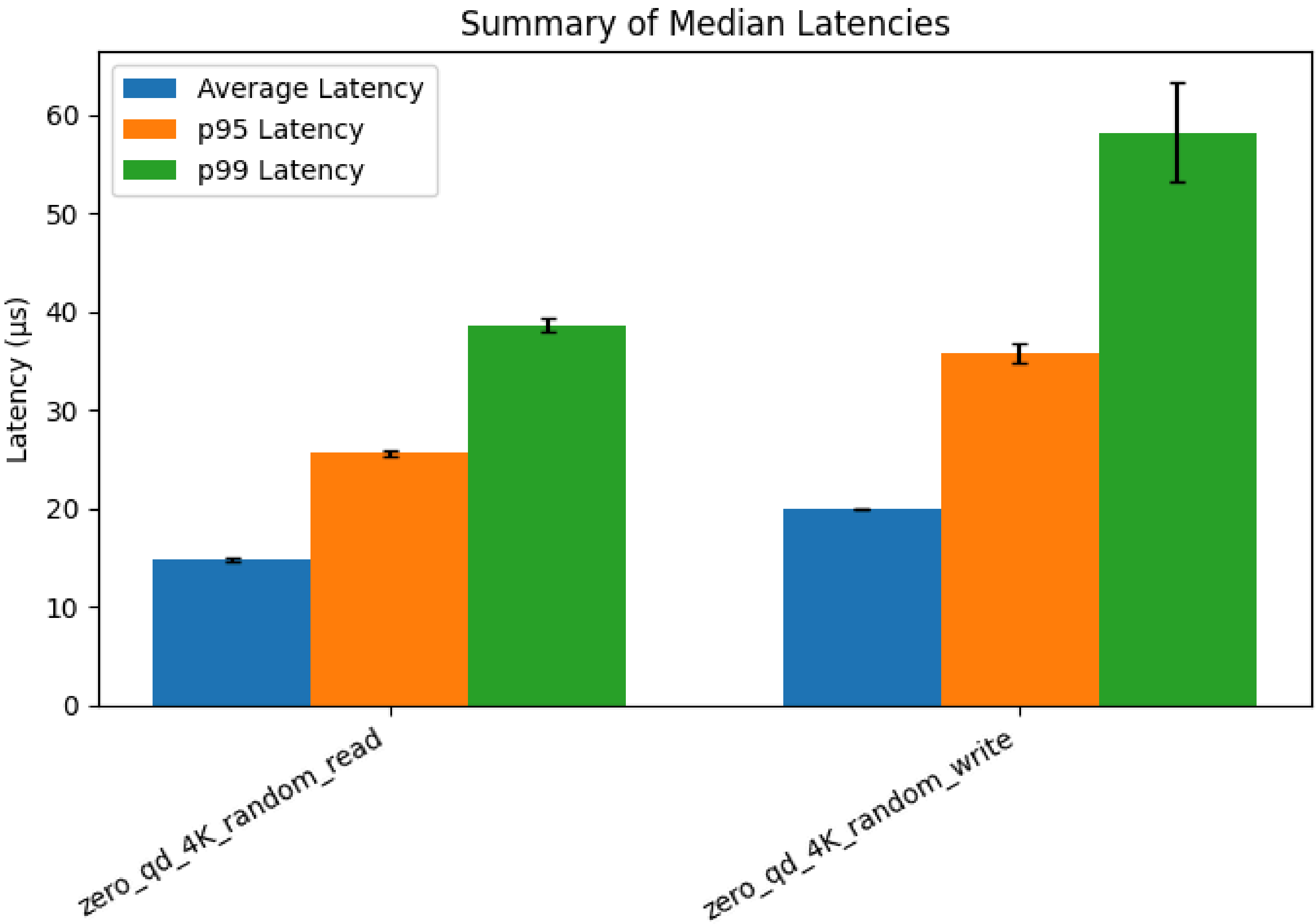
- Throughput increases with QD, latency increases slowly until saturation
- Knee of curve around QD 32-64
- Useful for identifying operating points balancing latency and throughput



Summary Overview

Median latency (avg/p95/p99) across experiments.

- Confirms reproducibility across three runs
- Provides quick reference for comparative analysis
- Shows variance across patterns and workloads


Summary of Median Latencies
Average Latency
p95 Latency
p99 Latency
Latency (µs)
zero_qd_4K_random_read
zero_qd_4K_random_write