Data Science of QoS

Dave Holmstrom
SSD System Architecture
Micron Technology, Inc.
Disclaimer

- This is about data
- *Not workloads*
- *Not SSDs / HDDs*
- *Not success criteria*
QoS
Quality of Service

- The term Quality of Service (QoS) is a measure of latency (quicker is better) at a specific consistency level.
- Example: A pizza company says they can deliver in 30 minutes or less. If out of 100 deliveries, one of them is later than 30 minutes, the QoS for the pizza delivery is 99/100, or 99% below 30 minutes.
QoS Traditional View

- In this case, 1,000,000 IOs are measured, and a very large portion (999,000) are complete under 44.3ms.
- There are 1,000 I/Os are outliers taking longer than 44.13ms.
- QoS is 99.9% (999,000/1,000,000) = 44.13ms. This is often referred to as “Three nines at 44 milliseconds”.

![Graph showing QoS distribution with 44.13 ms as 3 9s, 62 ms as 4 9s, 153.1 ms as 5 9s, 167.7 ms as 6 9s, and Max as 171.9 ms.](image)
Improving QoS Measure
Ensure Consistency

- Device at steady state for the current workload and test platform
- IOPS level should be stable across the measurement

In this example, the first interval shows higher performance and will shift QoS.
Percentiles should be based on good data

- Good data requires
  - Repeatability
  - Reliable data up to the level of percentile of interest
  - Understanding the measured workload

- Reliable data cannot stand on any one single datapoint. Individual datapoints can be unreliable and can represent random events.
Sample Size MATTERS

- Sufficient samples are needed for mathematical significance in a QoS measurement
- According to the Central Limit Theorem, if a large uniform distribution of sample data is not a “normal distribution”, statistically, the minimum number of samples should be well beyond “just enough”
- For storage latencies, at least 100x more samples beyond the base requirement is needed to ensure the measurement is accurate and repeatable (i.e. 1,000,000 points are needed to yield a reliable 4 9s)
Sample size effect on resolution

Repeatability (5 runs - ~3.2 million samples each run)

10%
Calculating percentiles

- Default percentile calculations can return values that were not measured
- Use the next *higher* latency when no latency is within a calculated percentile

7.8M I/Os latency histogram

---

A better view of QoS
Waterfall View: QoS is more than a number

- Traditional views of histograms and tables don’t tell the whole story
- Implementing a 1-CDF waterfall view illustrates continuous QoS levels
- 1-CDF clarifies driving factors for levels of 9s
- This view, when combined with histograms, gives insights into the innerworkings of a device
Combined view
Multiple runs reveal variations

Displacement indicates one run had an event occur

Horizontal at 7.9s level leads to variability

Low sample count near maximum shows variations
Run Time QoS
QoS in real time

- Calculating QoS on an interval returns point in time of service levels
- The service levels can be aggregated across the test interval
- Output indicates variability and expose periodic long latency events
Previously unseen behavior emerges
A New View of QoS
Questions?