QLC storage

Balancing power, cost and performance

Sumit Gupta

Meta Platforms Inc.

Challenges in estimating storage demand

- Data center workloads are all over the place.
- Data temperature : Throughput / Bytes-used (MB/s/TB)

- Workloads on HDD clusters
  - Supply: 5-7 MB/s/TB
  - Demand: 1 – 100+ MB/s/TB (extreme bin packing)

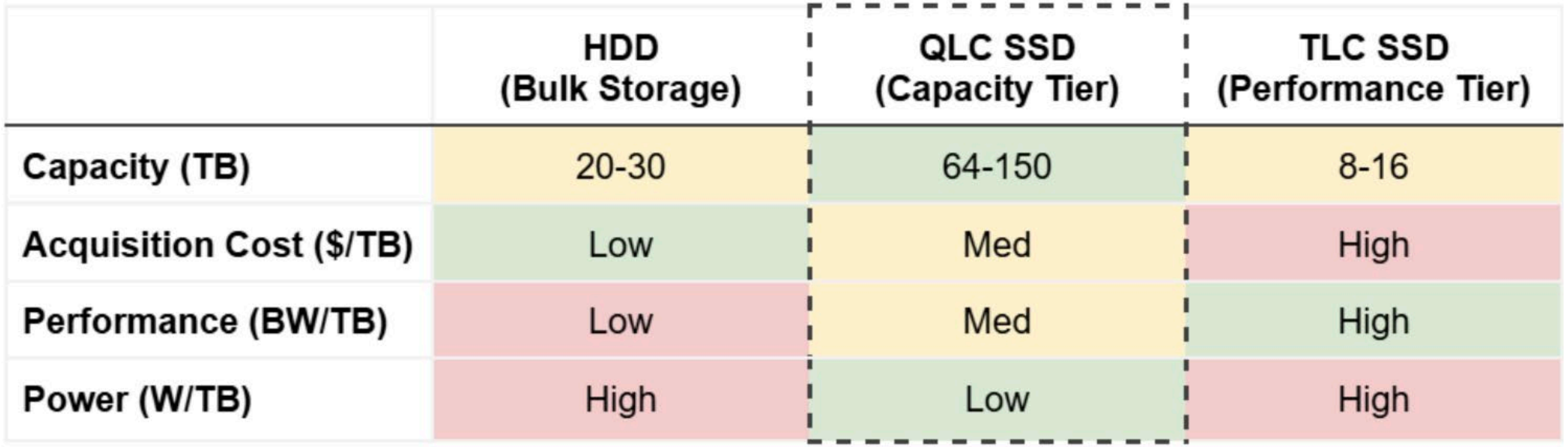
- Workloads on SSD clusters
  - Supply: limited by power, ~100 MB/s/TB
  - Demand: 20 – 120 MB/s/TB (limited bin packing)

- Massive delta in supply, we end up buying storage for I/O

Challenges in estimating storage demand – GenAI

- GenAI storage demands are unpredictable.
  - Worst case estimates are often IO bound, even on SSDs.
  - Typical use cases are more space bound. Cluster sizes are in Exabytes, still we are often low of byte capacity
  - IO patterns are very erratic and bursty and the data temperature is higher than what HDDs can supply.
- Power budgets for infra are shrinking as GPUs need all the power.
- HDDs are getting denser (and thus colder) but TLC SSDs demand more power.

Enter QLC flash



	HDD (Bulk Storage)	QLC SSD (Capacity Tier)	TLC SSD (Performance Tier)
Capacity (TB)	20-30	64-150	8-16
Acquisition Cost ($/TB)	Low	Med	High
Performance (BW/TB)	Low	Med	High
Power (W/TB)	High	Low	High

- A middle QLC tier can scale to much higher density per rack than TLC and can significantly lower the W/TB footprint.

QLC@Meta – Starting points

- QLC racks are much denser, 10+ PB per rack.
- QLC offers much higher read BW, for now we scale it to 4x the write BW.
- Usable TBs are assumed to be 90% of capacity exposed to Linux.
- Performance is measured around WAF of 2.0
- Expected performance (power constrained) is given by the formula:

R + 4W >= 32 MB/s/usable-TB

QLC@Meta – issues

- Write performance is very low.
- Read performance is not getting fully utilized.
  - Still deciding on workloads placement beyond GenAI.
  - R + 6W <= 48 MB/s/TB ?
- High server density demands very high throughput i.e. still higher power consumption than what we like.
- Cost ($/TB) is still high.
  - Handling hotter workloads by HDD byte stranding vs. moving to QLC.

QLC@Meta – Future directions

- Reduce power further
- Go beyond 90% fill
- Better utilize read BW
- Grow the footprint.