Boosting QLC SSD performance and endurance for data Centers

Orit Wasserman
Principal Architect
Lightbits Labs

A little bit about Lightbits Labs and me

- Lightbits is a hyperscale software defined storage startup with offices in Israel and San Jose, CA
- Doing cool things with NVMe and NVMe-oF
- Inventors of NVMe/TCP

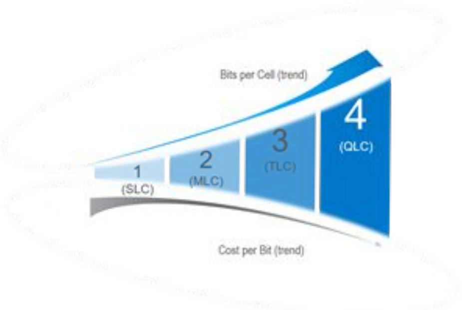
- Me: Principal Architect at Lightbits Labs
- Ceph RGW core developer, KVM/Qemu hypervisor, clouds and storage

QLC

Cost/Capacity optimized SSD

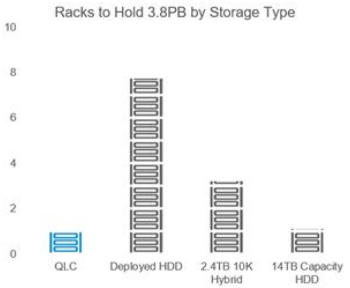
Lower TCO

Lower $ per GB of SSD storage



More Capacity

4 bits per cell, 33% more capacity on the same number of cells than TLC



Smaller Footprint

Less rack space

QLC Lower Endurance

The lower P/E cycles (~1000) results in lower endurance of the disk .QLC is estimated to wear out 3.4x-4.5x faster than TLC.

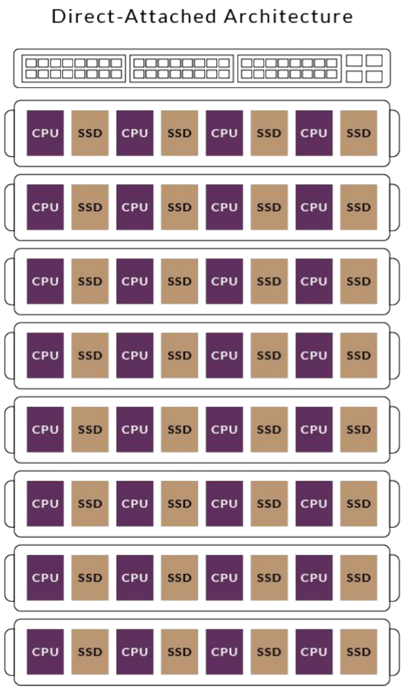
	Intel P4510 (8T TLC)	Intel P4320 (8T QLC)
DWPD for random workload	0.9	0.2 (4.5x)
DWPD for sequential workload	3.0	0.88 (3.4x)

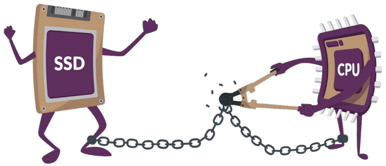
QLC Lower Performance

The higher error rates with QLC require more ECC (Error Correction Code) computation cycles on the read and write paths, resulting in an overall slowing down of I/O operations

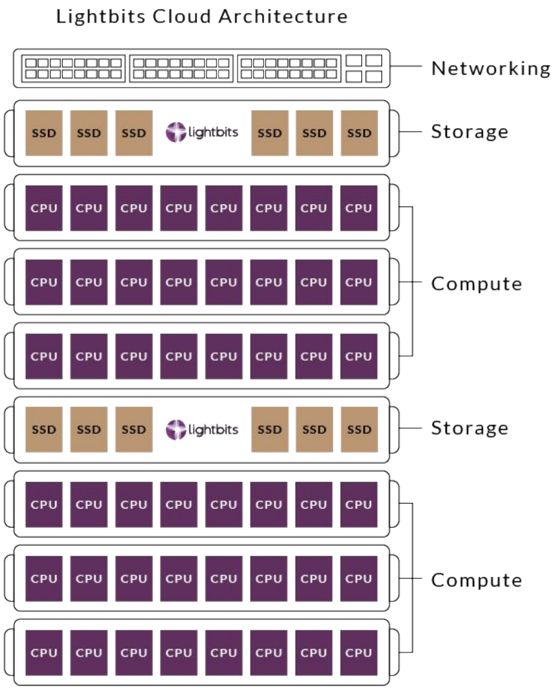
	Intel P4510 (8T TLC)	Intel P4320 (8T QLC)
Random 4KB Read (IOPS)	642000	427000
Random 4KB Write (IOPS)	135000	36000
128K Sequential Read (MB/S)	3200	3200
128K Sequential Write (MB/S)	3000	1000
4K Random Latency (typ.) R/W	100/30 μs	138/30 μs
4K Sequential Latency (typ.) R/W	10/12 μs	10/12 μs

From direct-attached storage to disaggregated storage servers





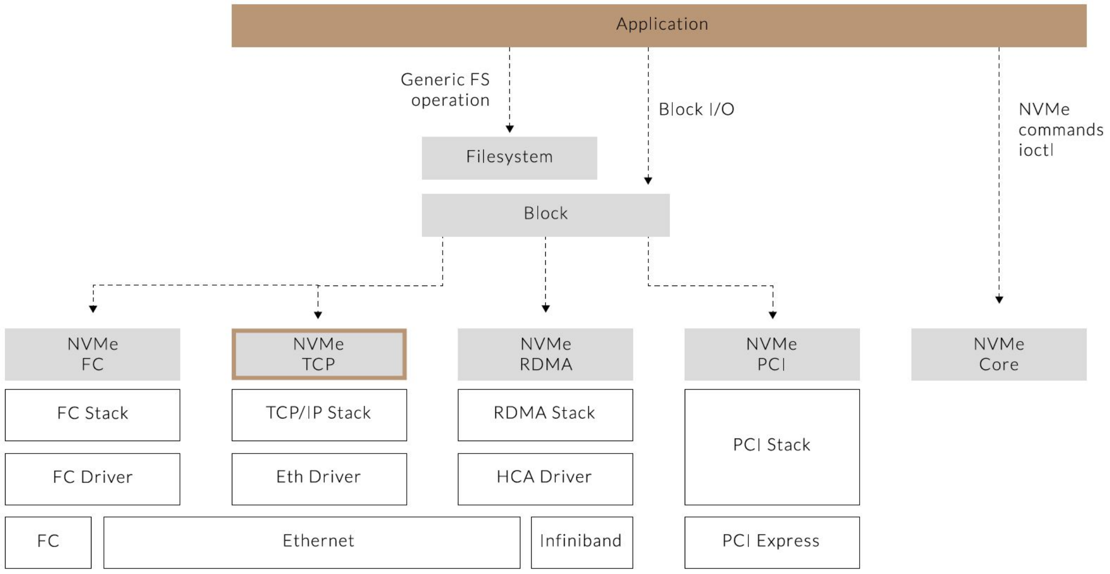
- Maximize utilization
- Reduce TCO
- Easy to maintain & scale
- Better user experience
- Support more users



NVMe/TCP

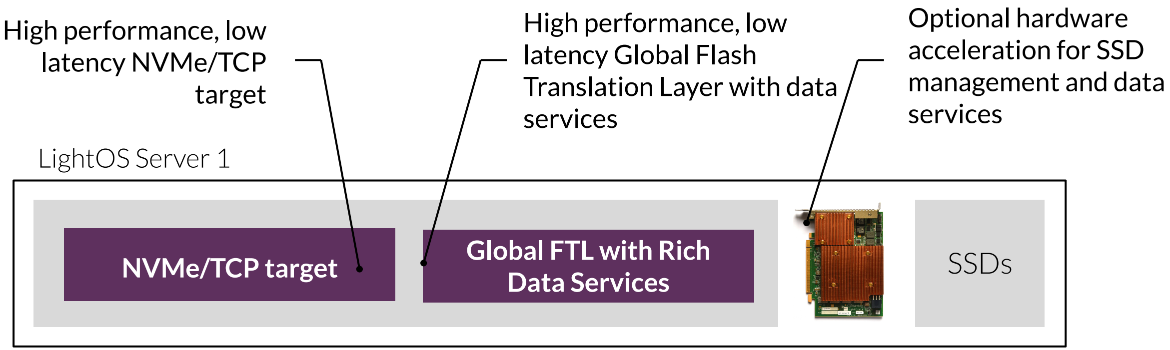
- Standard ratified Nov, 2018
  - Fastest time to ratification
- Supports remote NVMe SSDs with minimal additional latency compared to local SSDs
- Same NVMe model: sub-systems, controllers namespaces, admin queues, data queues

- Lightbits invented NVMe/TCP
  - Lead author of the NVMe/TCP standard, maintainer of Linux drivers



Lightbits LightOS

- The Lightbits NVMe/TCP target
  - The First commercial available, production grade NVMe/TCP target
  - Open storage platform
  - High performance, consistent low latency, QoS, flow control, ...



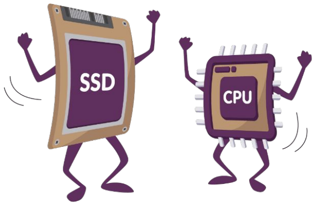
Lightbits LightOS

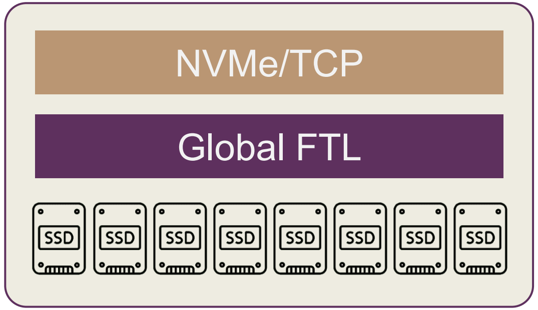
Disaggregated storage for the core and edge data centers

- Increase Availability
- Up to 50% lower TCO
- No changes to network infrastructure
- Hyperscale & software defined
- Secure
- Consistent low latency
- Scalable high performance
- Enable new applications
- Automated, API driven & designed for Cloud
- Agile, standard servers and SSDs

Fast ACK

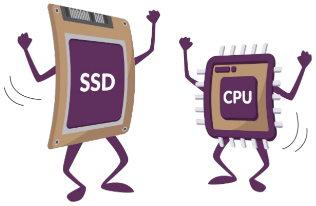
- Frontend writes first to NVRAM, then moves the data to the SSDs in the background

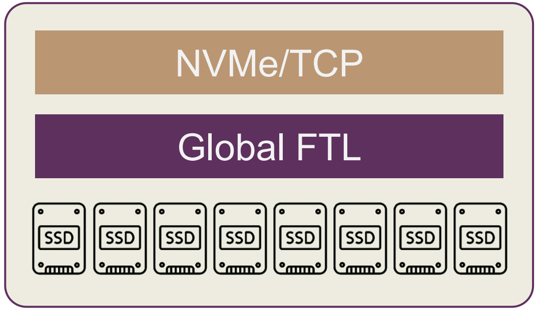




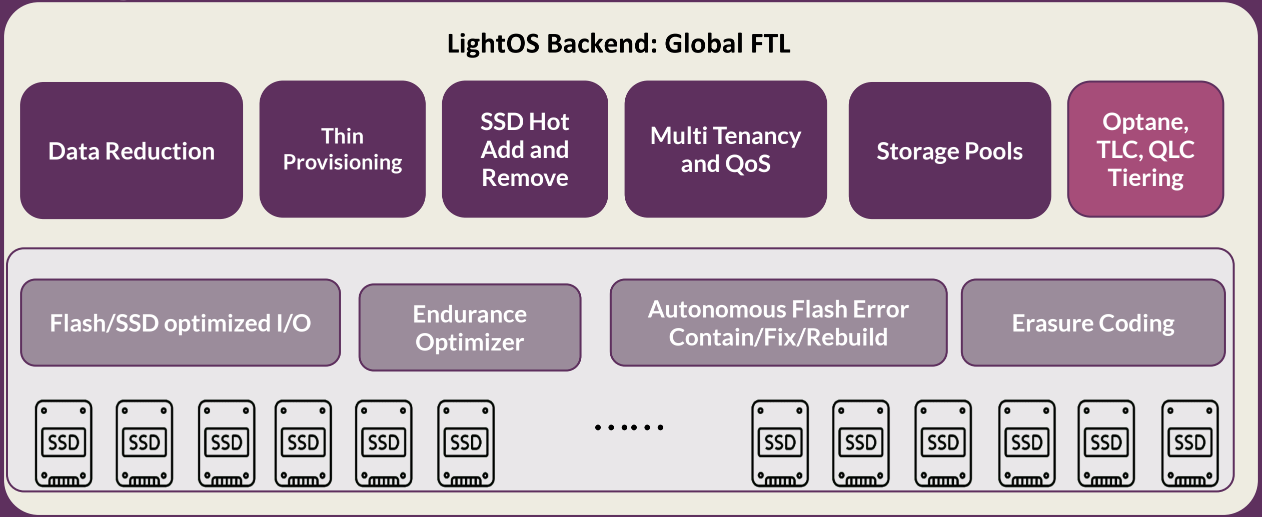
Fast ACK: Improving QLC write performance

Write latency and throughput do not depend on the underlying media (assuming data set that can fit in NVRAM)



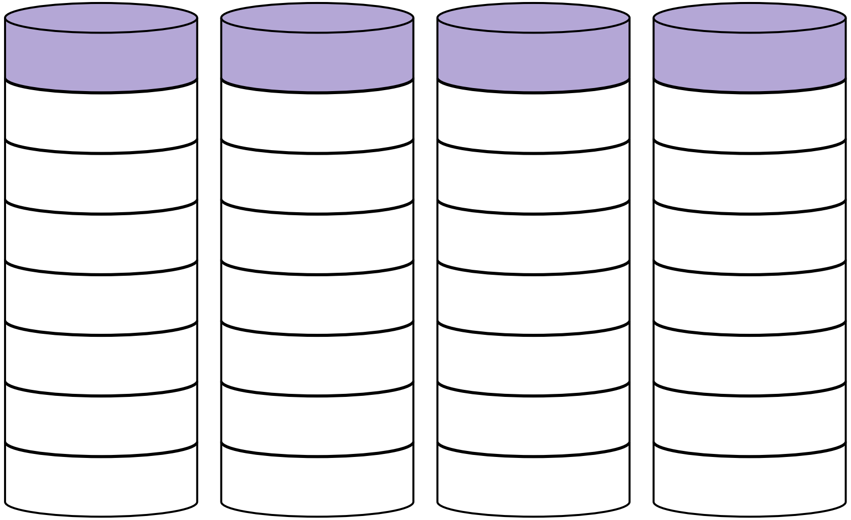


LightOS Global FTL (GFTL)



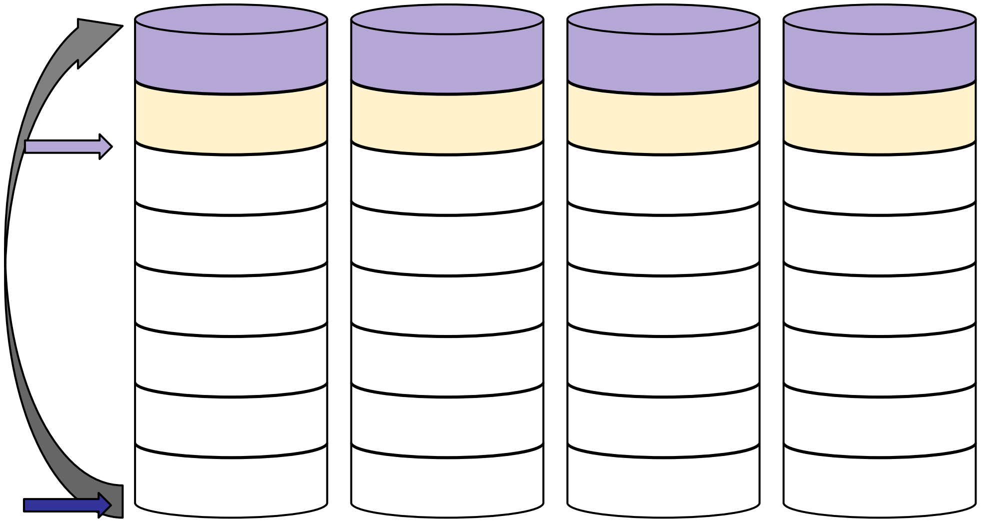
LightOS GFTL

- Accumulate writes + sequential writes
- Fill complete stripe
- Thick stripes
- Meta Data



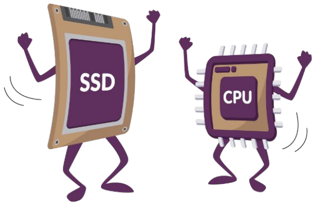
LightOS GFTL

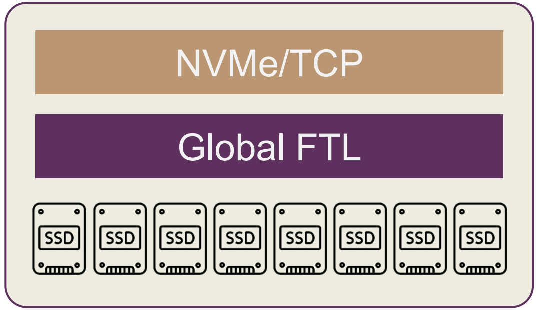
- Accumulate New writes + Rewrites
- Write another stripe
- Cyclic, Pointers



GFTL: Improving QLC

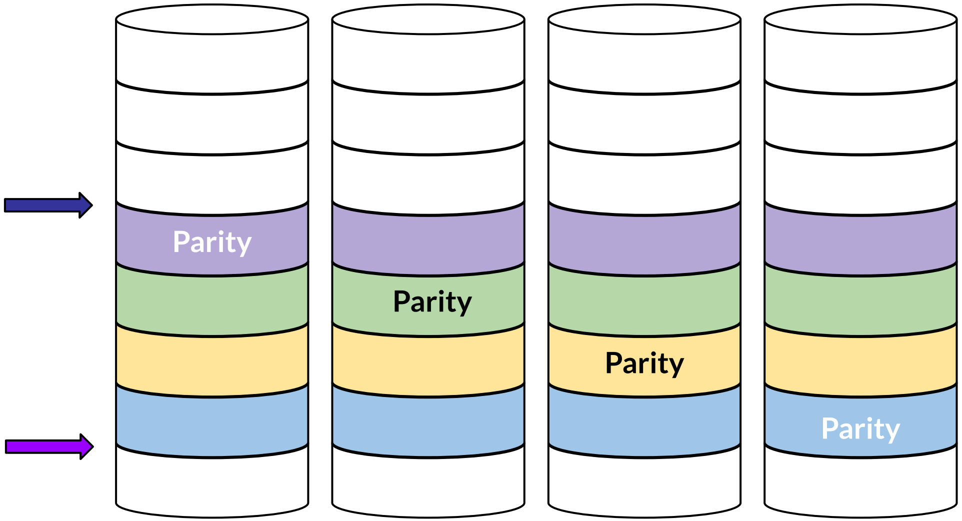
- Append only and sequential writes to the SSDs reducing write amplification and performance
- Writes are balanced across all SSDs, no SSD hot spots to wear out sooner
- Software GC





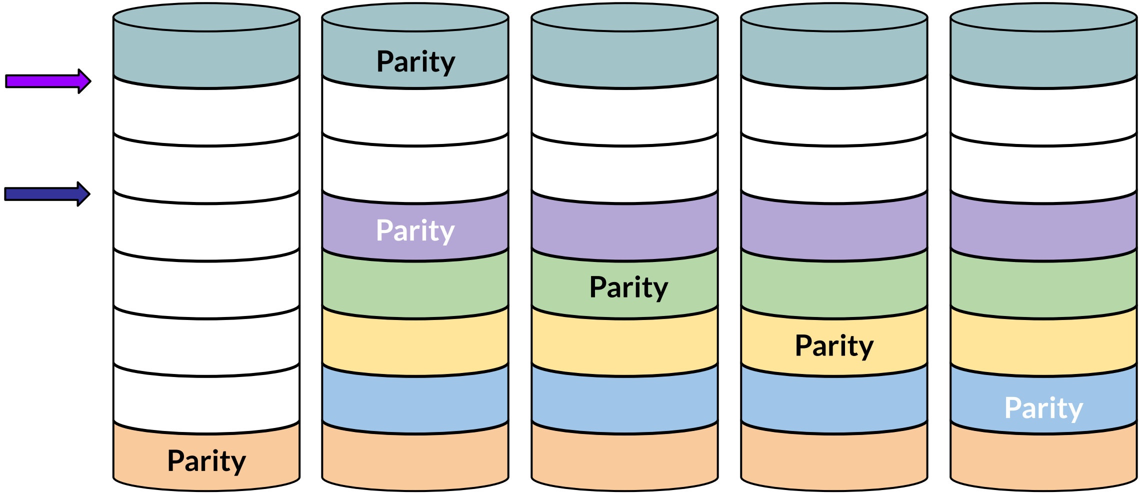
Erasure Coding

- Default: RAID5-like parity with append-only (no RMW)
- Can also support RAID6, other schemes
- Stripe optimization



Erasure Coding

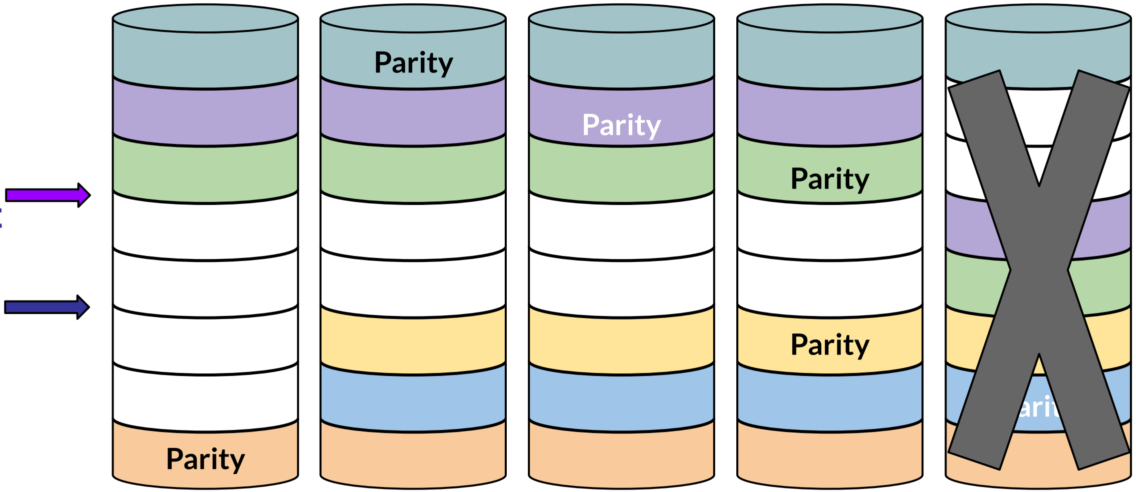
- Adding SSD
- Variable stripe width
- GC will gradually fix



Erasure Coding

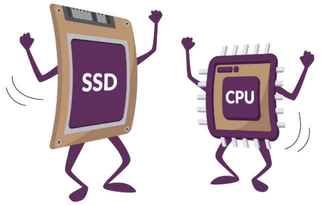
- Losing SSD
- Variable stripe width
- GC will aggressively rebuild
- Lower negative rebuild impact

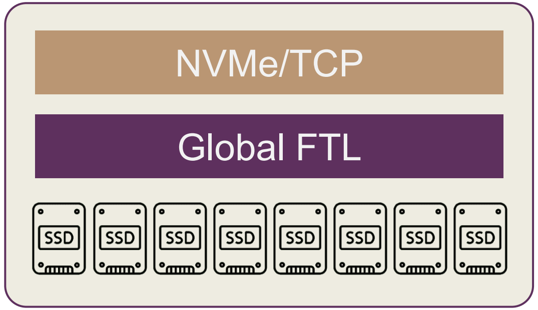
- SSD resets / transient failures handled by reducing stripe size and doing “read reconstruct”



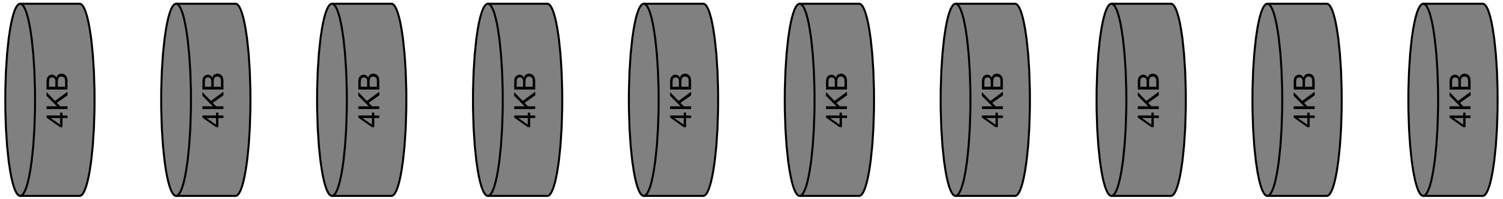
EC: Improving QLC

- Enables quick & transparent recovery from SSD failure without any performance cost
- Uses append only writes

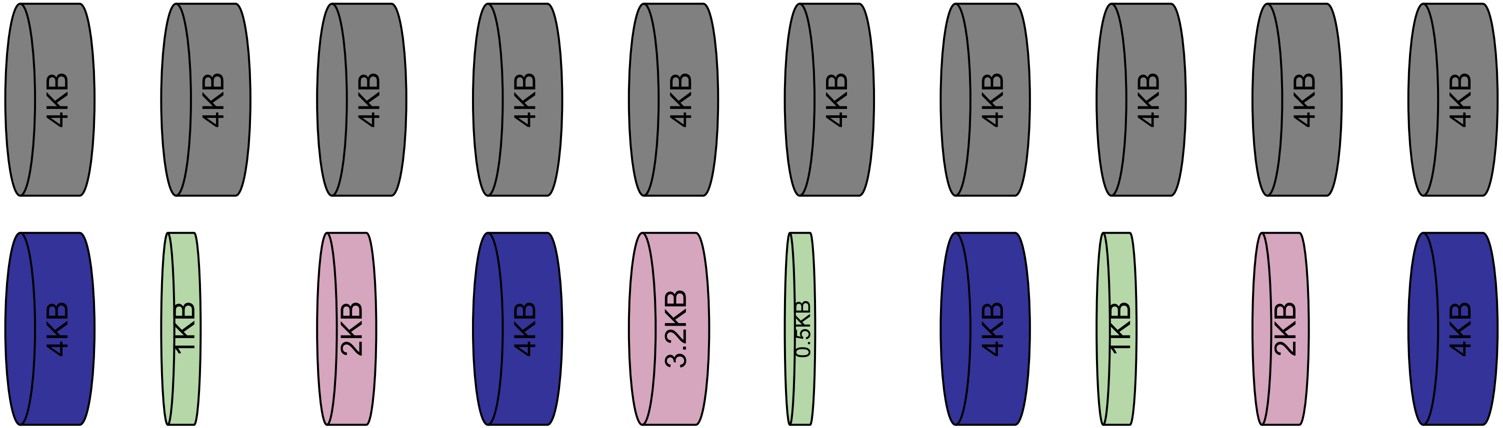




Compression



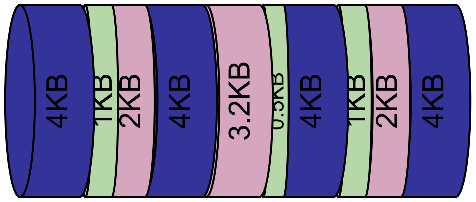
Compression



Flash Memory Summit

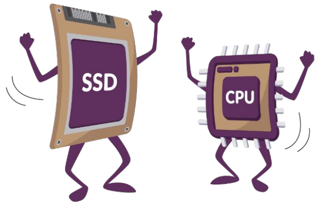
Compression

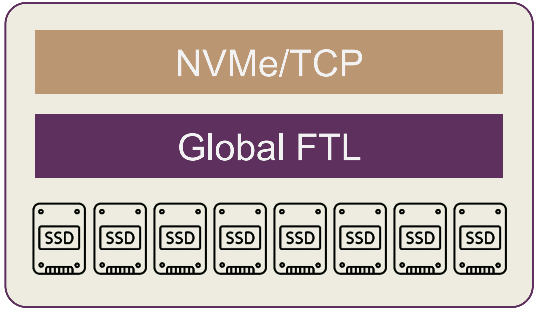
- Meta-data address alignment - 32 Bytes
- Optimal space utilization
- Integrated with the GC without any fragmetation



Compression: Improving QLC

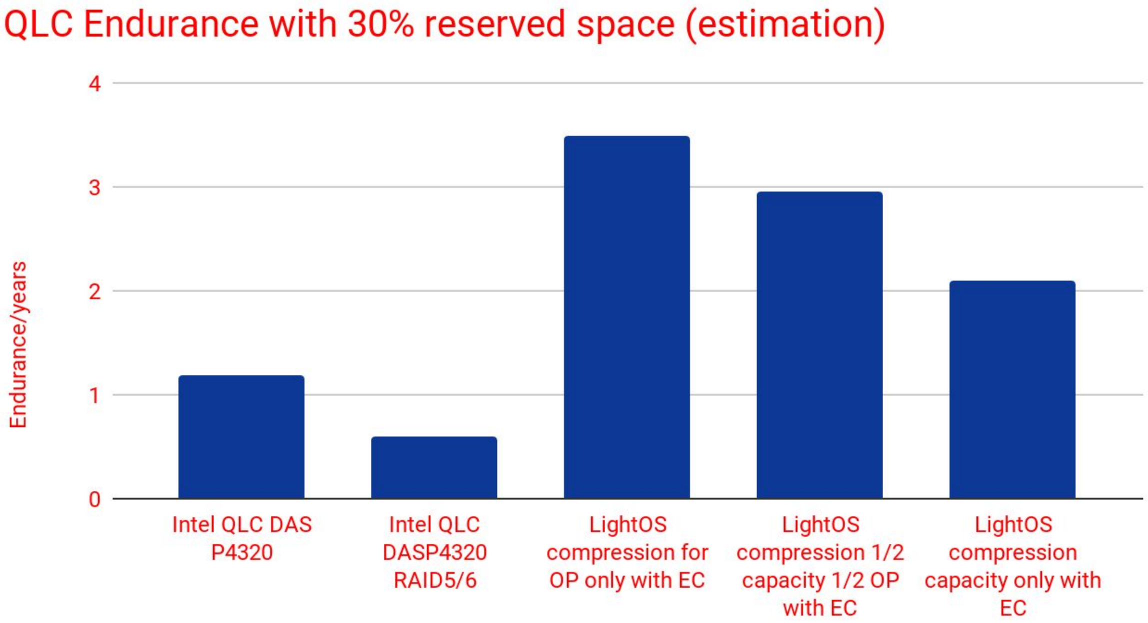
- Reduces the overall amount of data written to the SSDs
- Increasing performance and endurance





Endurance/OP

- Compression rate 50%
- Used saved space for endurance or capacity
- User can choose a scheme depending on his workload
- Adaptive scheme



16K or larger page

- QLC higher density results in bigger SSDs.
- In order to keep the translation page table in the control memory the page size has to increase.
- For smaller writes than this page size, like common 4K the device will need to do Read/Modify/Write cycle.
- This affects write performance and mixed workloads. The extra reads will increase the SSD read disturbance reducing its endurance.

16K or larger page

Lets estimate the performance of 4k write on 16k page QLC:

Each write cost an additional read:

1x4k write = 1x16k read + 1x 16k write

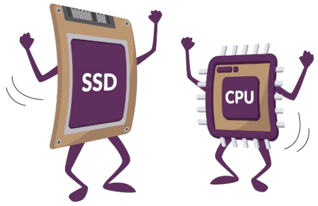
Random 4k writes with 16k page: 8793 IOPS

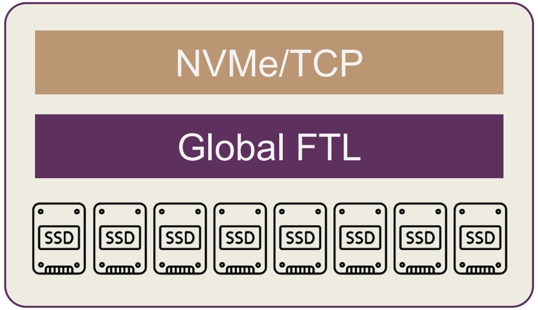
4x slower!

	Intel P4320 (8T QLC)
Random 4KB Read (IOPS)	363000
Random 16KB Read (IOPS)	199000
Random 4KB Write (IOPS)	35000
Random 16KB Write (IOPS)	9200

Improving 16k page

- Append only sequential write
- Thick stripes
- No read/modify/write
- No performance penalty when with 4k random writes on 16k page SSD





Summary

- Lightbits can get more from QLC SSDs:
  - GFTL
  - EC
  - Compression
- Visit our partner booth #848 - International Computer Concepts to see a demonstration of LightOS NVMe/TCP
- Hear more on NVMe/TCP from Sagi Grimberg in the Panel “NVME-202B-1: Leveraging NVMe-oF for Existing and New Applications”
- Come see Alex’s talk “An NVME/TCP Software-Defined Platform for Guaranteed QoS” tomorrow

Contact information

https://www.lightbitslabs.com/

@oritwas
orit@lightbitslabs.com