A Global FTL Architecture to Drive Multiple SSDs

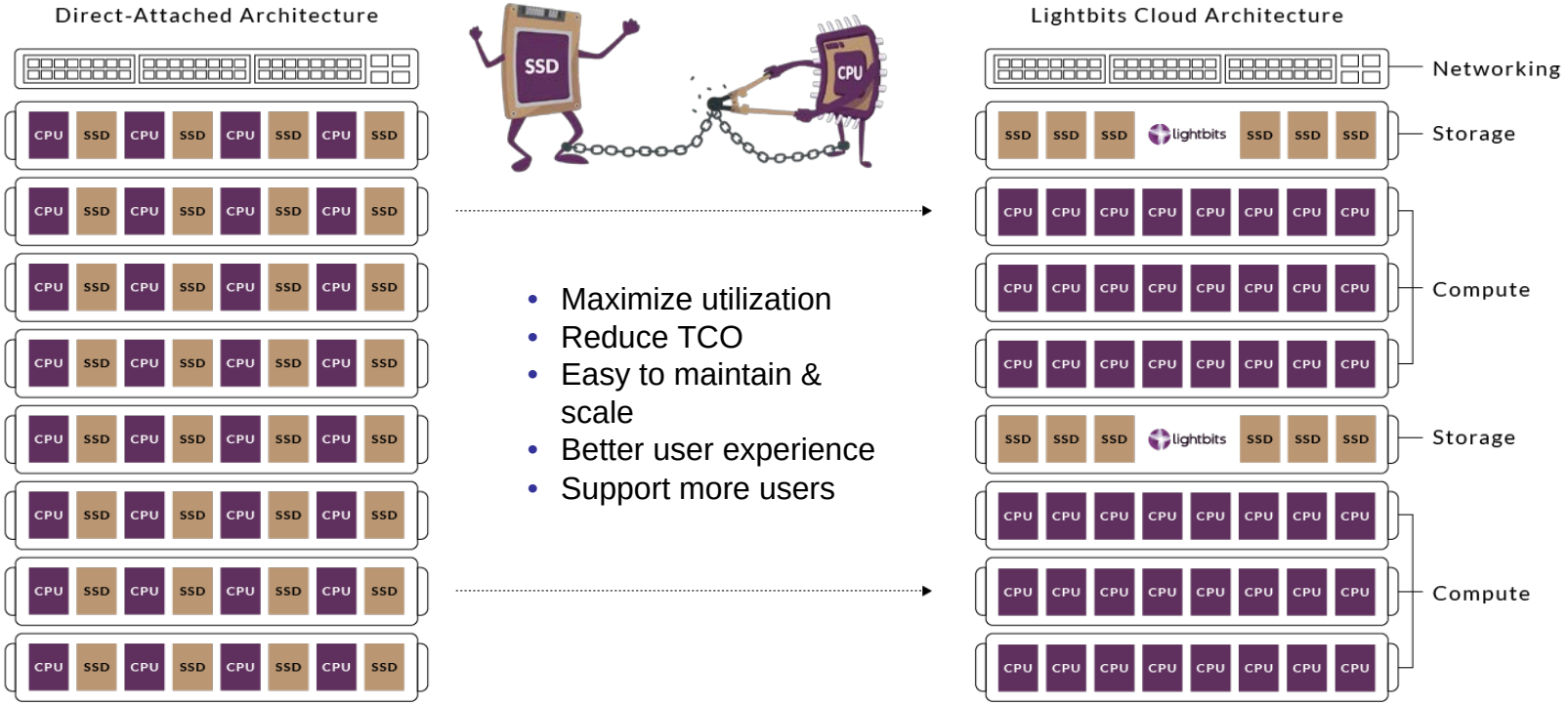
Roy Shterman

Lightbits Labs

Agenda

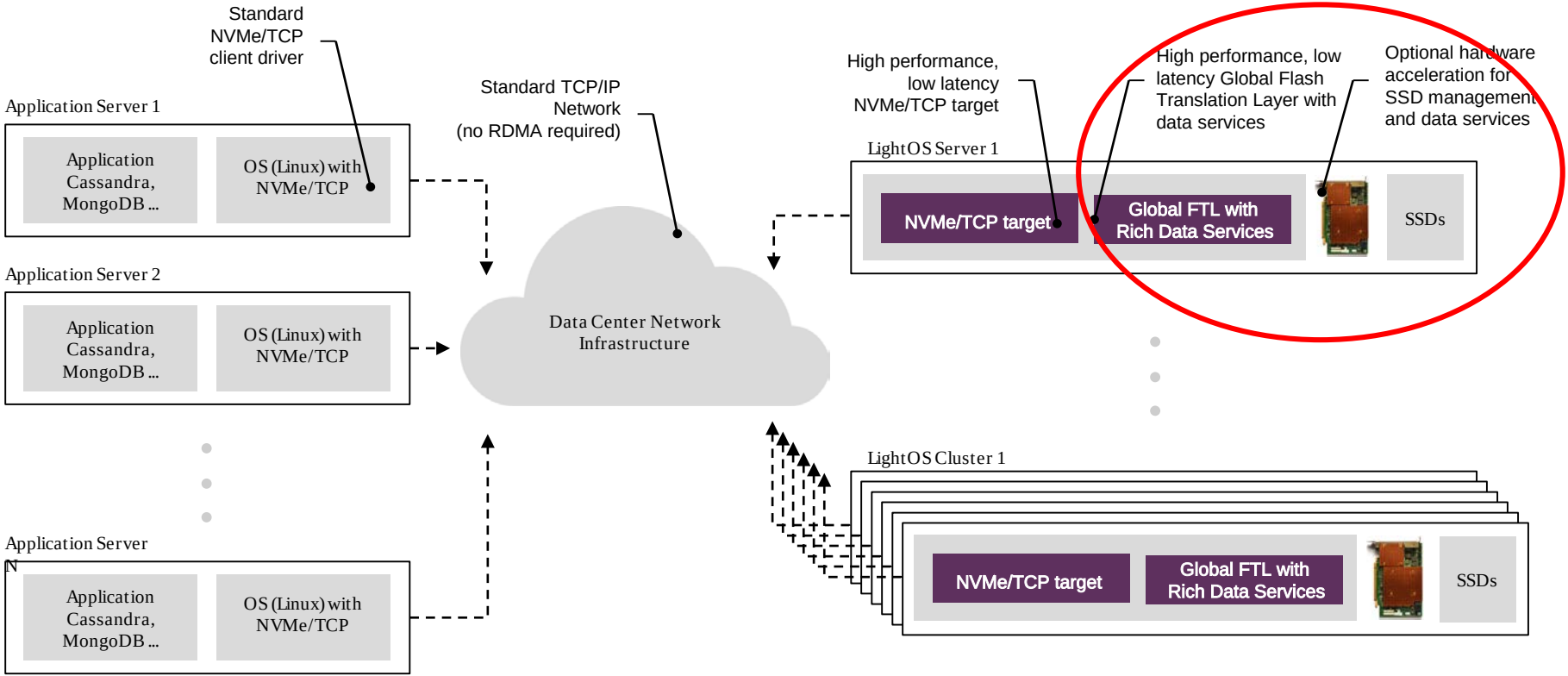
1. Disaggregated Storage - Why and How?
2. Lightbits LightOS(R) in a nutshell
3. Global Flash Translation Layer (GFTL™)
4. Data Services - Performance, Endurance and more.
5. Performance

From DAS to Disaggregated Storage



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

Lightbits LightOS solution building blocks



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

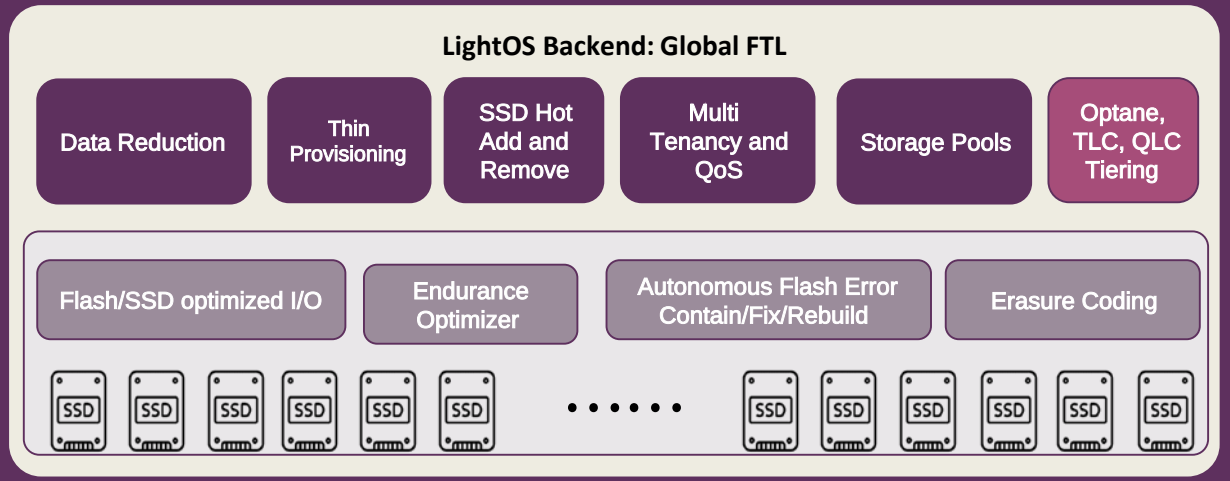
Scalable high performance

Enable new applications

Automated, API driven & designed for Cloud

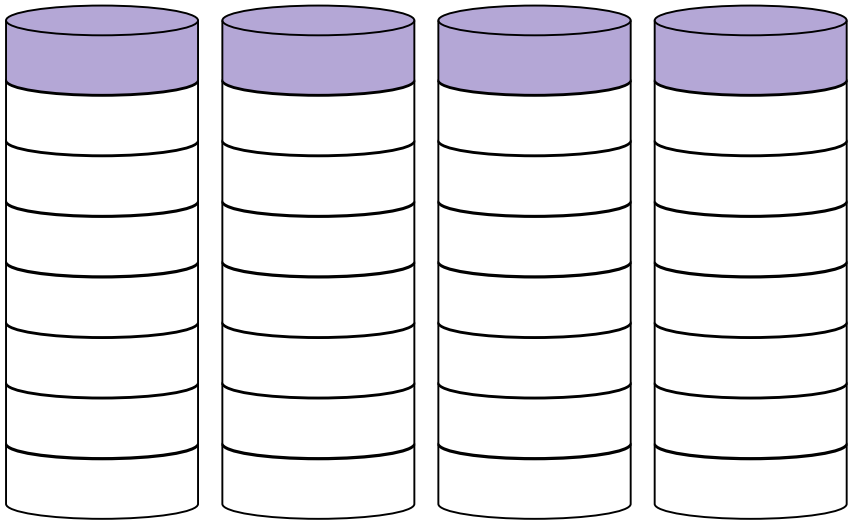
Agile, standard servers and SSDs

LightOS Global FTL (GFTL)



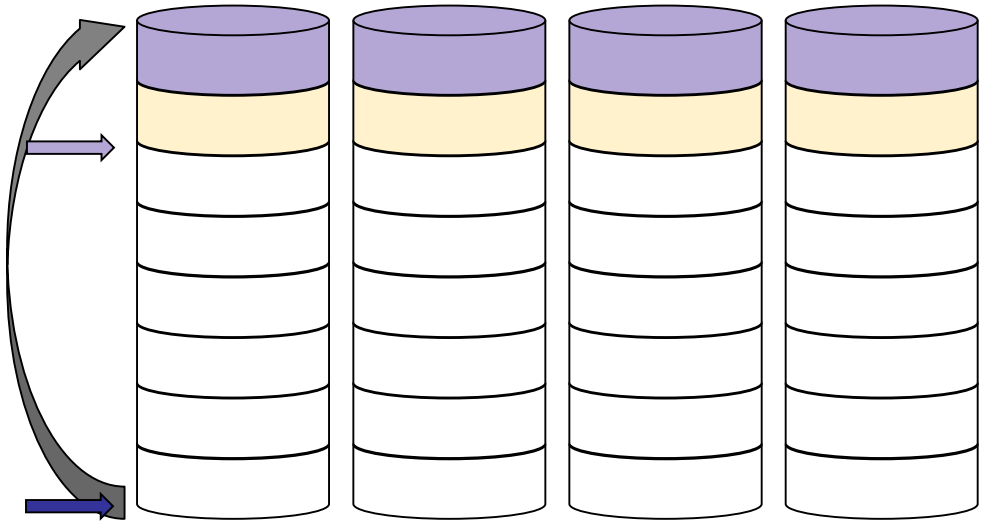
LightOS GFTL: Write Strategy

- Accumulate writes + sequential writes
- Fill complete stripe
- Thick stripes
- Metadata



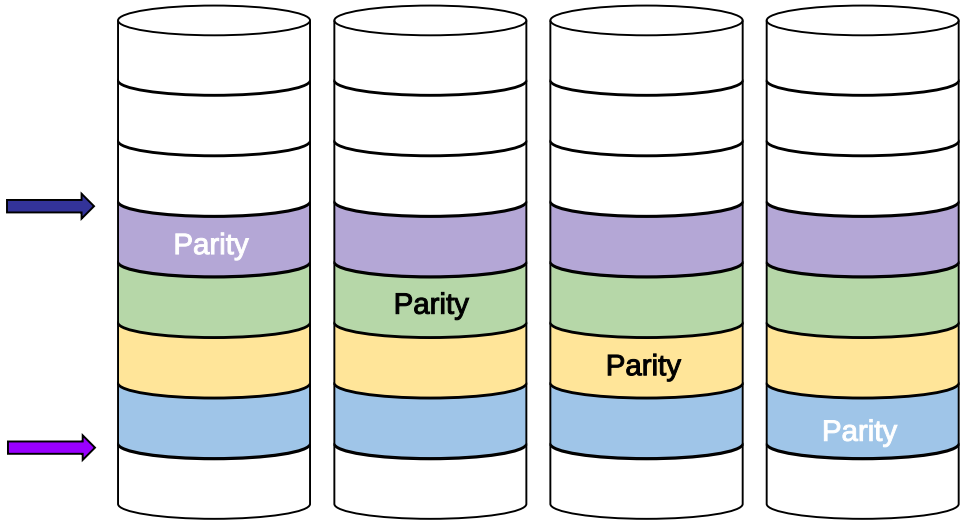
LightOS GFTL: Write Strategy

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



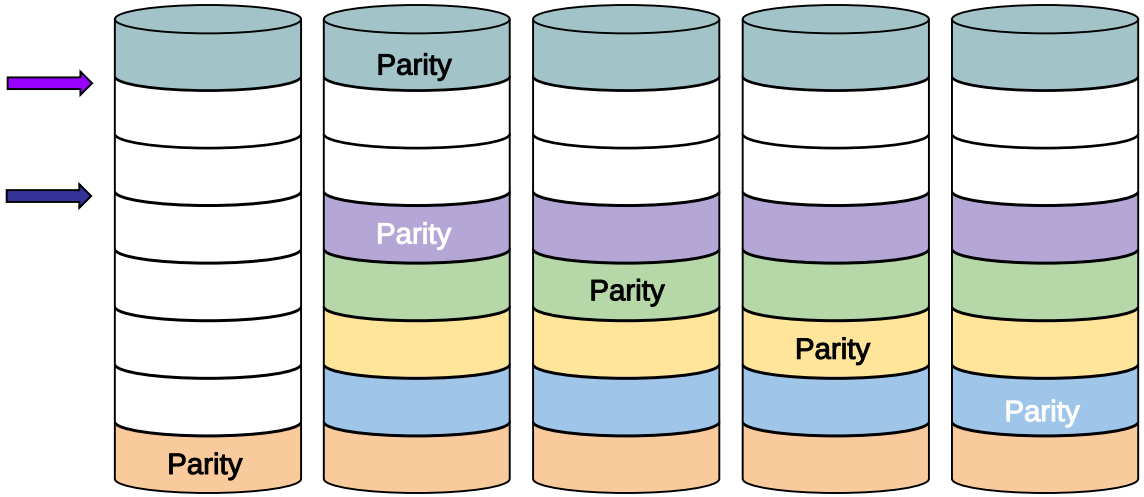
Erasure Coding

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



NVMe Drive Pooling

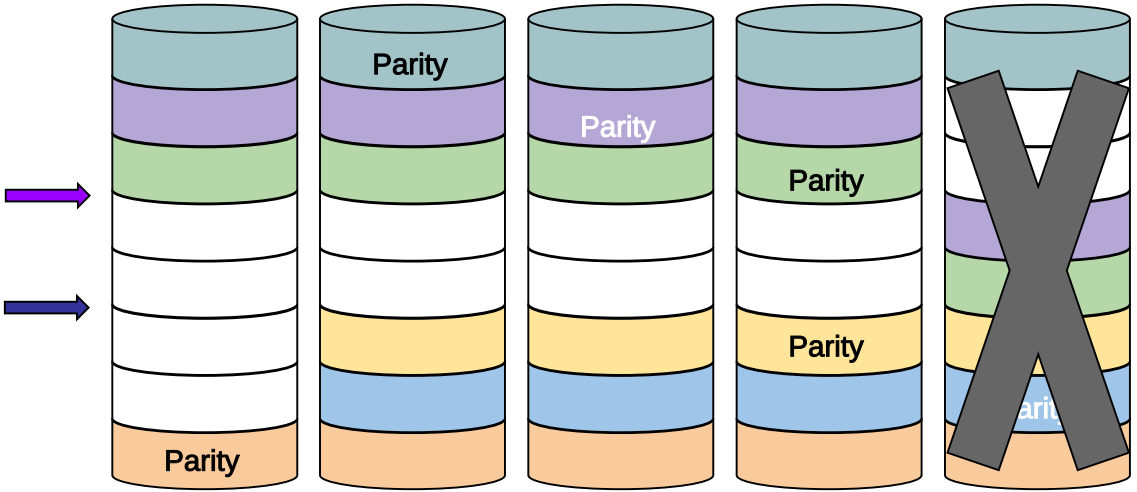
- Adding SSD
- Variable stripe width
- GC will gradually fix



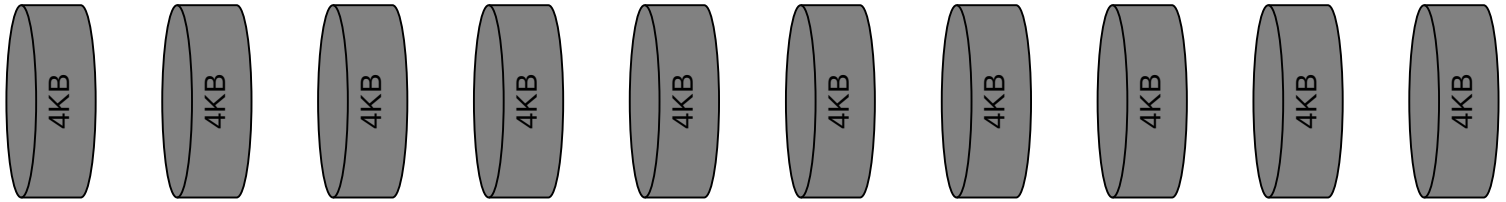
Drive Failure

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

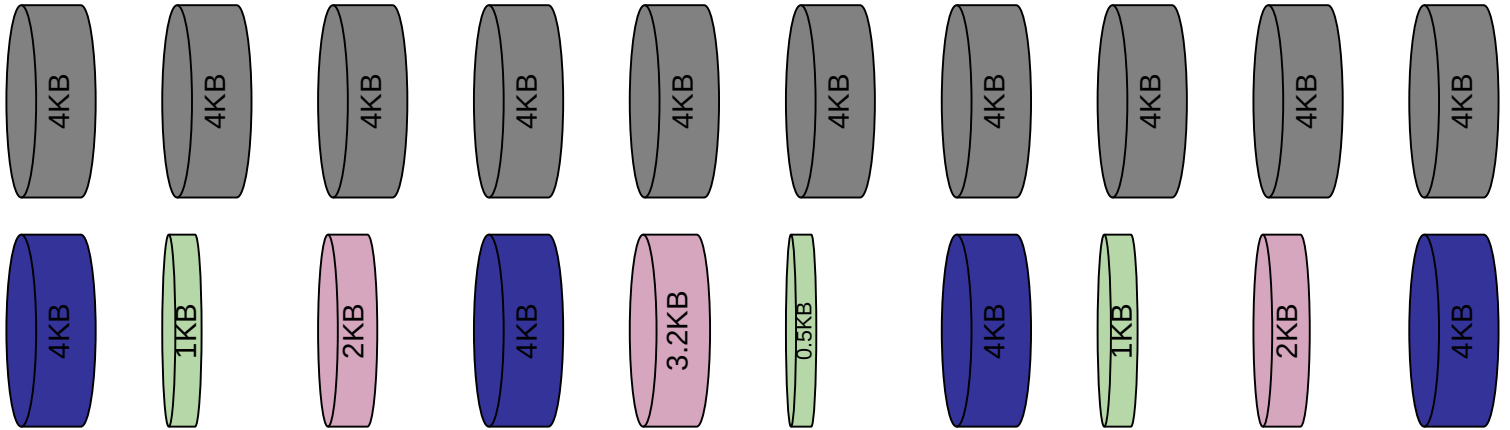
- SSD resets / transient failures handled by reducing stripe size and doing “read reconstruct”



Compression

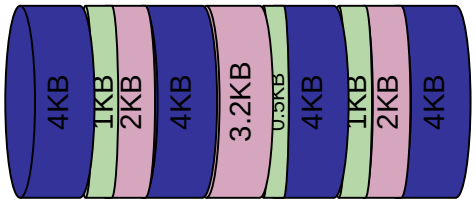


Compression



Compression

- Meta-data address alignment - 32 Bytes
- Optimal space utilization



Performance and Latencies (random 4k)

LightOS + LightField with 2:1 compression and EC data protection

	Read/Write: 70/30	Read/Write: 50/50	Write Only	Read Only
Max IOPS (M)	5M	3.8M	2M	5M
Typical IOPS (M)	3M	2.6M	1.6M	3M
Read avg latency typical (usecs)	240	221	-	242
Write avg latency typical (usecs)	89	72	48	-

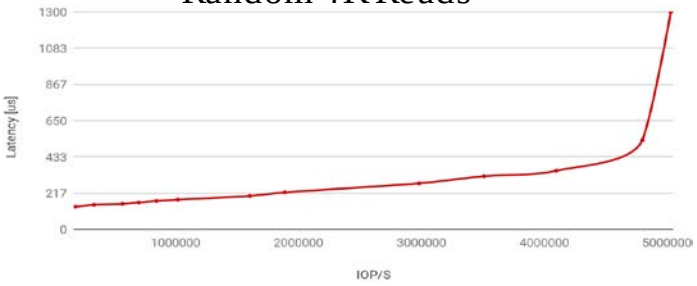
Storage Server
Single Dell 740XD
16x Intel P4510 8TB SSDs
Intel Xeon 6154 dual socket CPU
2x 100GbE ports

Clients
12x clients
Intel E5 2620 v4 CPU
25GbE port per client

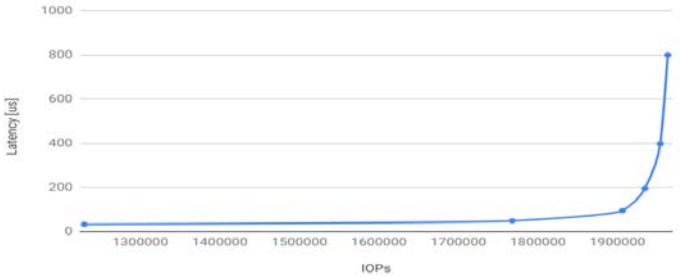
Performance and Latencies (random 4k)

LightOS + LightField with 2:1 compression and EC data protection

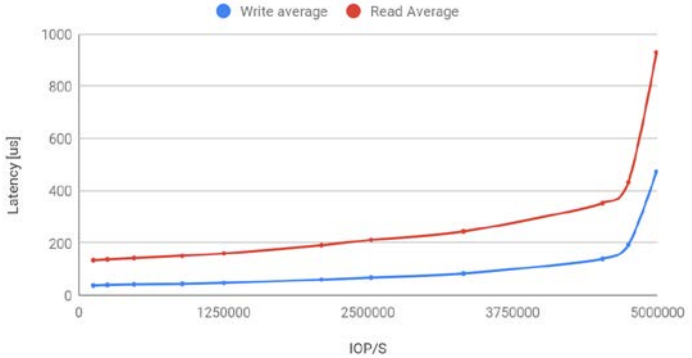
Random 4K Reads



Random 4K Writes



Random 4K 70/30 RW



Random 4K 50/50 RW

Data Protection

DAS RAID5 single 4KB write:

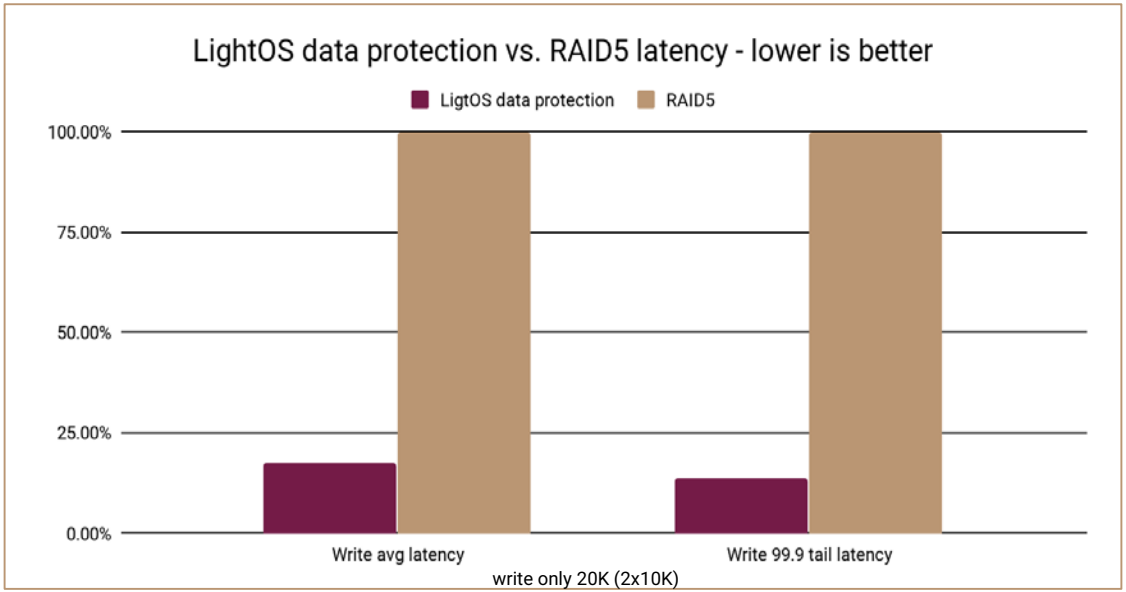
- Requires 2x4KB reads (old data and old parity)
- Requires 2x4KB writes (new data, new parity)
- Requires additional write to journal to avoid RAID5 write hole

LightOS GFTL with data protection single 4KB write:

- Requires single 4KB write (+ parity as the number of SSDs, e.g 1/8)
- No additional read or write IOPs
- Same latency as no data protection
- Same endurance as no data protection

LightOS GFTL enables data protection with no latency and no endurance penalty

Latency: FIO RAID5 vs. LightOS with data protection



Even at very low IOPs, LightOS with data protection has significantly lower latencies than than RAID5

Summary

- LightOS Global FTL offsets the inherent cost of NVMe/TCP by driving multiple SSDs together
- LightOS Global FTL provides data services such as compression, erasure, coding, others
- LightOS Global FTL beats Linux hands-on on performance and latencies
- LightOS Global FTL makes NVMe/TCP better than DAS

Contact information

https://www.lightbitslabs.com/

roys@lightbitslabs.com