A Global FTL Architecture to Drive Multiple SSDs

Roy Shterman
Lightbits Labs
Agenda

1. Disaggregated Storage - Why and How?
2. Lightbits LightOS\textsuperscript{(R)} in a nutshell
3. Global Flash Translation Layer (GFTL\textsuperscript{TM})
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

Application Server 1
- Application: Cassandra, MongoDB ...
- OS (Linux) with NVMe/TCP
- Standard TCP/IP Network (no RDMA required)

Data Center Network Infrastructure

Application Server 2
- Application: Cassandra, MongoDB ...
- OS (Linux) with NVMe/TCP

Application Server
- Application: Cassandra, MongoDB ...
- OS (Linux) with NVMe/TCP

LightOS Server 1
- High performance, low latency NVMe/TCP target
- Global FTL with Rich Data Services
- Optional hardware acceleration for SSD management and data services

LightOS Cluster 1
- NVMe/TCP target
- Global FTL with Rich Data Services
- SSDs
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)

- Data Reduction
- Thin Provisioning
- SSD Hot Add and Remove
- Multi Tenancy and QoS
- Storage Pools
- Optane, TLC, QLC Tiering

- Flash/SSD optimized I/O
- Endurance Optimizer
- Autonomous Flash Error Contain/Fix/Rebuild
- Erasure Coding
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

4KB 4KB 4KB 4KB 4KB 4KB 4KB 4KB 4KB
4KB 4KB 4KB 4KB 4KB 4KB 4KB 4KB 4KB
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

<table>
<thead>
<tr>
<th></th>
<th>Read/Write: 70/30</th>
<th>Read/Write: 50/50</th>
<th>Write Only</th>
<th>Read Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max IOPS (M)</td>
<td>5M</td>
<td>3.8M</td>
<td>2M</td>
<td>5M</td>
</tr>
<tr>
<td>Typical IOPS (M)</td>
<td>3M</td>
<td>2.6M</td>
<td>1.6M</td>
<td>3M</td>
</tr>
<tr>
<td>Read avg latency typical (usecs)</td>
<td>240</td>
<td>221</td>
<td>-</td>
<td>242</td>
</tr>
<tr>
<td>Write avg latency typical (usecs)</td>
<td>89</td>
<td>72</td>
<td>48</td>
<td>-</td>
</tr>
</tbody>
</table>

Storage Server
Single Dell 740XD
16x IntelP4510 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