OPEN POSSIBILITIES.

SmartFTL SSDs
SmartFTL SSDs

Chris Sabol, SSD Architect, Google
Smriti Desai, Storage Engineering Director, Google
Storage@Google

- **Growing** exponentially
- **Breadth** of use cases
- Use **HDD & SSD based storage systems**

Today’s focus ⇒ **SSDs at Google**
Google’s path to SmartFTL SSD

• History of SSDs at Google
• What is SmartFTL
  • How it works
  • The benefits
• Call to Action
## Google SSDs over 11 years

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Beginning: Simple PCIe SSD</strong></td>
<td>● Major use case: memory offload for Search. Think Google Instant Search.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Host software stack did pretty much everything (FTL, mapping table)</td>
</tr>
<tr>
<td>2</td>
<td><strong>Evolution of Internal SSD</strong></td>
<td>● Improvements were made to incrementally offload functionality to the drives</td>
</tr>
<tr>
<td>3</td>
<td><strong>3rd Party SSD</strong></td>
<td>● Benefit from industry optimizations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● We see a gap on GC efficiency; Close this gap through SmartFTL</td>
</tr>
</tbody>
</table>

---

**OPEN POSSIBILITIES.**
Google SSD Trends

- Host Offload
  - Kernel Bypass
  - FTL
- Growing NAND Complexity
  - ECC
  - 3D NAND
  - Bifurcation
- Performance continues to scale
  - Plane Growth
  - Independent Plane Read
  - Program Suspend

OPEN POSSIBILITIES.
Current industry architecture half works

Industry
- RAID across flash dies for error reduction
- Legacy HDD commands
- Single workload optimized

Google Use Case
- No data position control
- No application hints
- Highly parallel workloads

Result: Higher WAF

Cluster Filesystem

Local SSD
- Need low error rates
- Standard interface needed
- Lower parallelism

OPEN POSSIBILITIES.
Optimizing GC – Saves Bytes!

• Write Amplification Factor (WAF) is a function of workload and flash over provisioning (OP).
  • WAF is a measure of the extra writes you need to do for GC purposes.
  • With a WAF of 2.5 for every 1MB of write requested from the drive, 2.5MB must be written to flash; since extra writes are GC, the drive must do 1.5MB of extra reads as well.

Example: with random 4KiB writes, ~28% OP, and greedy GC algorithm, can expect WAF of ~2.5

<table>
<thead>
<tr>
<th>WAF Reduction from 2.5 to 1.25 can</th>
<th>Quantified Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduce OP / Higher Usable Capacity</td>
<td>18% Capex Savings</td>
</tr>
<tr>
<td>Enable 2x drive size with the same application write density</td>
<td>7% Capex / 15% Opex Savings</td>
</tr>
<tr>
<td>Double effective drive lifetime</td>
<td>0-35% Capex Savings</td>
</tr>
<tr>
<td>Enable 2x application write rate</td>
<td>Performance</td>
</tr>
</tbody>
</table>
Introducing SmartFTL – Google’s Solution
SmartFTL Goals

Baseline SSD

Drive Responsibility
- Flash management
- Write data positioning
- GC decisions

 DRIVE RESPONSIBILITY
- Flash management
- Write data positioning
- GC decisions

 Application Responsibility
- Write data positioning
- GC decisions

SmartFTL SSD

- Align control with information/knowledge and contain complexity at its source
- NAND management complexity owned by drive
- Layout/GC complexity comes from workload; owed by application

OPEN POSSIBILITIES
What is SmartFTL?

• Positioning IO Directive extension to NVMe
• Gives host control of data position down to the die level & access to multiple append points per die
• Data is still addressed by LBA; there are no LBA restrictions.
• Host GC Mode: Optional additional control of GC selection, targeting, & timing.
• Two new modes of operation
  • Pure positioning mode
  • Host managed GC mode
• Can coexist existing NVMe mode of operation
SmartFTL Performance Profile

- The application can tradeoff read/write performance and WAF.
- Pure random 32MB write WAF does not capture benefit of 64-128x stream isolation.
SmartFTL – Under the Hood
SmartFTL Positioning Mode Life of a Write

- Positioning mode is hugely valuable by itself.
- If you are going to get to WAF 1.0, you need to do it with optimal initial positioning.
• GC mode the host keep its own map
• Keeps drive from doing GC copies by managing block occupancy
• Allows full flexibility at host of host CPU and DRAM
• Drives typically have something like a RAID5 stripe across 64 dies with a 8KiB - 32 KiB stride
• Writes must append to the stripe; no ability to pick block.
• So if you deallocate / overwrite 1MiB, instead of freeing up 5% of a block, you free up 0.08% of 64 blocks.
• This matters a lot for objects in the 512KiB to 1GiB size range.
• Correlated lifetimes at the block level are critical for low WAF
SmartFTL is useful if your workload…

- Has erasure coding above the drive layer
- Data deletion is correlated at sizes between 16KiB and 1GiB
- 64KiB to 100MiB single object read performance doesn’t have to achieve full drive read throughput
- For example: a cluster filesystem workload
Call to Action

Standardize & Enable SmartFTL adoption
- Co-developing w/ SSD vendors - Samsung, Kioxia & Intel
- Standardize SmartFTL interface through NVMe
- Open source a smartftl library to provide useful application hooks / abstractions

Engage in OCP Storage workgroup & SSD specs

We hope other SSD users will find this valuable, believe it addresses a common challenge, and would love to hear your feedback.
Open Discussion