Tunable (and Flexible) Flash Translation Layer Improves Storage System Performance Behavior

Chris Bergman - Burlywood
Overview

- Intended Audience:
  - **Data center and cloud storage system designers and architects**
- Current Configuration Options
  - Capacity, OP%, and Endurance
- Why not have more options?
  - GC and Wear Leveling Schemes => Where and how data is placed on the media.
  - Performance Optimizations
  - Data Integrity
- How you evaluate your options is important!
  - Benchmarks, standard tests, and data sheets can be misleading
- There is opportunity!
  - Lower Total Cost of Ownership, Improved drive life
  - Better and more consistent performance
Why do we need an FTL?

- Media Granularities (Erase >> Write >= Read), Sequential Programming
- Endurance, read disturb, retention, power loss handling, defect handling
- It’s not just flash => Shingled Magnetic Recording HDD, Storage Class Memory

It’s the Storage Media Properties!
What does an FTL do about these issues?

<table>
<thead>
<tr>
<th>Feature</th>
<th>Side Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation Tables</td>
<td>Memory Cost (1:1000), Performance</td>
</tr>
<tr>
<td>Garbage Collection</td>
<td>Performance, Drive Life [Write Amp]</td>
</tr>
<tr>
<td>Wear Leveling</td>
<td>Performance, Drive Life [Write Amp]</td>
</tr>
<tr>
<td>Data Integrity (ECC/RAID)</td>
<td>Capacity, Performance, Drive Life</td>
</tr>
<tr>
<td>Background Scanning</td>
<td>Data Integrity, Performance</td>
</tr>
<tr>
<td>Read/write priority (QoS)</td>
<td>Performance</td>
</tr>
<tr>
<td>Overprovisioning</td>
<td>Media Cost, Performance</td>
</tr>
</tbody>
</table>

Necessary evils!
How is it all related?

Lower cost (OP) = lower performance and shorter drive life.

Normalized Random Write Performance vs. Write Amp

Drive Life vs. Write Amp
Based on 3 yrs, 7% OP

2.5 GB/s
325 MB/s (13%)
Imagine the flexibility to optimize based on the application and use model

Requires knowledge of the workload at the drive

Traditional FTL’s are statically configured, one size fits all
  - Pick a point on the graphs and that’s what you get
  - Designed for least common denominator (4K random write)
  - One Firmware update away from trouble

If you’re not the least common denominator you’re sacrificing something!
Examples

- Workload complexity
- Read, Write Mixed Workloads, Consistency and QoS
- Good intentions = Not so good results
- Data integrity, ECC optimization
Workloads Are Complex

Even the simplest scenario can be complex.

Jetstress(2) + virtualization, RAID, snapshots, etc.
Visualize the workload

Periodic, heavy write activity followed by very little write activity. Always mixed read/write.
Tuning Results

- Scenario: Optimize garbage collection selection and timing

Tuning improved overall performance and consistency.
Highly sequential workload may not lead to better behavior.
Drive Snapshots

Standard FTL

Optimized FTL
Tuning Results

- Scenario: 16 TB drive, 7% nominal OP, 3 year drive life
- Tuned: Weight OP to small random area + table optimization

<table>
<thead>
<tr>
<th>Feature</th>
<th>Std FTL</th>
<th>Tuned FTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effective WA</td>
<td>4-7</td>
<td>~1</td>
</tr>
<tr>
<td>Performance</td>
<td>14%-25% of FOB*</td>
<td>~100% of FOB</td>
</tr>
<tr>
<td>Drive Life†</td>
<td>up to 3.3-5.7 years</td>
<td>up to 20+ years</td>
</tr>
<tr>
<td>DRAM</td>
<td>16 GB</td>
<td>&lt; 1 GB (or used for other purposes)</td>
</tr>
</tbody>
</table>

* Fresh Out of Box, † Relative to 100% Random Write Workload

Tuned = significant benefit. Standard configuration is highly susceptible to design choices.
Summary

- One size fits all is likely costing you something
- Knowledge is **KEY** and using that knowledge can lead to
  - Lower Total Cost of Ownership
  - Better and more consistent performance
  - Improved drive life
  - Proper evaluation of your storage solution
- Benchmarks and “standard” workloads don’t tell the whole story
- This is even more important in data center applications where inefficiencies can be amplified by 100x, 1000x, 10000x, …
References & Contributors

1) Write Amplification Calculation
   ▪ [http://www.ece.neu.edu/groups/nucar/NUCARTALKS/WriteAmplification.pdf](http://www.ece.neu.edu/groups/nucar/NUCARTALKS/WriteAmplification.pdf)

2) Jetstress Workload Emulation

Thanks to those who contributed time and effort to this presentation:
Nate Koch, Ed Daelli, John Slattery, Mike Tomky, Tod Earhart, John Murphy, & the entire Burlywood team!
Understand the system requirements

Redundancy at many levels
Tune accordingly

- Scenario: 4 TB drive, 7% nominal OP
- Eliminate or reduce protection on drive, redirect to OP

<table>
<thead>
<tr>
<th>Config</th>
<th>Metric</th>
<th>LUN Protection</th>
<th>Single Plane Protection</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>128 planes/stripe [quad plane]</td>
<td>Perf</td>
<td>13% of FOB*</td>
<td>20% of FOB</td>
<td>21% of FOB</td>
</tr>
<tr>
<td></td>
<td>Life†</td>
<td>3 years</td>
<td>4.0 years</td>
<td>4.2 years</td>
</tr>
<tr>
<td>128 planes/stripe [dual plane]</td>
<td>Perf</td>
<td>13% of FOB</td>
<td>16% of FOB</td>
<td>17% of FOB</td>
</tr>
<tr>
<td></td>
<td>Life</td>
<td>3 years</td>
<td>3.4 years</td>
<td>3.6 years</td>
</tr>
</tbody>
</table>

* Fresh Out of Box, †Relative to 100% Random Write Workload
Jetstress IO Breakdown

\[ \leq 32K \text{ WRs: } 75\% \text{ of the traffic, but only } \sim 12\% \text{ of capacity} \]
Jetstress IO Breakdown

Heatmap view of the same data