GRIFFIN

Extending SSD Lifetimes with Disk-Based Write Caches

Gokul Soundararajan
Vijayan Prabhakaran
Mahesh Balakrishnan
Ted Wobber

FAST 2010
# Flash Drives

## Intel X25-M

<table>
<thead>
<tr>
<th></th>
<th>Fast Reads</th>
<th>Slow Random Writes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequential read rate</strong></td>
<td>250 MB/s</td>
<td>70 MB/s</td>
</tr>
<tr>
<td><strong>Random 4 KB reads</strong></td>
<td>35000</td>
<td>3300</td>
</tr>
<tr>
<td><strong>Sequential write rate</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Random 4KB writes</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# MLC vs SLC

<table>
<thead>
<tr>
<th></th>
<th>Single Level Cell</th>
<th>Multi Level Cell</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Density</strong></td>
<td>16Mbit</td>
<td>32Mbit / 64Mbit</td>
</tr>
<tr>
<td><strong>Endurance</strong></td>
<td>100,000 cycles</td>
<td>10,000 cycles</td>
</tr>
<tr>
<td><strong>Cost (128 GB)</strong></td>
<td>$1200</td>
<td>$300</td>
</tr>
</tbody>
</table>
SSD Write Lifetime

- Ideal: 300 TB
- Micron C200: 42 TB
- Intel X-25M: 37 TB
Write Cache

Benefit - Coalesce Overwrites to Increase Lifetime
Hybrid design Options

RAM
- Fast
- Not persistent

HDD
- Good Sequential Write Speed
- Cheap, Large Capacity

Flash
- Lower Power
- Higher Cost
Hybrid design Options

RAM
- Fast
- Not persistent

HDD
- Good Sequential Write Speed
- Cheap, Large Capacity

Flash
- Lower Power
- Higher Cost
Coalesced Writes

- **Buffer Cache**
- **Coalesced Write**
- **Evicted**

- **W:100 R:100**
- **W:100 W:100**
- **W:100 R:100**
Ideal Savings

Potential Reduction:
36-64% - Desktops       94% - Web server    62% - Linux Desktop
WAW and RAW

- **Write After Write (WAW)**
  
  Time between two consecutive writes to the same block

- **Read After Write (RAW)**
  
  Time between a write and a subsequent read to the same block
WAW/RAW Intervals

![Graph showing cumulative percentages of WAW and RAW intervals over time.]

- **WAW**: Red line, increases sharply in the initial seconds, plateaus around 360 seconds, then continues to rise slowly.
- **RAW**: Green line, rises gradually over time, reaching a cumulative percentage of 60% at 3,600 seconds.

**X-axis**: Histogram Buckets - Seconds
**Y-axis**: Cumulative %
Griffin Write

- LOG: 1
- W: 100

- HDD
- SSD
Griffin Write

Blockmap

<table>
<thead>
<tr>
<th>Block</th>
<th>Log</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1</td>
</tr>
</tbody>
</table>

Griffin

HDD

SSD
Griffin Read
What to cache and How long to Cache

• Metrics:

• Write Savings - Percentage of writes that don’t reach the SSD

• Read Penalty - Percentage of reads serviced by HDD

• Fault Tolerance - HDD failures
What to Cache – Selective Caching

- Maintain overwrite ratio for each block
- Higher overhead but lower read penalty

<table>
<thead>
<tr>
<th>Blockmap</th>
<th>Blockmap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block</td>
<td>Log</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Block</td>
<td>Overwritten?</td>
</tr>
<tr>
<td>100</td>
<td>✓</td>
</tr>
</tbody>
</table>
What to Cache – Write Savings

Selective Caching (overwrite threshold = 0.25)  21% average
Full Caching  51% average
What to Cache – Read Penalty

Average Read Penalty – 20%
How long to Cache

- Timeout Trigger
  - Ex: Migrate every 5 minutes
  - Bounds data lost due to failure

- Read-Threshold Trigger:
  - Ex: Read Penalty should not exceed 5%
  - Ensures performance is reasonable

- Migration Size Trigger:
  - Ex: Migrate if cache size is > 100 MB.

- Hybrid scheme – Combine all triggers
How long to cache – Write Savings

SSD write savings (percentage of total writes)

- Migration timeout 3600 s
- Migration timeout 1800 s
- Migration timeout 900 s

Traces:
- D-1A
- D-1B
- D-1C
- D-1D
- D-2A
- D-2B
- D-2C
- D-2D
- D-3A
- D-3B
- D-3C
- D-3D
- S-EXCH
- S-PRXY1
- S-SRC10
- S-SRC22
- S-STG1
- S-WDEV2
How long to cache – Read Penalty

- Migration timeout 3600 s
- Migration timeout 1800 s
- Migration timeout 900 s

HDD read penalty (percentage of total reads)

Traces:
- D-1A
- D-1B
- D-1C
- D-1D
- D-2A
- D-2B
- D-2C
- D-2D
- D-3A
- D-3B
- D-3C
- D-3D
- S-EXCH
- S-PRXY1
- S-SRC10
- S-SRC22
- S-STG1
- S-WDEV2
Handling Failure

- Power Failures
  - Recovery similar to log-structured and journaling file systems

- Device Failures
  - HDD: Additional point of failure in storage stack
  - Full Caching - Most recent writes are lost
  - Selective Caching - More complex data recovery
Latency measurements

Relative to MLC-based SSD without Write Cache.
Discussion

• File system level vs Block level
  • File systems: More aware of what files to cache
  • Block Device: No modifications to software stack

• Power Consumption due to HDD

• Failure handling without caching all writes?

• Adoption of Griffin given price and technology changes

• Phase Change Memory vs Flash
## Comparison

<table>
<thead>
<tr>
<th></th>
<th>DRAM</th>
<th>Phase Change Memory</th>
<th>MLC NAND</th>
<th>HDD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Granularity</strong></td>
<td>Bit</td>
<td>Bit</td>
<td>Block</td>
<td>Sector</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>~W/GB</td>
<td>100-500mW /die</td>
<td>~100 mv/die</td>
<td>~10W</td>
</tr>
<tr>
<td><strong>Write Bandwidth</strong></td>
<td>~GB/s</td>
<td>1-100+ MB/s/die</td>
<td>~10 MB/s/die</td>
<td>200-400 MB/s</td>
</tr>
<tr>
<td><strong>Write Latency</strong></td>
<td>20-50 ns</td>
<td>~1 µs</td>
<td>~800 µs</td>
<td>~10 ms</td>
</tr>
<tr>
<td><strong>Read Latency</strong></td>
<td>50 ns</td>
<td>50-100 ns</td>
<td>25-50 µs</td>
<td>~10 ms</td>
</tr>
<tr>
<td><strong>Endurance</strong></td>
<td>∞</td>
<td>$10^8$</td>
<td>$10^4$</td>
<td>∞</td>
</tr>
</tbody>
</table>