ZFS, Cache, and Flash

Adam Leventhal
Sun Microsystems, Fishworks
blogs.sun.com/ahl
• Initial release of ZFS
• Vision: enterprise grade storage from commodity components
• Designed for manageability, scale, performance, etc.
Performance c. 2005

- **Enterprise storage**
  - 15K RPM FC or SAS disks, NVRAM
- **ZFS**
  - 7200 RPM SATA disks, no NVRAM
- **Result**
  - Though designed for performance
  - Hampered by commodity components
- **Note:** drives relied upon to deliver capacity, throughput, IOPS
Storage Disruption: Flash

- **Flash memory**
  - Non-volatile
  - Fast for writes (300µs)
  - Faster for reads (50µs)
  - 2001 (birth of ZFS) cost as much as DRAM
  - 2009 less than 10th the cost of DRAM

- Still much more expensive than disk
- Inappropriate as a general purpose replacement
Hybrid Storage Pool

- Use flash to complement the storage hierarchy (DRAM → disk)

- Forms a completely new tier for price/performance
- New ZFS features to enable HSP
- ZIL slogs: separate log devices
  - Props to Neil Perrin
- L2ARC: second level cache
  - Props to Brendan Gregg
- Both available in OpenSolaris today
Ideal HSP Devices

- **ZIL slog devices**
  - Very low-latency, high-IOPS for writes
  - 10GB capacity is plenty

- **L2ARC devices**
  - Low-latency, high-IOPS for reads
  - High capacity (enough to cache the full working set)
  - Low $/GB
SSDs 2008

• SSDs designed as drive replacements
  – e.g. for your laptop
• Pretty good for ZIL slog, L2ARC
• For example, Intel X25-E
  – 32GB, 3K write IOPS, 35K read IOPS, $15/GB
  – ZIL slog: want more write IOPS
  – L2ARC: want larger capacity
HSP Example

4 Xeon 7350 Processors (16 cores)
32GB FB DDR2 ECC DRAM
OpenSolaris with ZFS

Configuration A:
(7) 146GB 10,000 RPM SAS Drives

Configuration B:
(1) 32G SSD ZIL Device
(1) 80G SSD Cache Device
(5) 400GB 4200 RPM SATA Drives
HSP Results

- Hybrid Storage Pool (DRAM + Read SSD + Write SSD + 5x 4200 RPM SATA)
- Traditional Storage Pool (DRAM + 7x 10K RPM 2.5”)

<table>
<thead>
<tr>
<th>Category</th>
<th>Hybrid</th>
<th>Traditional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read IOPs</td>
<td>3.2x</td>
<td>11%</td>
</tr>
<tr>
<td>Write IOPs</td>
<td></td>
<td>4%</td>
</tr>
<tr>
<td>Cost</td>
<td>4%</td>
<td>4.9x</td>
</tr>
<tr>
<td>Storage Power (Watts)</td>
<td></td>
<td>2x</td>
</tr>
<tr>
<td>Raw Capacity (TB)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
HSP in the SS 7410

• Sun Storage 7410
• ZIL device: Logzilla
  – 10K write IOPS
  – Scales with more devices
• L2ARC device: Readzilla
  – 20K read IOPS
  – 6 x 100GB
SS 7410 HSP Results

Sun Storage 7410

Read IOPS increased by 500% with the L2ARC

http://blogs.sun.com/brendan/entry/l2arcScreenshots
HSP Impact

- Traditional storage pool uses drives for all aspects of capacity and performance
- HSP breaks these apart
  - Disks for capacity, throughput
  - Read-optimized SSD for read IOPS
  - Write-optimized SSD for write IOPS
- Optimal use, optimal economics
- Scale to fit the application
ZFS Commands

• Create a pool with log and cache devices
  
zpool create pool <vdevs ...> log <logzilla> cache <readzilla>

• Add log and cache devices to a pool
  
zpool add pool cache <readzilla>
zpool add pool log <logzilla>

• Cheap SSDs today are sufficient for testing
Conclusions

- Hybrid Storage Pool enabled by ZFS
- Best use of resources: DRAM, flash, disk
- Uses flash seamlessly as a new tier in the storage hierarchy
- HSP well positioned to use cheaper, less-reliable MLC flash for L2ARC
- SSDs designed for the HSP in 2009
- HSM without the M
Q&A

Adam Leventhal
Sun Microsystems, Fishworks
blogs.sun.com/ahl

Links:
http://blogs.sun.com/ahl/entry/hybrid_storage_pools_in_cacm
http://blogs.sun.com/brendan/entry/l2arcScreenshots
http://blogs.sun.com/brendan/entry/test