EMC Disk Tiering Technology Review

Tony Negro
EMC Corporation

Wednesday, February 06, 2013 12:15 PM
Session Number 13154
Agenda

• Basis for FAST
• Implementation Characteristics
• Operational Considerations
• Planning for Performance
Basis for FAST

- With information growth trends, all Fibre Channel (FC) configurations will:
  - Cost too much
  - Consume too much energy
  - Take up too much space
- FAST helps by leveraging disk drive technologies
- What makes FAST work in real-world environments?
  - **Skew**: At any given time, only a small address range is active – the smaller the range, the better
  - **Persistence**: If an address range is active (or inactive), it remains so for a while – the longer the duration, the better

80% of IO’s on 20% of capacity
Wide striping and short stroking are common practice

- The vast majority of ‘online-apps’ workloads enjoy high cache-hit percentages, but service levels are dictated by read-misses during transitional periods like market open.

60 TB of RAID 5 3+1 Storage
Read misses per second per Device type

<table>
<thead>
<tr>
<th># of drives</th>
<th>Read Misses per sec (small blocks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>146 15K (588)</td>
<td>105840</td>
</tr>
<tr>
<td>300 15K (296)</td>
<td>53280</td>
</tr>
<tr>
<td>300 10K (296)</td>
<td>38480</td>
</tr>
<tr>
<td>450 15K (196)</td>
<td>35280</td>
</tr>
<tr>
<td>600 15K (148)</td>
<td>26640</td>
</tr>
<tr>
<td>600 10K (148)</td>
<td>19240</td>
</tr>
<tr>
<td>1TB (88)</td>
<td>6160</td>
</tr>
<tr>
<td>2TB (48)</td>
<td>3360</td>
</tr>
</tbody>
</table>

$ per GB, footprint, power
Measuring Asymmetry in Access Patterns – Skew

Data Extents – Sorted by Activity Level

Enterprise Flash Drive targets
FC targets
SATA II targets

I/Os per Second

Devices: total I/Os per sec (Max)
IMPLEMENTATION CHARACTERISTICS
## 2008 – Fundamental Storage Media Shift

<table>
<thead>
<tr>
<th>Disk</th>
<th>Cost/GB</th>
<th>IOPS/GB</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 K RPM</td>
<td>1</td>
<td>1</td>
<td>6 ms</td>
</tr>
<tr>
<td>Serial ATA (SATA) 7,200 RPM</td>
<td>1/3</td>
<td>1/6</td>
<td>12 ms</td>
</tr>
<tr>
<td>Enterprise Flash Drive (EFD)</td>
<td>8</td>
<td>30</td>
<td>&lt; 1 ms</td>
</tr>
</tbody>
</table>
Data Movement Granularity Trade-offs

- Larger granularity
  - Uses EFD ineffectively

- Smaller granularity
  - Uses EFD effectively
  - Requires more system resources to maintain statistics

- There is a sweet spot that maximizes the benefits through better use of EFD and reasonable system resource use
## The impact of a larger extent size on configuration cost

<table>
<thead>
<tr>
<th>Disk</th>
<th>Cost GB</th>
<th>No. of TB</th>
<th>Capacity Distrib</th>
<th>$Cost</th>
<th>Capacity Distrib</th>
<th>$Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 K RPM (SATA) 7,200 RPM</td>
<td>1/3</td>
<td>24</td>
<td>40%</td>
<td>$13.3</td>
<td>40%</td>
<td>$13.3</td>
</tr>
<tr>
<td>Flash Drive (EFD)</td>
<td>8</td>
<td>4</td>
<td>7%</td>
<td>$32.0</td>
<td>21%</td>
<td>$100.8</td>
</tr>
<tr>
<td></td>
<td>60TB</td>
<td>-</td>
<td></td>
<td>$98.3</td>
<td></td>
<td>$137.5</td>
</tr>
</tbody>
</table>

A configuration with 3X more Flash is 40% more expensive
Storage Elements

- **Symmetrix Tier** – a shared storage resource with common technologies (Virtual Pools)
- **FAST Policy** – manage Symmetrix Tiers to achieve service levels for one or more Storage Groups
- **FAST Storage Group** – logical grouping of thin devices for common management
FAST VP Time Windows

- **Performance time window** defines when statistics are collected and decayed
- **Workload Analysis Period**
  - Affects decay rates
  - Affects time to respond to changes
  - Default of 7 days (168 hours)
- **Data movement time window** defines when FAST VP is allowed to move data
  - Windows can be customized, but recommendation for initial implementation is 24x7
FAST VP Implementation

- Performance data collected by the system
- **Intelligent Tiering** algorithm generates movement requests based on performance data
- **Allocation Compliance** algorithm generates movement requests based on capacity utilization
- Data movements executed by VLUN VP data movement engine
- Algorithms continuously assess I/O statistics and capacity use, and make decisions for promotion and demotion
FAST VP Hierarchy

• Extent Group
  – 10 Track Groups (thin device extents)
  – 7.5 MB FBA / 6.8 MB CKD
  – Data movement unit

• Track Group (Thin Device Extent)
  – 768 KB FBA / 680 KB CKD
  – VP allocation unit

• I/O rates are collected during the “open” performance time window
  – Read Miss (RM)
  – Write (W)
  – Prefetch (P)

• Rates are updated every 10 minutes changing the ‘score’ of the Extent Group Set
FAST VP – Score Analysis

- Extents Group Sets are grouped in a histogram according to the Score

- Start from left, and fill in tiers to determine promotion thresholds

- Allocated Storage

- Prioritized Short Term Score Density

- Low Score Density to fill SATA tier

- Medium Score Density to fill FC tier

- High Score Density to fill EFD tier

- FC Threshold

- EFD Threshold
Movement Decisions Promotion

- Movement decisions are made at extent group level
- The extents with a high I/O density are candidates for EFD
- The extent is promoted to EFD if
  - EFD is not overloaded
  - There is potential for Response Time improvement
- Otherwise the extent will be promoted to the next best tier
Control Mechanisms

- Storage Group Prioritization – useful when multiple SG’s are associated to the same policy and competing for the same resources

Data Movement Speed

- High priority is moved to FLASH first

FRR: Fast Relocation Rate

FRR = 1 is the fastest
Virtual LUN VP Mobility and Pinning

- Using VLUN, you can control the location of devices
  - Move all extents of a device to a desired pool

- FAST VP allows pinning all extents of a device in their current locations
  - When pinned, FAST VP does not promote or demote extents of the device

- Leverage these two features to **override FAST VP**
  - For example, move previously active volumes to EFD and pin them in preparation for end-of-quarter processing
FAST VP SRDF Support

- SRDF Integration enables predictable performance during failover
  - Full RDF Awareness to FAST VP
  - R2 system reflects promotion and demotion decisions of the R1 system

FAST VP R2 statistics are merged with the R1 to reflect the R1 Read Miss ratio

Enabled per Storage Group
Requires R2 devices to also be under FAST VP control
FAST VP “SMALL”

Virtual Pool Compression

$/IOP
OPTIMIZED

MAXIMIZE IO DENSITY

$/TB
OPTIMIZED

MAXIMIZE TB EFFICIENCY

UP TO
~3:1
• Done Well, Tiered Storage has Great Potential…
Planning for Performance
Tier Advisor

- Disk modeling tool to:
  - Identify Skew & Workload
  - Find best price vs. performance combination
  - Ensure that Tiers have the capabilities to support current load and growth

Tier Advisor is recommended for planning. It is not required for FAST VP operations.
Example of a Multi-tiered MF Configuration

% of total capacity in each Tier

<table>
<thead>
<tr>
<th>Tier</th>
<th>Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>EFD</td>
<td>2-3%</td>
</tr>
<tr>
<td>FC/SAS</td>
<td>57%</td>
</tr>
<tr>
<td>SATA</td>
<td>40%</td>
</tr>
</tbody>
</table>

- The EFD and FC/SAS Tiers are expected to capture more than 90% of the System I/Os

% of load expected in each Tier for skews near 80%

<table>
<thead>
<tr>
<th>Tier</th>
<th>Load</th>
</tr>
</thead>
<tbody>
<tr>
<td>EFD</td>
<td>20-30%</td>
</tr>
<tr>
<td>FC</td>
<td>60-70%</td>
</tr>
<tr>
<td>SATA</td>
<td>&lt;2%</td>
</tr>
</tbody>
</table>

- The FC/SAS tier is expected to have a higher I/O density
Choose the Right Disk Type to the FC/SAS Tier

- As a general rule of thumb consider smaller and faster FC/SAS Disks in the mid-Tier because of the expected I/O density
- But note that we cannot be prescriptive...
  - Best recommendation still is to do the proper analysis to find the drive type that matches the workload
  - Ex: larger drive sizes can be a good match for light workloads

The number of IOPS per GB on a 300 GB 15K drive is ~3X higher than in a 600 GB 10K
Ensure that Tiers have the Capability to Support Current Load AND Growth

Configuration 1 is 10% less expensive than Configuration 2

This is only an example
**CURRENT CONFIG OF A MF ARRAY**

<table>
<thead>
<tr>
<th>Name</th>
<th>Rel Cost</th>
<th>Rel RT</th>
<th>Rel Pwr</th>
<th>Rel Cap</th>
<th>Log Alloc</th>
<th>Disks</th>
<th>EFD 95G</th>
<th>FC 15K 279G</th>
<th>SATA 1827G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>44.60 TB</td>
<td>320</td>
<td>0</td>
<td>320</td>
</tr>
</tbody>
</table>

**Configuration: Base**

- **Group**: FULLARRAY
- **Cost**: 100.0%
- **Rel RT**: 0%
- **Log Alloc**: 44.600 TB
- **Policy**: FC Only

**Policies**

<table>
<thead>
<tr>
<th>Policy</th>
<th>Rel Cost</th>
<th>Rel RT</th>
<th>Flash 100GB 7R5</th>
<th>15K300 Mir</th>
<th>Sata 2T 6R6</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC Only</td>
<td>0%</td>
<td>0%</td>
<td>100 / 100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Disk Utilization**

- Skew: 72.4% / 27.6%

**Disk Utilization**

- 320 Disk
- FC 15K 279G
- Utilization

**Policy Relative Response Time**

- Response Time (X)
- Rel ± %
- Capacity TB
- Avg+Std

**I/O Sec**: 52000
- BE I/O Sec: 17230
- BE Writes (%): 61.3%
## Potential for improvement

<table>
<thead>
<tr>
<th>Name</th>
<th>Rel Cost</th>
<th>Rel ST</th>
<th>Rel Pwr</th>
<th>EFD</th>
<th>FC</th>
<th>SATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Multi-Tiered</td>
<td>-10%</td>
<td>-24%</td>
<td>-32%</td>
<td>3%</td>
<td>60%</td>
<td>37%</td>
</tr>
</tbody>
</table>

**Policies:**

<table>
<thead>
<tr>
<th>Group</th>
<th>Cost</th>
<th>Rel RT</th>
<th>Log Alloc</th>
<th>Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>FULLARRAY</td>
<td>100.0%</td>
<td>-24%</td>
<td>44.600 TB</td>
<td>3Tier</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Policy</th>
<th>Rel Cost</th>
<th>Rel RT</th>
<th>100G8 7R5</th>
<th>15K300</th>
<th>Sata 2T ER6</th>
</tr>
</thead>
<tbody>
<tr>
<td>3Tier</td>
<td>-10%</td>
<td>-24%</td>
<td>32.48/ 2.9</td>
<td>67.4/ 60.1</td>
<td>0.1125/ 37</td>
</tr>
</tbody>
</table>

**Disk Utilization:**

- Skew: 73.5% / 20.5%
- Response Time

**Policy Relative Response Time**

![Graph showing response time and policy relative response time](image)
Applications/workloads have different needs

<table>
<thead>
<tr>
<th>Name</th>
<th>Rel Cost</th>
<th>Rel ST</th>
<th>Rel Pwr</th>
<th>EFD</th>
<th>FC</th>
<th>SATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Multi-Tiered</td>
<td>-10%</td>
<td>-24%</td>
<td>-32%</td>
<td>3%</td>
<td>60%</td>
<td>37%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group</th>
<th>Rel. RT</th>
<th>Capacity</th>
<th>Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB High Prty</td>
<td>-26%</td>
<td>8 TB</td>
<td>3 Tiers</td>
</tr>
<tr>
<td>DB Very High</td>
<td>-48%</td>
<td>3 TB</td>
<td>EFD + FC</td>
</tr>
<tr>
<td>BATCH</td>
<td>-6%</td>
<td>17 TB</td>
<td>FC Only</td>
</tr>
<tr>
<td>Other</td>
<td>-10%</td>
<td>15 TB</td>
<td>FC and SATA</td>
</tr>
</tbody>
</table>
Summary

• FAST VP is a policy-based system that promotes and demotes data at the sub-volume, and more importantly, *sub-dataset* which makes it responsive to the workload and efficient in its use of control unit resources

• FAST VP introduces active performance management, a revolutionary step forward in storage management

• FAST VP delivers all these benefits without using any host resources
For more Virtual/Tiered storage information:

- Other SHARE sessions
  - **12174**: Getting Even More (and a lot less) From Storage Virtual Provisioning and Automated Storage Tiering – Tue @ 1:30 in Golden Gate 7
  - **12945**: DB2 for z/OS With EMC Storage Tiering: FAST VP – Wed. @ 8AM in Golden Gate 8
  - **12708**: What’s New With EMC Symmetrix VMAX and Enginuity? – Wed. @ 4:30PM in Golden Gate 7
  - **12317**: Less=More with Thin Provisioning and Linux on System z – Thur. @ 3:00PM in Franciscan D

- EMC.COM Mainframe Page
  - [http://www.emc.com/storage/mainframe.htm](http://www.emc.com/storage/mainframe.htm)

- Follow EMC Symmetrix
  - @EMCSymmetrix
  - [www.facebook.com/Symmetrix](http://www.facebook.com/Symmetrix)
EMC Disk Tiering Technology Review

Tony Negro
EMC Corporation

Wednesday, February 06, 2013 12:15 PM
Session Number 13154

Thank You