EMC Disk Tiering Technology Review

Tony Negro

EMC Corporation

Wednesday, February 06, 2013 12:15 PM

Session Number 13154

SHARE
in San Francisco
2013

Agenda

- Basis for FAST

- Implementation Characteristics

- Operational Considerations

- Planning for Performance
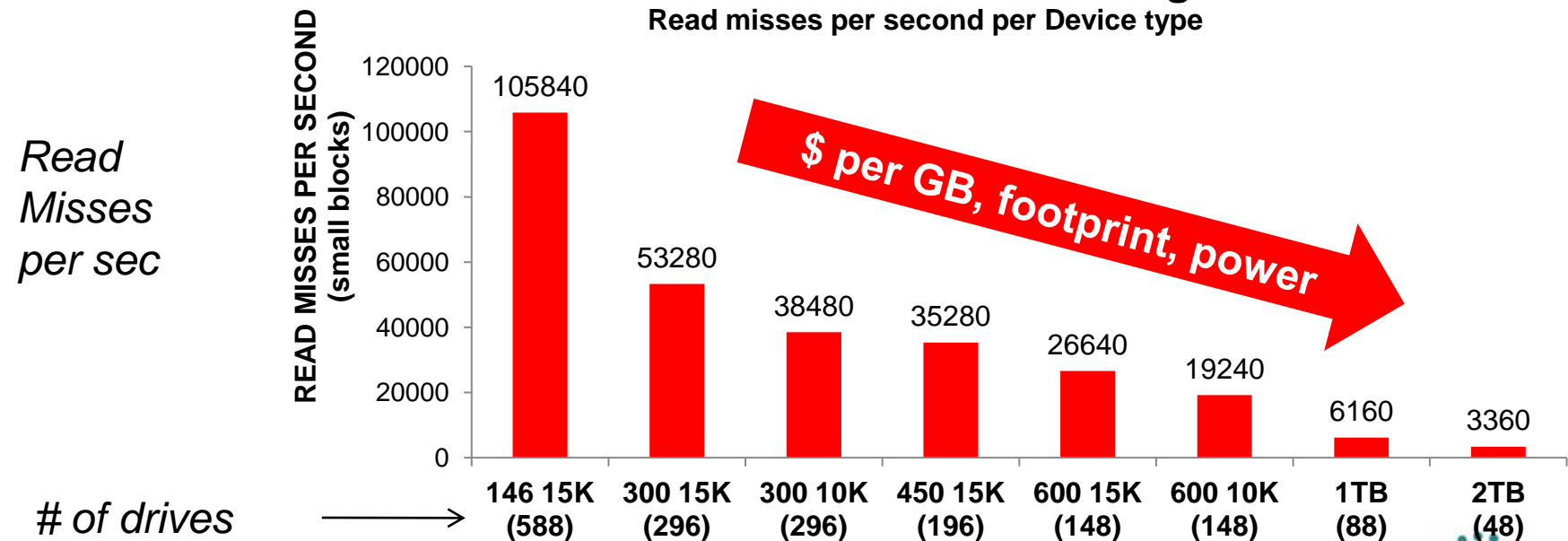
Basis for FAST

- 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 active (or inactive), it remains so for a while – the longer the duration, the better

EFD   **FC**   **SATA**

Workload

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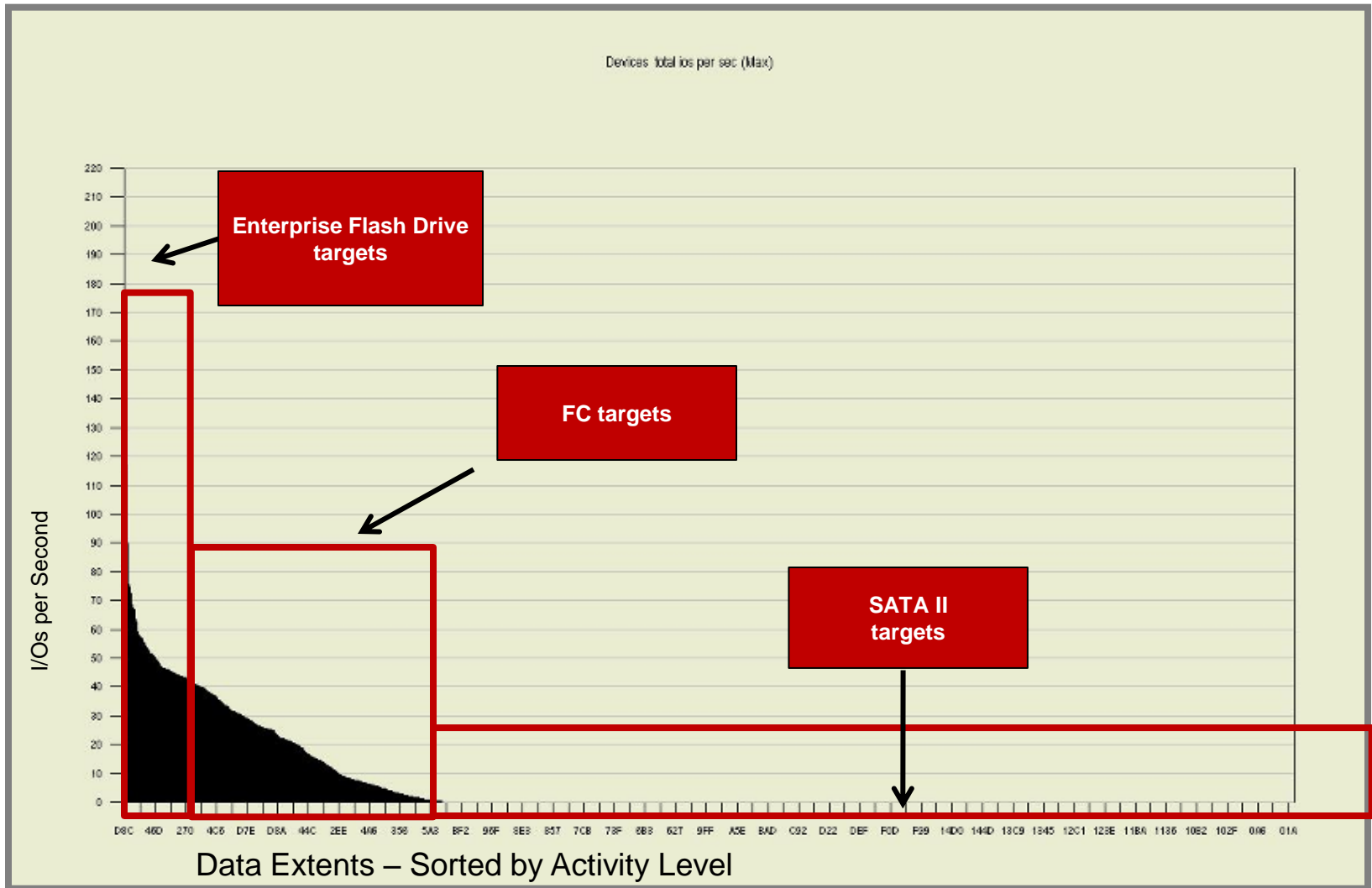
80% of IO's on 20% of capacity
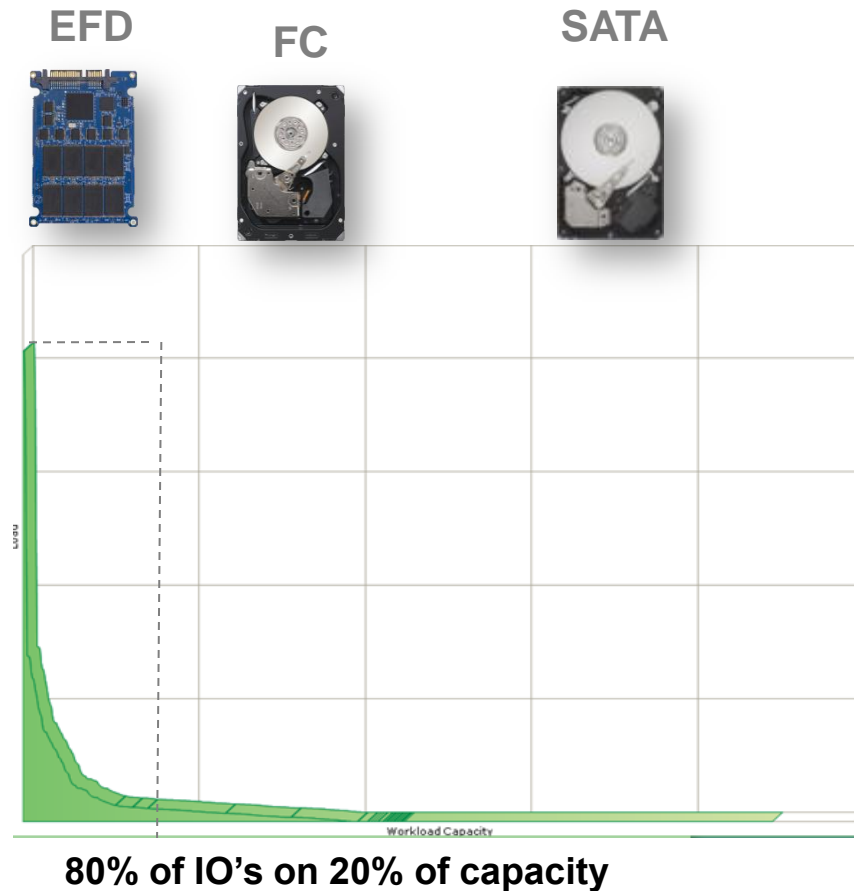
Wide striping and short stroking are common practice

- 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

Read Misses per sec

$ per GB, footprint, power

# of drives →	146 15K (588)	300 15K (296)	300 10K (296)	450 15K (196)	600 15K (148)	600 10K (148)	1TB (88)	2TB (48)
READ MISSES PER SECOND (small blocks)	105840	53280	38480	35280	26640	19240	6160	3360

Measuring Asymmetry in Access Patterns – Skew

Devices total ios per sec (Max)

Enterprise Flash Drive targets

FC targets

SATA II targets

I/Os per Second

Data Extents – Sorted by Activity Level

IMPLEMENTATION CHARACTERISTICS

2008 – Fundamental Storage Media Shift

Disk	Cost/GB	IOPS/GB	Response Time
15 K RPM	1	1	6 ms
Serial ATA (SATA) 7,200 RPM	1/3	1/6	12 ms
Enterprise Flash Drive (EFD)	8	30	< 1 ms

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

% of EFD capacity needed to capture majority of I/Os in the system

SHARE
in San Francisco
2013
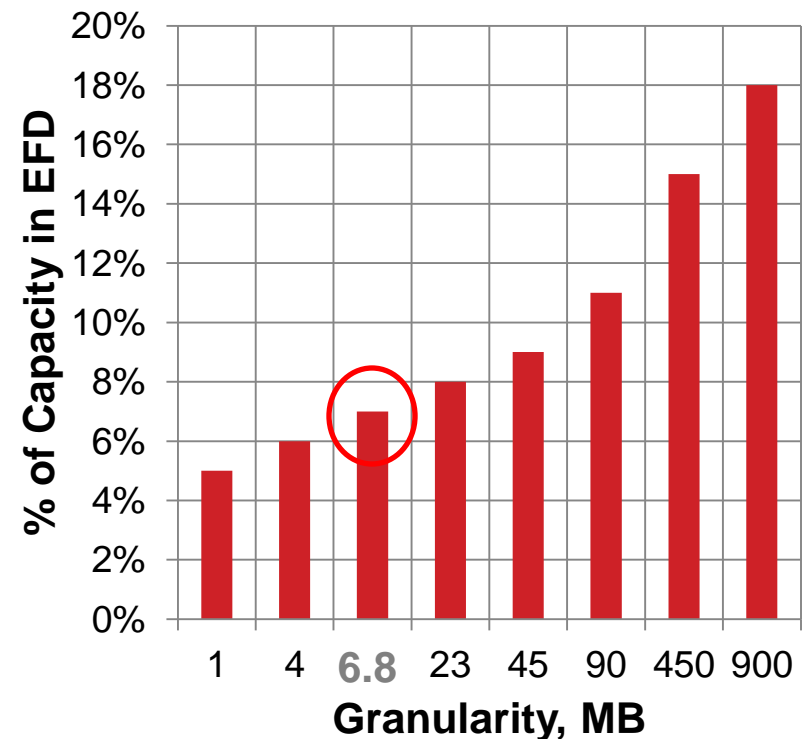
The impact of a larger extent size on configuration cost

			6.8MB EXTENTS		1 GB EXTENTS	
Disk	Cost GB	No. of TB	Capacity Distrib	$Cost	Capacity Distrib.	$Cost
15 K RPM	1	32	53%	$53.0	39%	$23.4
(SATA) 7,200 RPM	1/3	24	40%	$13.3	40%	$13.3
Flash Drive (EFD)	8	4	7%	$32.0	21%	$100.8
		60TB	-	$98.3	-	$137.5

40%

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
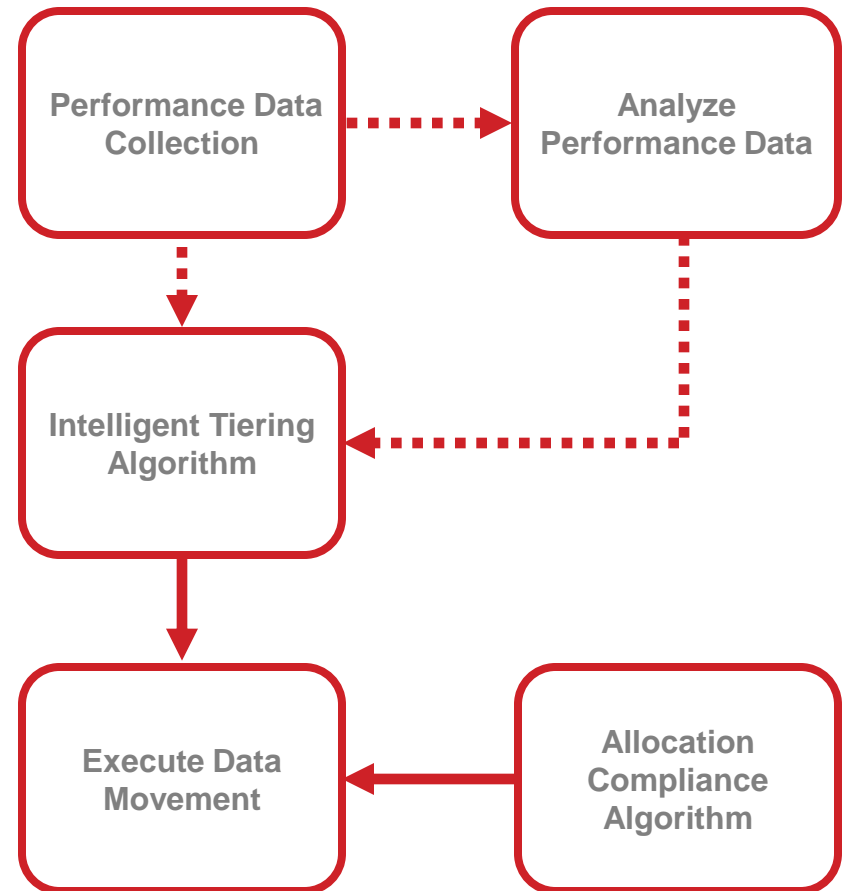- **FAST Storage Group** – logical grouping of thin devices for common management

FAST Storage Groups — VP_Prod_DB2, VP_QA_DB2

FAST Policies — Automatic (<100%, 100%), Custom (x%, y%, z%)

Symmetrix Tiers — R53_EFD_200GB (200 GB EFD RAID 5 (3+1)), R1_FC_450GB (450 GB 15K FC RAID 1), R66_SATA_1TB (1 TB SATA RAID 6 (6+2))
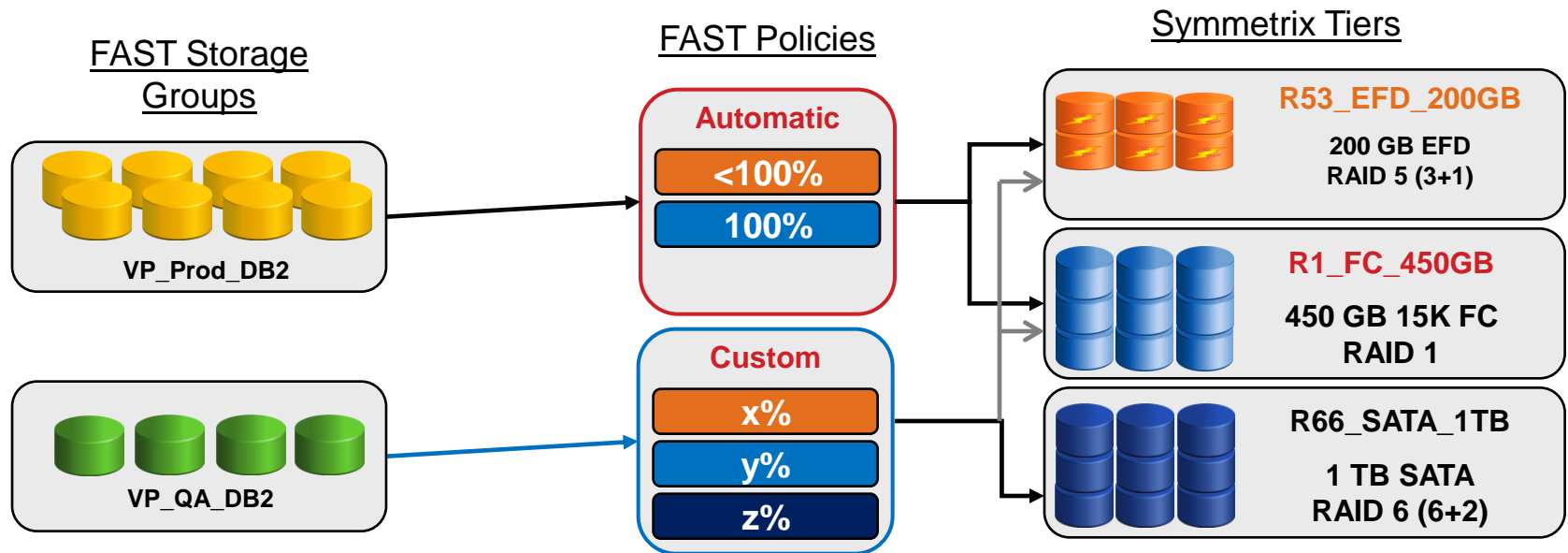
FAST VP Time Windows

- **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

```
Performance Data          Analyze
Collection        ----->  Performance Data
     |                          |
     |                          |
     v                          v
Intelligent Tiering  <---------
Algorithm
     |
     v
Execute Data      <------  Allocation
Movement                   Compliance
                           Algorithm
```

SHARE
in San Francisco
2013

FAST VP Hierarchy

TDEV

- **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)

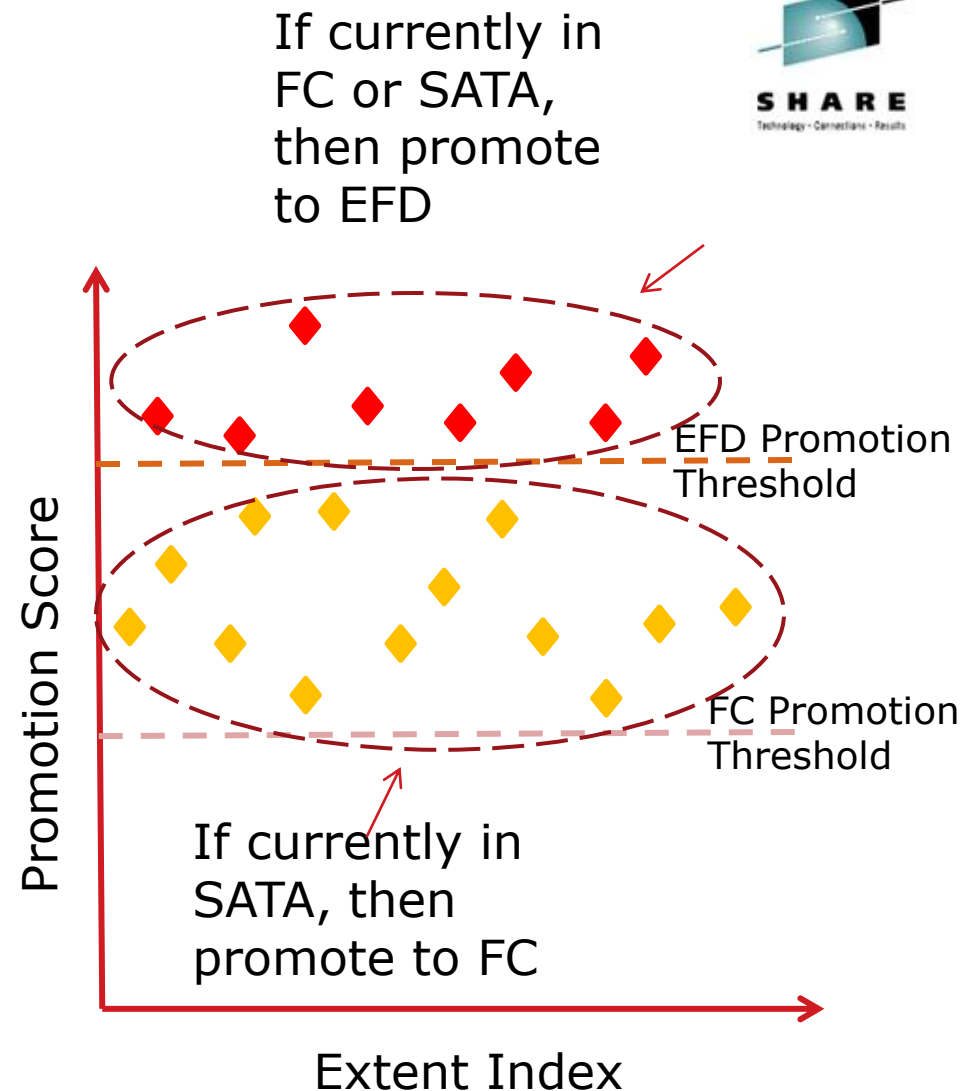- Rates are updated every 10 minutes changing the 'score' of the Extent Group Set

SHARE in San Francisco
2013

FAST VP – Score Analysis

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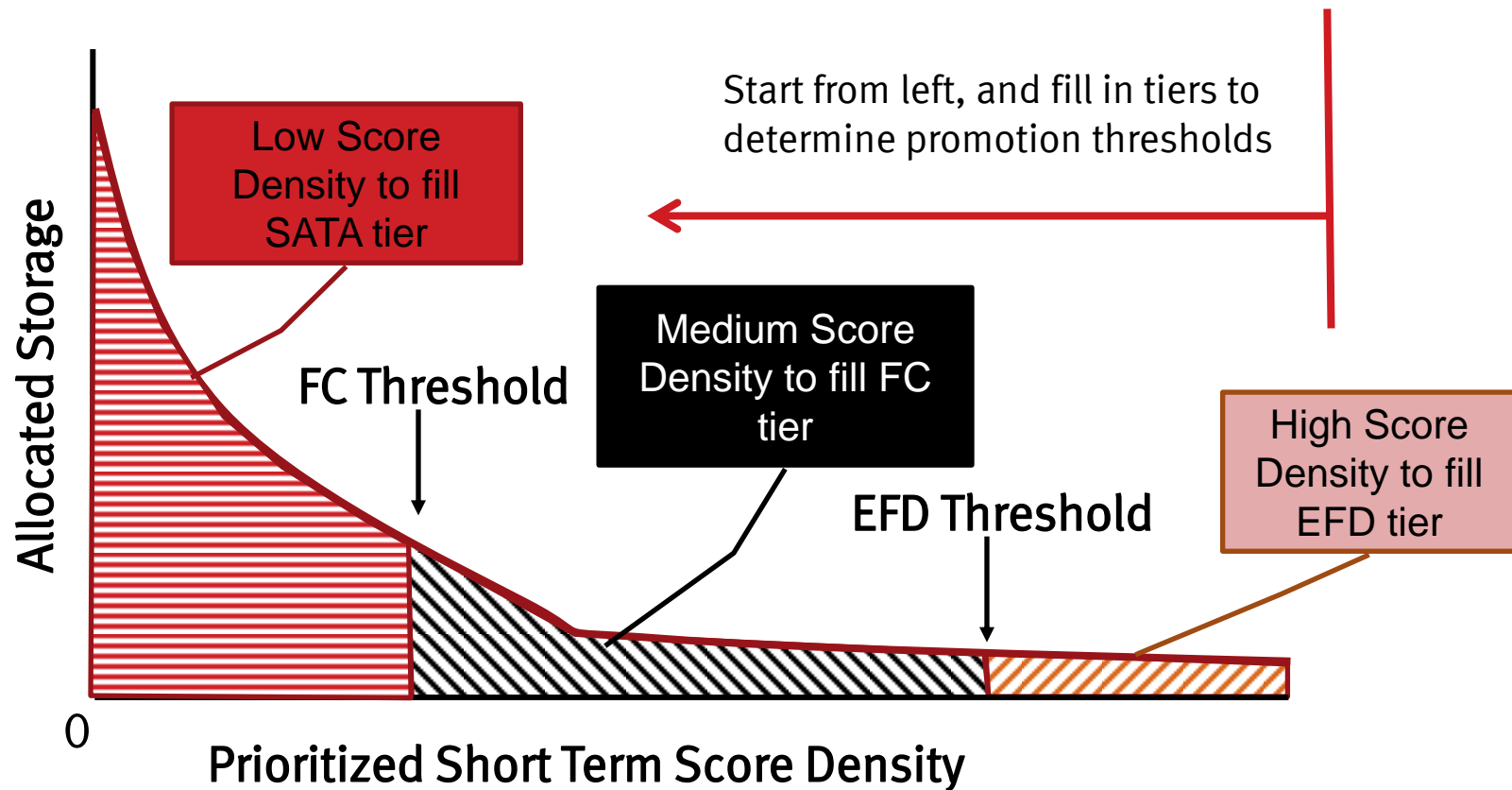
Movement Decisions Promotion

- 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

If currently in FC or SATA, then promote to EFD

EFD Promotion Threshold

Promotion Score

FC Promotion Threshold

If currently in SATA, then promote to FC

Extent Index

SHARE in San Francisco

2013

Control Mechanisms

- Storage Group Prioritization – useful when multiple SG's are associated to the same policy and competing for the same resources
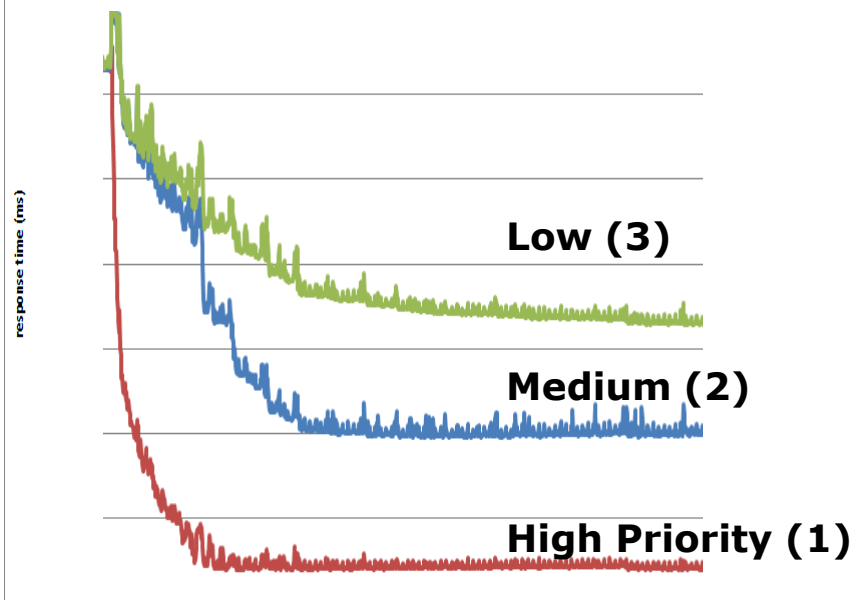
High priority is moved to FLASH first

Data Movement Speed
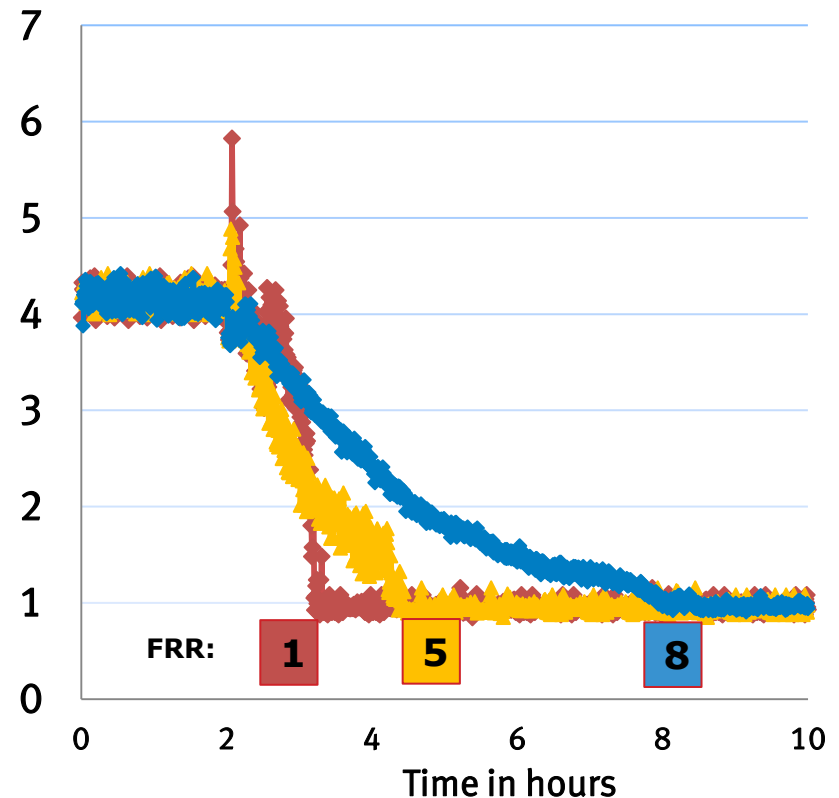
FRR = Fast Relocation Rate
1 is the fastest

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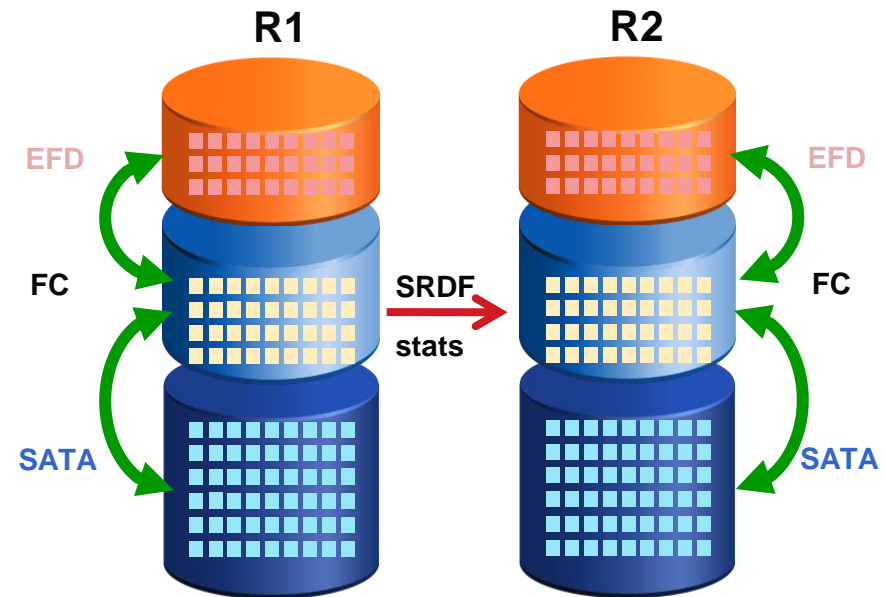
FAST VP R2 statistics are merged with the R1 to reflect the R1 Read Miss ratio

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

FAST VP "SMALL"

$/IOP OPTIMIZED

MAXIMIZE IO DENSITY

$/TB OPTIMIZED

MAXIMIZE TB EFFICIENCY

UP TO ~3:1

FAST
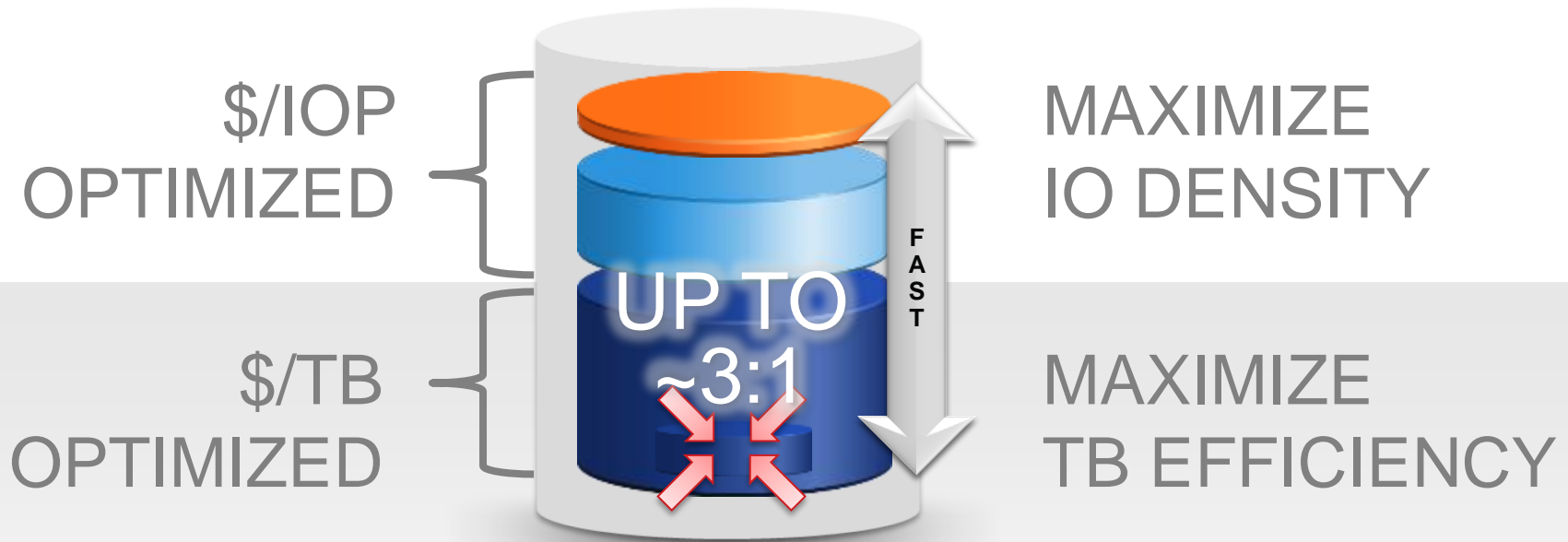
Virtual Pool Compression

SHARE in San Francisco 2013

- 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
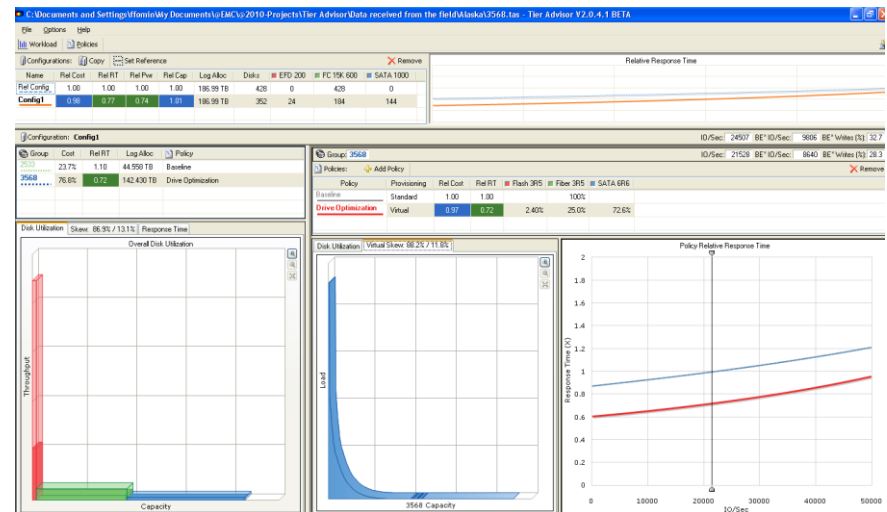
Tier Advisor is recommended for planning.
It is not required for FAST VP operations.

Example of a Multi-tiered MF Configuration

% of total capacity in each Tier

2-3%	57%	40%
EFD	FC/SAS	SATA
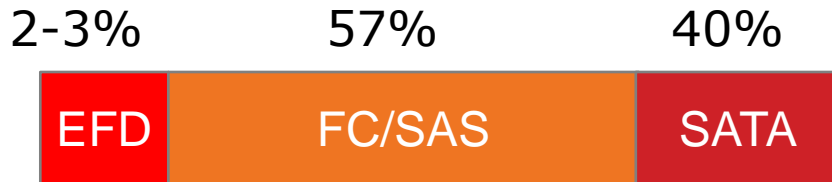
EFD	FC	SATA
20-30%	60-70%	<2%

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

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

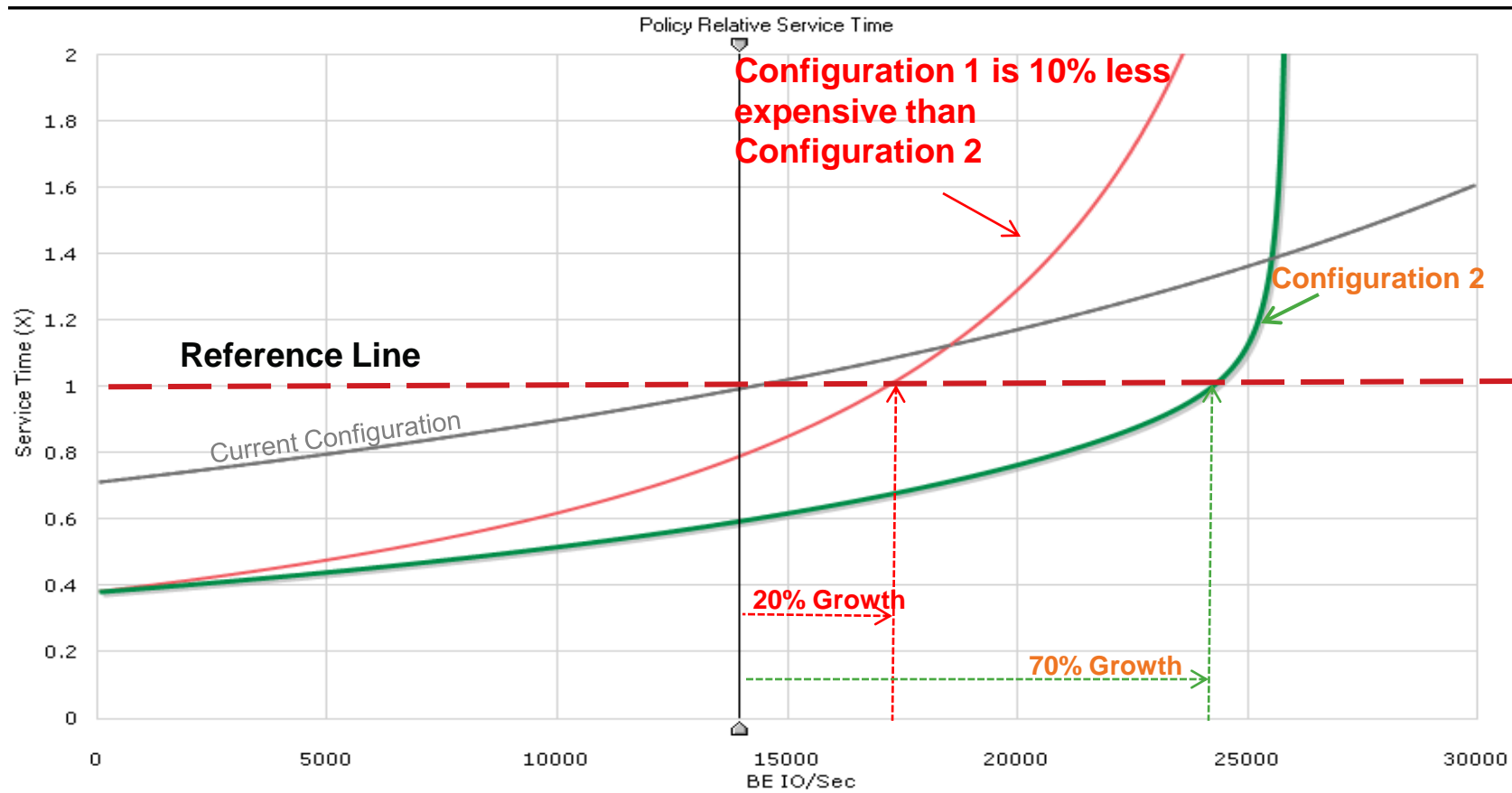- The FC/SAS tier is expected to have a higher I/O density

SHARE in San Francisco
2013

Choose the Right Disk Type to the FC/SAS Tier

Bubble size indicates I/Os per sec per GB

EFD 200 GB

$/GB

600 GB 10K

300 GB 15K

SATA 2 TB

$/IOPS

Log base 10
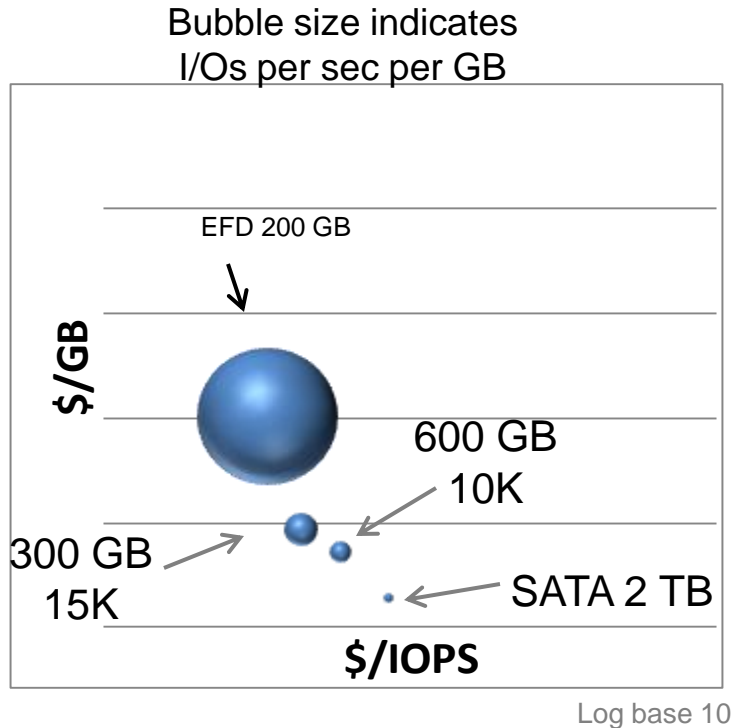
The number of IOPS per GB on a 300 GB 15K drive is ~3X higher than in a 600 GB 10K

- 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

Ensure that Tiers have the Capability to Support Current Load <u>AND Growth</u>

This is only an example

CURRENT CONFIG OF A MF ARRAY

Potential for improvement
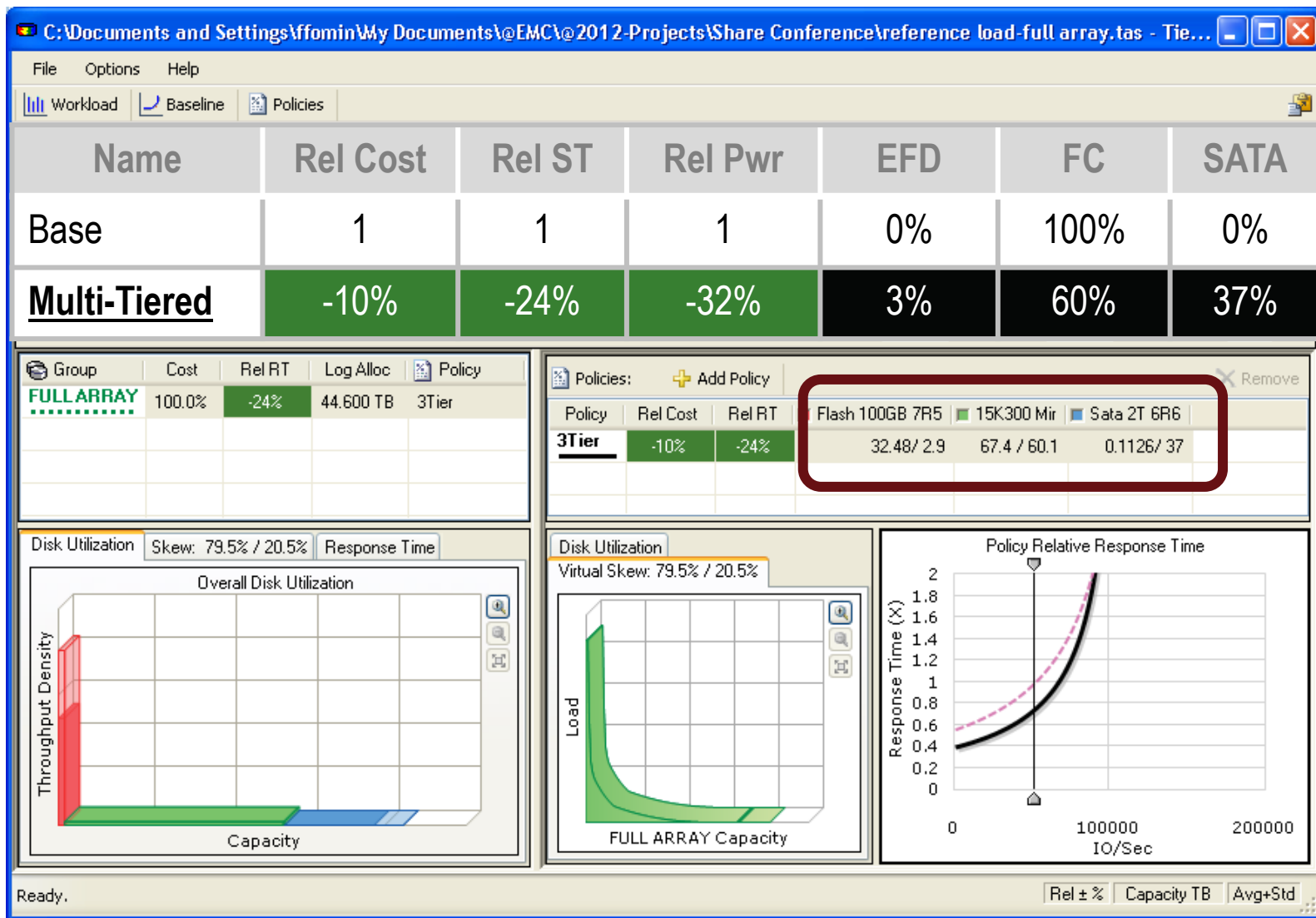
Applications/workloads have different needs

C:\Documents and Settings\ffomin\My Documents\@EMC\@2012-Projects\Share Conference\reference load.tas - Tier Advisor ...

File   Options   Help

Workload | Baseline | Policies

Name	Rel Cost	Rel ST	Rel Pwr	EFD	FC	SATA
Base	1	1	1	0%	100%	0%
Multi-Tiered	-10%	-24%	-32%	3%	60%	37%

Group	Rel. RT	Capacity	Policy
DB High Prty	**-26%**	**8 TB**	**3 Tiers**
DB Very High	**-48%**	**3 TB**	**EFD +FC**
BATCH	**-6%**	**17 TB**	**FC Only**
Other	**-10%**	**15TB**	**FC and SATA**

Ready.   Rel ± %   Capacity TB   Avg+Std

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
- Follow EMC Symmetrix
  - @EMCSymmetrix
  - 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