#SHAREorg

SHARE
Technology · Connections · Results

Less=More with Virtual Provisioning and Linux on System z

Gail Riley
EMC Corporation

February 7, 2013
Thursday @ 3:00pm
Session Number 12317

SHARE in San Francisco 2013

Agenda

- Introduction to Virtual Provisioning
- Virtual Provisioning features
  - FBA
  - CKD
- Virtual Provisioning Benefits
- Fully Automated Storage Tiering for Virtual Pools (FAST VP) Overview

Virtual Provisioning = Thin Provisioning

- From wiki:

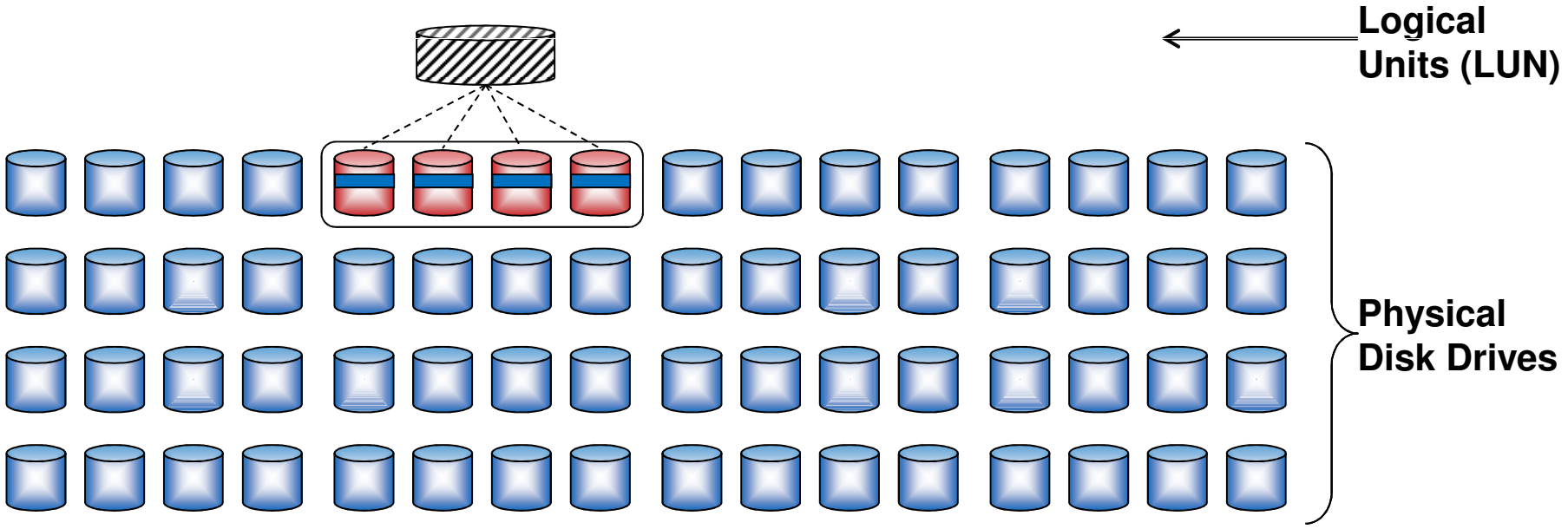
  "**Thin provisioning** is the act of using virtualization technology to give the appearance of having more physical resources than are actually available."

  "**Thin provisioning**[1] is a mechanism that applies to large-scale centralized computer disk storage systems, SANs, and storage virtualization systems. Thin provisioning allows space to be easily allocated to servers, on a just-enough and just-in-time basis."

- Virtual Provisioning is the EMC term for thin provisioning

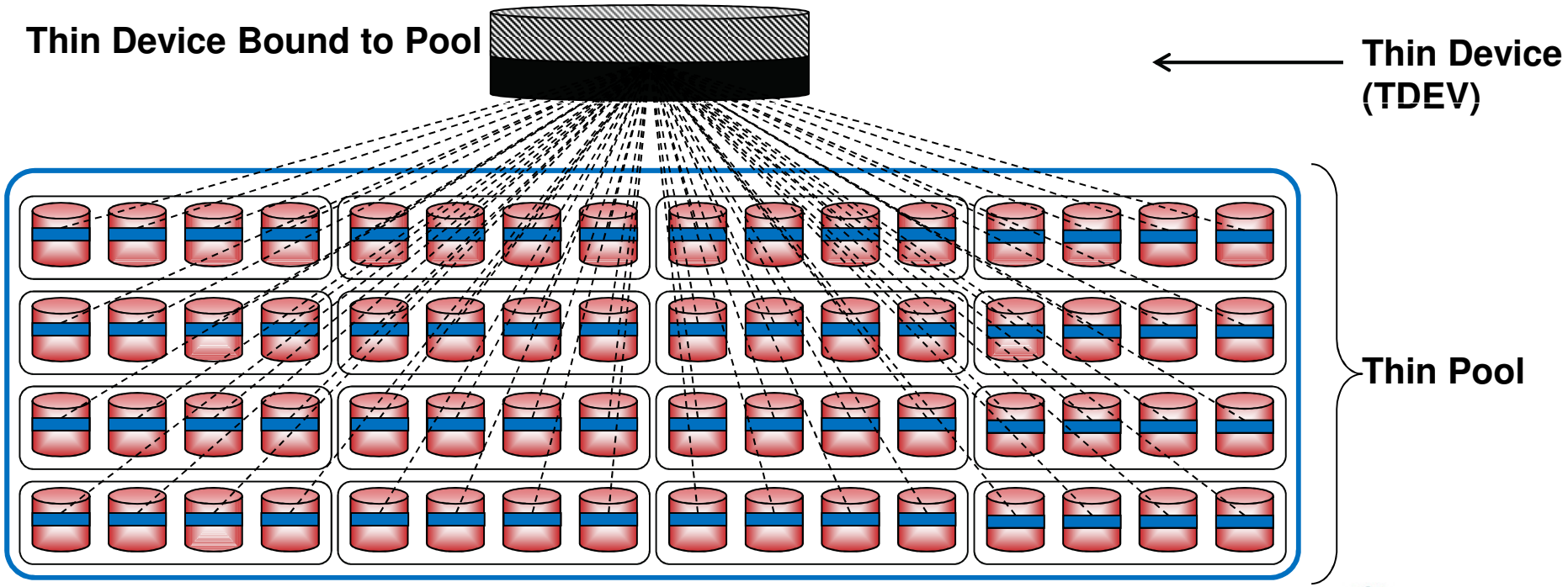
Data Layout – disk device

- Capacity for a disk device is allocated from a group of physical disks
  - Example: RAID 5 with striped data + parity
- Workload is spread across multiple physical disk



Data Layout – Pool-based Allocation Virtual Provisioning

- Storage capacity is structured in pools
- Thin devices are disk devices that are provisioned to hosts



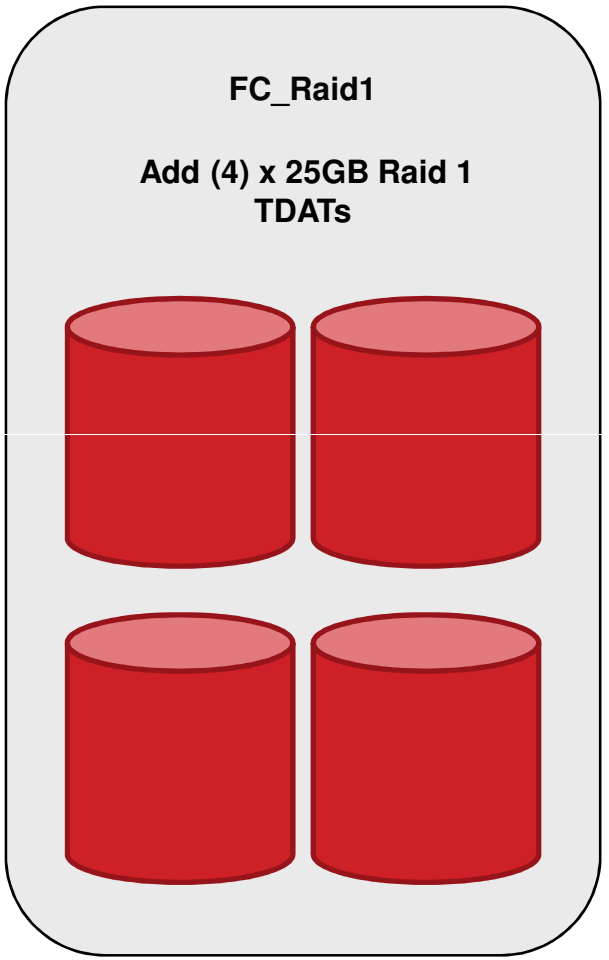
Storage Requirement: Performance

- Storage Layout

- Goal is to spread workload across all available system resources
  - Optimize resource utilization
  - Maximize performance
- Three approaches:
  - RAID data protection
  - Meta Devices (Symmetrix)
  - Virtual Provisioning

VP Components

- Thin Data Device (TDAT)
  - An internal, non-addressable device
  - Provides the physical storage for a thin device
  - Multiple RAID protection types
    - RAID 1, RAID 5, RAID 6
- Thin Pool
  - a shared, physical storage resource of a single RAID protection and drive technology
  - the first TDAT added determines the protection type

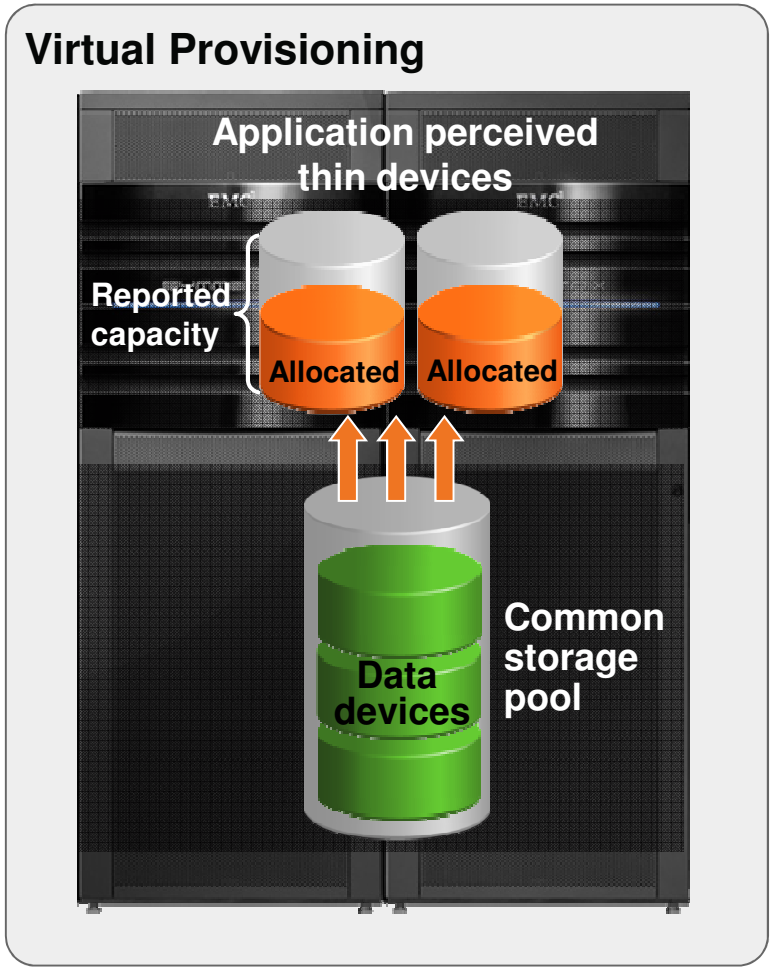


VP Components

- Thin Device (TDEV)
  - Host-addressable, cache only device
  - bound to a thin pool and provisioned to hosts
  - Seen by the operating system as a “normal” device
  - Used in the same way as other host-addressable devices
    - Can be replicated both locally and remotely
  - Physical storage need not be completely allocated at device creation
  - Physical storage is allocated from a thin pool of DATA devices
- Thin Device Extent
  - unit of allocation from a thin pool when a host writes to a new area of a thin device
  - 12 Symmetrix tracks, 768 KB (aka track group)

Virtual Provisioning for FBA (SCSI) devices with Linux on System z

VP Concepts for FBA as a SCSI LUN



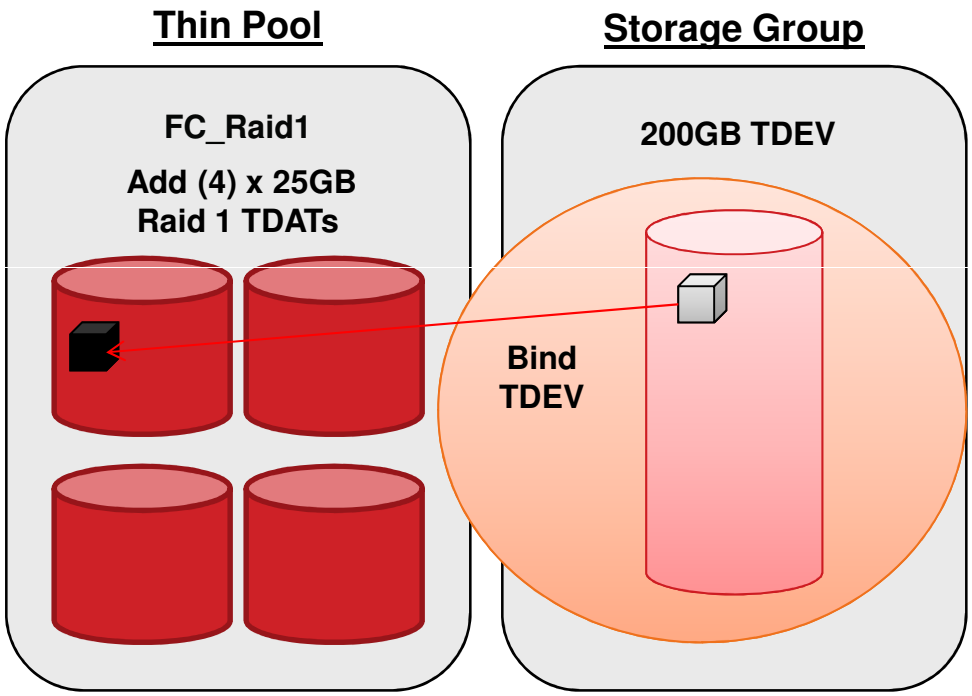
- Thin Provisioning - SCSI
  - Space efficient technology
  - Data storage never 100% full
  - Present **thin device** to Linux
  - Only consumes storage as the host writes to the **thin device**
  - Physical storage allocated from a shared pool
- Over Subscription
  - **Thin device** capacity > pool

Binding a Thin Device

- A thin device must be **bound** to a pool in order to be allocated any storage
- One extent is allocated from the pool when it's bound
- Any write to a new area of a thin device will trigger an extent allocation from the pool the device is bound to
  - New allocations are performed using a round robin algorithm to spread extents across all of the enabled data devices in the thin pool

Virtual Provisioning Bind

bind allocates initial extent in thin pool



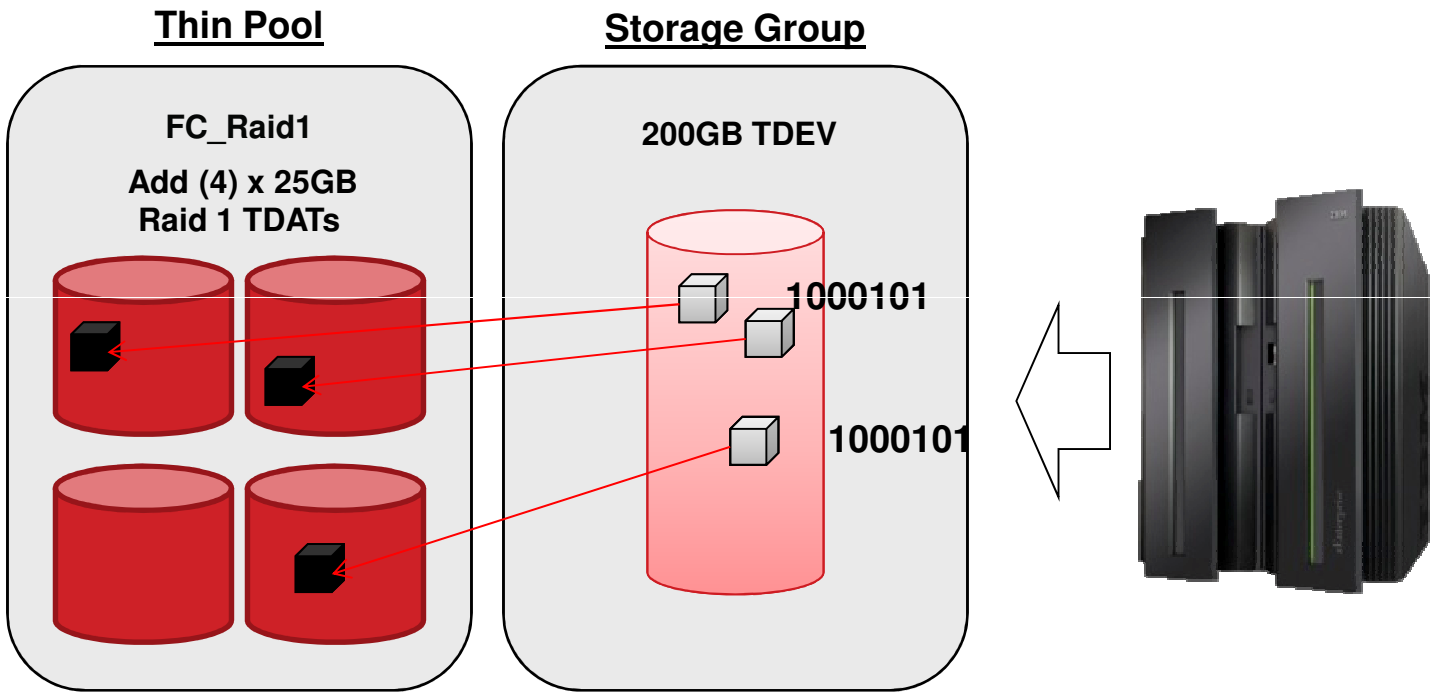
100GB of Storage Capacity in Pool

Thin Pool is Oversubscribed 2:1

Host sees 200GB Device (Ready)

Virtual Provisioning Writes

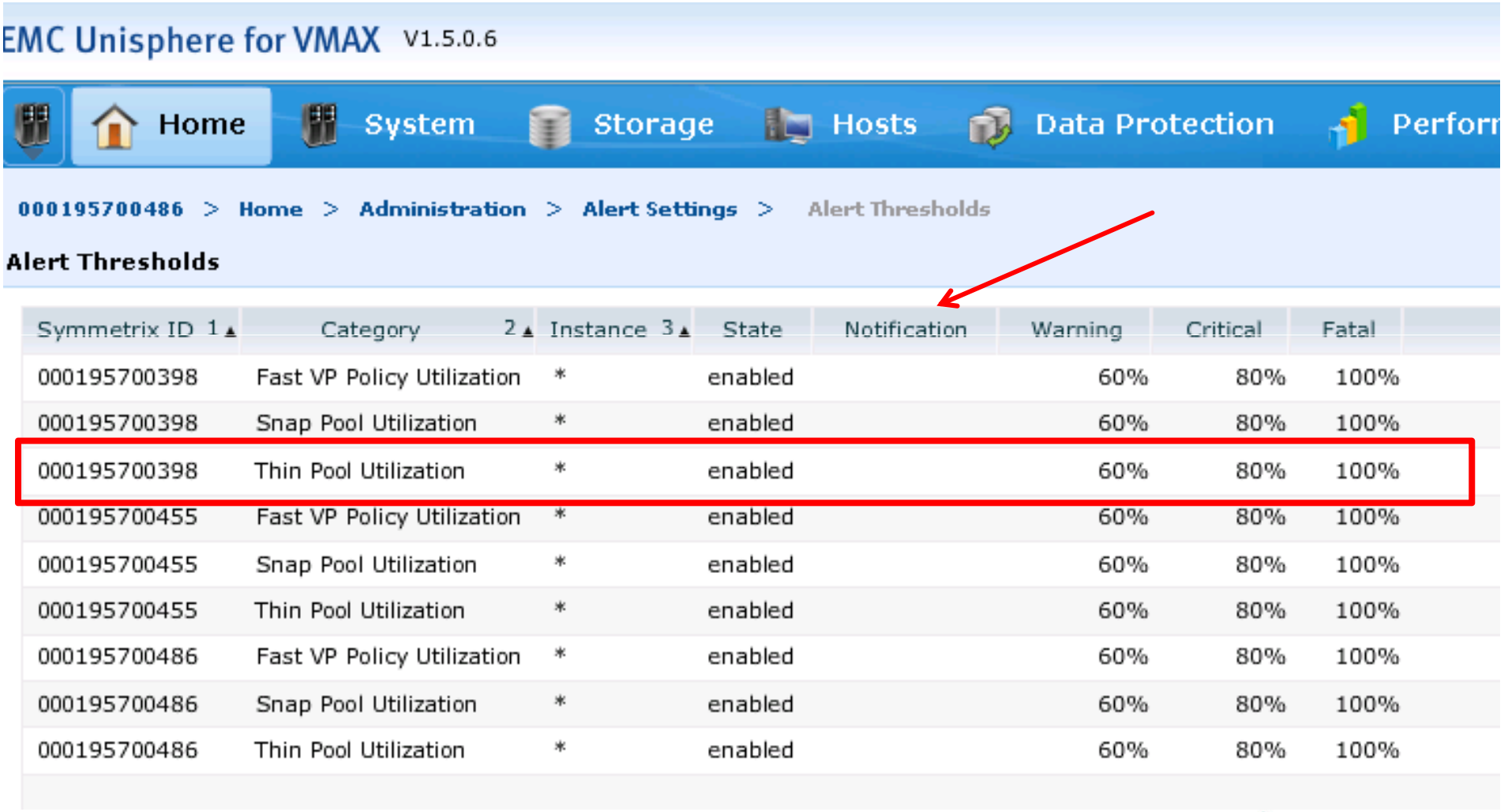
Write to new area of tdev will allocate extents in thin pool



Host Reads from Thin Devices

- Thin devices are cache only devices that contain pointers to the allocated extents on the data devices
- When a read is performed to a thin device, the data is retrieved from the appropriate data device
- Reading from a previously unallocated logical block address will:
  - return a block containing all zeros
  - not trigger an allocation of a new extent

VP Threshold Settings



Symmetrix ID 1	Category 2	Instance 3	State	Notification	Warning	Critical	Fatal
000195700398	Fast VP Policy Utilization	*	enabled		60%	80%	100%
000195700398	Snap Pool Utilization	*	enabled		60%	80%	100%
000195700398	Thin Pool Utilization	*	enabled		60%	80%	100%
000195700455	Fast VP Policy Utilization	*	enabled		60%	80%	100%
000195700455	Snap Pool Utilization	*	enabled		60%	80%	100%
000195700455	Thin Pool Utilization	*	enabled		60%	80%	100%
000195700486	Fast VP Policy Utilization	*	enabled		60%	80%	100%
000195700486	Snap Pool Utilization	*	enabled		60%	80%	100%
000195700486	Thin Pool Utilization	*	enabled		60%	80%	100%

Over Subscription with SCSI devices

- A thin pool can be over subscribed
  - Provision more space than exists in the pool
- A thin device's entire configured capacity counts against the bound pool's maximum subscription percentage
  - Even if the device remains thin (or all of its allocated extents are promoted/demoted to other pools by FAST VP)

Extended Pool Functions and Attributes

- Pool Rebalancing
  - Rebalancing Variance % - controls whether a data device (TDAT) will be chosen for a possible rebalance
  - Maximum Rebalance Scan Device Range – the maximum number of data devices (TDATs) to concurrently balance at any one time
- Attributes (for FBA as a SCSI device)
  - Maximum Subscription % - controls whether a pool can be over subscribed (allocated)
  - Pool Reserve Capacity (PRC) – pools enabled capacity to be reserved for allocating new extents for the bound devices in the pool

Space Reclamation Use Case

- Some migration methods between regular and thin devices will leave the target thin devices fully allocated
- Extents that are allocated on the thin devices may be eligible to be returned to the thin pool
  - Some extents may never have been written to by a host
  - Some extents may contain all zero data
- Available capacity in the thin pool can be maximized by returning unneeded extents back to the pool
- Space Reclamation is an extension of the existing Virtual Provisioning space de-allocation mechanism

Space Reclamation Feature

- Reclamation operations are run against individual thin devices
- Enginuity* will examine all of the allocated groups on specified thin device
  - All tracks will be examined to see if they contain all-zero data
- If all tracks in an extent contain all-zero data, the extent will be de-allocated
  - Tracks that are marked Never Written By Host (NWBH) do not need to be examined by Enginuity
- Space Reclamation is a slow running process
  - Enginuity does not reclaim space at the expense of host performance

*Enginuity is the EMC Symmetrix Storage Operating environment

Thin Provisioning “cleanup”

- Terms are used loosely which can be confusing
- SCSI standard (t10.org) - T10 Technical Committee on SCSI Storage Interfaces
- Host Based SCSI commands for thin devices
  - SCSI unmap
  - SCSI write same with unmap
- Support for these SCSI commands are
  - kernel dependent – Linux vendor and release
  - Storage array dependent

- Any new technology should be tested and fully understood before being put into production!

Check the vendor’s documentation and support matrix for requirements and/or restrictions

Thin Provisioning “cleanup” Terminology

- Unmap
  - SCSI command
  - Sent to thin device to unmap (or deallocate) one or more logical blocks
- Write Same (with unmap flag)
  - SCSI command to write at least one block and unmap other logical blocks
- fstrim – executable, batch command used on filesystems
- Discard
  - option on mount and mkfs command for ext4 and xfs filesystems
  - controls if filesystem supports the SCSI unmap command so thin devices can free specific blocks

Filesystem mount discard option

- Linux Releases supporting the discard option on the filesystem mount command
  - SLES 11 SP2*
  - RHEL 6.2 with a hot fix and ext4
  - RHEL 6.3 and ext4

- Storage Array
  - EMC VMAX @ Enginuity level 5876*
  - Other?

*Check the vendor's support matrix for the specific details

Verification of discard support

- Thin device must be mapped and masked to Linux
- Examine file(s) to verify discard support for the device

/sys/bock/<device>/queue / discard_max_bytes

cat discard_max_bytes
25165824

from kernel.org:

"The discard_max_bytes parameter is set by the device driver to the maximum number of bytes that can be discarded in a single operation. Discard requests issued to the device must not exceed this limit. A discard_max_bytes value of 0 means that the device does not support discard functionality."

Create ext4 filesystem with discard

- ext4 filesystem created with discard first discards blocks on thin device, then creates filesystem

```
# mke2fs -F -t ext4 -E discard -vvv /dev/sdb
mke2fs 1.41.12 (17-May-2010)
fs_types for mke2fs.conf resolution: 'ext4', 'default'
Discarding device blocks: done
Discard succeeded and will return 0s  - skipping inode
  table wipe
.....
```

mount ext4 with discard

- **Filesystem mounted with the discard option**
  - **Frees up space on thin device at time of file deletion**

  ***And* when the array receives the actual write request**

  - NOTE: there is overhead associated with active discard so this should be tested in your own environment

mount -o discard -t ext4 /dev/sdb /thin_mount

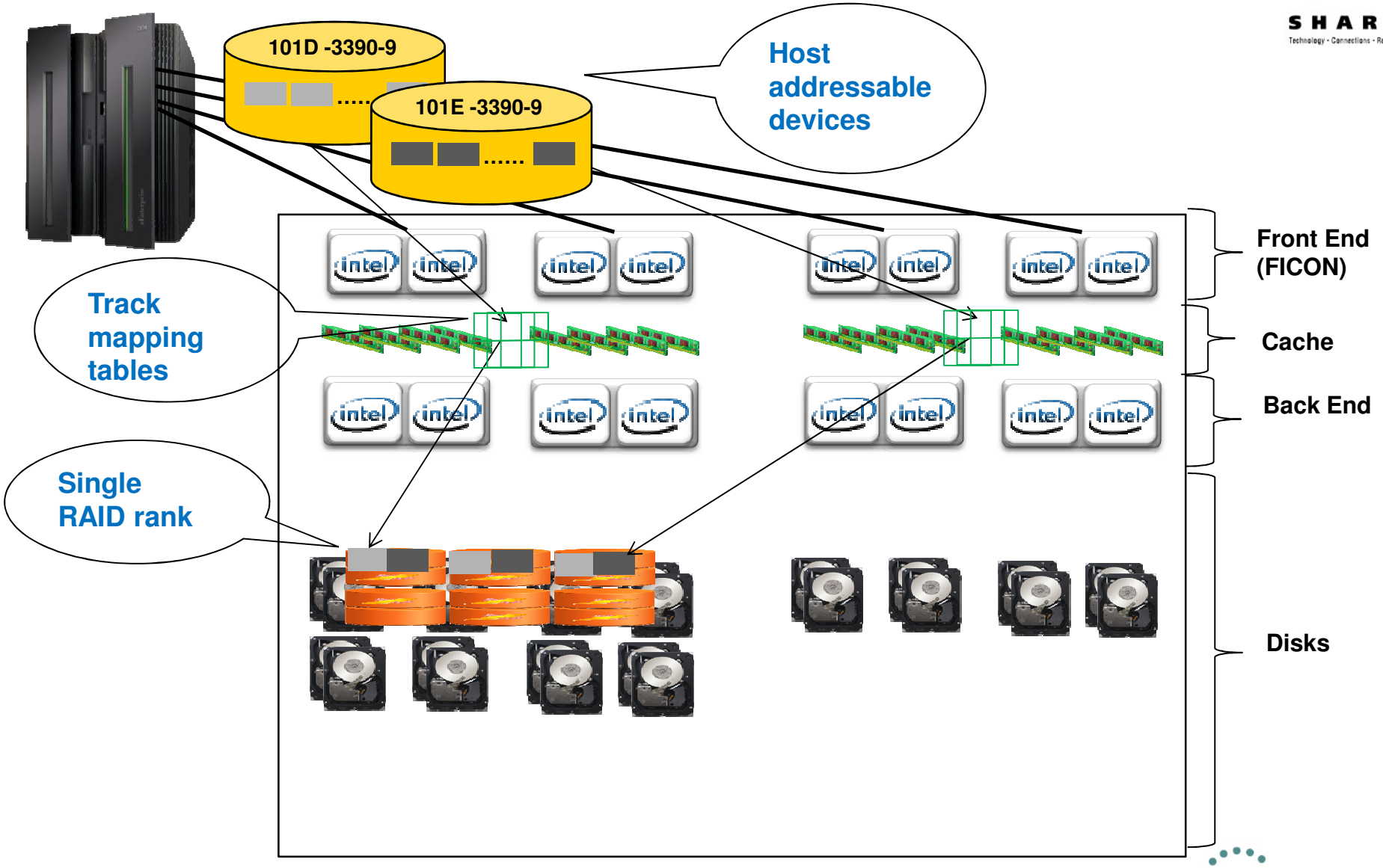
```
# mount
/dev/sdb on /thin_mount type ext4 (rw,discard)
```

fstrim

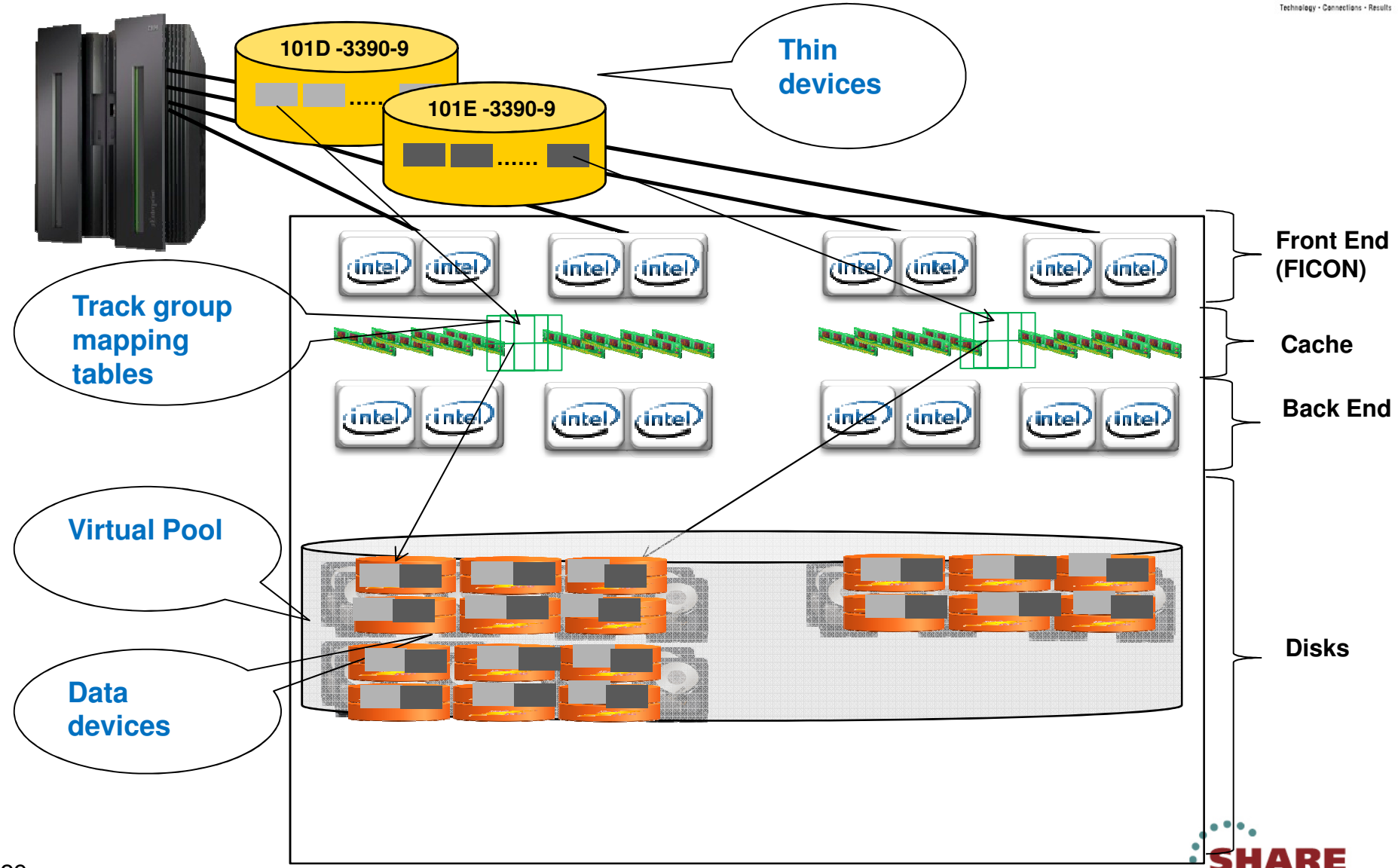
- mount ext4 filesystem without *discard* mount option
- Filesystem mounted without the *discard* option
  - Does not frees up space on thin device at time of file deletion
- You may free up space on a filesystem, where files were previously deleted, on a thin device with fstrim
- fstrim is executed against a filesystem and it's underlying thin device
- Linux support - release and vendor dependent. Check vendor's support matrix for proper support requirements

Virtual Provisioning for CKD devices with Linux on System z

Standard Provisioning Concept (CKD)



Virtual (thin) Provisioning Concept



VP Components for CKD

- CKD VP components are same for CKD as they are for FBA:
  - Thin Pool – a shared, physical storage resource of a single RAID protection and drive technology
  - Data Device (TDAT) – RAID protected devices that provide the actual storage for a thin pool
  - Thin Device (TDEV) – cache only devices that are bound to a thin pool and provisioned to hosts
  - Thin Device Extent – allocation unit from a thin pool when a host writes to a new area of a thin device
    - 12 Symmetrix tracks, 768 KB (aka track group)

VP for CKD with Linux on System z

- Present thin CKD device to z/VM and/or Linux on z
- Thin CKD device must be fully provisioned for z/VM and Linux
- Initial format of thin CKD device fully allocates device
  - cpfmtxa
  - dasdfmt
- Benefits
  - Wide striping
  - EMC FAST – Fully Automated Storage Tiering

Common Functions of VP for CKD and FBA

- Underlying VP technology is the same for FBA and CKD therefore certain management activities are also the same
  - Rebalancing
  - Drain
  - Fully Automated Storage Tiering (FAST)

Rebalancing

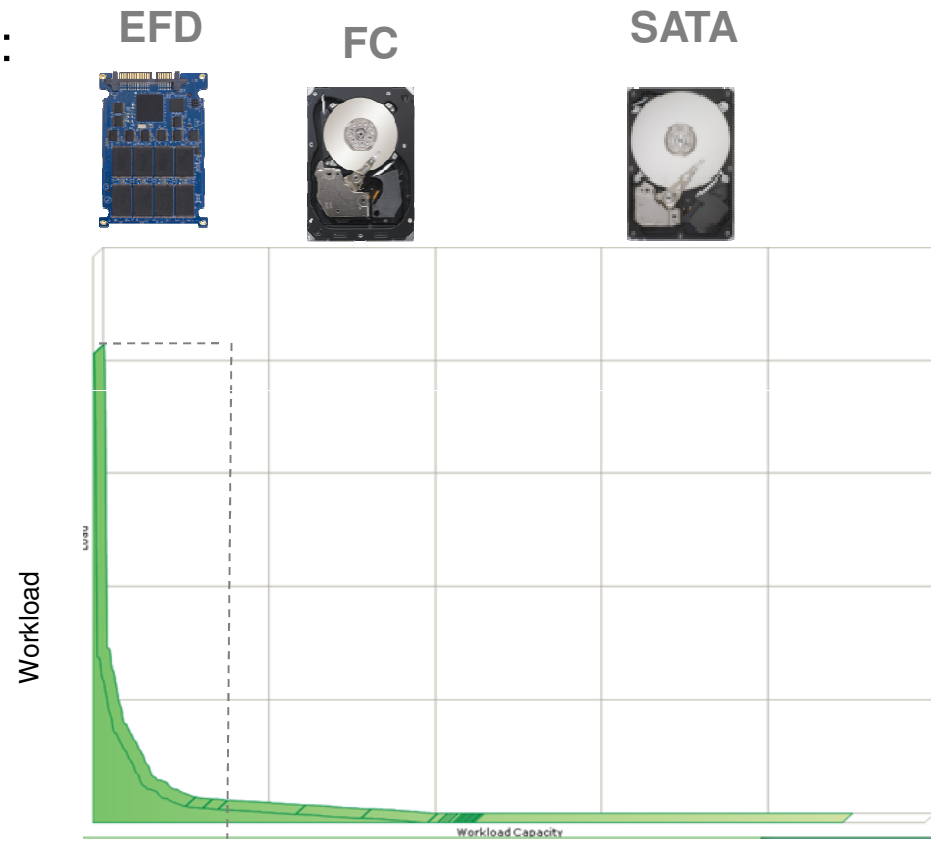
- Should be started after adding new TDATs to an existing pool
- Runs at a very low priority
- Can be influenced by two extended pool attributes:
  - Rebalancing Variance %
    - controls whether a data device (TDAT) will be chosen for a possible rebalance
  - Maximum Rebalance Scan Device Range
    - the maximum number of data devices (TDATs) to concurrently balance at any one time

VP Benefits

- Improved capacity utilization (with VP LUNs and Linux)
  - Reduces the amount of allocated but unused physical storage
  - Avoids over-allocation of physical storage to applications
- Efficient utilization of available resources
  - Wide striping distributes I/O across spindles
  - Reduces disk contention and enhances performance
  - Maximizes return on investment
- Ease and speed of provisioning
  - Simplifies data layout
  - Lowers operational and administrative costs
- Basis for Automated Tiering (FAST VP)
  - Active performance management at a sub-volume, sub dataset level

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

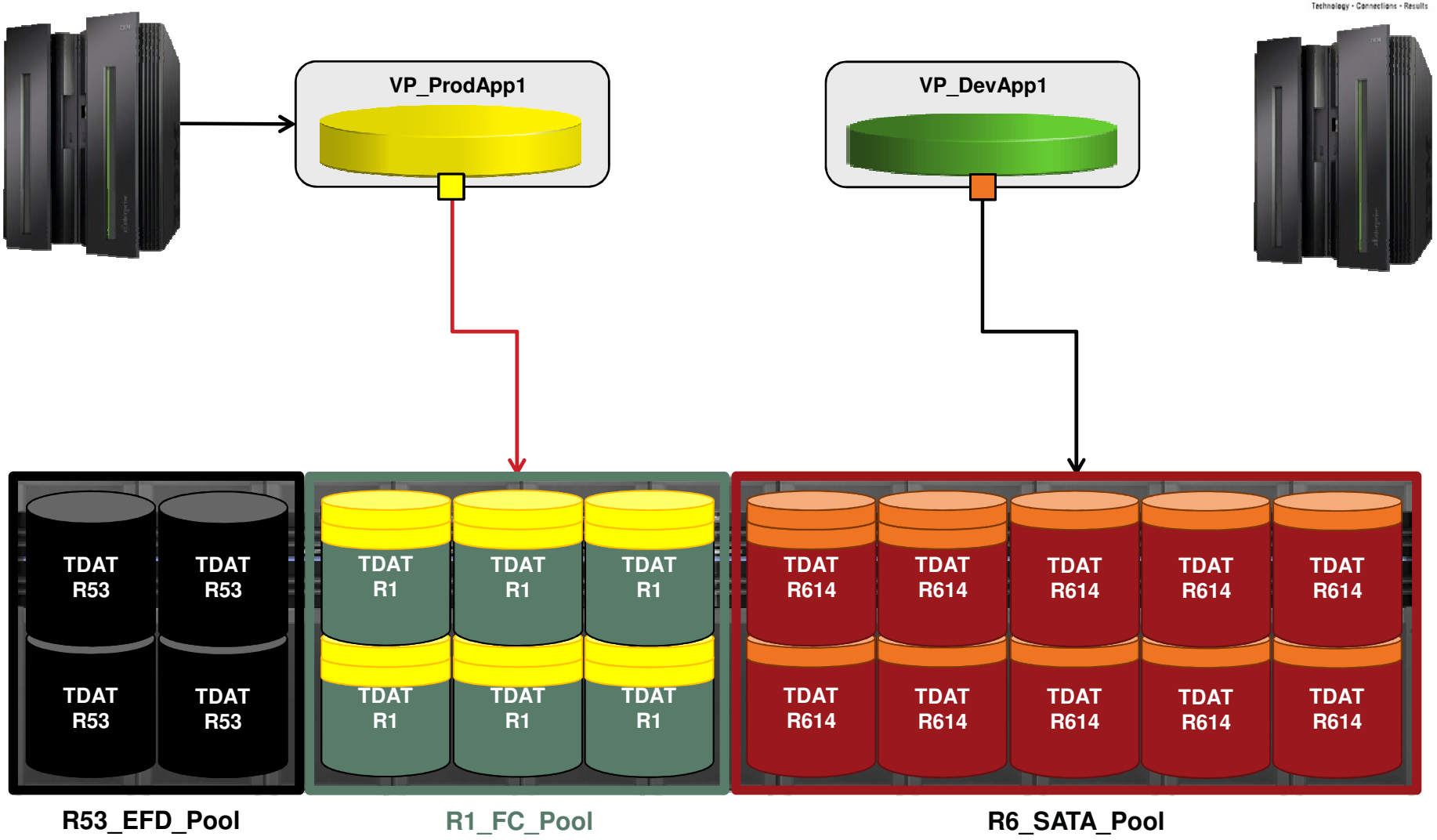


80% of IO's on 20% of capacity

Fully Automated Storage Tiering VP

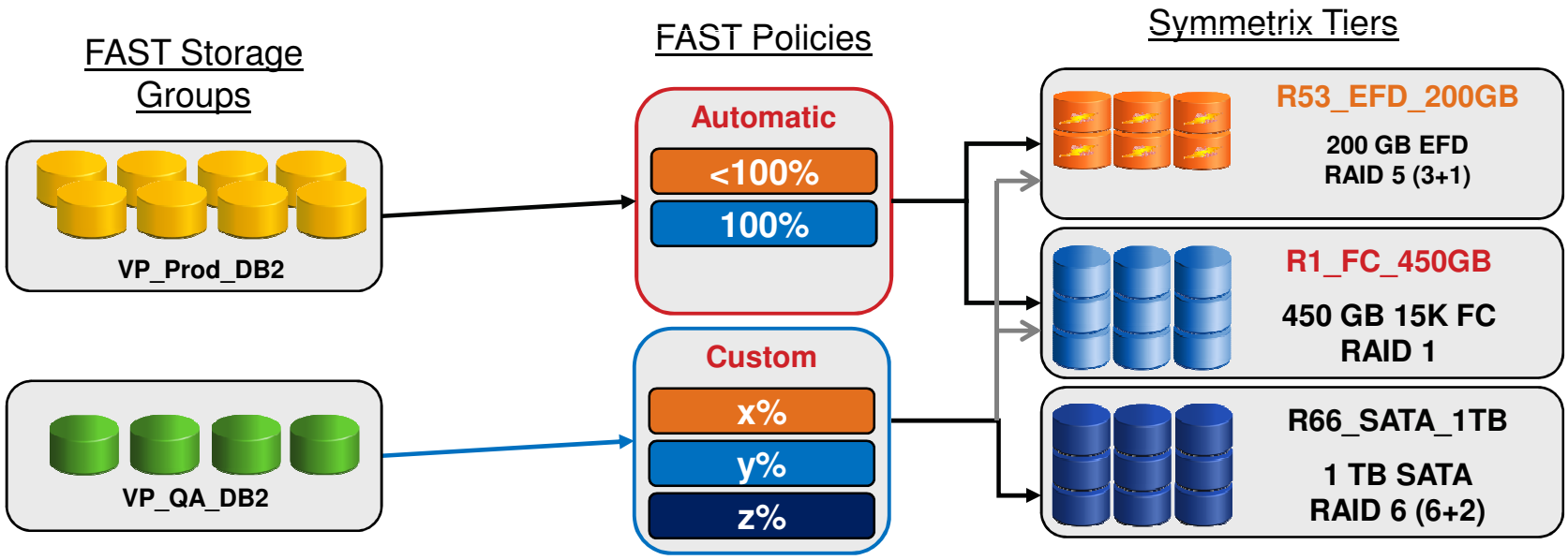
- FAST VP is a policy-based system that promotes and demotes data at the sub-volume, and more importantly, *sub-dataset/sub-lun* which makes it responsive to the workload and efficient in its use of control unit resources
- Performance behavior analysis is ongoing
- Active performance management
- FAST VP delivers all these benefits without using any host resources

Virtual Provisioning with Tiers



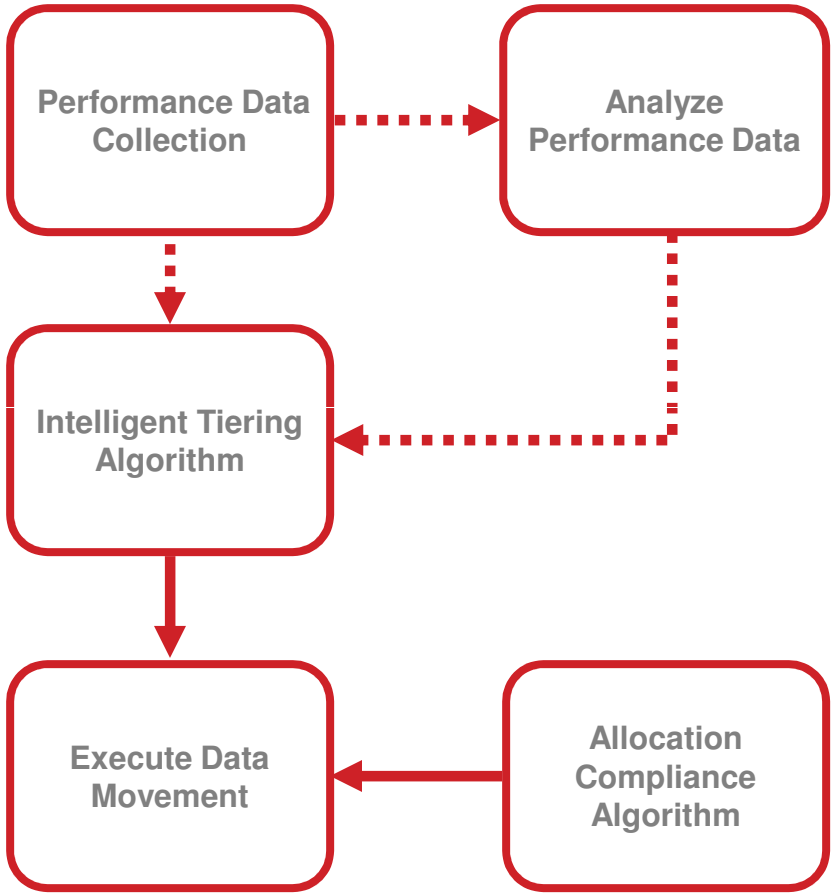
Storage Elements

- 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 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
- Algorithms continuously assess I/O statistics and capacity use, and make decisions for promotion and demotion



Summary

- Virtual Provisioning = Thin Provisioning
- Available for FBA/SCSI and CKD devices
- FBA as SCSI devices
  - Space is allocated as needed
  - Over subscription
  - Cleanup of unused space via space reclamation or T10 SCSI command standards
  - Linux and Storage array dependent
- CKD
  - Fully allocated
- Wide Striping
- FAST VP – Fully Automated Storage Tiering VP
  - Active performance management

THANK YOU

Gail Riley
EMC Corporation
Gail.Riley@emc.com