Thin Provisioning and Storage Reclamation

Anirban Mukherjee
Senior Software Engineer
Symantec
Kirubakaran Kaliannan
Principal Software Engineer
Symantec

Agenda

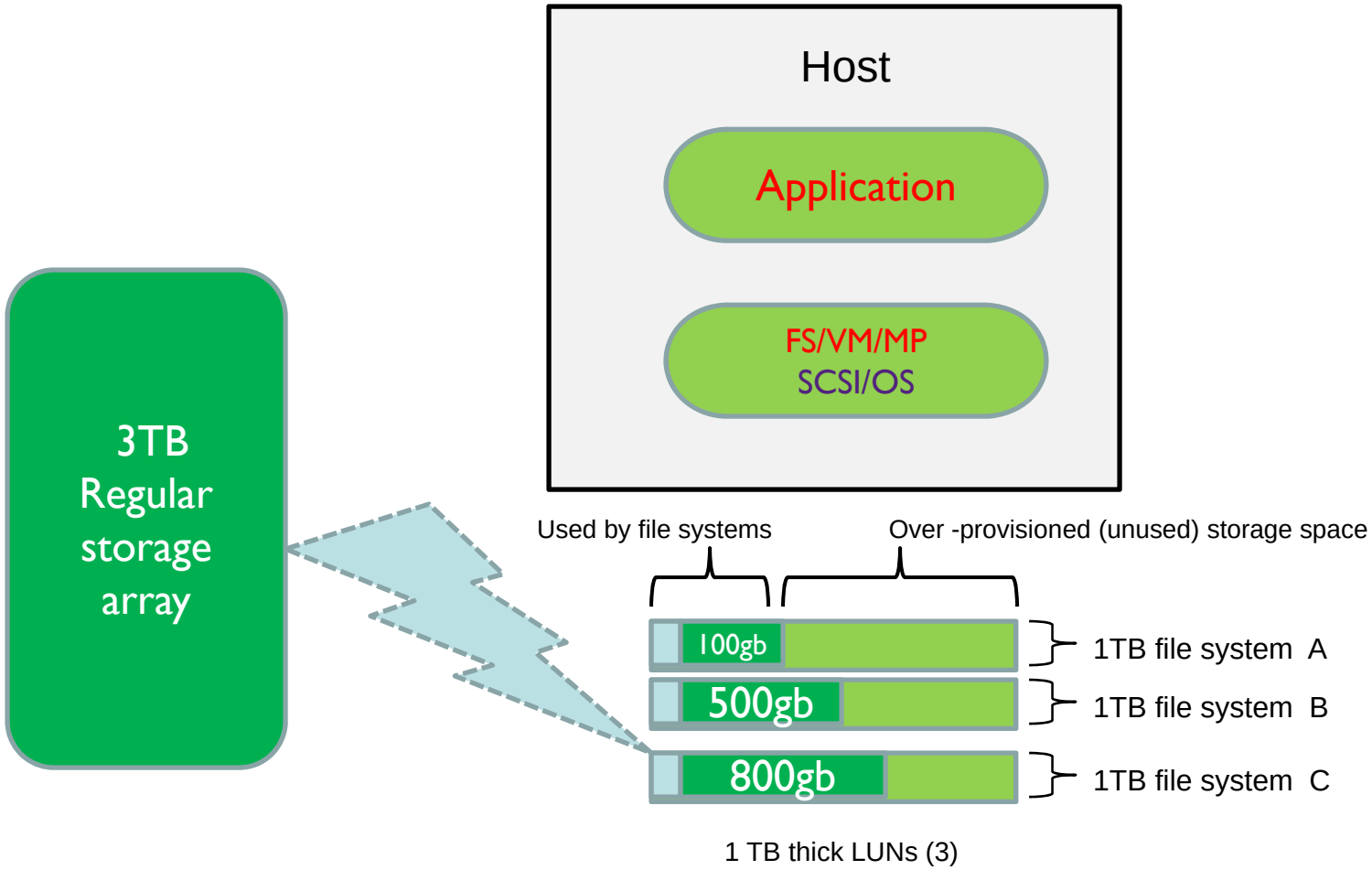
- Definitions
- Thin provisioning
- Storage reclamation
- Solution in Symantec's Storage Foundation
  - Acronyms used in Storage Foundation
  - SmartMove
    - Use cases
  - Thin Provisioning
  - Storage reclamation
    - Reclamation policies
  - Usability
- Challenges faced in Storage Foundation
- Advantage of Thin Provisioning in Storage Foundation
- Reclaiming Challenges

Definitions

- Thin Provisioning
  - Storage virtualization at the hardware layer. A feature of some disk arrays in which an array allocates physical storage to a LUN from a common storage pool only when data is written.
- Thin Array
  - A storage array that supports thin provisioning
- Thin LUN
  - Virtual LUN with no backing physical storage exported to the host
- Storage Reclamation
  - The process of removing unused storage from a LUN and returning it to the common pool
- Thick LUN
  - Fully provisioned LUN exported by a disk array

Thin Provisioning

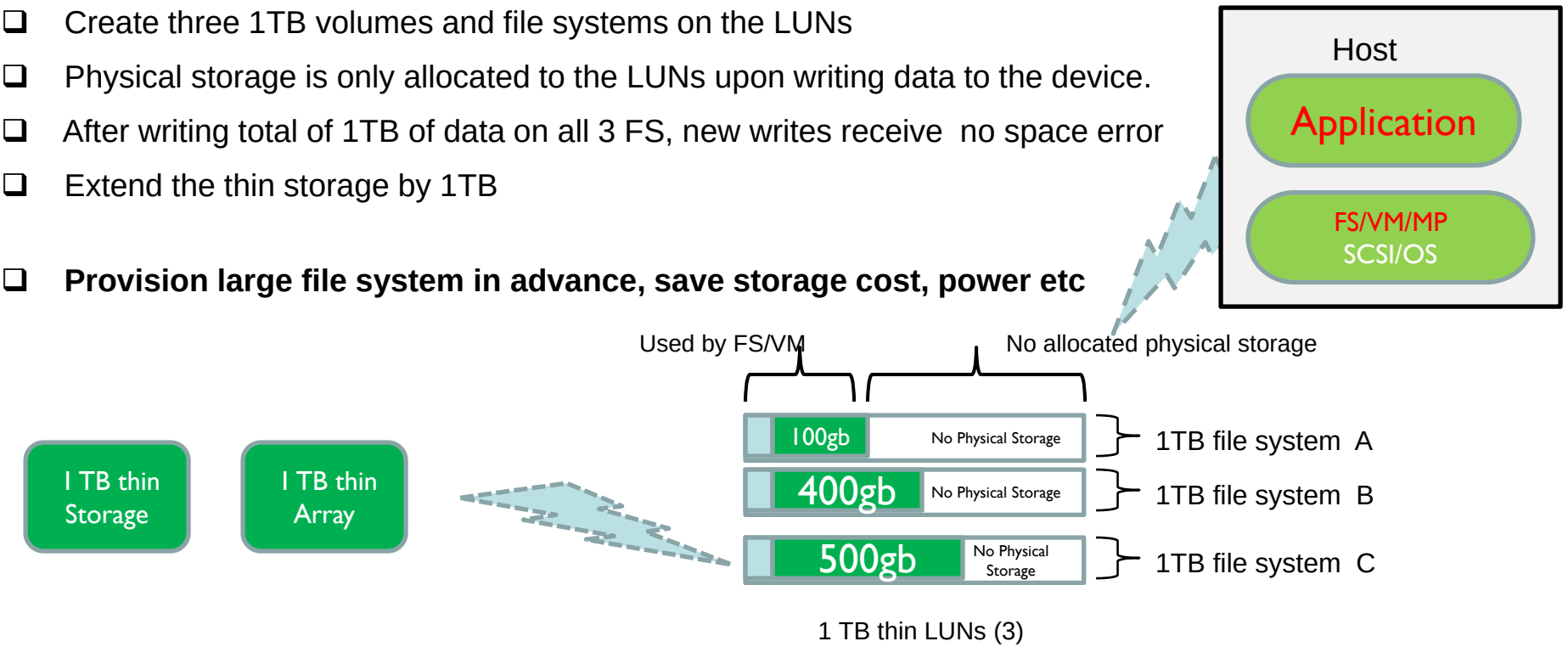
Thin Provisioning (Non-thin array)



Thin Provisioning

- 1TB thin storage array
- Create thin LUNs (size 1TB each), with zero physically allocated storage
- Attach the thin LUNs to the host
- Hosts don't see any functional difference between the thin and thick LUNs.
- Create three 1TB volumes and file systems on the LUNs
- Physical storage is only allocated to the LUNs upon writing data to the device.
- After writing total of 1TB of data on all 3 FS, new writes receive no space error
- Extend the thin storage by 1TB

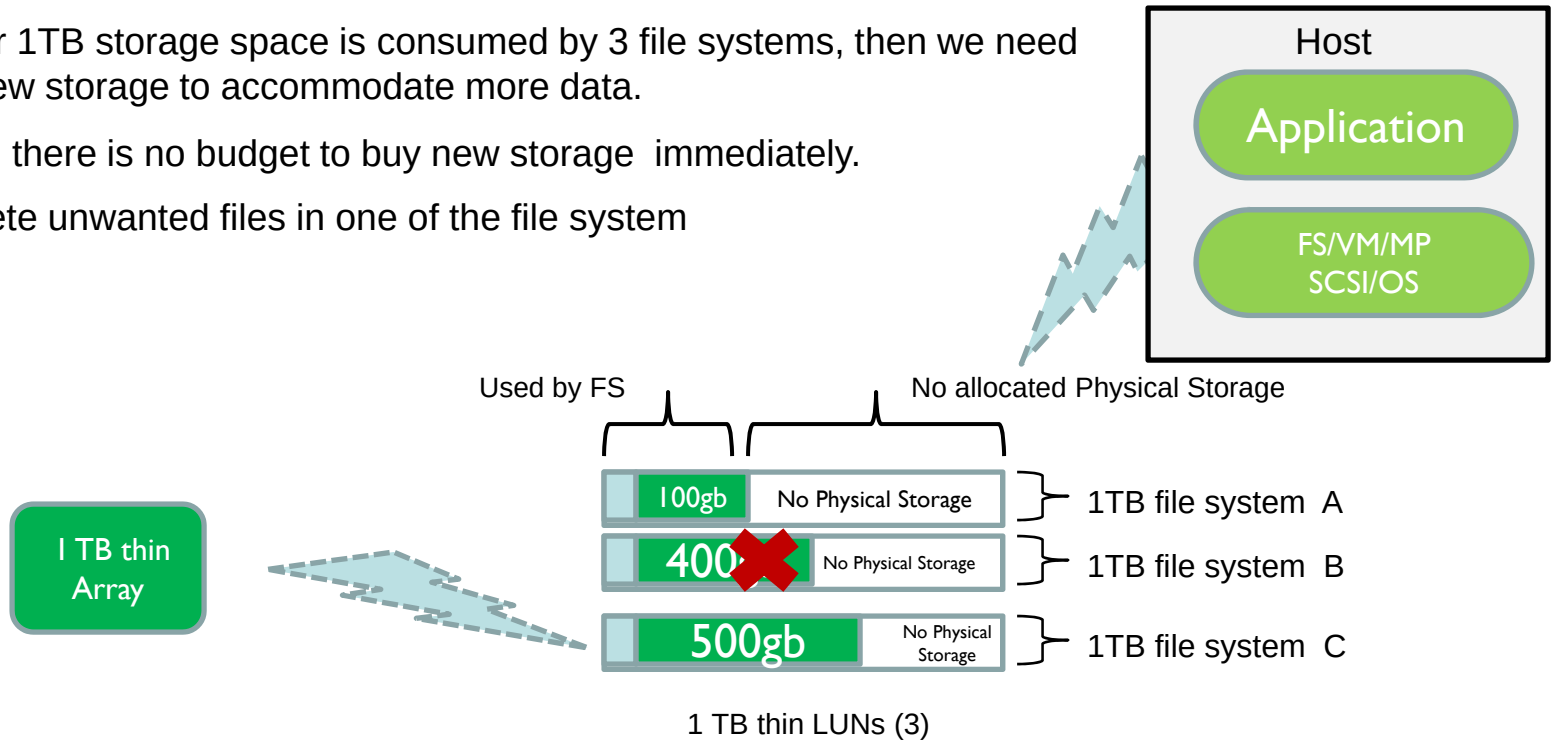
- **Provision large file system in advance, save storage cost, power etc**



Storage Reclamation

Storage Reclamation

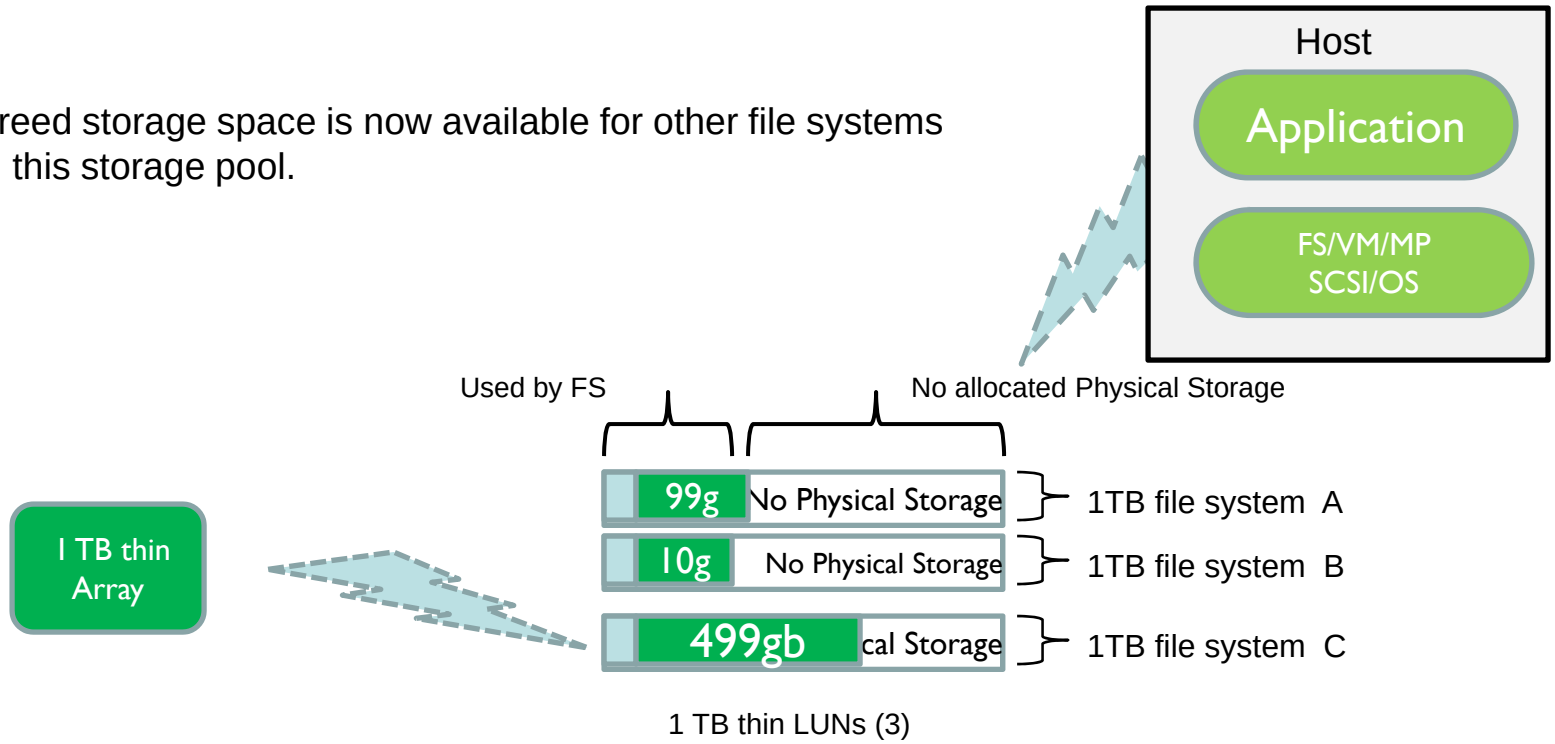
- 1TB thin storage array
- 3 1TB file system is created on 3 thin LUNs. 1TB of physical storage is shared across 3 1TB file system.
- After 1TB storage space is consumed by 3 file systems, then we need to buy new storage to accommodate more data.
- Say, there is no budget to buy new storage immediately.
- Delete unwanted files in one of the file system



Storage Reclamation

- Reclaim the deleted storage space from file system using the array supported reclaim API.
- Reclaimed storage in the file system goes back to free storage pool.

- The freed storage space is now available for other file systems created in this storage pool.



Storage Reclamation in general

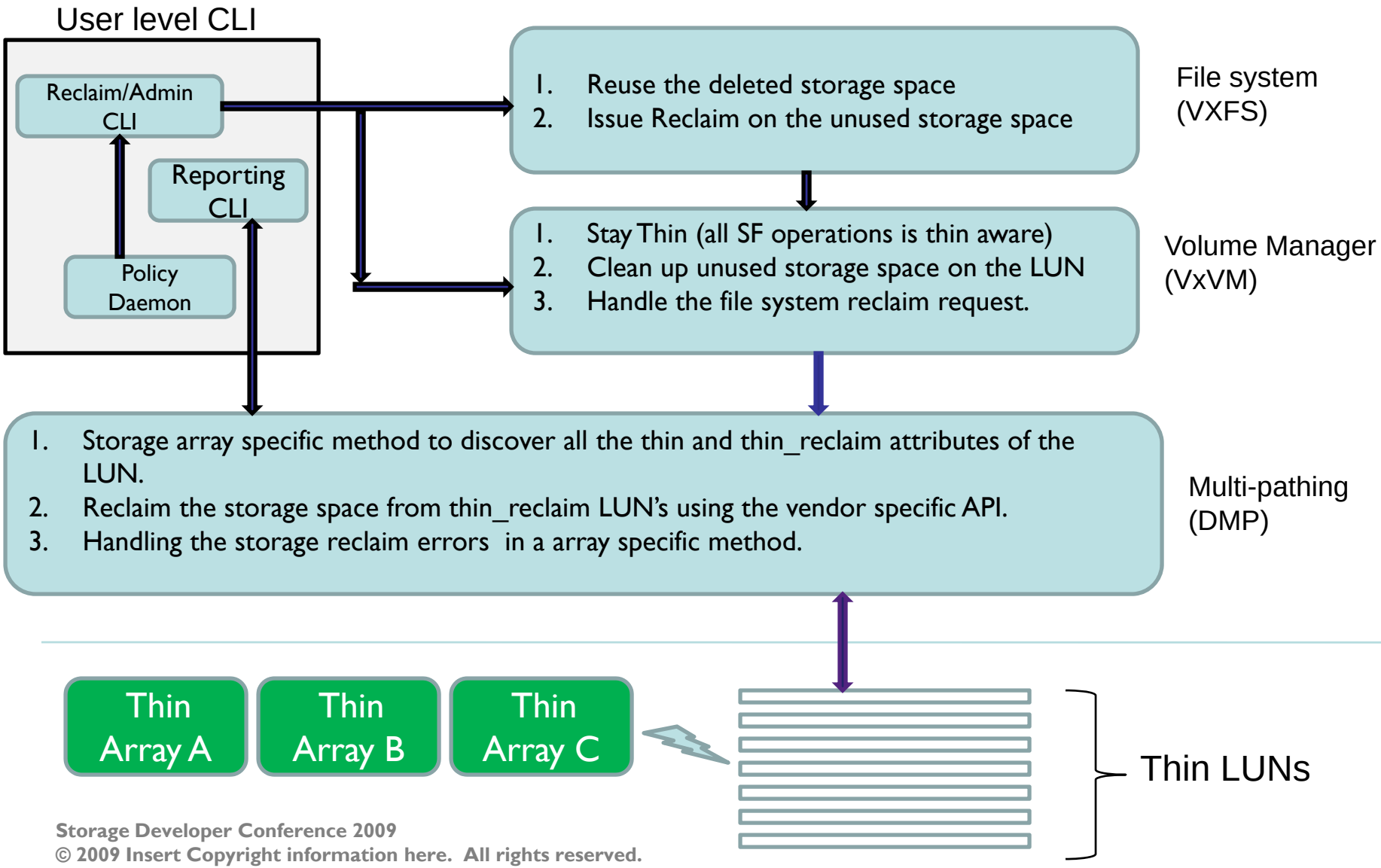
- Hardware supports, 2 different types of reclaim
  - Zero page reclaim (scrubbing zero filled pages)
  - Instant reclaim
- Storage array, provides API to reclaim the storage space
- Reclamation is transparent to application. (no down time)
- Software solution is required to effectively reclaim the unused storage pages.
- Different policies can be built on reclamation
  - Immediate reclaim of storage space after deleting the files
  - Aggressive reclaim
  - Delayed reclaim
  - Manual reclaim

Thin Provisioning and Storage Reclamation in Symantec's Storage Foundation

Acronyms

- SF – Symantec's Storage Foundation, consisting of:
  - VxVM – VERITAS Volume Manager (makes "volumes" from LUNs)
  - VxFS – VERITAS File System
- VxVM components
  - DMP – VERITAS Dynamic Multi-Pathing
  - ASL – Array Support Library (user-mode array specific modules)
  - APM – Array Policy Module (Kernel-mode array specific modules)
- SF feature
  - SmartMove – intelligent data movement tool (to improve administeative I/O performance)

Overview of Thin Provisioning in Storage Foundation



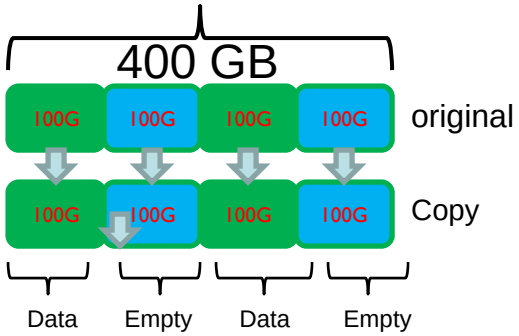
SmartMove

SmartMove in Storage Foundation

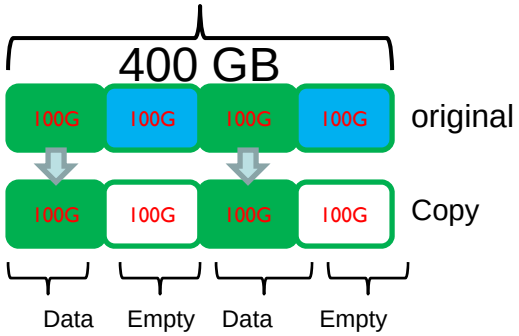
Storage Mirror operation

- Conventional: copies full file system block space
  - VM creates duplicate volume and initiates the copy
  - Block-by-block of all 400G allocated to file system
- SmartMove: copies only “live” data
  - VxFS file system on VxVM volume (volume may span multiple disks)
  - File system metadata identifies used/unused blocks
  - VM queries file system about each block and copies only “live” file system data.
  - VM copies data to correct duplicate volume location
- SmartMove example (right)
  - copies only 200GB of data from a 400GB file system
  - 50% performance improvement

Conventional copy/mirror



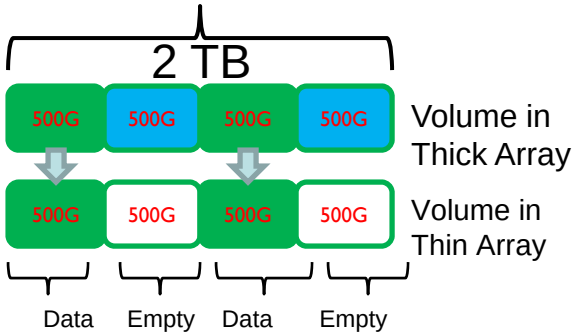
SmartMove copy/mirror



Smart Move Use case 1
Performance

- The following are common storage management operations:
  - Mirroring a volume
  - Volume/disk backup
  - Snapshot
  - Array migration
  - Replacing the disk
- All require copying data from one location to another
- With SmartMove:
  - All operations copy only actual data on the volume
  - Performance of all these SF operations are improved (*Improvement is inversely proportional to the percentage of file system space occupied by data*)

Thick to thin array migration



Smart Move Use case 2

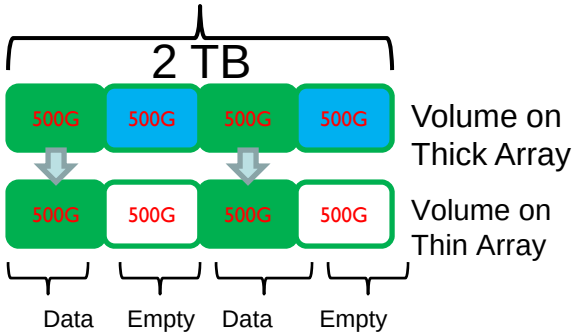
Array Migration

- Thick Array (migration source)
  - Contains thick LUNs (full capacity allocated)
  - In most cases only 30%-50% of storage contains actual data
- Thin Array (migration target)
  - Contains LUNs with virtually allocated storage space
  - To migrate: create large LUNs with less pre-allocated storage space and provision large file systems

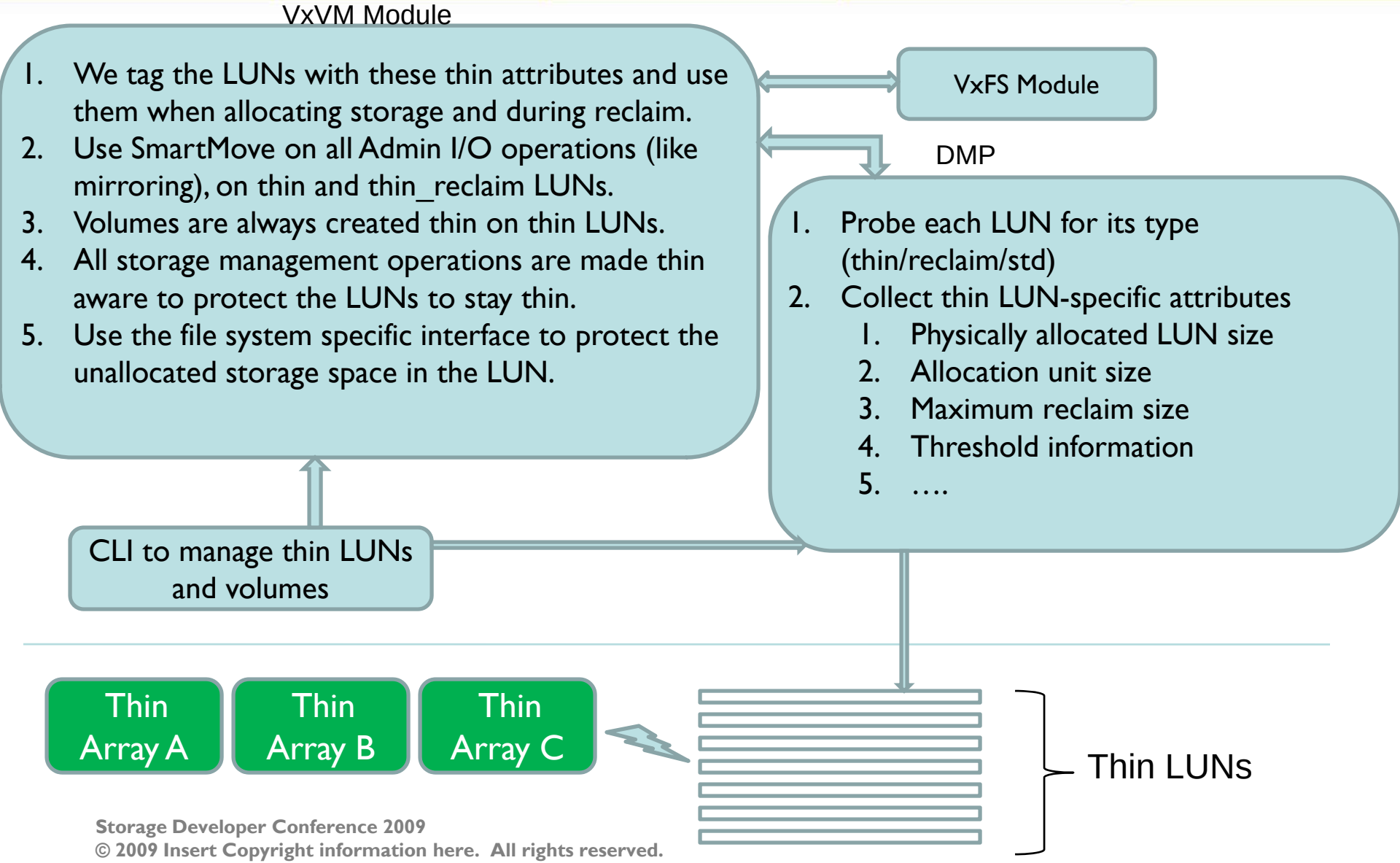
- Thick to thin array migration with SmartMove
  - From each source volume, SmartMove copies regions containing data

- In the example; 2 TB volume migration
  - Only 1 TB of data is copied
  - Only 1 TB of storage is allocated in thin array
  - Savings compared to conventional thick array: 1 TB

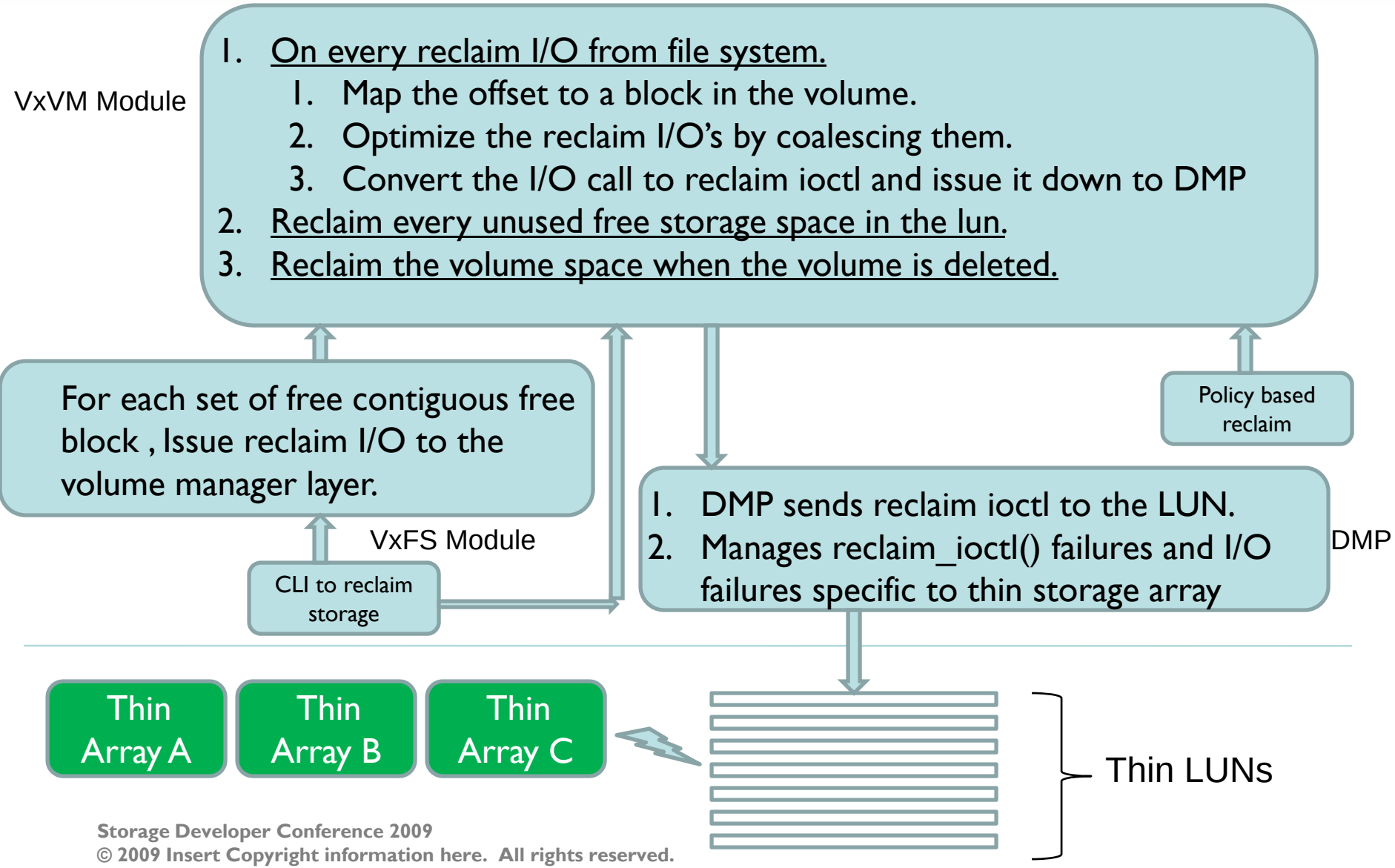
Thick to thin array migration



Thin Provision in Storage Foundation



Storage Reclamation in Storage Foundation



File system reclaim

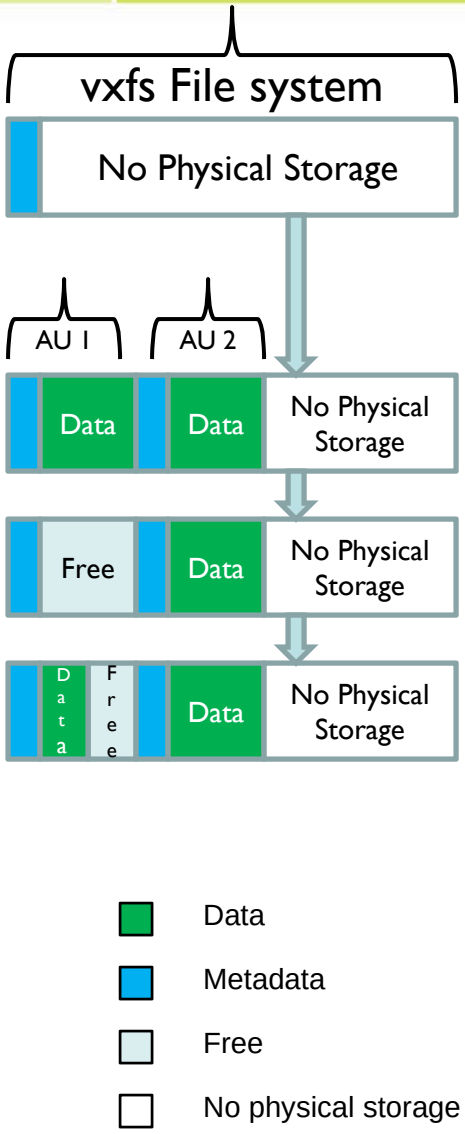
VxFS

Reclaim Request

- ❑ Extent-based file system and creates metadata only when needed.
- ❑ Thin friendly FS. (Reuses the deleted files storage space effectively)
- ❑ File system regions are well managed for reclaim code not to affect the regular IO performance.
  - ❑ Set exclusion zone over a range of blocks (or) regions.
  - ❑ Generate a map of free blocks within the exclusion zone.
  - ❑ For each set of contiguous free blocks, invoke VxVM API's to reclaim the storage.
  - ❑ Remove the exclusion zone.
- ❑ Policies are built in for efficient storage reclamation.

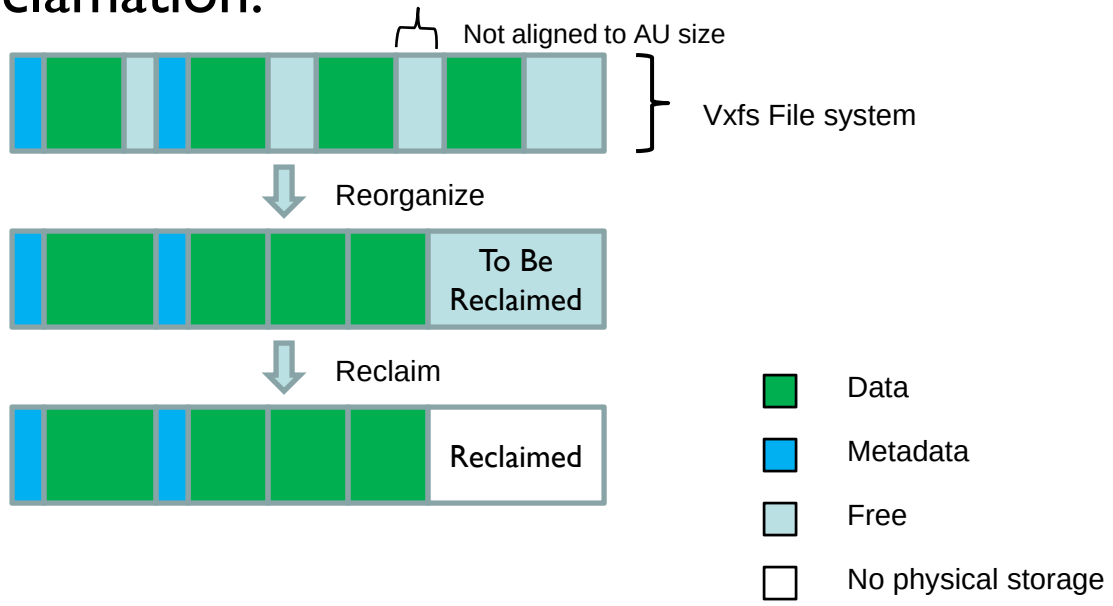
Thin Friendly File System

- Create a VxFS file system
  - VxFS divides file system into regions called Allocation Units (AUs)
- Create new files in the file system.
- Delete some files in the file system.
- Again add new files to the file system.
  - FS uses partially filled AU's before using free AUs
  - New files are created on the deleted region of files (or) in other words the physically allocated storage space is reused by the file system, instead of allocating storage.
- The unallocated physical storage space is not used until required.



File system aggressive reclaim

- Problem: Poor reclaim efficiency due to file system fragmentation.
  - Contiguous free space is smaller than array Allocation unit size (42MB for Hitachi, 16KB for 3PAR).
  - Not aligned properly to LUN allocation unit size
- Reorganize files to enlarge contiguous free space.
- Proceed with normal reclamation.

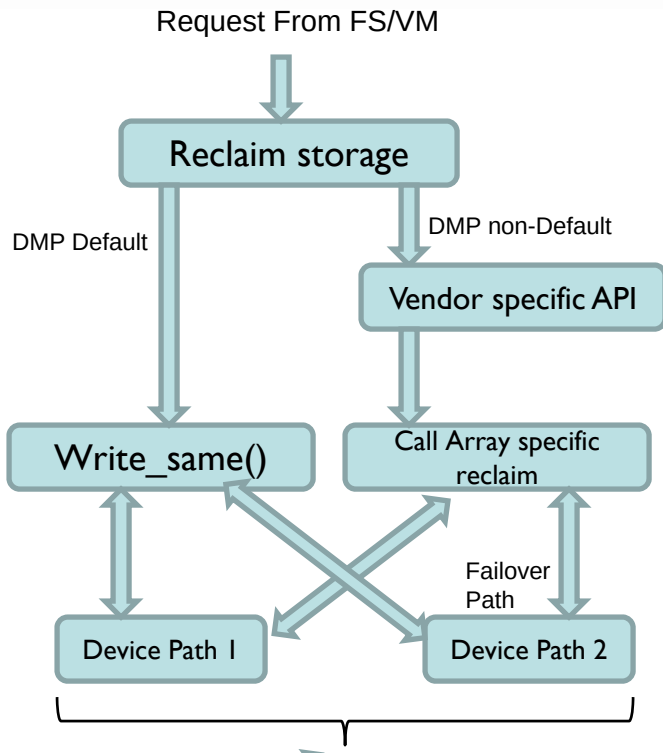


DMP's role in storage reclamation

DMP

Reclaim Request

1. If the LUN supports T10 Standard reclaim request, then DMP send the write_same() reclaim request to the LUN.
2. If T10 standard is not supported by the storage, then the DMP redirects the reclaim request using the unique API provided by the storage vendor.
3. DMP handles following reclaim ioctl failures
   1. Path failure
   2. Array busy failures
4. DMP handles the following I/O failures which are unique to thin array
   1. No physical storage space I/O error.
   2. Handling the storage space reaching the threshold limits.



Thin Array A

Thin Array B

Thin Array C

DMP - Array Specific Module

Thin LUNs

Usability

Command to report the Thin LUN usage

```
root>vxdisk -o thin list
DEVICE          SIZE(mb)     PHYS_ALLOC(mb)  GROUP          TYPE
3pardata0_65    51200        79              tcrundg        thinrclm
3pardata0_66    51200        549             tcrundg        thinrclm
3pardata0_67    51200        1056            tcrundg        thinrclm
3pardata0_68    51200        417             tcrundg        thinrclm
3pardata0_69    51200        544             tcrundg        thinrclm
3pardata0_70    51200        11184           tcrundg        thinrclm
3pardata0_71    51200        1187            tcrundg        thinrclm
3pardata0_72    51200        79              tcrundg        thinrclm

root>
```

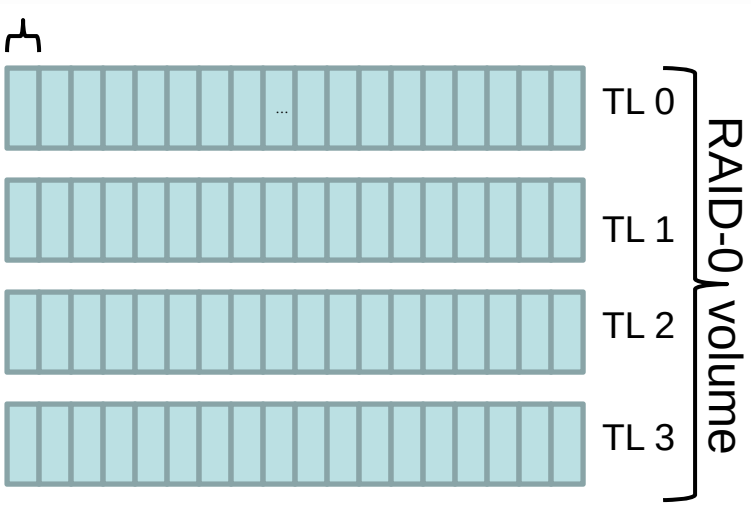
Command to reclaim the storage in the array

```
root>vxdisk reclaim 3pardata0
Reclaiming thin storage on:
Disk 3pardata0_7 : Skipped. Not a thin reclaimable disk.
Disk 3pardata0_48 : Skipped. Diskgroup, deported or imported on different host.
Disk 3pardata0_59 : Reclaimed full. Disk not in any disk group.
Disk 3pardata0_62 : Skipped. Diskgroup, deported or imported on different host.
Disk 3pardata0_63 : Done.
Disk 3pardata0_64 : Done.
Disk 3pardata0_65 : Skipped. No VxFS file system found.
Disk 3pardata0_67 : Done.
Disk 3pardata0_73 : Skipped. Disk Not Valid.

root>
```

Challenges in Storage Reclamation

- **Coalesce Reclaim request on RAID-0**
  - RAID-0 volume is striped across multiple LUNs.
  - Default stripe size is 64KB
  - A reclaim request for 1GB creates 1024 * 1024 / 64 = 16,384 separate reclaim call for each 64KB block in the stripe.
  - VxVM merges these 16,384 reclaim requests into 4 (one for each LUN)



Challenges in Storage Reclamation

- Align reclaim command offset and length to storage allocation unit size
  - Each array has unique allocation unit size
  - Volumes created using different array types add code complexity
  - Shift in LUN start offset
- Interlocking reclaim and transaction code
  - Transaction and reclaim are mutually exclusive
  - Interlocking them using locks is not possible
- Deleted volume reclaim
  - Deleted volumes are not reclaimed immediately
  - Protects data and avoids overloading the system with reclaim requests
  - Solution: policy based reclaim and efficient space reuse

Advantages of Storage Foundation thin provisioning

- Hardware independent
  - Discovers thin array attributes
  - Reclaim operation
  - Maintains error and disk usage statistics
- Any thin provision array can be supported using a array independent strategy:
  - ASL
  - APM
- VxFS file system is "thin friendly"
- VxVM SmartMove complements thin provisioning

Remaining challenges

- Pre-scan to report how much space can be reclaimed before reclaiming
- Report physical storage used by file system on different LUN types
- Optimize already reclaimed storage space
- Prioritize reclaim I/Os among other I/Os
- Threshold based policy for automatic reclamation
- Reclamation on raw volumes with Oracle database

Thank you for your attention