

# What Every DB2 for z/OS Professional Needs to Understand About IBM Disk and Tape Virtualization

DB2Night Show

November 13, 2015

John Iczkovits  
iczkovit@us.ibm.com

# Disclaimer

- **© Copyright IBM Corporation 2015. All rights reserved. U.S. Government Users Restricted Rights - Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.**
- **THE INFORMATION CONTAINED IN THIS PRESENTATION IS PROVIDED FOR INFORMATIONAL PURPOSES ONLY. WHILE EFFORTS WERE MADE TO VERIFY THE COMPLETENESS AND ACCURACY OF THE INFORMATION CONTAINED IN THIS PRESENTATION, IT IS PROVIDED “AS IS” WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED. IN ADDITION, THIS INFORMATION IS BASED ON IBM’S CURRENT PRODUCT PLANS AND STRATEGY, WHICH ARE SUBJECT TO CHANGE BY IBM WITHOUT NOTICE. IBM SHALL NOT BE RESPONSIBLE FOR ANY DAMAGES ARISING OUT OF THE USE OF, OR OTHERWISE RELATED TO, THIS PRESENTATION OR ANY OTHER DOCUMENTATION. NOTHING CONTAINED IN THIS PRESENTATION IS INTENDED TO, NOR SHALL HAVE THE EFFECT OF, CREATING ANY WARRANTIES OR REPRESENTATIONS FROM IBM (OR ITS SUPPLIERS OR LICENSORS), OR ALTERING THE TERMS AND CONDITIONS OF ANY AGREEMENT OR LICENSE GOVERNING THE USE OF IBM PRODUCTS AND/OR SOFTWARE.**
- **IBM, the IBM logo, ibm.com, DB2 are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at “Copyright and trademark information” at [www.ibm.com/legal/copytrade.shtml](http://www.ibm.com/legal/copytrade.shtml)**

- **MYTH: As a DB2 professional, there is no reason for me to understand how disk and tape virtualization works**
  
- **FACT: Your DB2 data resides on disk and probably tape as well**
  - Are your data sets allocated with the best HA (High Availability) in mind?
  - Are your data sets allocated with the best recovery in mind?
  - Are your data sets allocated with the best performance in mind?
  - Do you fully understand what is available for your data sets and the best practices for allocation and recovery?

- **MYTH: My Storage Administrator understands what DB2 requires, so there is no need for me to worry about disk or tape**
- **FACT: You can have the best Storage Administrator in the world, but the vast majority do not know how DB2 operates or what it requires.**
  - Your Storage Administrator is working with dozens of products, many their own, such as dss, hsm, SMS, rmm, etc. Do not assume they understand database requirements.
  - DB2, IMS, CICS and other products have specific product related requirements that are not all the same.
  - Do not assume that your Storage Administrator knows what boot strap data sets, active and archive logs, image copy data sets, etc. are, nor their specific requirements. They generally do not.
    - Case in point – some customers have ALL of their DB2 data sets described above, plus the DB2 Catalog and Directory, user data sets, and DB2 sort data sets on the same set of volumes causing a single point of failure and performance problems.

- **MYTH: As a DB2 professional, I have no influence over the disk and tape environment. I take what I get.**
  
- **FACT: Storage Administrators do their best to allocate DB2 data sets based on their understanding of requirements.**
  - Be specific about DB2 requirements. Keep in mind, most Storage Administrators do not know the difference between an image copy data set and a DB2 user data set. Educate your Storage Administrator on DB2 and its requirements.
  - At many sites, data sets are placed on less than the most efficient devices. At times your Storage Administrator has special technology, but has no idea DB2 can benefit from it.
    - Case in point – you have heavily random DB2 data sets that reside on HDD (regular drives), but your company has lots of new SSD (Solid State Devices) with little or nothing allocated on them). Your Storage Administrator does not know that these specific DB2 data sets can have 15 – 60x performance improvements by allocating the data sets on SSD devices instead.
    - Not all DB2 related tape data sets should reside on virtual tape. Some DB2 data sets may be much better off on manual devices.
  - Your company may have a much better technology to allocate your DB2 data sets, but it does not help if your Storage Administrator does not know your requirements.

- **MYTH: Knowing my disk channel speed does not help me understand DB2 performance**
  
- **FACT: Customers have different disk channel speeds:**
  - ESCON (some customers with old EMC devices still run ESCON)
  - FICON Express 2 (FEx2)
  - FICON Express 4 (FEx4)
  - FICON Express 8 (FEx8)
  - FICON Express 8S (FEx8S)
  - FICON Express 16S (FEx16S)
  - zHPF (High Performance FICON)
  
- **Channel speed and whether zHPF is used influence data transfer time and therefore relate to the DB2 performance**
  
- **Discuss the different technological advantages for DB2 with your Storage Administrator**

- **MYTH: The amount of disk cache on the disk boxes has nothing to do with the performance of my DB2 applications**
  
- **FACT: Think of disk cache as DB2 buffer pools**
  - There is an unofficial hand shaking between disk cache and DB2 buffer pools. In concept they work similarly to accomplish the same task
  
  - Disk cache comes in different sizes.
    - When disk boxes are purchased, it is not uncommon to purchase less cache than required.
    - The price is cheaper and the performance consequences not yet known
  
  - Does increasing the size of your buffer pools help your application because the data is re-referenced?
    - If not, increasing the size of disk cache will probably not help
    - Otherwise, would adding disk cache increase performance for your DB2 application?
    - For IBM disk, Storage ATS can help determine if purchasing more cache would benefit DB2 performance.

- **MYTH: I do not need to know the type of RAID my DB2 data resides on**
  
- **FACT: RAID (Redundant Arrays of Inexpensive Disks) allows for different options for recovery of physical devices**
  - The most common mainframe options are RAID 5, 6, and 10
  - Different RAID options provide different levels of performance and protection from device failures
    - Performance – Disk can be allocated based on different RAID options. For example, you may have RAID 5 for your DB2 environment, but RAID 10 for your non DB2 data. Discuss with your Storage Administrator which RAID option is most suitable for performance. Discuss RAID performance penalties as trade offs in relation to the total cost of your disk.
    - Protection from failure – Discuss with your Storage Administrator your companies threshold for failure. Most customers implement RAID 5 technology. Does your current RAID implementation adequately protect you from multiple physical failures?



## RAID (Redundant Arrays of Inexpensive Disks) Technology- Disk Type Dependent

- RAID 0 - data striping without parity
- RAID 1 - dual copy
- RAID 2 - synchronized access with separate error correction disks
- RAID 3 - synchronized access with fixed parity disk
- RAID 4 - independent access with fixed parity disk
- **RAID 5 – data striping, independent access with floating parity (tolerates loss of 1 physical volume (DDM))**
- **RAID 6 – data striping, dual redundancy with floating parity (tolerates loss of 2 physical volumes (DDMs))**
- **RAID 10 (DS8000 and some ESS) - RAID 0 + RAID 1, mirrored, data striping but with no parity**

Parity is additional data, “internal” to the RAID subsystem, that enables a RAID device to regenerate complete data when a portion of the data is missing. Parity works on the principle that you can sum the individual bits that make up a data block or byte across separate disk drives to arrive at an odd or even sum.

# Storage Resource Summary

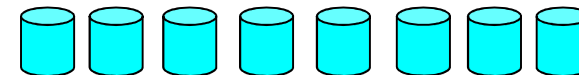
- **Disk**

- Individual DDMs (Disk Drive Module) – physical drive. Depending on the hardware, each DDM is only 2.5 to 3.5 inches wide.



- **Array Sites**

- Pre-determined grouping of DDMs of same speed and capacity (8 DDMs for DS8000; 4 DDMs for DS6000)



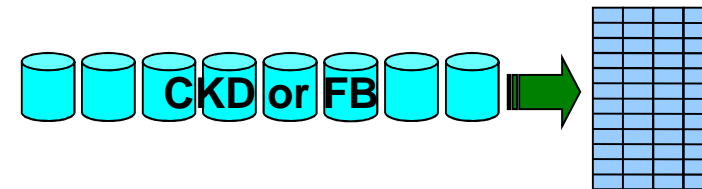
- **Arrays**

- One 8-DDM Array Site used to construct one RAID array (DS8000)



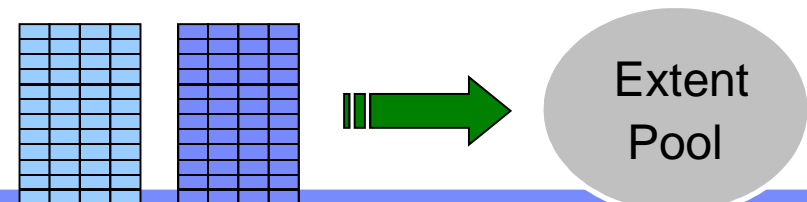
- **Ranks**

- One Array forms one CKD or FB Rank (8 DDMs for DS8000; 4 or 8 DDMs for DS6000)
- No fixed, pre-determined relation to LSS (Logical Sub System)



- **Extent Pools**

- 1 or more ranks of a single storage type (CKD or FB)
- Assigned to Server0 or Server1

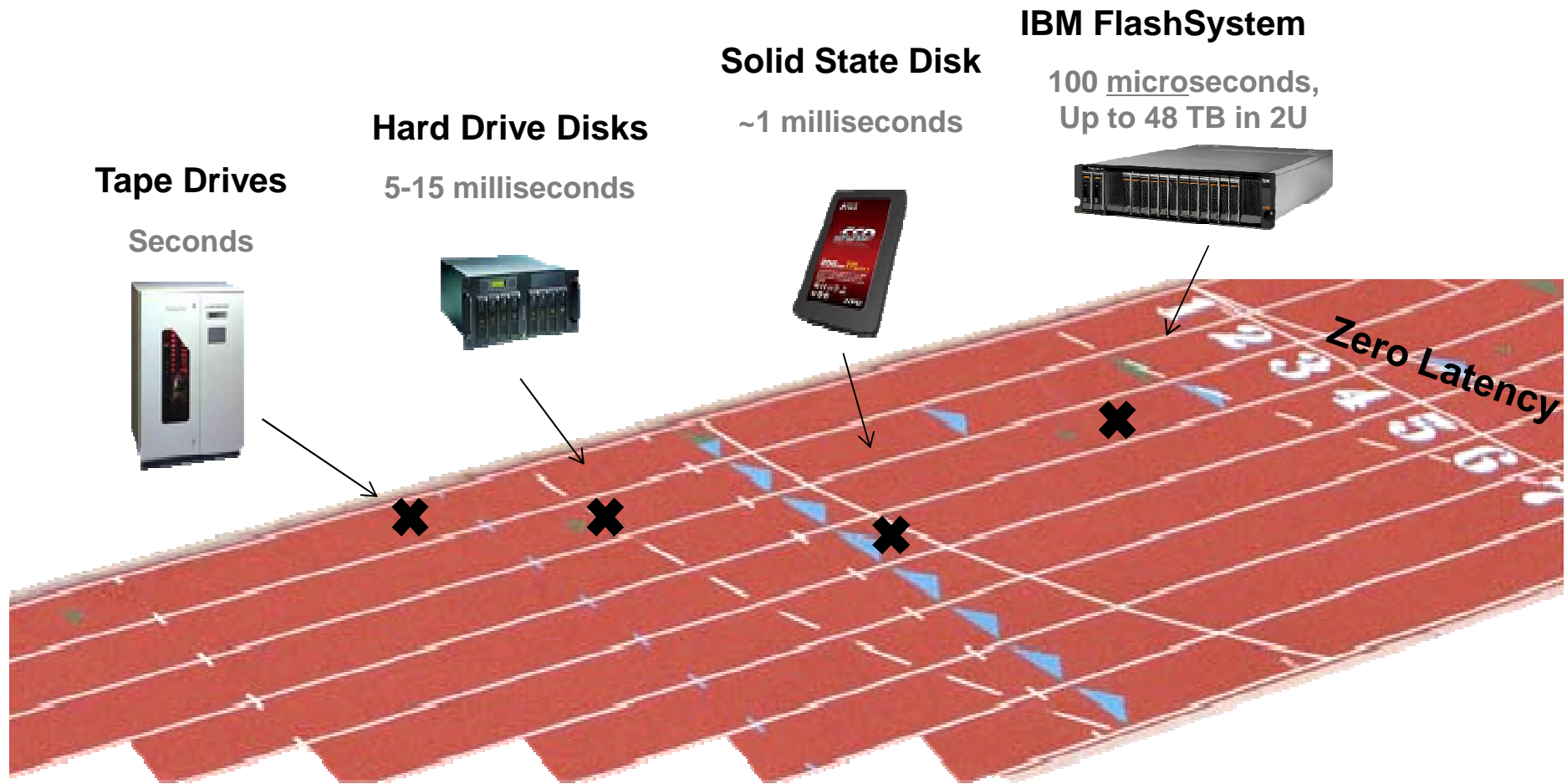


# What is a DDM (Disk Drive Modules) physical drive?

- **Three different Fibre Channel (FC) DDM types (available in both non-encrypted and encrypted versions):**
  - 146 GB, 15K RPM drive
  - 300 GB, 15K RPM drive
  - 600 GB, 10K RPM drive
  - 1.2 TB, 10K RPM drive
- **One Serial Advanced Technology Attachment (SATA)/Nearline DDM drive type:**
  - 1 TB, 7.2K RPM drive
  - 2 TB, 7.2K RPM drive
  - 4 TB, 7.2K RPM drive
- **Flash/Solid State Drive (SSD) types:**
  - 400 GB Flash
  - 200/400/800 GB SSD
  - See next page regarding for the DS8800, which includes the new 300 GB drives
- **Disks are 2.5 inch SAS (Serial Attached SCSI) for DS8800, otherwise 3.5 inches FC**
- **Each DDM (physical disk) can hold many logical volumes. For example, a mod 3 device is 2.83 GB. Many mod 3s can reside on one DDM. When you lose one DDM, how many logical volumes are you really losing? Think about storage pool striping in the next few slides as well.**
- **Three types of drives:**
  - HDD (Hard Disk Drive) Fibre Channel. Almost all mainframe disk is HDD.
  - SATA. Too slow for most mainframe uses. Some companies will use SATA drives for Data Warehousing if slower performance is acceptable, or for HSM Level 2, or backups, such as image copies. Much slower performance must be acceptable when putting data on SATA drives. Generally SATA drives are not used for DB2 environments.
  - SSD. VERY expensive and not generally used for DB2 environments. Very useful when access is random. Random access is better on SDD vs. HDD or SATA.

# Race to Zero Latency

With each new gen. of storage, comes performance gains in order of magnitude



## Disk Space Numbers - 3390 Emulated Volumes

Model	Cylinders	Tracks	Bytes/Volume	Bytes/Track **
3390-1	1113	16695	946 MB	56664
3390-2	2226	33390	1.89 GB	56664
3390-3	3339	50085	2.83 GB	56664
3390-9 *	10017	150255	8.51 GB	56664
3390-27 *	32760	491400	27.84 GB	56664
3390-54 *	65520	982800	55.68 GB	56664
3390-A <b>EAV</b>	1,182,006	* * *	1TB	56664

-EAVs now supports 1TB volumes with z/OS 1.13 and DS8K LIC (Release 6.2).

-z/OS 1.12 supports 1TB volumes with required PTFs

\* Storage Administrators refer to 3390 mod 9, 27, and 54 as large-volume support 10017 cylinders, 32760 cylinders, and 65520 cylinders respectively. These large volumes are all considered mod 9s with varying number of cylinders.

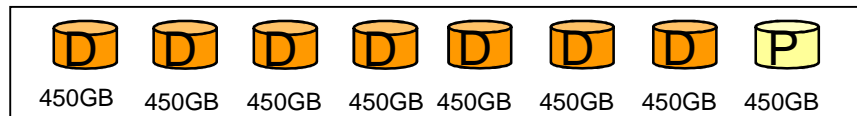
\*\* Bytes per track refers to the device capacity only. The actual allocation for a DB2 LDS data set is 48 KB, not the total of 56 KB. This allows for 12 - 4K DB2 pages to be allocated per track.

Why larger disk sizes?

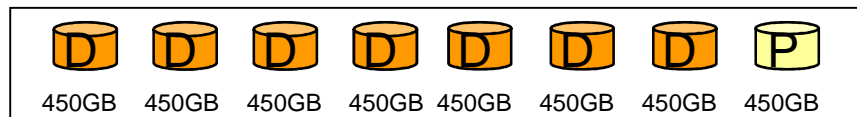
- reduces issues with MVS addresses (maximum number of devices met)
- simpler management of storage

## Storage Pool Striping

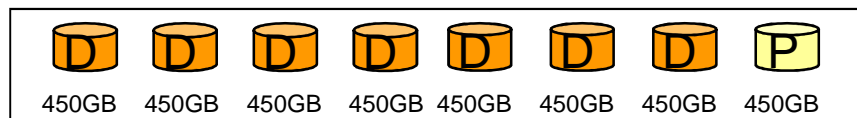
- **Example of using storage pool striping with rotating extents across 3 ranks with RAID 5 (7+P) for a mod 3 device (2.83GB - 3,339 cylinders) – 450GB DDMs**



**Rank 1 – 1,113 cylinders  
Logical volume - .94GB**



**Rank 2 – 1,113 cylinders  
Logical volume - .94GB**



**Rank 3 – 1,113 cylinders  
Logical volume - .94GB**

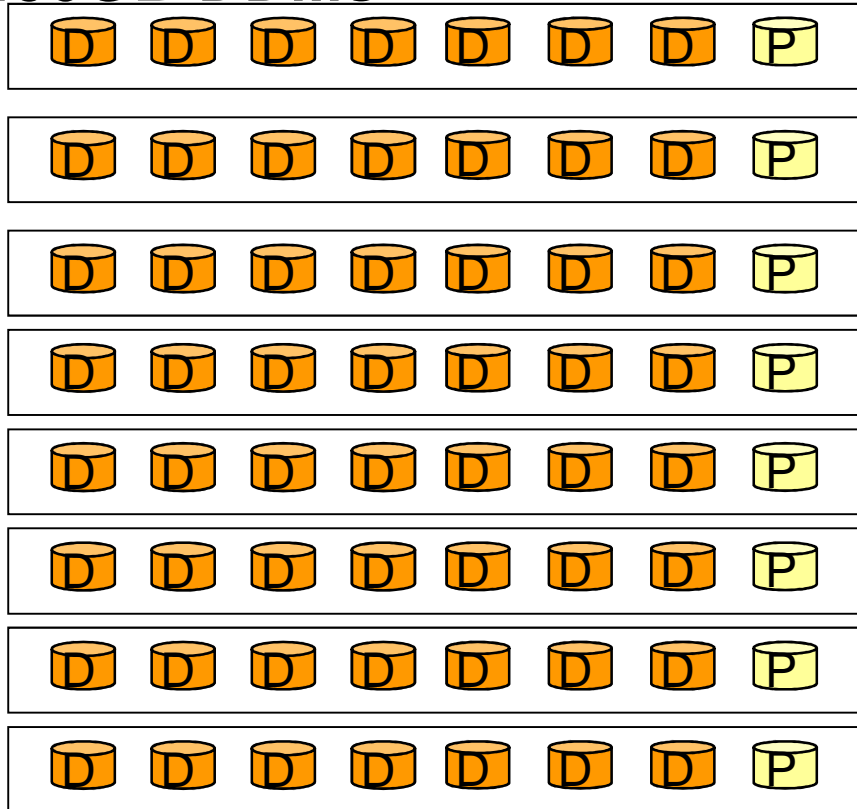
**This one logical mod 3 volume with 3,339 cylinders really resides on 24 physical devices (DDMs)**

**In this scenario when you backup or Flash Copy one logical volume you are really accessing 24 physical volumes.**

- Parity is striped across all disks in array but consumes capacity equivalent to one disk

- Example of using storage pool striping with rotating extents across 8 ranks with RAID 5 (7+P) for a mod 9 device (8.51GB - 10,017 cylinders) -

450GB DDMs



Rank 1 – 1,113 cylinders  
Logical volume - 1.88GB

Rank 2 – 1,113 cylinders  
Logical volume - .94GB

Rank 3 – 1,113 cylinders  
Logical volume - .94GB

Rank 4 – 1,113 cylinders

Rank 5 – 1,113 cylinders

Rank 6 – 1,113 cylinders

Rank 7 – 1,113 cylinders

Rank 8 – 1,113 cylinders  
Logical volume - .94GB

This one logical mod 9 volume with 10,017 cylinders really resides on 64 physical devices (DDMs). Mod 9 devices require the equivalent of 9 ranks. Since only 8 ranks exist, the last 1,113 cylinders will be added back to rank 1. When adding to ranks. full ranks are skipped.

## IBM System Storage DS8800 Architecture and Implementation SG24-8886

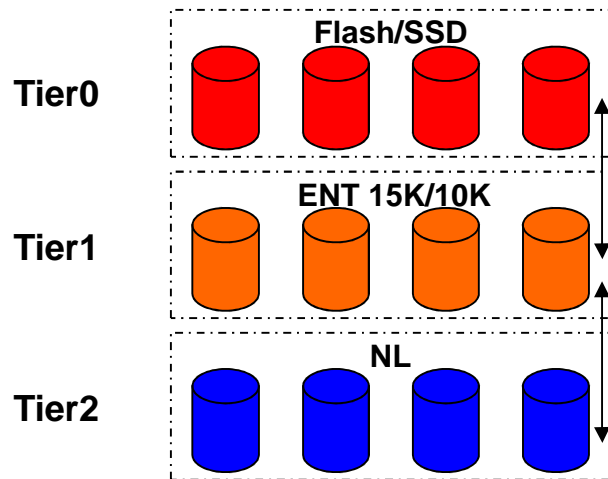
**“Storage Pool Striping can enhance performance significantly, but when you lose one rank (in the unlikely event that a whole RAID array failed due to a scenario with multiple failures at the same time), not only is the data of this rank lost, but also all data in this Extent Pool because data is striped across all ranks. To avoid data loss, mirror your data to a remote DS8000.”**



# Easy Tier Support for High Performance Flash Enclosure

- **Easy Tier VI (DS8870 R7.3)**

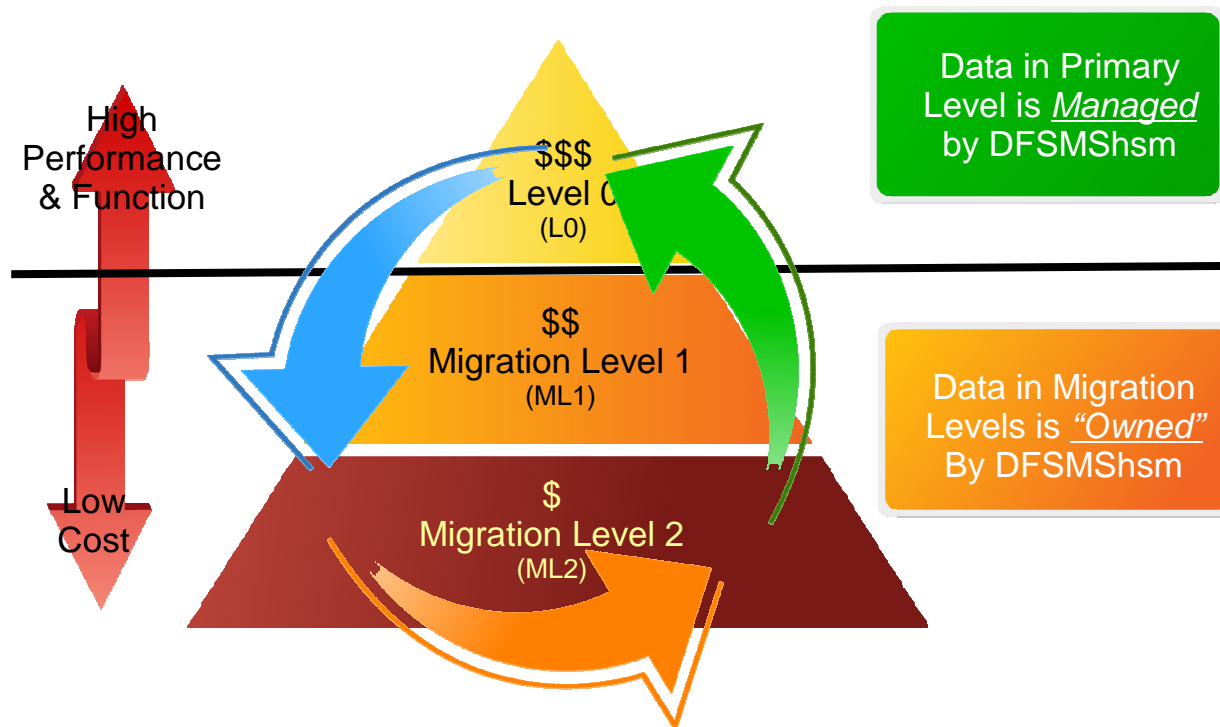
- High Performance Flash enclosure with Flash cards can provide higher IOPs capacity compared with traditional SSDs
  - Response times are similar to DS8870 SSD performance so Easy Tier considers both Flash and traditional SSDs as Tier0
- Easy Tier Auto Rebalance within Tier0
  - Will distribute higher proportion of the IOPs workload to High Performance Flash
- All other existing Easy Tier functions work the same as Easy Tier V



Configuration	Traditional SSD and Flash cards	ENT (10 K / 15 K SAS or FC)	NL (NL - SAS or SATA)
Flash/SSD, ENT, and NL	Tier 0	Tier 1	Tier 2
Flash/SSD and ENT	Tier 0	Tier 1	N/A
Flash/SSD and NL	Tier 0	N/A	Tier 1
ENT and NL	N/A	Tier 1	Tier 2

## Storage Tiers: V2R1 Background

### The *classic* DFSMS storage hierarchy (35 years old)



## Storage Tiers: V2R1 Background

- **The Space Management Environment is evolving**
  - Typical configurations have changed to leave data on Level 0 longer and then migrate directly to ML2
    - ★ When ML2 is a VTS, the VTS disk cache replaces the ML1 tier
      - Eliminates MIPS required for software compression to ML1
      - Eliminates DFSMSHsm ML1->ML2 processing
    - ★ Tape storage controller functionality replacing DFSMSHsm duplexing
  
- **Shortcomings of today's DFSMS functionality**
  - ✓ No policy-based automation for moving data within the Primary Storage Hierarchy (Level 0)
  - ✓ No policy-based management of Active (open) data



## Storage Tiers: V2R1 Overview

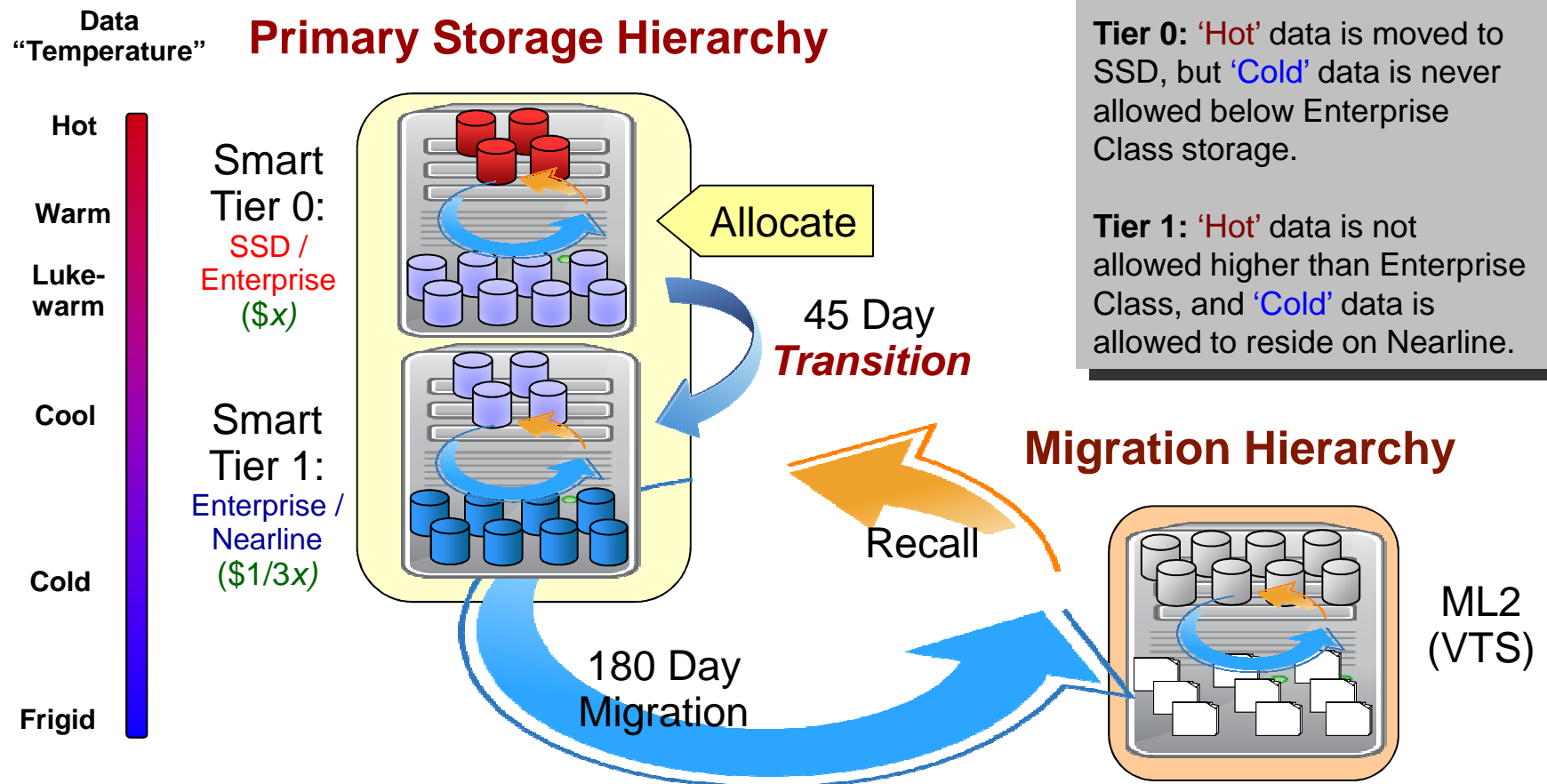
### **z/OS V2R1 DFSMS provides the Storage Tiers Solution...**

***Automated, policy-based space management that moves data from tier to tier within the Primary (Level 0) Hierarchy***

- ✓ Automated movement provided via the existing DFSMShsm Space Management function
  - *Movement is referred to as a 'Class Transition'*
  - *Data remains in its original format and can be immediately accessed after the movement is complete*
- ✓ Policies implemented via the existing Class Transition policies and updated Management Class policies
- ✓ Enhanced support for DB2, CICS and zFS data
  - - *Open data temporarily closed to enable movement*

# Storage Tiers: V2R1 Overview

This example environment shows data being “transitioned” from ‘Smart Tier 0’ to ‘Smart Tier 1’ after 45 days and then becoming eligible for migration after 180 days of inactivity.



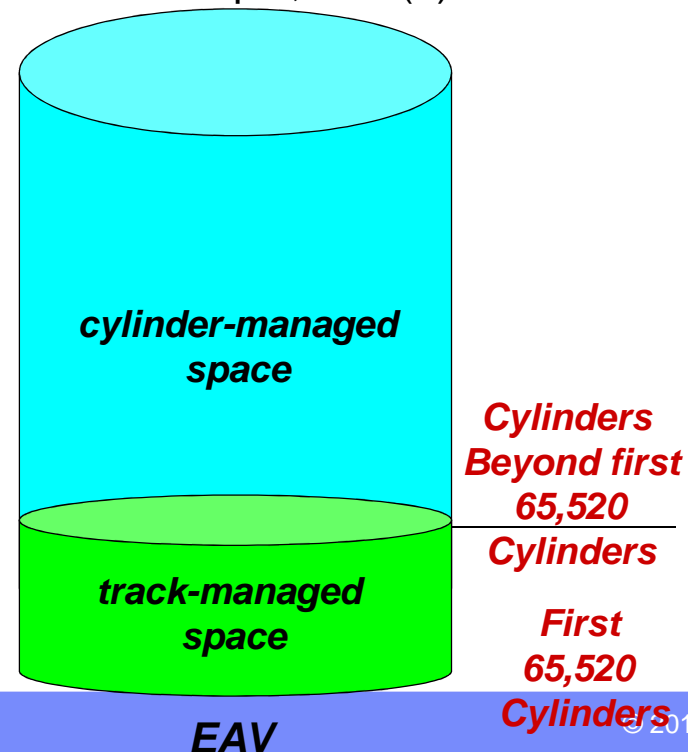
**Tier 0:** ‘Hot’ data is moved to SSD, but ‘Cold’ data is never allowed below Enterprise Class storage.

**Tier 1:** ‘Hot’ data is not allowed higher than Enterprise Class, and ‘Cold’ data is allowed to reside on Nearline.

- **MYTH: My data always resides on the physical volser they were allocated on**
- **FACT: IBM's Easy Tier product can move "data" between different types of disk or the same disk**
  - Customer can have a mix of two or all three types of IBM disk. Easy Tier will move "data" which is a volume extent (not a data set extent) to the best fit disk for performance
  - Extents are moved dynamically without an outage
  - Customer example – owns HDD and SSD volumes. DB2 data is mostly random residing on HDD. Easy Tier can move the volume extent to an SSD volume.
    - No outage required!
    - VOLSER is the same, however the volume extent resides on different physical volumes and different type of disk
    - DB2 is not aware of the move
    - When the data set is REORGed and no longer flagged as random, placement will probably be on HDD and the process has to start over again.
  - IBM disk competitors have similar products for their disk
  - Keep in mind that when z/OS states that a data set resides on one volser, it can actually reside on different tiers in the hardware.
  - **Warning!** Even if you only have HDD, extents can be moved laterally to any rank in the disk box in its extent pool, from a hot rank to cool rank. Make sure that the BSDS and active log data sets are excluded from East Tier or equivalent product to avoid them residing in the same extent pool. Alternatively pin the volumes to avoid movement.

- **MYTH: The 3390 disk emulation my DB2 data resides on does not matter**
  
- **FACT: 3390 logical volsers emulate mod 1, 2, 3, 9, 27, 54, and EAV devices**
  - How much data do you need to manage for DB2?
  - How large are the data sets that you need to manage?
  - How many logical volsers are you willing to manage?
    - One 64GB data set would require about 35 mod 3s, or two to three 64GB data sets can reside on one EAV (based on 223GB per EAV)
      - EAVs now supports 1TB volumes with z/OS 1.13 and DS8K LIC (Release 6.2). z/OS 1.12 supports 1TB volumes with required PTFs
    - For 4GB data sets, one data set would require two mod 3s

- EAVs work differently depending if the data is allocated on the track managed portion or the cylinder manager portion
  - Track managed (1<sup>st</sup> ¼ of the volume – mod 54 equivalent) business as usual
  - Cylinder managed portion (EAS – Extended Addressing Space) works differently because CCCC (cylinder location) maxes out at 'FFFF' which is 65535 cylinders (slightly greater than a mod 54)
    - Allocations in the EAS not divisible by 21 cylinders are promoted to 21 cylinders
      - > Allocation is 10 cylinders in the EAS, allocation automatically promoted to 21 cylinders. Allocation is 100 cylinders in the EAS, automatically promoted to be divisible by 21 cylinders (105 cylinders total).
      - > If the track managed portion is full and there is space in the EAS, allocations in tracks are promoted to 21 cylinders. For example, TRK(1) on the EAS is converted to CYL(21).





IBM 3850 data cartridge

and

IBM 3336 disk pack



When reviewing CCCCHHHH (Cylinder Head) information, MVS still views technology when multiple disk platters were used. Above is not a 3390 depiction, but will do for illustration.

- CCCC is the cylinder HHHH is the read/write head. CCCCHHHH is the location of 1 track.
- Each track is either on its own platter or top and bottom of a platter. Each platter has either a dedicated or shared read write head.
- Cylinder and head information is off by 1 in LISTCAT and IEHLIST output
- CCCCHHHH '00000000' translates to cylinder 1, head 1 (cylinder 1 track 1)
- CCCCHHHH '00000001' translates to cylinder 1, head 2 (cylinder 1 track 2)
- CCCCHHHH '0000000E' translates to cylinder 1, head 15 (cylinder 1 track 15)
- When 1 cylinder is read, all read/write heads read data in unison

**NOTE – this format is no longer reality with disk arrays. It has not been reality for two decades, however MVS still reports disk allocation with the old cylinder and head formats**

## Data set location

- **Data sets no longer reside where LISTCAT or IEHLIST display**
  - Generally multi platters are no longer used. At times four platters were used, but generally they are single platter disks
  - With modern technology we allocate 4 tracks at a time on a single physical volume
  - 15 contiguous tracks (1 cylinder) of older technology meant that data was read vertically, not horizontally. Reading 1 cylinder meant that all read/write heads read data in unison.
  - 15 contiguous tracks (1 cylinder) with the newer disk array technology means that 4 tracks are contiguous only. 1 cylinder read requires reading several separate physical disk volumes. The number of physical volumes is dependent of the RAID implementation.
    - With SSD, tracks and cylinders do not exist, and the cells containing the data are not contiguous. A 3 track data set in 1 extent on SSD really means that data is placed in non-contiguous cells. The cell equivalent of tracks 1, 2, and 3 are nowhere near each other.
  - In today's world, 1 cylinder contiguous no longer means all data is contiguous on one disk with multiple read write heads.
  - Keep in mind, each physical volume houses dozens, if not hundreds of logical volumes. Each DDM is only 2.5 to 3.5 inches wide.

- **MYTH: I do not need to know if my DB2 data resides on PAV devices**
  
- **FACT: PAV (Parallel Access Volumes) allows for multiple reads and writes for a volser so long as those sets of tracks are not being updated at that time**
  - With older technology IOSQ (I/O Sequential Queuing) was a major performance inhibitor
    - DB2 thread 1 reads the first 100 cylinders from volume DB200A
    - While the first 100 cylinders are still being read in, thread 2 asks to read 5 tracks at cylinder 3000 from the same volume – DB200A
    - While the first 100 cylinders are still being read in, and thread 2 is still waiting, thread 3 asks to read 30 tracks at cylinder 1500 from the same volume – DB200A
    - Thread 2 and 3 wait (I/O Sequential Queuing) until the first 100 cylinders were read for thread 1.
      - All read/write arms are reading in the 100 cylinders and we do not do an I/O interrupt for thread 2 or 3.
      - Once thread 1 is complete, then thread 2 will process, once that is complete, thread 3 will process
  - Newer technology does not work with data on just one volume with several platters, and the access arms do not need to move very far as the entire disk is only 2.5 to 3.5 inches wide.
    - Several physical disks are read from at once
    - In the above example, 1 base UCB is used for DB200A and 2 aliases, 3 UCBs in total and the reads are done simultaneously and not queued
  - Same volser, several different MVS addresses (UCBs)
    - Logical volsers can have several different MVS addresses and several channel/paths
  - A heavily used DB2 logical volume can have dozens of MVS addresses
  - Three different types of PAV:
    - Static – original PAV. Specific number of PAVs for logical volumes
    - Dynamic (most common) – Use of WLM to manage logical volsers across a SYSPLEX. Logical volumes only acquire additional UCBs when required
    - Hyper (becoming more common) - Logical volumes only acquire additional UCBs when required, however does not have the WLM overhead.
  - With newer disk, IOSQ is totally eliminated or greatly reduced
  - Some customers do not have PAVs or enough PAVs. Validate the setup to ensure DB2 does not queue and wait for a device.

## How many DDMs does my single volume data set reside on?

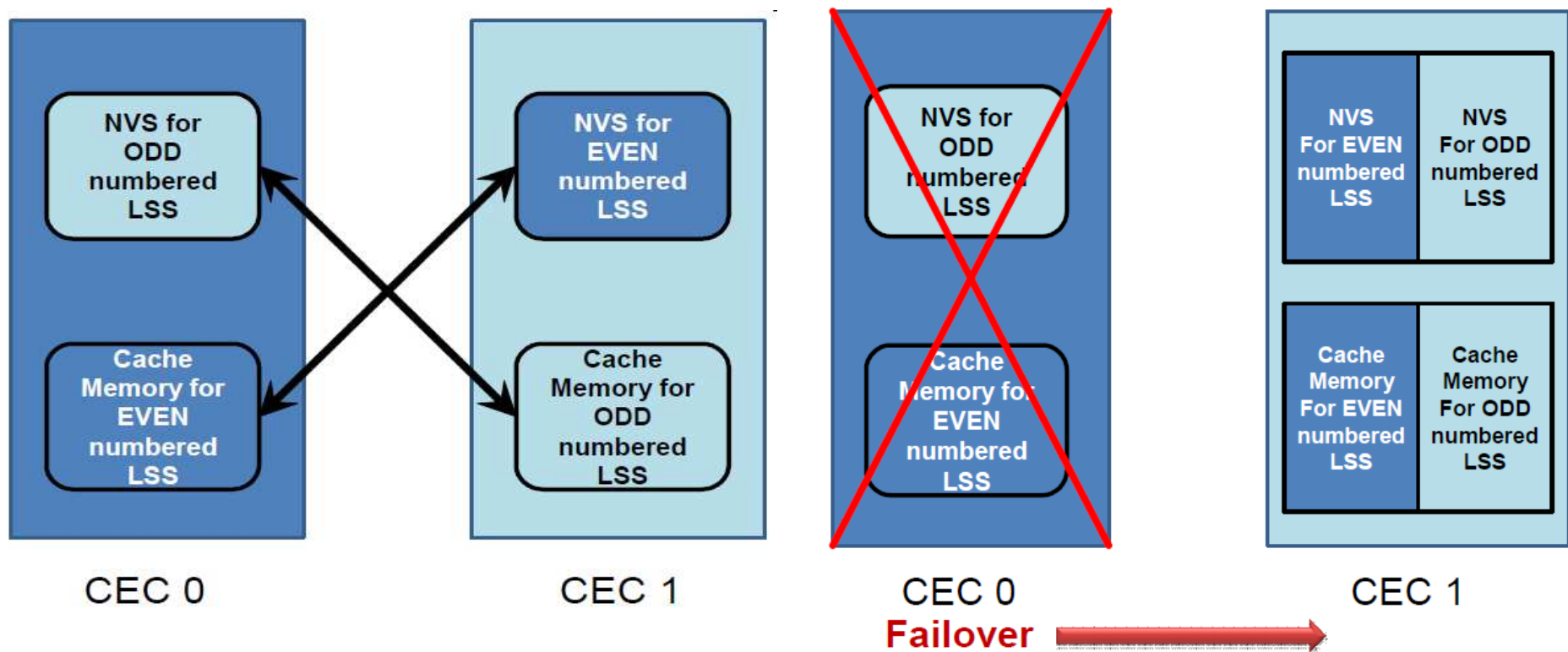
- From a VTOC perspective an allocation on a single volume will display one (logical) volume. In reality the data set resides on several different physical volumes.
- For our examples we are using a RAID 5 7+P configuration with at least 3 ranks
- Data set is on a single logical volume with 2,500 cylinders allocated. Can be 1 or multiple extents, in this case it does not matter. In this scenario the data set is allocated on 24 DDMs (3 ranks \* 8 DDMs).
- Data set is on a single logical volume with 25 cylinders allocated in 1 extent. Allocation can be on 8 DDMs (1 rank), or 16 DDMs (2 ranks). Since the VTOC is not aware of ranks, it is possible there is a chunk of space that spans ranks that they are both used. For something as small as 25 cylinders, 3 ranks would not be used as they would not exceed 1 extent on 2 ranks.
- Data set has 2 extents on a single logical volume, each extent is 1 track. It depends. For example, extent 1 was allocated on 1 array, extent 2 is not allocated until a month later. Extent 2 could be on any rank with sufficient space. It is possible that extent 1 resides on an array in one rank, and that extent 2 resides on a different array in a different rank.

- **MYTH: Extents always matter for my DB2 data**
  
- **FACT: When MVS opens a data set it creates pointers to all of the extents.**
  - More extents means longer data set open
  - In previous tests, 1 extent vs. 100 extents resulted in a 2x longer data set open time for the 100 extents
    - SELECT \* for all the cylinders took the same time
      - Once the data was opened, execution time was the same 1 extent vs. 100
  - Data set open and pseudo open are extremely expensive.
    - Be careful about opens and ZPARM values for PCLOSEN and PCLOSET
    - Verify open times with Accounting Reports
  - Apply APARs PM00068, PM17542, PM18557 for faster data set opens
    - requires z/OS 1.12

- **MYTH: It does not matter if I multi stripe my data sets, but they are never striped anyway**
  
- **FACT: With RAID 5, 6, and 10 all data sets are automatically single striped, but not multi striped**
  - VSAM and sequential data sets can be multi striped
  - Striping is only eligible for data sets residing on disk. Virtual tapes that use disk do not count as they are still seen by MVS as tape.
  - Stripe heavily sequential data sets only
    - Active logs can be striped
    - Stripe the archive log data sets (starting in DB2 9 NFM) only if write time needs to decrease due to them not offloading fast enough
    - Striping utility data sets may decrease elapsed time
  - Test and retest. Make sure that the stripes are across different ranks for best performance
    - Storage Administrator should be able to place logical volumes across different ranks

- **MYTH: I never lose disk volumes, but if I lose one volume, my entire disk box is lost**
  
- **FACT: DDMs are lost frequently.**
  - It is not unusual that very large customers with lots of DDMs to lose a DDM every day
    - Many mid size customers lose a DDM once a week, with smaller customers losing a DDM once a month
  - At times DDMs with DB2 data are lost.
    - With RAID technology, recovery is automatic and a volume is rebuilt, unless the number of failed volumes in a rank exceed the RAID implemented limit.
    - Generally when just one DDM housing DB2 data is lost, you will not be aware of it, and will probably not notice the performance difference when the volume is being rebuilt.
  - Even if the number of failed volumes exceed the implemented RAID limitation, only the rank or possibly the extent pool is lost, not the entire disk box.
    - The DS8000 contains two separate CECs. One side has its cache and the failover for the other CEC's NVS (Non Volatile Storage) and visa versa. Even if the entire CEC (one side) fails, it fails over to the other CEC.
  - When all else fails, customers that install HyperSwap have an additional line of defense to failover to another set of volumes rather than going to DR.

## Write data when CECs are dual operational and failover





- **MYTH: SMS maps to my physical environment**
  
- **FACT: With very few exceptions, SMS maps to the logical disk level, not the physical.**
  - Almost nothing that happens in the disk subsystem is known to SMS or MVS
  - We see our data through the logical volser's VTOC. The VTOC is unaware that DDMMs, ranks, arrays, and rotate extents vs. rotate volumes exists
    - SMS works using the logical volume concept, not the reality of the disk array
    - Very little is known of the physical environment, almost everything we see is from a logical volume perspective
    - Newer disk boxes are extremely sophisticated and cannot be micro managed. We tell the disk box what we require and it will decide on how to configure its resources.

- **MYTH: So long as my data is separated onto different logical volumes, job complete. I am protected.**
- **FACT: Tens or hundreds of logical volumes can reside on the same rank – same set of physical disks**
  - Availability – you have done your due diligence, BSDS1 and the LOGCOPY1 data sets reside on DB2001, and BSDS2 and LOGCOPY2 data sets reside on DB2002. If DB2001 and DB2002 are on the same rank or extent pool and a failure beyond the RAID implementation occurs, you have lost both copies of the BSDS and active logs. Best practices dictate that the pairs be split at a minimum on different ranks or extent pools.
    - Placing the BSDS and active log data sets in the SMS Separation Profile data set assures that the data sets do not reside on the same logical volume, but they do not protect against the data sets residing on the same physical rank or extent pool.
  - Performance – if your DB2 data resides on 100 logical volumes, best practices dictates that the data is spread across different ranks or extent pools. This would help avoid some potential hot spot issues as well.
    - Having data sets on different logical volumes, but on the same rank may eliminate VTOC and VVDS reserve issues, but it does not mitigate the availability issue
    - Hot spots can still occur, but are much less frequent. Even with rotate extents, if by chance two heavily competing data sets reside on the same DDMMs, hot spots can still occur.

- **MYTH: I do not need to know what type of tape my DB2 data resides on**
  
- **FACT: Several types of tape technology exists:**
  
- **Virtualized tape – TS7700 family, VTS, TS7680. Use of disk with logical tape volumes that may create data sets on physical tape volumes depending on product installed.**
  - Virtual tape can be used together. For example, a TS7720 and TS7740 can be used together almost as an hsm ML1 and ML2 approach. Data sets can be allocated to the TS7720 tapeless unit and then later migrated to the TS7740.
  
  - Interesting fact – the repository that manages tapes in IBM virtual tape systems is DB2 for LUW!
  
- **ATL – Automated Tape Libraries. Not front ended by disk or using disk. Real physical tapes are used with robotics instead of manual intervention.**
  
- **Manually mounted tapes. No disk used.**
  
- **VTFM – virtual tape, but allocated on native disk (DS8000, etc.)**
  
- **TMM (Tape Mount Management) – tape data set allocations are converted to disk allocations**

# IBM Backup & Archive Portfolio



**TS1050 (LTO5)**  
**TS1140 (Jaguar)**

## Tape Drives

- LTO5 tape drive
  - Encryption capable
  - 1.5TB Native capacity cartridge
  - Up to 140 MB/sec throughput
  - LTF5 support
  - Dual ported drive
- TS1140 tape drive/controller
  - Fourth generation tape drive
  - Controller supports FICON & ESCON
  - Tape drive data encryption
  - Up to 4TB cartridge capacity
  - Up to 1200 MB/sec throughput
  - Auto Virtual Backhitch



**TS3100 (3573)**  
**TS3200 (3573)**  
**TS3310 (3576)**  
**TS3500 (3584)**

## Tape Libraries

- TS3100 tape library (up to 19.2TB)
- TS3200 tape library (up to 38.4TB)
- TS3310 tape library (up to 316.8TB)
  - Stackable modular design
  - LTO Tape drives
- TS3500 tape library (up to 30PB with LTO5 or up to 180PB with TS1140)
  - Linear, scalable, balanced design
  - High Availability
  - High Density
  - Fastest robotics in industry
  - LTO and TS1100 tape drive
  - Connect up to 15 libraries for over 300,000 slots using shuttle complex



**TS7720 (tapeless)**  
**TS7740 (Hydra)**  
**TS7680 (DeDup)**  
**TS7650 App (DeDup)**  
**TS7610 (DeDup)**  
**TS7650G (DeDup)**

## Virtualization

### Mainframe Virtual Tape

- TS7720 (Virtual Tape)
  - Tapeless
  - Up to 1000 MB/s throughput
  - Up to 440 TB native cache
  - Standalone or GRID (PtP)
  - Up to 6-way Grid
  - Hybrid Grid support
- TS7740 (Virtual Tape)
  - Up to 1000 MB/s throughput
  - Up to 28 TB native cache
  - Standalone or GRID (PtP)
  - “Touchless” with Export option
  - Up to 6-way Grid
  - Hybrid Grid support
- TS7680G (Dedup)
  - 600Mb/s INLINE
  - Up to 1PB Repository
  - 100% Data Integrity
  - Data / Disk Agnostic
  - Native Replication

### Open Systems Virtual Tape

- TS7610 App Express (Dedup)
  - 80Mb/s INLINE
  - 4TB & 5.4TB Useable capacity
  - 100% Data Integrity
  - Data Agnostic
  - Native replication
  - Many to one replication support
- TS7650 Appliance (Dedup)
  - 500Mb/sec INLINE
  - 7TB to 36TB Useable capacity
  - 100% Data Integrity
  - Data Agnostic
  - Many to one replication support
  - High Availability (36TB option)
- TS7650G (Dedup)
  - 1GB/s (Cluster) INLINE
  - Up to 1PB Useable capacity
  - 100% Data Integrity
  - Data / Disk Agnostic
  - Native replication
  - Many to one replication support
  - High Availability

## Where do your tape data sets reside? Some things to think about

### ■ **Manual tapes:**

- How long does it take to allocate and then mount and read back the data? Is that time frame acceptable to DB2 or will it cause resource unavailable or other problems?
- Manual cartridges can house 4TB of data. How much of your data resides on one physical tape? Having many of your data sets on one physical tape still requires the data to be serialized – no parallel processing for your data sets on tape. Because tapes are serialized, parallel processing more than one data set, or more than one job requesting the same data set on the same physical tape is not permitted.
- Tape data sets need to be duplicated. What happens if a tape snaps or is lost?

### ■ **Automated tapes:**

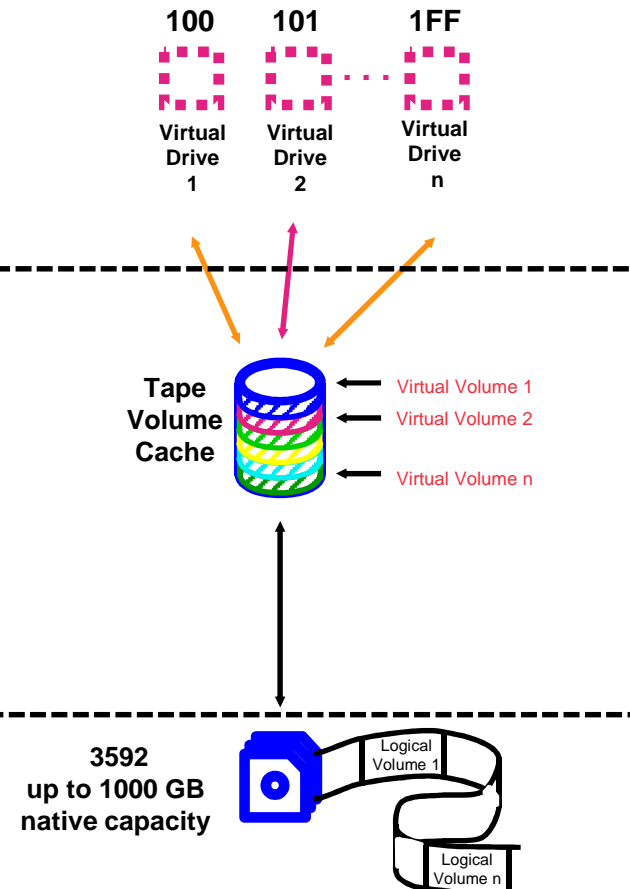
- Same serialization issues as manual tapes
- Same duplication requirements as manual tapes

## Benefits of TS7700 Virtualization Engine

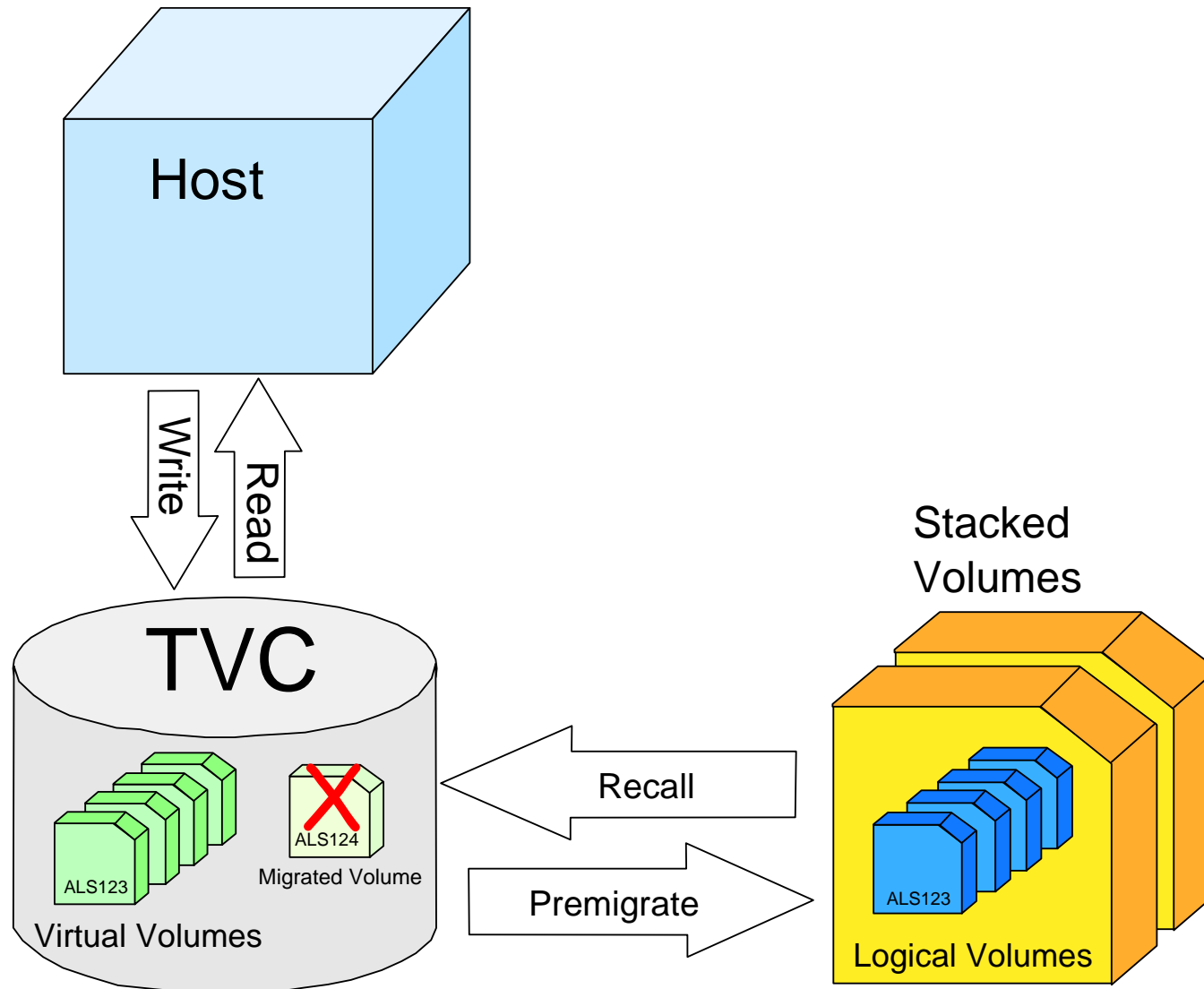
- Can fulfill all tape requirements. DASD Dumps as well as very small volumes.
- Fills large capacity cartridge. Putting multiple virtual volumes into a stacked volume, TS7740 uses all of the available space on the cartridge.
- DFSMSHsm, TMM, or Tape Management System is not required. TS7700 is a hardware stacking function; no software prerequisites are required.
- More tape drives are available. Data stacking is not the only benefit; TS7700 emulates 256 virtual drives per cluster, allowing more tape volumes to be used concurrently.
- There is little management overhead. The stacking and copying of data is performed by the TS7700 management software, without host cycles.
- Provides off site storage of physical tapes through Copy Export.

# Virtual Tape Concepts

- Virtual tape drives
  - Appear as multiple 3490E tape drives
  - Shared / partitioned like real tape drives
  - Designed to provide enhanced job parallelism
  - Requires fewer real tape drives
  - TS7700 offers 256 virtual drives per cluster
- Tape Volume Caching (TVC)
  - All data access is to disk cache
  - Removes common tape physical delays
  - Fast mount, positioning, load, demount
  - Up to 28 TB / 440 TB of cache (uncompressed)
- Volume stacking (TS7740)
  - Designed to fully utilize cartridge capacity
  - Helps reduces cartridge requirement
  - Helps reduces footprint requirement



# Logical Volumes and Stacked Physical Volumes

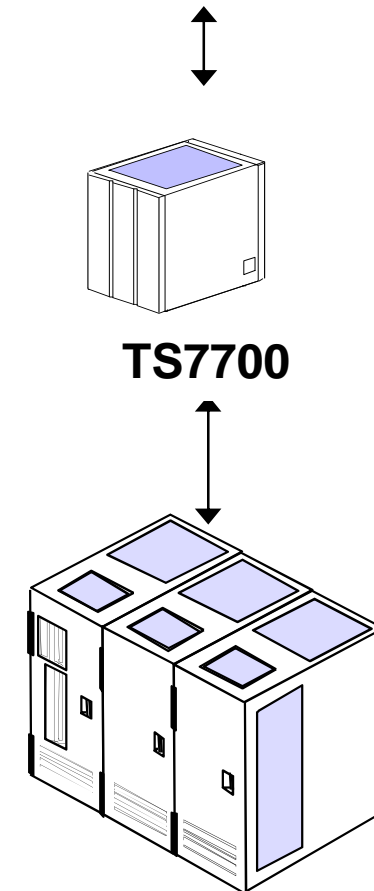




## TS7700 - Capabilities

- Tape Volume Cache
  - TS7740
    - RAID 5
    - 1 to 28.17 TB cache (3 – 84.5 TB @3:1 compression)
  - TS7720
    - RAID 6
    - Up to 1 PB without compression
- 256 Virtual Tape Devices
- 2 Million Logical Volumes
- Advanced Policy Management
- 4 to 16 3592 Physical Drives (TS7740)
- 3584 Physical Library Support (TS7740)

### Mainframe Attachment



## Benefits of tape virtualization

- **High Availability and Disaster Recovery configurations**
- **Fast access to data through caching on disk**
- **Utilization of current tape drive, tape media, and tape automation technology**
- **Capability of filling high capacity media 100%**
- **Large number of tape drives available for concurrent use**
- **No additional software required**
- **Reduced Total Cost of Ownership (TCO)**

## Where do your tape data sets reside? Some things to think about

### ■ Virtual tapes

- Do your data sets use the right size virtual tape?
  - 400, 800, 1000, 2000, 4000, or 6000 MB – set via SMS Data Class
  - 4GB data set can reside on one virtual tape volume or 10
- Do you stack/append your data sets or waste valuable logical volsers?
- Are your tapes part of a grid/PtP implementation, or duplexed another way?
- Same serialization problem as manual and ATL tapes
  - This is true as well, even if the “tape” data set resides in the TVC (disk) portion of the tape unit. MVS tells DB2 the data set resides on tape, even though it resides on disk and therefore DB2 must serialize all “tape” requests.

## Virtual Tape VOLSER vs. Physical Tape VOLSER

- **The tape VOLSER we read from and write to is the virtual VOLSER, not the physical VOLSER.**
  - Keep in mind, we allow up to 2,000,000 virtual volumes, however only a few thousand physical volumes.
  - From a DB2 perspective, at no time are the physical VOLSERs externalized. We do not know that the physical tapes or tape drives exist.

## Some key differences between ATLS and virtualized tape

- **ATLS do not contain disk drives to house data, therefore no virtualization is possible.**
- **ATL VOLSERS are the actual physical tape VOLSER, while only the logical tape is known for virtualized tape.**
- **ATL tape UCBs are the physical tape drives, while virtualized tapes will only externalize the logical drive.**
- **Virtual tape allows for copying of data to another location via VTS PtP or the TS7700 grid system, ATLS do not allow for this type of copy mechanism.**
- **ATLS do not automatically stack data sets on tape as virtualized tapes do. Software is required to stack tapes on ATLS.**
- **Only virtualized tape will emulate an IBM 3490 Enhanced Capacity (3490E) tape drive of a specific size: 400, 800, 1000, 2000 or 4000 MB. When using ATLS, you are not restricted by these sizes, rather the size of the physical tape.**

## Media for DB2 related data (archive logs, image copies, large sort work data sets, some other assorted data sets)

- **Disk – if parallelism is one of the key factors, create your data sets on disk. There are some alternatives:**
  - create data sets on tape, then use an MVS utility to recreate the data on disk when required.
  - Use hsm to migrate the data to tape, then recall the data. This is a more automated approach than the previous alternative. Both alternatives require additional time to bring your data sets back to disk.
  - Disk is also a good alternative if you need to recover hundreds or thousands of objects. For example, tape might not be the best alternative for an SAP customer requiring hundreds or thousands of tape image copy data sets for recovery.
- **ATL**
  - If you are sending tapes offsite use the ATL instead of virtualized tape.
  - Extremely large data sets. Because of data transfer time, ATLs may be a better alternative than virtualized tape. Benchmark testing should be executed.
- **VTS/TS7700**
  - When it is imperative to have a second copy of a tape data set automatically created at an alternate site.
  - When stacking methods are not used or you do not have many files to create and you want an automated way of stacking your data instead of making them a multi file tape.

- **MYTH: It does not matter if my archive log data sets reside on tape or disk**
  
- **FACT: Archive logs on disk only take up one slot in the BSDS as they must be cataloged. Tape allocations can be cataloged or uncataloged, therefore each volser takes up one BSDS slot for multi-volume tapes if not cataloged and one slot total if cataloged – as with disk.**
  - An archive log data set residing on three disk volumes consumes one BSDS slot – only the first volser is recorded
  - An archive log data set residing on three tape volumes that is not cataloged consumes three BSDS slots – one for each volser
  - Review ZPARM MAXARCH, ARCRETN, and CATALOG (default NO) settings
  
- **MVS serializes access to all DB2 data sets on tape. Archive log data sets residing on tape can only be serialized, even when residing on the disk portion of the tape unit.**
  - Archive logs residing on disk can be parallelized

- **MYTH: When I see a tape mount, a physical tape is always mounted**
- **FACT: Seeing a take mount, unmount, request for additional tape volumes, etc. does not mean tapes are real**
  - Virtual tapes still issue the mount/unmount, etc. commands as you would see on a manual or ATL tape. You cannot tell strictly by the mounts you see in the JESLOG what type of tape you are dealing with without knowing UCB addresses, etc.
  - When the hosts requests a volume that is still in cache, the volume will be virtually mounted, no physical mount is required.
  - Access to data is at disk speed. Tape commands such as space, locate, rewind, and unload are mapped into disk commands that are completed in tens of milliseconds rather than the tens of seconds required for traditional tape commands.
  - Mounting of scratch tapes is also virtual and does not require a physical mount.
  - The TS7740 Node manages the physical tape drives and physical volumes in the tape library and controls the movement of data between physical and virtual volumes.
  - Multiple, different, emulated 3490E volumes can be accessed in parallel because they physically reside in the tape volume cache. **(A single virtual volume cannot be shared by different jobs or systems at the same time, even though the media is really disk.)**



- **MYTH: For virtual tapes, my Storage Administrator does not need to know how long my data set should reside on the disk vs. tape portion.**
  
- **FACT: The Storage Administrator can influence the length of time your data set resides in the TVC vs. physical tape**
  - Discuss your requirements with your Storage Administrator
  - Do your image copy data sets really need to reside in the TVC for a long time or can they quickly be moved to physical tape?
  - How about your archive log data sets, etc.?
  - TVC space is limited. Most data sets do not require long term placement of data in the TVC.

- **MYTH: My Storage Administrator knows how to separate my data on physical tape volumes**
  
- **FACT: Most Storage Administrators do not know how archive logs and image copies are used**
  - Even for the Storage Administrators that know what archive logs and image copies are used for, they do not typically know the naming convention you are using for dual pair data sets
  - If you have dual pair of archive logs and/or image copies on virtual tape, let the Storage Administrator know that they must ultimately reside on different physical tapes.
  - The Storage Administrator can pool specific data sets onto different physical tape volumes.
  - This recommendation goes along with duplicating tapes in case one snaps or has other media failure

- **MYTH: It is a good idea to compress a data set that will reside on disk, and then to compress it again on tape**
  
- **FACT: Although disk subsystems no longer supports compression, you can CPU compress certain data sets residing on disk**
  - For example, starting with DB2 9 NFM, you can compress the archive log data sets, potentially saving a significant amount of disk space
  - Consider not compressing the data set if it moves from disk to hardware compressed tape
    - Customers have run into problems compressing on disk and then trying to re-compress on tape
  - CPU starved customers need to weigh compression vs. CPU costs
  - Generally, it is inefficient to compress very small data sets

- **MYTH: BLKSIZE does not matter if my data set resides on disk or tape**
  
- **Fact: BLKSIZE can matter, especially for large data sets with higher blocking factors going to tape**
  - Some customers see as much as a 50% elapsed time reduction for the COPY utility when writing to tape using the LBI (Large Block Interface) with 256K
  - Archive log data sets are still capped at 28K
    - Best practices dictates that archive log data sets should be allocated at 24K. If archive log data sets are moved to disk for parallel recoveries, 24K would allow for two records per track, vs. 28K which would only allow one, thereby wasting about 40% disk space.
    - There is not a significant performance difference when using 24K block size on tape rather than 28K.

# Tape Issues

- **Tape is an excellent medium to place your DB2 related data on. Consider tape for your:**
  - Archive logs
  - Image copies
  - Very large Sortwork data sets
  - Other assorted data sets
- **There are some challenges using tape in a DB2 environment:**
  - VSAM data sets, PDSs and PDSEs must reside on disk
  - No concurrency or sharing within a data set or volume, therefore parallelism does not exist.
  - Tapes perform best using pure sequential access. Other access types may result in some performance degradation.
  - Data sets residing on tape cannot be striped
- **Discuss your tape and data requirements with your Storage Administrator. Understand your company's tape solutions and how to best manage your DB2 related data.**