IMS and Modern Storage Subsystems

Rich Lewis, Bob Magid, and Frank Ricchio
IBM

Information Management software

IMS and Modern Storage Subsystems

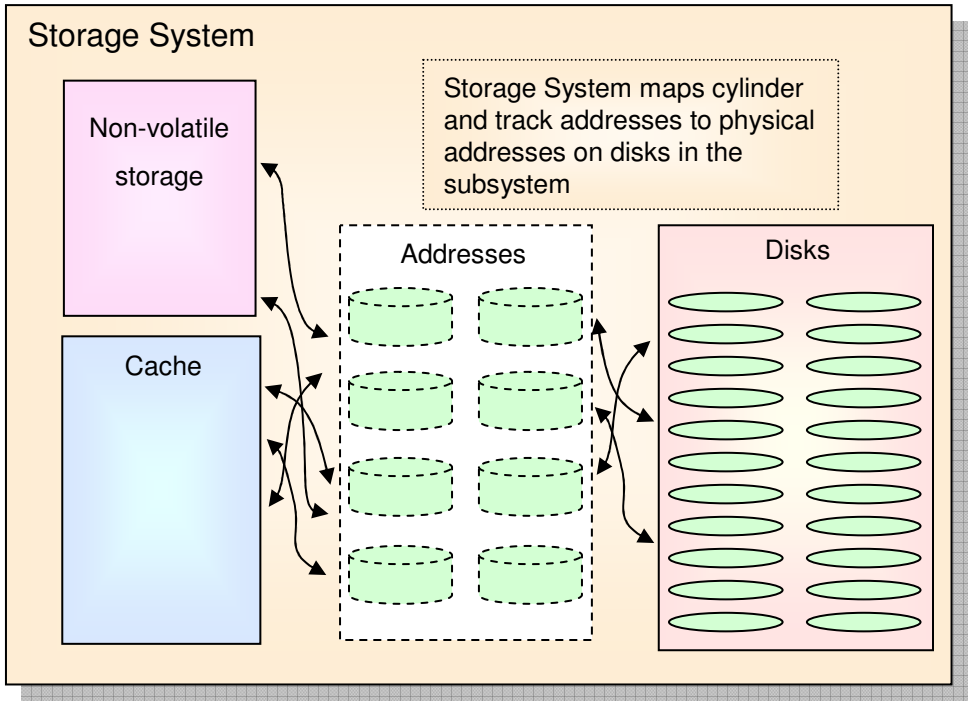
- Storage Systems
  - Storage systems overview
  - RAID architecture
  - Mirroring
  - Caching and cache algorithms
  - Solid State Devices
- Parallel Access Volumes and Multiple Allegiance
- IOS Queue time, Pend time, Connect time, and Disconnect time
- I/O times
- Caching and IMS buffer pools
- Disk volume sizes
- IMS 10 Large Data Set Support
- Data set allocation
- FlashCopy
- Extended Address Volumes (EAV)

Storage Systems

- IBM DS8300
  - 1.1 TB to 1024 TB capacity
  - Support for
    - FlashCopy®
    - Global Mirror
    - Metro Mirror
    - Metro/Global Mirror
    - Global Copy
  - High-capacity Serial ATA (SATA) drives
  - Solid-state drives
  - High Performance FICON® for System z®
  - Extended Address Volumes
  - HyperPAV
  - Extended Distance FICON
  - Cooperative Caching
  - Self encrypting disk drives

Storage Systems

- Address mapping
  - Cylinder and head (CCHH) addresses are mapped to physical disks in the system
- Cache
  - Large cache holds data
    - Recently referenced data
    - Prefetched data
- Non-volatile storage
  - Holds updates which have not yet been written to disks



RAID Architecture

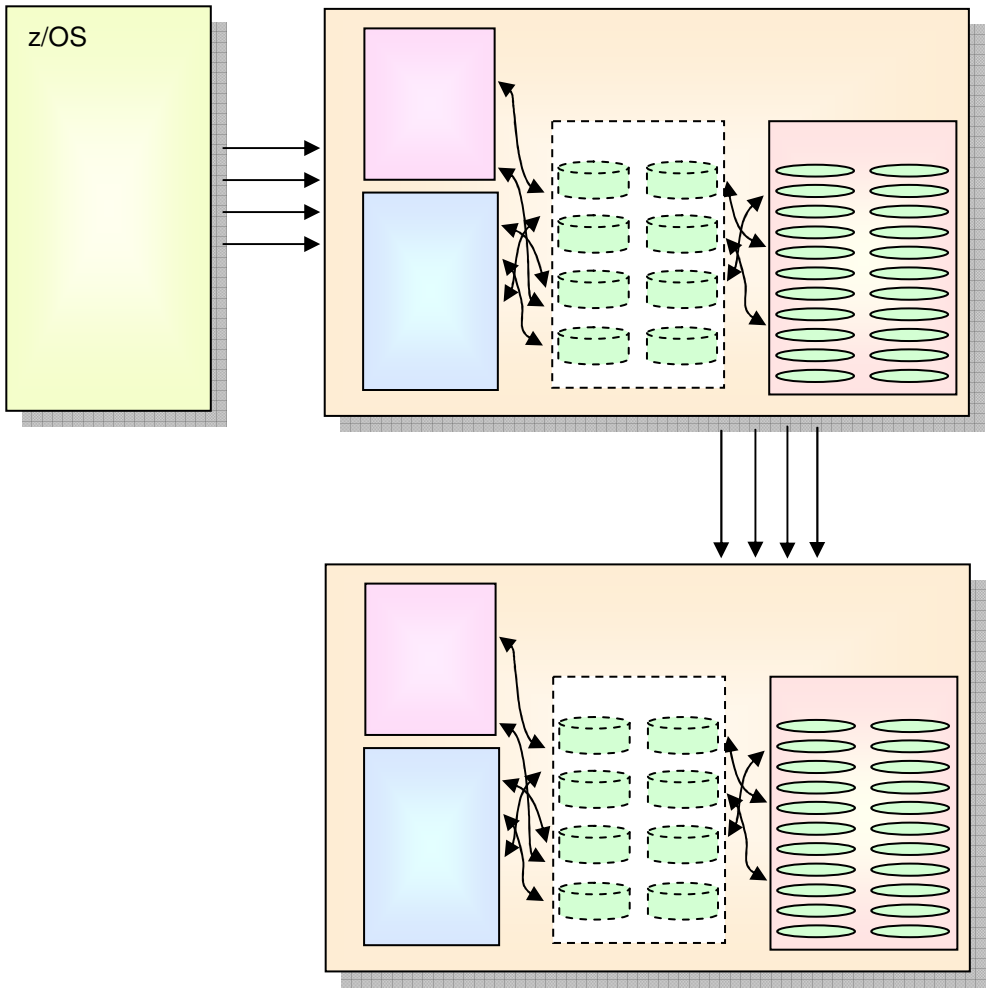
- Redundant Array of Independent (or inexpensive) Disks
  - Data for a track is written to multiple physical tracks and devices
  - Any device may be lost and the data can be rebuilt
- Good news
  - We very rarely lose individual volumes
    - For IMS this means we rarely do full database recoveries
    - Our database recoveries are timestamp recoveries for application or operational errors
- Bad news
  - If we lose data, we probably lose a whole storage system!

RAID Architecture

- All addresses are virtual
  - The address known to a z/OS system is virtual
    - CCHH is not a real cylinder and track
  - Data is mapped to some other address on some disk
    - The subsystem dynamically maps addresses to physical drives.
- Data placement is not physical placement
  - Placement of data sets in not so critical
  - You don't need to separate data sets to address "arm movement"

Storage Systems

- Mirroring (remote copy)
  - Data written to one storage system is then written to another storage system
  - Typically done for disaster recovery
- Synchronous mirroring
  - Writes to remote system must complete before host is notified of write completion
    - Writes are elongated
- Asynchronous mirroring
  - Writes to remote system may be done after host is notified of write completion
    - Writes are not elongated



Storage System Caching

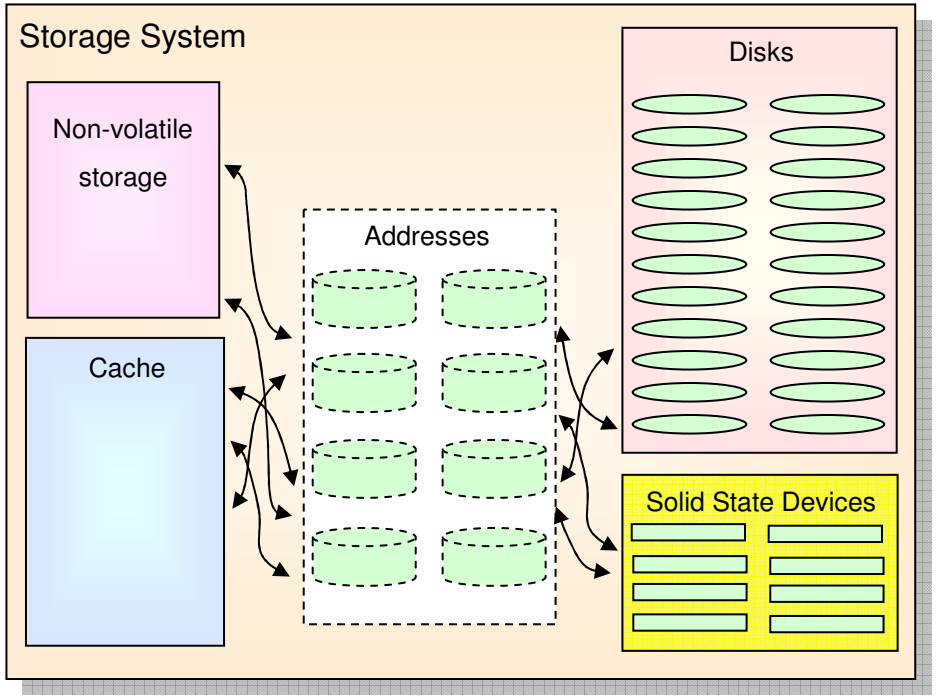
- Storage systems have very large caches
  - Up to 256GB in DS8300
  - Frequently accessed tracks are kept in cache
    - Like an IMS buffer pool
  - Anticipatory caching is done
    - Like OSAM sequential buffering
- Sophisticated caching algorithms are used
  - May stage block, partial track, full track, cylinder, or next cylinder
  - Adaptively monitors and adjusts caching
    - Not just a simple "least recently used" algorithm

Storage System Caching

- Caching is used to avoid disk reads
  - Data is kept to satisfy reads from z/OS
- Non-volatile storage is used to shorten write times
  - Data does not have to be written to disk before "end of write" signal is sent to z/OS
    - Write to disk is done asynchronously
- Caching reduces DISCONNECT times
  - Avoids the access to the actual physical disks

Solid State Devices (SSDs)

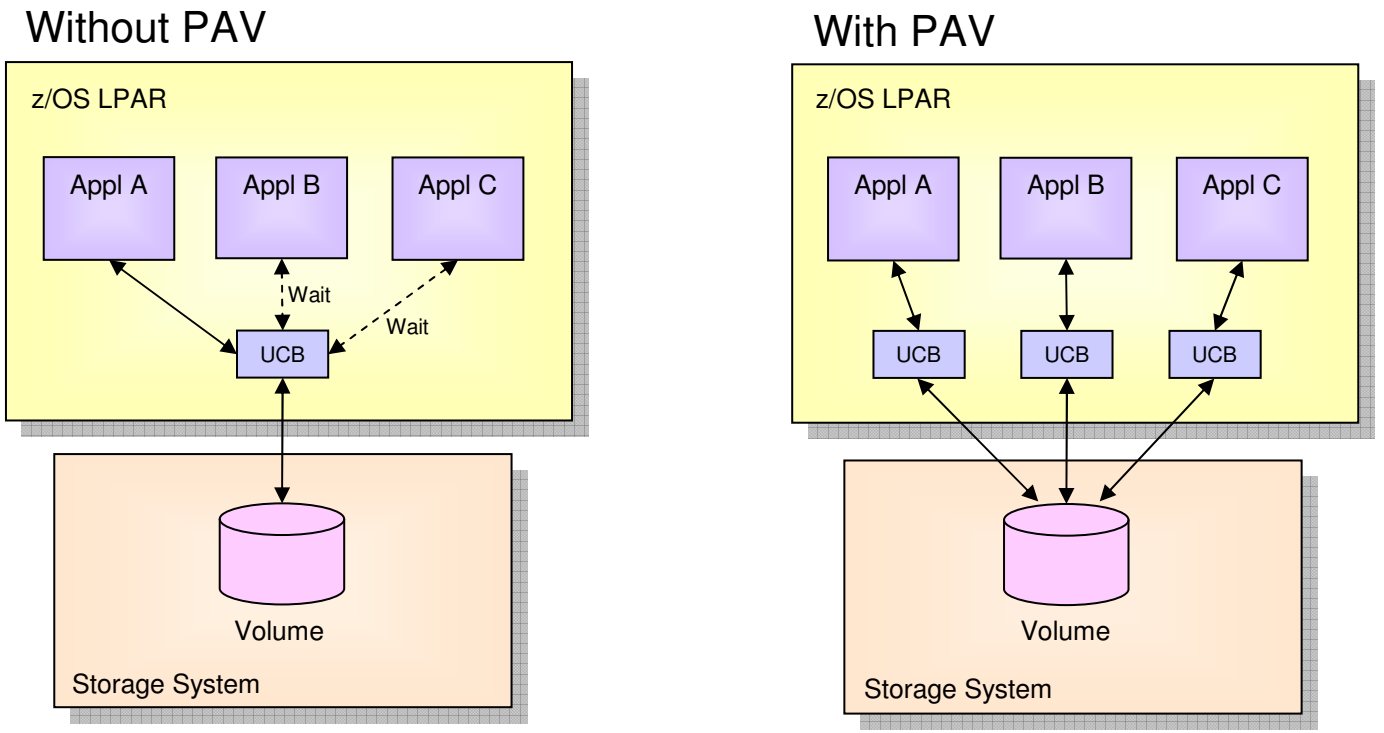
- DS8300 now supports SSDs
  - Alternative to disks
  - System has mix of disks and SSDs
- SSD properties
  - Flash technology
  - Faster than disks
    - No seeks
    - No rotational delays
  - More expensive than disks
- DFSMS is aware of SSDs
  - May be assigned to storage pools
  - Used by data classes
- Enhanced reporting
  - SMF 42-6 record includes Data Set Read-Only Disconnect Time
  - Tool to report good candidates using these SMF records



- SSDs reduce DISCONNECT times
  - No seeks
  - No rotational delay

PAV (Parallel Access Volumes)

- Multiple UCBs (unit control blocks) per volume
  - Allows concurrent I/Os from the same z/OS system

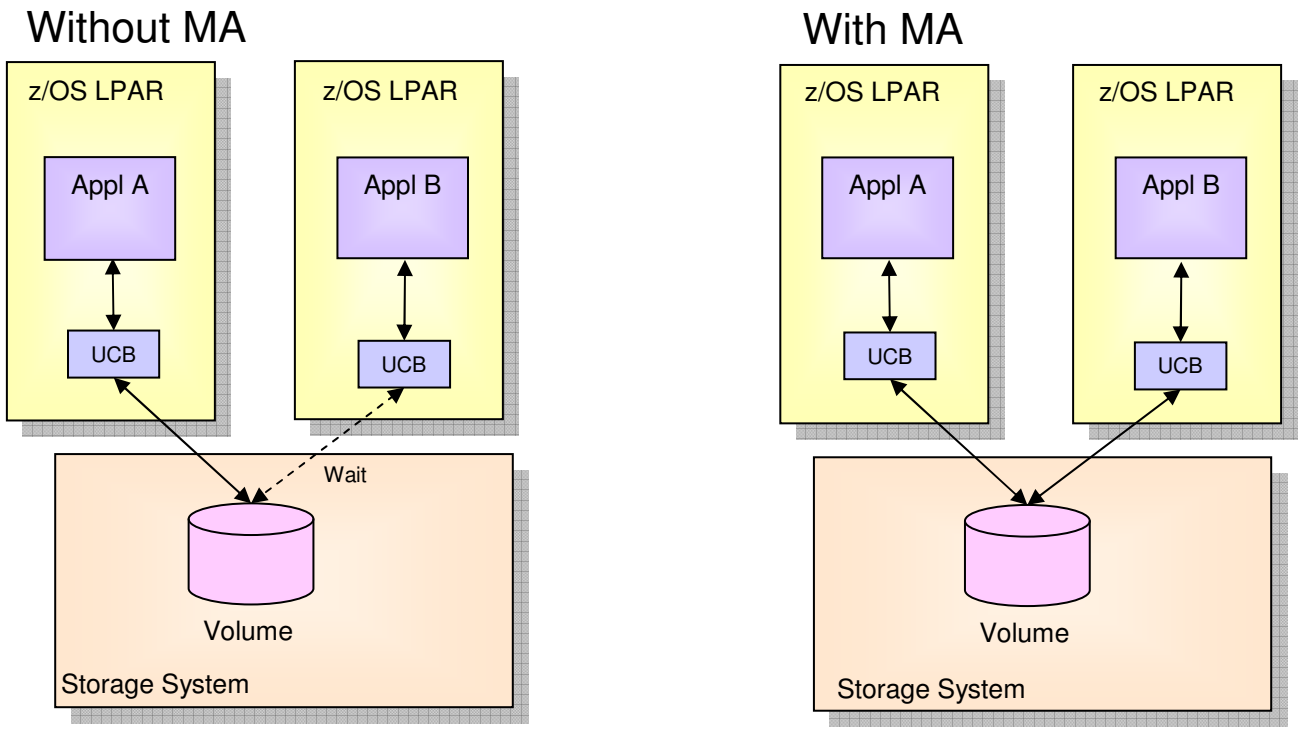


PAV (Parallel Access Volumes)

- Multiple UCBs per volume
  - PAVs allow simultaneous access to volumes by I/Os from one system
    - Reads are simultaneous
    - Writes to different domains are simultaneous
    - Writes to same domain are serialized
- High I/O activity to a volume benefits from PAV
- Static PAVs (fixed number of UCBs per volume)
- Dynamic PAVs adjusts the number of UCBs per volume
  - Requires WLM goal mode and coordination across a sysplex
- Hyper PAVs with DS8000
  - Does not use WLM or require coordination across a sysplex
- PAV sharply reduces IOSQ time
  - IOSQ time is time waiting for a UCB

Multiple Allegiance

- Multiple Allegiance allows concurrent I/Os from different LPARs



Multiple Allegiance

- Multiple Allegiance
  - Compatible I/Os (no extent conflict) from different LPARs may run in parallel
  - Incompatible I/Os are queued in the storage system
  - Storage system guarantees data integrity

- Multiple Allegiance reduces PEND time (device busy)

PAV and MA Extent Conflicts

- Reads
  - There is no restriction on concurrent reads using PAV and/or MA
  - Multiple reads of the same data may occur concurrently
- Writes
  - There are some restrictions on concurrent writes or writes with reads
  - A write cannot occur concurrently with a read or another write to the same *domain*
    - The *domain* is set by the DEFINE EXTENT in the channel program
    - Typically, for IMS the *domain* is an allocation extent for a data set

Elements of I/O Times

- Storage administrators use RMF DASD Reports to analyze I/O response times
  - IOSQ time - Device is busy in this z/OS, UCB not available
    - Time waiting for the device availability in the z/OS operating system.
  - PEND time - Device is reserved by another system
    - Time from the SSCH instruction (issued by z/OS) till the start of the dialog between the channel and the I/O controller.
  - DISCONNECT time - Data not being transferred
    - Time when the I/O operation is already started but the channel and I/O controller are not in a dialog.
      - *Read cache miss, sync remote copy, PAV and MA write extent conflicts, …*
  - CONNECT time - Channel data and protocol transfer
    - Time when the channel is transferring data from or to the controller cache or exchanging control information with the controller about an I/O operation.

I/O Response Time = IOSQ time + PEND time + DISCONNECT time + CONNECT time

I/O Times

- Typical I/O times[1]

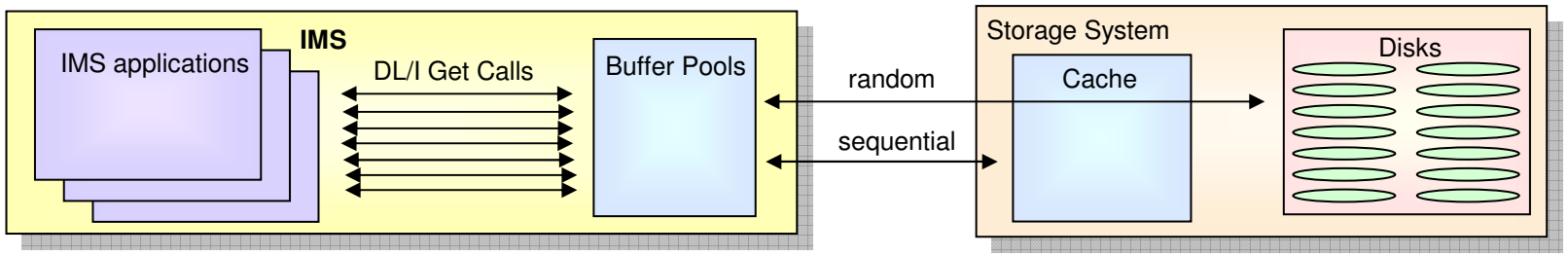
I/O type	3390	DS8300 Disk	DS8300 SSD
4K sequential I/O[2]	1.6 to 2 ms	0.035 to 0.060 ms (in DS8300 cache)	0.035 to 0.060 ms (in DS8300 cache)
8K sequential I/O[2]	3 to 4 ms	0.070 to 0.12 ms (in DS8300 cache)	0.070 to 0.12 ms (in DS8300 cache)
4K or 8K random read	20 ms	5 to 10 ms (not in DS8300 cache)	~0.75 ms (not in DS8300 cache)

[1]These are non-scientific observations from anecdotal evidence.
[2]Includes random reads with cache hit in DS8300

- Caching provides dramatically improved I/O times
- SSDs can be used to improve random reads
  - If today there are buffer and cache misses
  - If number or importance of I/Os justifies costs

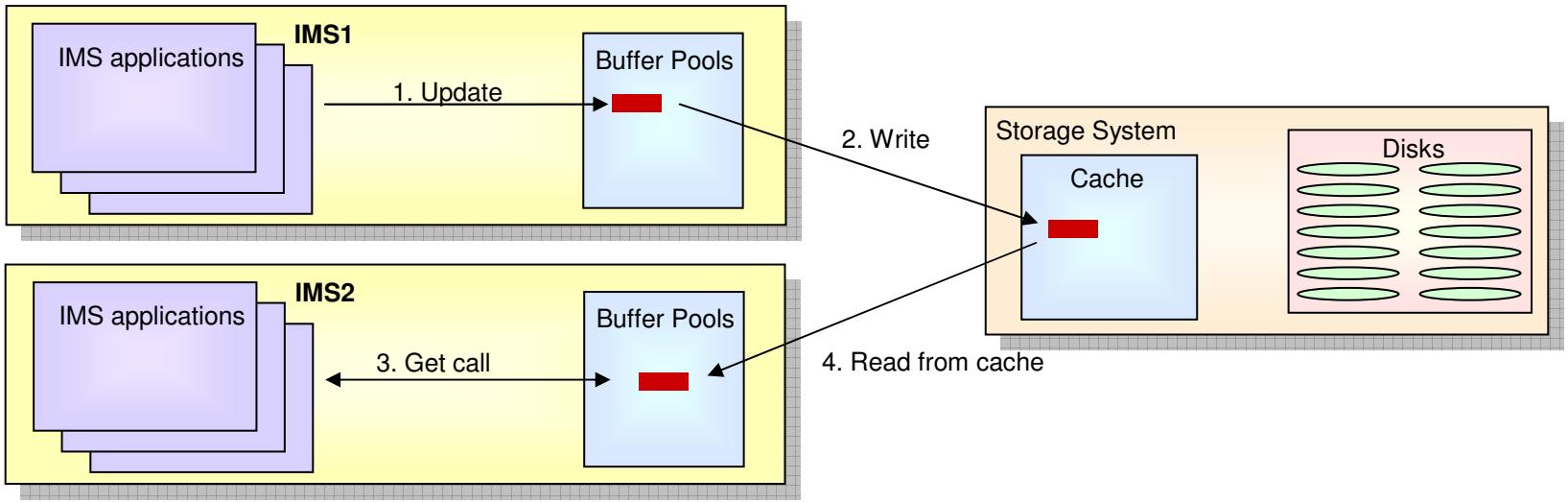
Caching and IMS Buffer Pools

- IMS full function buffer pools
  - Keep frequently referenced data in the pools
    - Avoids reads of recently referenced blocks/CIs
- Caching
  - Keeps frequently read data in cache
  - May prefetch blocks for sequential processing
- IMS data references
  - Not many hits in cache for reads
    - IMS only asks for infrequently referenced data
    - May get hit for sequential processing



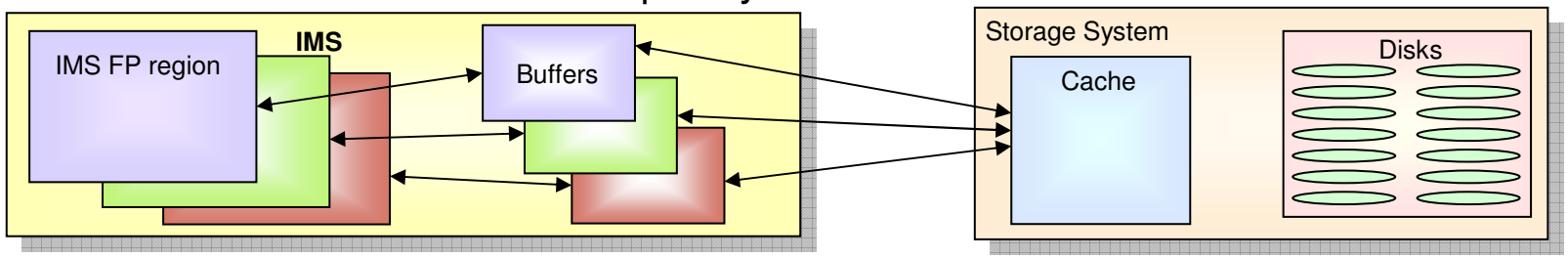
Caching, Buffer Pools, and Data Sharing

- Data sharing
  - Buffers may be invalidated by updates on another IMS
  - If data is needed again, IMS reads it from storage system
    - Or from cache structure in Coupling Facility
  - This data may be in storage system cache
    - Written by the updating IMS system
  - Data sharing systems may benefit from cache hits

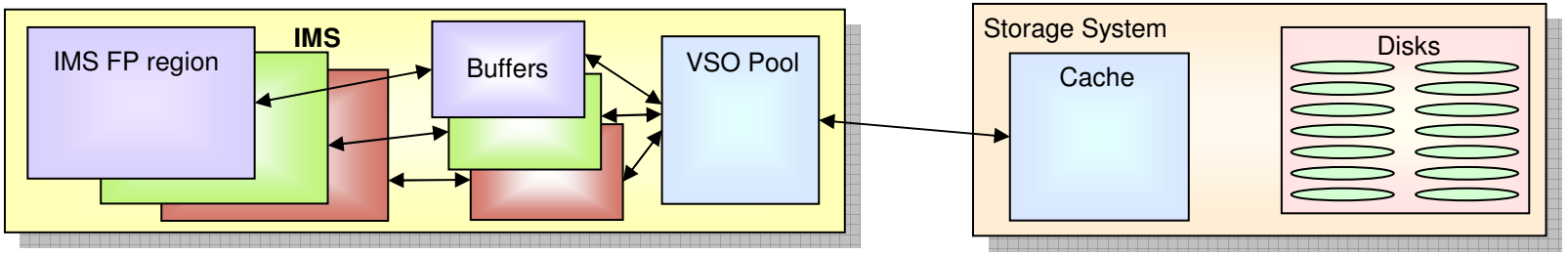


Caching and IMS Buffer Pools

- IMS Fast Path buffer pools
  - Without VSO,
    - Data is not kept in pool for successive transactions
    - Each region has its own buffers
    - Reads are required
    - Cache hits will occur for frequently referenced data



  - With VSO
    - Frequently referenced and preloaded data are kept in VSO pool
    - Not many hits in cache for reads



Disk Volume Sizes

- Storage systems use 3390 emulation
  - Track size is 56,664 potential bytes per track
    - Actual capacity depends on block sizes
- Cylinders per volume can be anything
  - 3390-1, 3390-2, 3390-3, and 3390-9
  - or
  - Custom volumes of any size
    - May be used to overcome difficulties with multi-volume data sets
    - Warning: access method may limit usefulness of large volumes for a single data set
- More than 65,520 cylinders per volume require EAV support
  - EAV - extended address volume

Disk Volumes Capacities

Model	Cylinders	Tracks	Bytes/Volume*	Bytes/Track*
3390-1	1113	16695	946,005,480	56664
3390-2	2226	33390	1,892,010,960	56664
3390-3	3339	50085	2,838,016,440	56664
3390-9	10017	150255	8,514,049,320	56664
3390-27**	32760	491400	27,844,689,600	56664
3390-54**	65520	982800	55,689,379,200	56664

*Bytes/volume and bytes/track are potential capacities.

**These are custom volumes.

Warning:

Disk people use	Software people use
GB = 1,000,000,000	GB = 1,073,741,824

- Block sizes determine actual capacities

Block size	2K	4K	8K	16K	24K	26K	32K
Bytes/track	42K	48K	48K	48K	48K	52K	32K
3390-9 capacity*	6.02 GB	6.88 GB	6.88 GB	6.88 GB	6.88 GB	7.45 GB	4.56 GB

*GB=1,073,741,824

Large Sequential Data Sets

- Before z/OS 1.7, sequential data sets were limited to 65,536 tracks per volume
  - Only 3GB with 4K, 8K, or 12K blocks
  - Data sets larger than 3GB required multiple volumes
- z/OS 1.7 DFSMS adds support for large sequential data sets
  - More than 65,535 tracks on one volume for a data set
  - Requires DSNTYPE=LARGE
- IMS 10 adds large sequential data set support
  - OSAM and GSAM/BSAM data sets
    - Includes
      - *OSAM database data sets*
      - *Logs*
      - *Trace data sets*
      - *Message queue data sets*
      - *GSAM/BSAM files*
      - *Image Copy and Change Accum data sets*

OSAM data set size is limited to 8GB

- 8GB data set with 4K or 8K block size requires 174,763 tracks = 11,651 cylinders

Allocation of Data Sets

- VSAM
  - Not SMS managed:
    - 255 extents per component
    - 123 extents per volume
  - SMS managed:
    - 255 extents per component
      - *More if extent constraint removal is specified in the data class*
    - Adjacent extents are consolidated
    - 123 extents per volume
  - Any number of tracks per volume
  - IMS limits VSAM database data sets to 4GB

Allocation of Data Sets

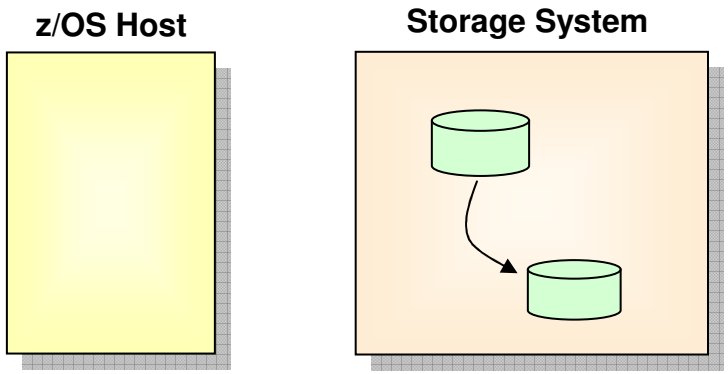
- OSAM
  - IMS 10 with DSNTYPE=LARGE
    - Any number of tracks per volume
  - IMS 8, IMS 9, or IMS 10 without DSNTYPE=LARGE
    - Maximum of 65,535 tracks per volume
  - Database data sets
    - Maximum of 60 extents per data set
    - Maximum of 16 extents per volume
    - Maximum of 59 volumes per data set
    - IMS limits OSAM non-HALDB database data sets to 8GB
    - IMS limits OSAM HALDB database data sets to 4GB
  - Message queue data sets
    - One volume without secondary extents
  - OLDS data sets
    - One volume without secondary extents

FlashCopy

- FlashCopy is a storage system capability to copy volumes or data sets
  - Copy is created in the same storage system
  - Almost instantaneous
    - Done by creating another "map" to the actual data on disks
  - Subsequent updates are written to different locations for the two data sets or volumes
    - New and old "maps" are used
- IMS 10 Image Copy 2 supports data set FlashCopy
  - Requires z/OS 1.8
  - IMS 10 Database Recovery can use these image copies
- IBM High Performance Image Copy supports data set FlashCopy
  - Does not require z/OS 1.8
    - Restriction: fuzzy copies of KSDSs are not allowed

FlashCopy vs. Concurrent Copy

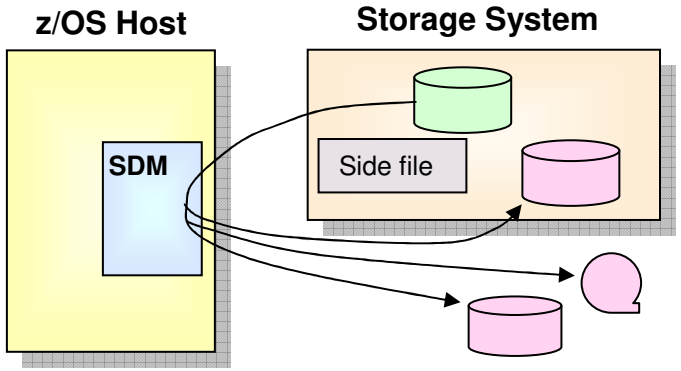
- FlashCopy is a storage system function
  - Copy must be to the same system
  - Single phase copy
  - Creates exact copy of data set
  - Invoked by IMS 10 IC2 or HPIC



FlashCopy vs. Concurrent Copy

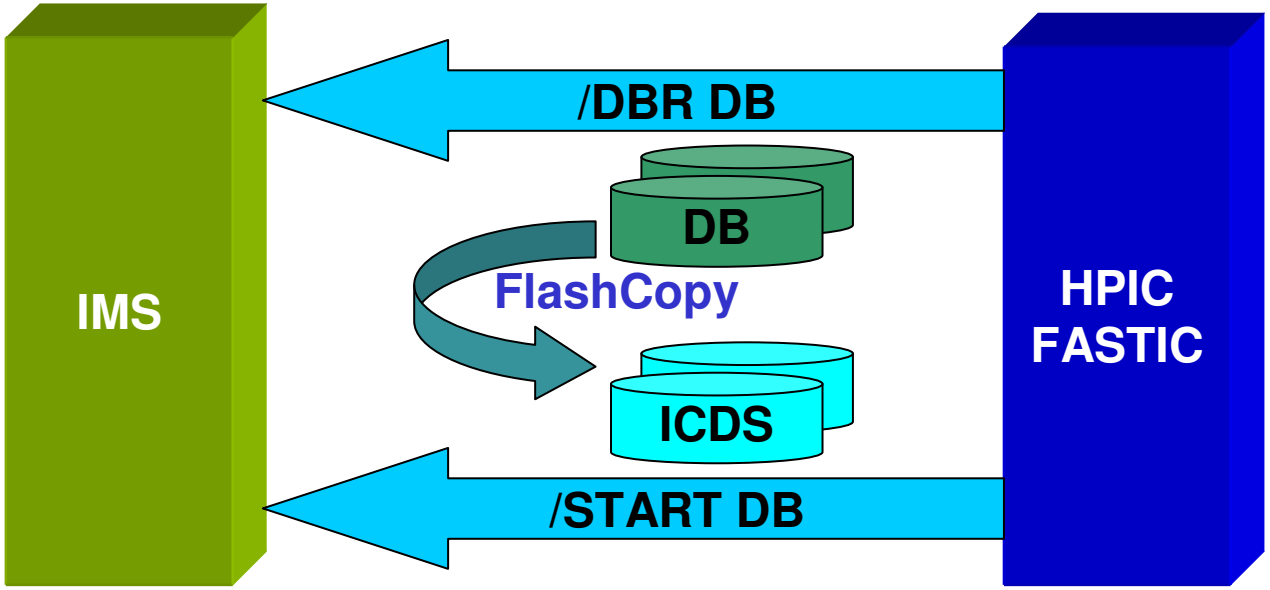
- Concurrent Copy uses storage system and the System Data Mover (SDM)
  - Copy may be to another system, including tape or disk
  - Two phase copy
    - Logical copy - sets up side file for updates during physical copy (very quick)
    - Physical copy - writes data using SDM and host resources
  - Copy format:
    - Exact copy (invoked by HPIC, but not by IMS IC2)

    or
    - In DUMP format (or in user format by using an exit routine)
  - Invoked by IMS 9, 10, or 11 IC2 or HPIC



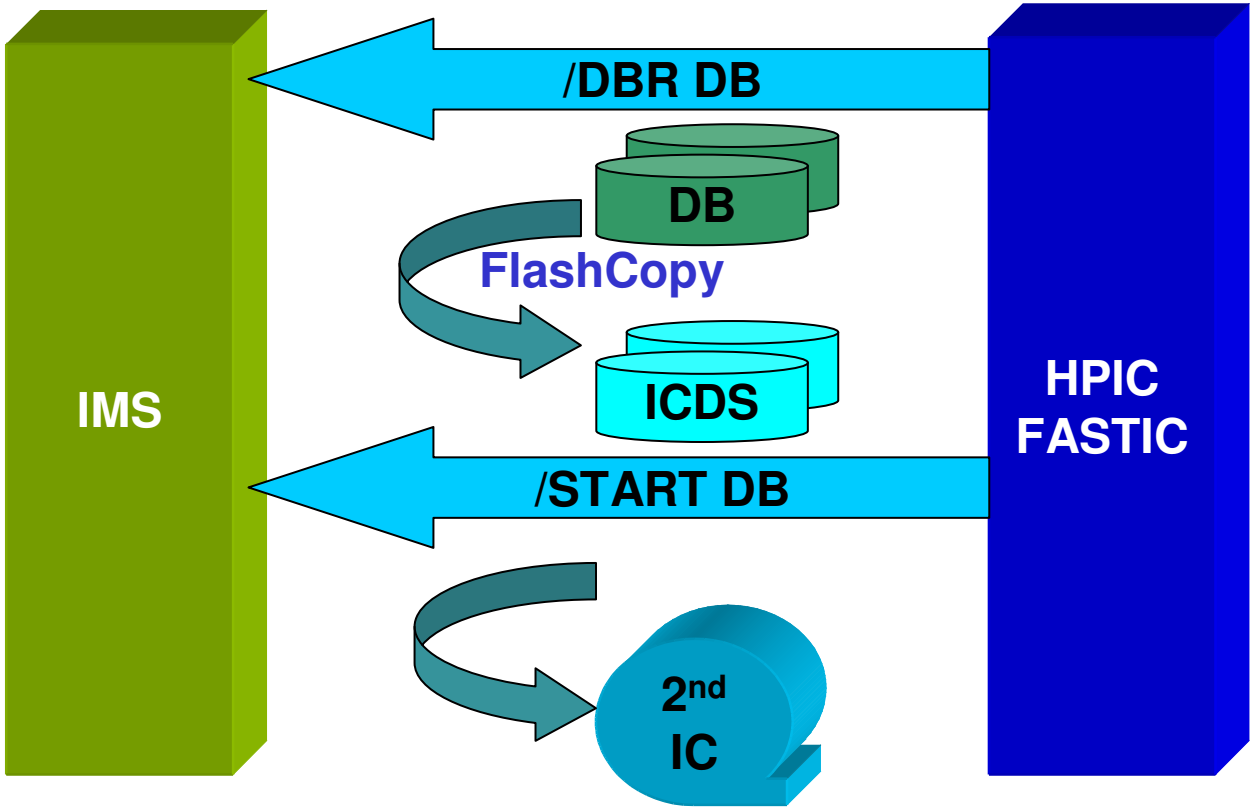
HPIC FlashCopy

- FlashCopy IC is available with IMS V9 or later
  - IC record in DBRC is SMSNOCIC/SMSCIC in IMS V9
  - DB can be recovered with HPIC or DRF
- TOIAUTO=Y
  - DB is deallocated from IMS by /DBR or /DBD command before taking image copy, and
  - DB is reallocated to IMS immediately after IC is created.



HPIC FlashCopy

- You can take dual IC data sets with HPIC. Secondary IC can be FlashCopy or Batch Format which can be written to tape
  - If you want to recover DB from secondary IC, you need to change IC type in DBRC to Batch or CIC.



Extended Address Volumes

Problem being addressed

- The continued rapid data growth on the z/OS platform is leading to a critical problem for some of our customers
  - They are are running out of z/OS addressable disk storage

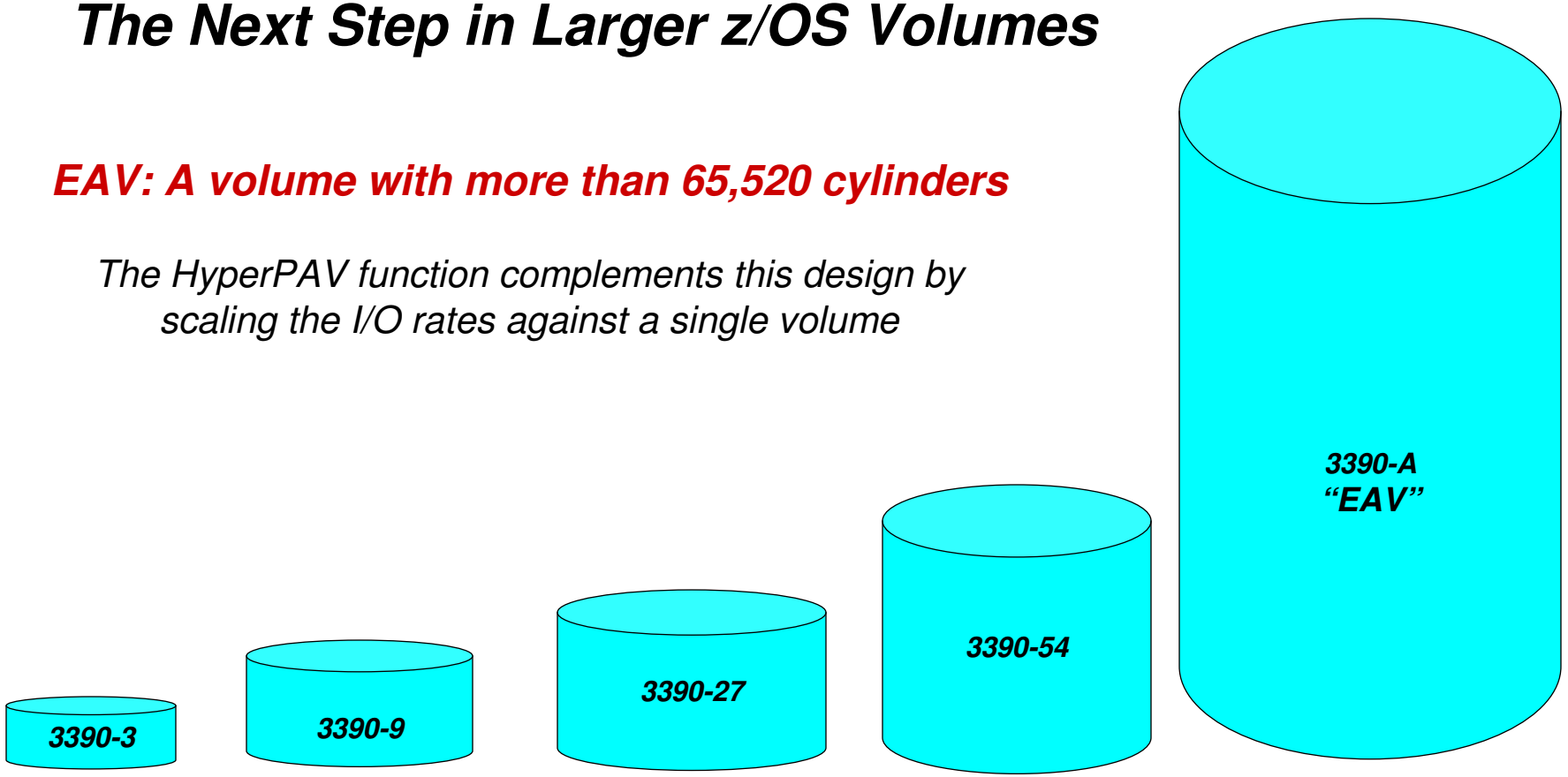
- IBM's solution to this problem is to:
  - Continue the direction started with the 3390-9 (3390 Model 9) of defining larger volumes by increasing the number of cylinders
    - Extend the number of cylinders per device to be >65,520

Extended Address Volume (EAV)

The Next Step in Larger z/OS Volumes

EAV: A volume with more than 65,520 cylinders

The HyperPAV function complements this design by scaling the I/O rates against a single volume

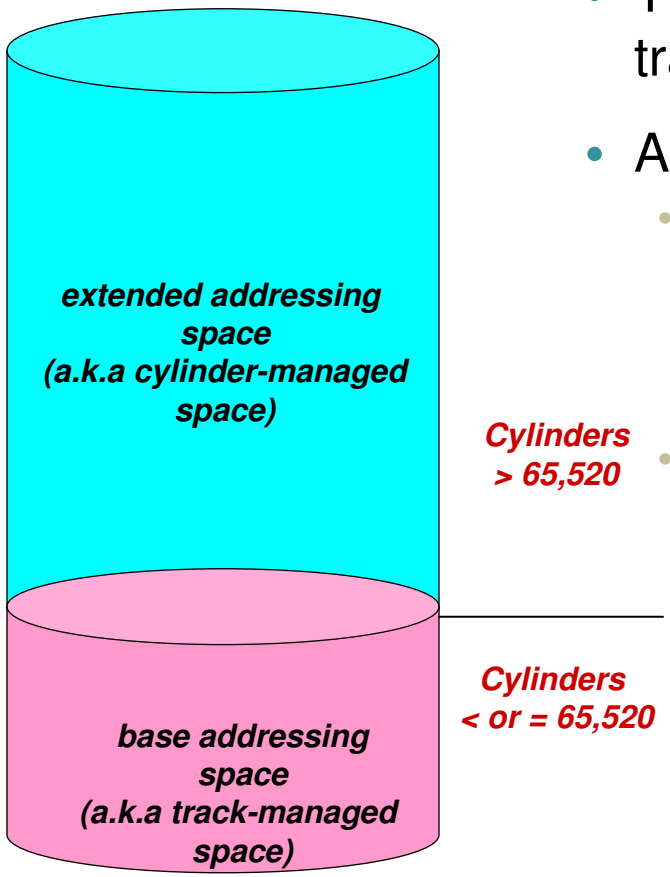


3 GB	9 GB	27 GB	54 GB	221 TBs
Max cyls: 3,339	Max cyls: 10,017	Max cyls: 32,760	Max cyls: 65,520	Max cyls: 268,434,453

Maximum Sizes

EAV Release 1 may limit the size to 236 GB (Max cyls 262,668) to reduce testing effort.

Extended Address Volume (EAV) – Key Design Points

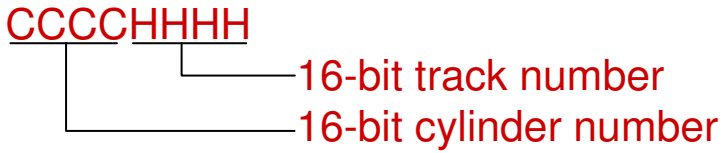


EAV

- The EAV volume format will maintain the 3390-9 track size, track image, and tracks per cylinder
- An EAV is partitioned into two regions:
  - **Base addressing space** (aka **track-managed space):** the area on an EAV located within the first 65,520 cylinders.
    - Space in **track-managed space** is allocated in track or cylinder increments (same as today)
  - **Extended addressing space** (aka **cylinder-managed space):** the area on an EAV located above the first 65,520 cylinders.
    - Space in **cylinder-managed space** is allocated in **multicylinder units**.
      - *A fixed unit of disk space that is larger than a cylinder. Currently, on an EAV, a multicylinder unit is 21 cylinders*
- New track address format for space on an EAV

Existing Track Addresses

- Track addresses
  - Existing track address format with 16-bit cylinder number



    - Today's supported maximum size volume is 65,520 cylinders, near the 16-bit theoretical limit of 65535

  - To handle cylinder numbers greater than 65,520, a new format for the track address is required

Track Addresses - New Format

- 28-bit cylinder numbers
  - Planned new cylinder-track address format

    CCCCcccH

    H is 4-t track number (0-14)

    ccc is high order 12 bits of 28-bit cylinder number

    CCCC is low order 16 bits of 28-bit cylinder number
  - Number
    - The cylinder number is in a non-linear form
    - This format preserves the 3390 track geometry
    - Track addresses for space in **track-managed space** will be compatible to today's track addresses
    - Track addresses for space in **cylinder-managed space** will NOT be compatible to today's track addresses
  - Normalized cylinder-track address (to be used only for printing)

    cccCCCC:H

    H is 4-bit track number (0-14)

    cccCCCC is 28-bit cylinder number in a linear form
    - The colon or other special character shows that it is a normalized address

IMS Support for DS8000 EAV Volumes

- Allows IMS customers to exploit the EAV available in z/OS 1.10.
  - This support allows IMS VSAM datasets to reside on these large volumes.
    - Datasets include DEDB databases, full function VSAM databases and RECON datasets
- Provides relief to customers running out of z/OS addressable disk storage due to the four-digit device number limit (actually 65,280 devices)
- Does not change the existing size limits on IMS databases
  - VSAM - 5G
  - OSAM - 8G
- Available in IMS V10 & V9 through the service process
  - V10 - PK72530 (PTF – UK43020)
  - V9 - PK72529 (PTF – UK43019)

IMS Tools Support for DS8000 EAV Volumes

- Available through the service process
  - IMS HP Pointer Checker 310 PK73041 (PTF-UK46126)
  - IMS HP Pointer Checker 310 PK88523 (PTF-UK48147)
  - IMS HP Load 210 PK84282 (PTF-UK46572)

Summary

- Modern storage systems
  - Virtualization
  - Availability
  - Performance
- Functions
  - Mirroring
  - FlashCopy
- IMS interfaces and uses
  - Caching and pools
  - Data set allocations