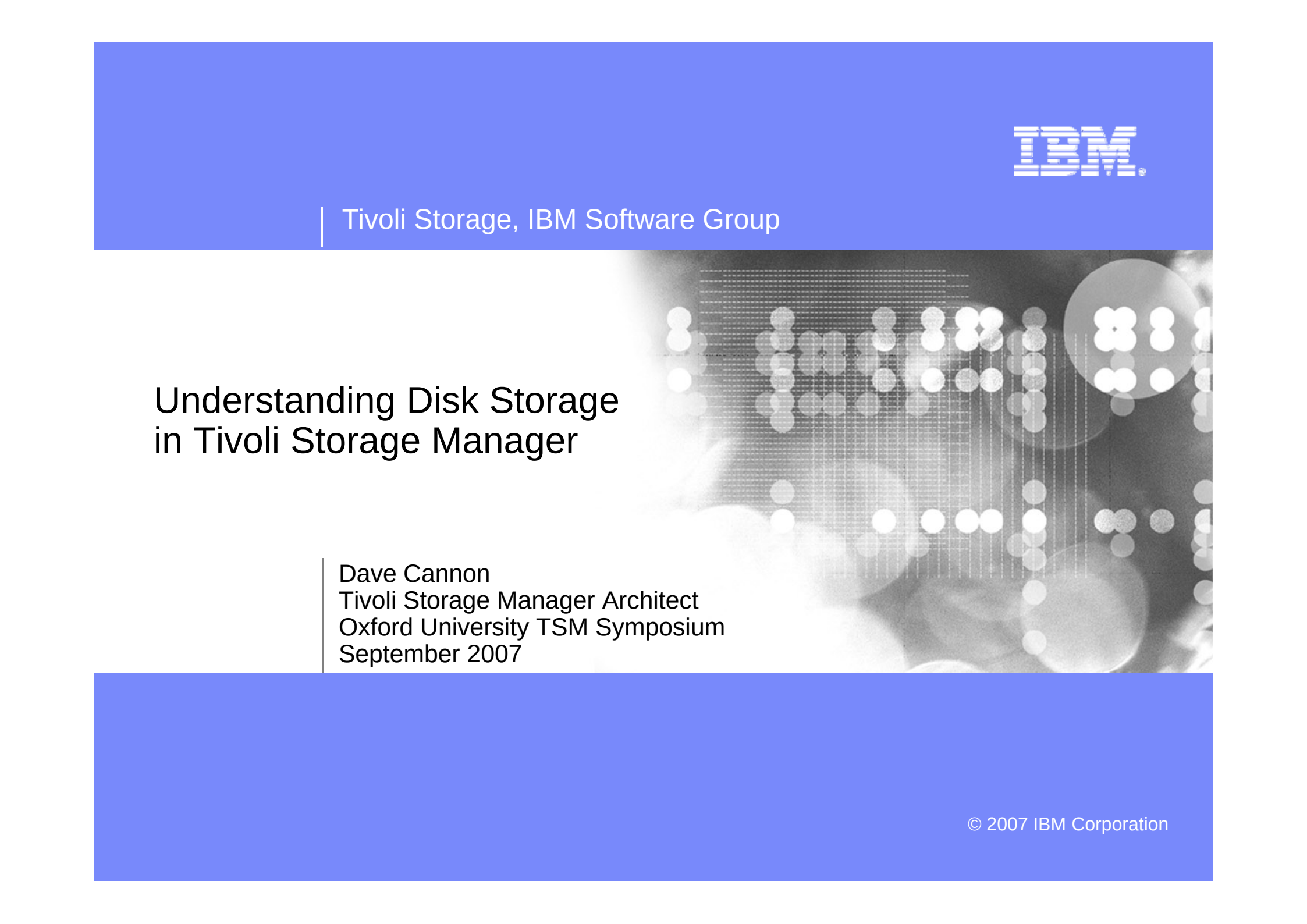


Understanding Disk Storage in Tivoli Storage Manager

Dave Cannon
Tivoli Storage Manager Architect
Oxford University TSM Symposium
September 2007

Disclaimer

- This presentation describes potential future enhancements to the IBM Tivoli Storage Manager family of products
- All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only
- Information in this presentation does not constitute a commitment to deliver the described enhancements or to do so in a particular timeframe
- IBM reserves the right to change product plans, features, and delivery schedules according to business needs and requirements
- This presentation uses the following designations regarding availability of potential product enhancements
  - Planned 5.5: Planned for delivery in TSM v5.5 (2007)
  - Next Release Candidate: Candidate for delivery in the next release after v5.5
  - Future Candidate: Candidate for delivery in future release

Agenda

- Background
- Random- and sequential-access disk in TSM
- Special considerations for sequential-access disk
- Potential future enhancements for disk storage

Disk vs. Tape

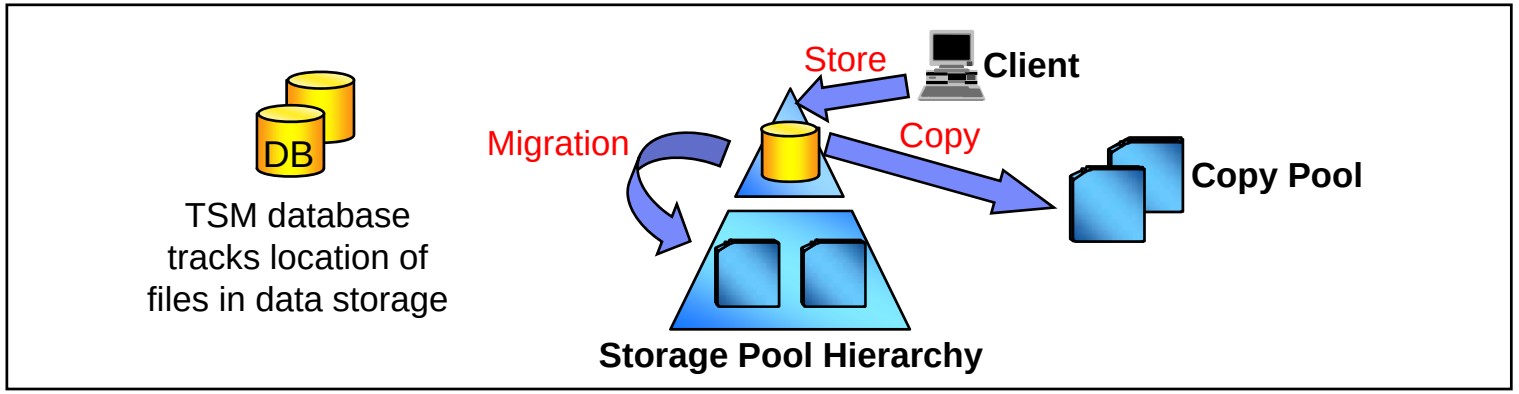
- Potential disk advantages
  - Faster access by avoiding delays for tape mounts and positioning
  - Reduced management cost (no tape handling)
  - Avoidance of errors introduced by media handling
- Potential tape advantages
  - Removability/portability for offsite storage (disaster recovery)
  - High-speed data transfer for large objects
  - Cost effectiveness (especially for long-term, offsite archiving)
- Tiered approach, with copies on offsite tape, exploits strengths of disk and tape

Industry Trend Toward Increasing Use of Disk

- Lower cost of disk storage (SATA)
- Promotion of disk-based appliances and solutions
- Virtual tape library (VTL) products comprised of preconfigured disk systems that emulate tape
- Disk-based technologies
  - Replication
  - Snapshots (point-in-time copies)
  - Continuous data protection (CDP)
  - Deduplication

TSM is Designed for Disk in a Storage Hierarchy

- Disk has been an integral part of the TSM data storage hierarchy since 1993
- Virtualization of disk volumes in a storage pool allows objects to be stored across multiple volumes and file systems
- Policy-based provisioning of disk storage pool space and allocation of that space during store operations
- Retention based on object-level policies rather than the tape used to store objects
- Automatic, policy-based migration to tape or other media types in tiered hierarchy
- Incremental backup of objects from primary disk pool to tape copy pool for availability or offsite vaulting
- Objects automatically accessed in copy pool if not available in primary storage pool



Disk Usage Trend in TSM

Traditional Disk Usage

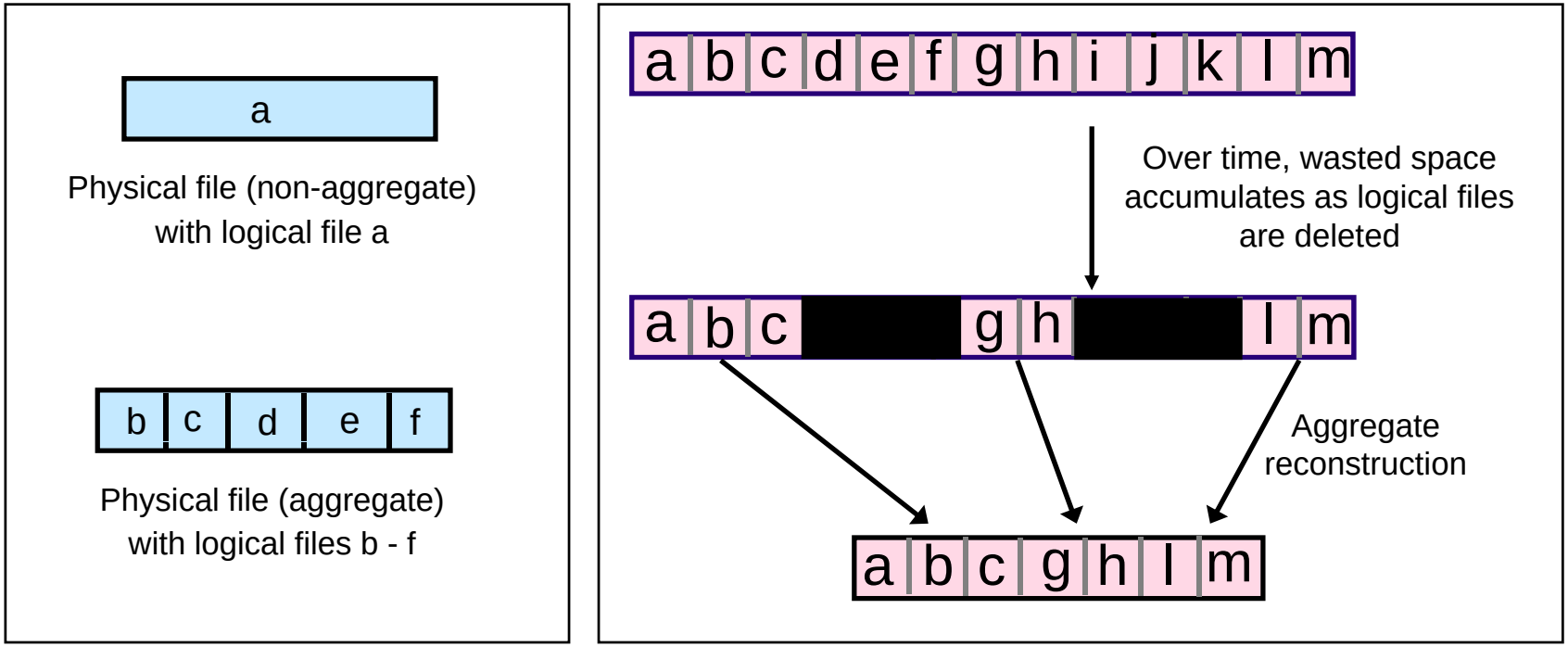
- LAN-based data transfer between client and disk storage
- Data initially buffered on disk to allow concurrent client backups without tape delays
- Backup from disk to tape copy storage pool for availability and disaster recovery
- Most data migrated to tape within 24 hours

Emerging Disk Usage

- Growing interest in LAN-free transfer between client and disk
- Data stored on disk to allow concurrent client backups without tape delays
- Backup from disk to tape copy storage pool (may be principal use for tape)
- Data may be stored on disk indefinitely for faster access

A Detour on File Aggregation

- TSM server groups client objects into aggregates during backup or archive
- Information about individual client objects is maintained and used for certain operations (e.g., deletion, retrieval)
- For internal data transfer operations (migration, storage pool backup), entire aggregate is processed as a single entity for greatly improved performance



Agenda

- Background
- Random- and sequential-access disk in TSM
- Special considerations for sequential-access disk
- Potential future enhancements for disk storage

Overview of Random- and Sequential-Access Disk

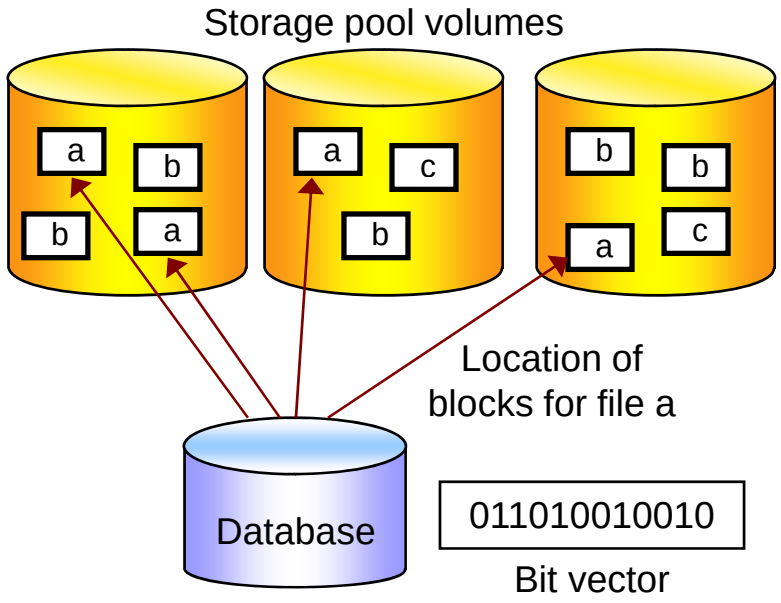
- TSM supports two methods for storing and accessing data on magnetic disk
  - Random-access storage pools (also known as DISK pools)
  - Sequential-access storage pools (also known as FILE pools)
- Random- and sequential-access disk pools differ in how TSM manages disk storage and the operations that are supported
- TSM development views sequential-access disk as strategic
  - Current functions on random-access disk supported for the foreseeable future
  - Future product enhancements involving disk storage may be offered only for sequential-access disk

Basics of Random- and Sequential-Access Disk

	Random-Access Disk	**Sequential-Access Disk**
Storage pool definition	Predefined device class DISK	Device class with device type of FILE
Pools spanning file systems	Supported	Supported
Storage pool volumes	Files or raw logical volumes	Files
Volume creation	▪Define Volume command ▪Space trigger	▪Define Volume command ▪Space trigger ▪Scratch volumes
TSM caching	Supported	Not supported
Collocation	Not applicable	Collocation by filespace, node or group of nodes
Use for copy storage pool	Not supported	Supported

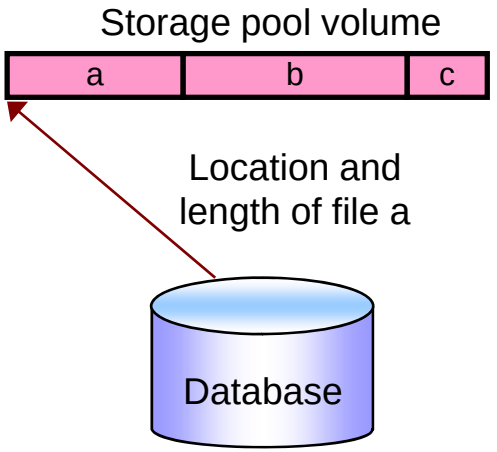
Space Allocation and Tracking

Random Access



- Space allocated in randomly located 4KB blocks
- TSM server tracks volumes and blocks on which each file is stored
- Bit vector in TSM database tracks allocated and free blocks for each volume
- Space allocation and tracking requires overhead
- May not scale well for extremely large files

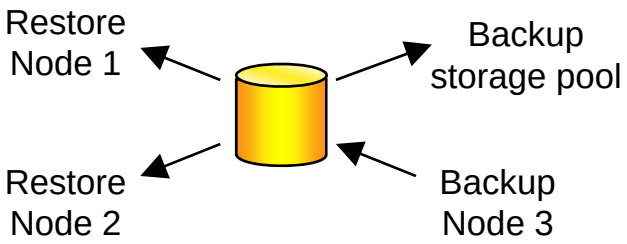
Sequential Access +



- Files written sequentially in FILE volume
- TSM database only tracks volume and offset at which each file is stored
- TSM has less overhead for space allocation and tracking

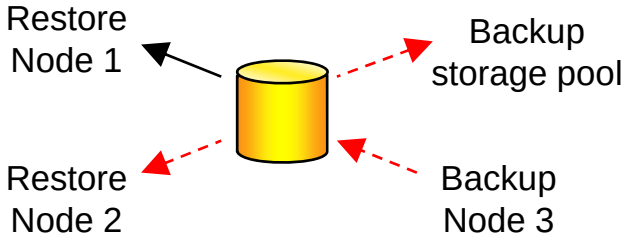
Concurrent Volume Access

Random Access + (based on 5.4 behavior)



- Multiple TSM sessions or processes can concurrently use the same disk volume
- However, individual I/O operations for each volume are serialized

Sequential Access



- Disk volume is locked by a single process or session using that volume
- Other operations cannot access the volume until the lock is released, usually when the locking operation has completed all work on the volume
- **Concurrent access (multiple read operations, one write operation) planned for 5.5**

To avoid volume contention, smaller volume sizes should be used for sequential-access disk as compared to random-access disk

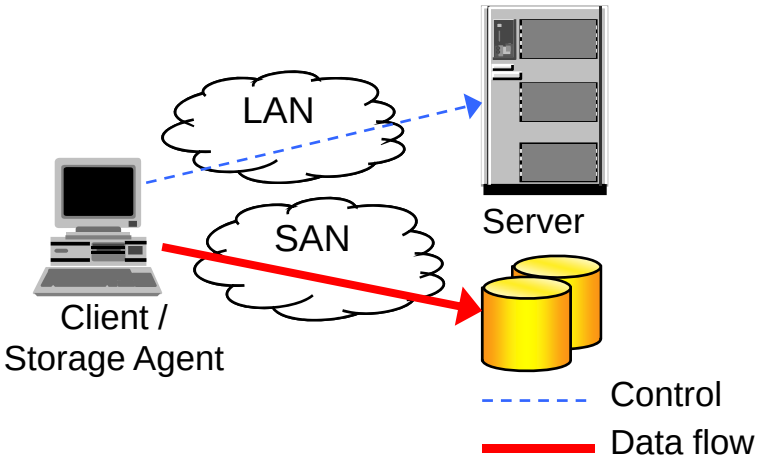
LAN-free Backup/Restore

Random Access

- Not supported

Sequential Access +

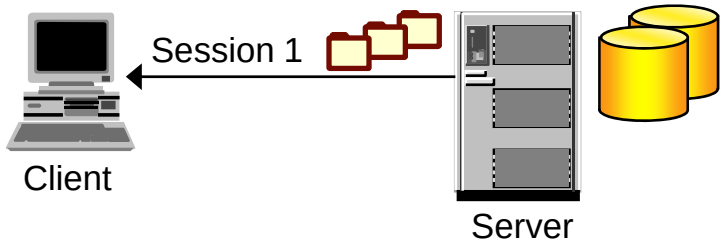
- Supported using SANergy to control shared access to sequential disk volumes
- Reduces CPU cycles on TSM server and moves network traffic from LAN to SAN



Alternative approach for LAN-free to disk would be a virtual tape library (VTL) appliance

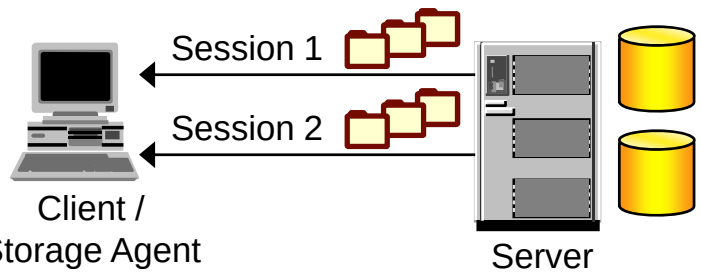
Multi-Session Restore

Random Access



Multi-session restore allows only one session for all random-access disk volumes

Sequential Access +

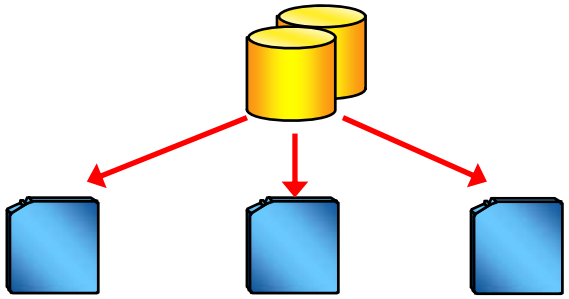


Multi-session restore allows one session per sequential-access volume

Multi-session restore is performed only for no-query restore (NQR) operations

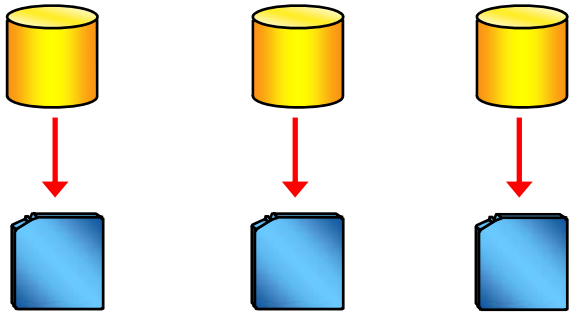
Migration

Random Access



- High/low migration thresholds based on percentage occupancy of the pool
- If node is grouped and target pool is collocated by group, parallel migration processes each work on a different group
- Otherwise, parallel migration processes each work on a different node
- Optimized for transfer by node and file space, making it an ideal intermediate buffer for transfer from non-collocated tape to collocated tape (e.g., restore from copy pool to collocated tape pool)

Sequential Access

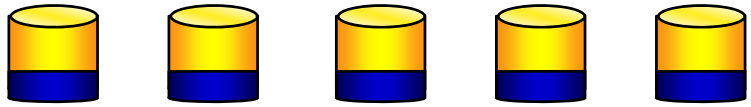


- High/low migration thresholds based on percentage of volumes containing data (behavior change planned for 5.5)
- Parallel migration processes each work on a different source volume, possibly dividing work more evenly among processes
- Collocated sequential disk can be used as a buffer for transfer from non-collocated tape to collocated tape

Migration Thresholds Example

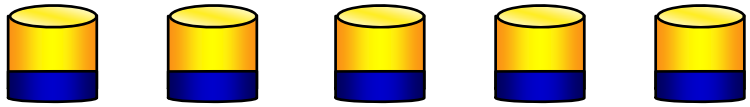
Sequential-Access Disk Today

- Migration from sequential-access disk is based on tape paradigm
- Migration begins when percentage of volumes containing data reaches the high migration threshold
- Example
  - High migration threshold is 80%
  - 5 volumes in pool, each 30% occupied
  - Percent migratable is 100%
  - Migration begins even though pool is only 30% occupied



Enhanced Sequential-Access Disk

- Migration thresholds for sequential-access disk similar to random-access disk
- Migration begins when percentage occupancy for the entire pool reaches the high migration threshold
- Example
  - High migration threshold is 80%
  - 5 volumes in pool, each 30% occupied
  - Percent migratable is 30%
  - Migration does not begin until the entire pool is 80% occupied

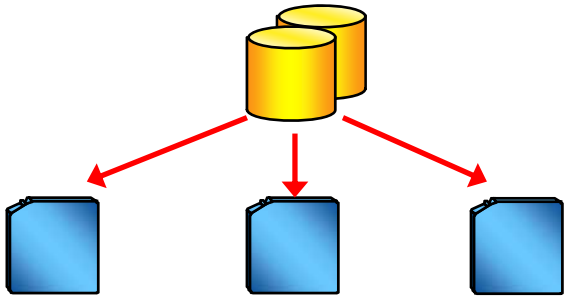


More data stored on sequential-access disk before migration

Planned 5.5

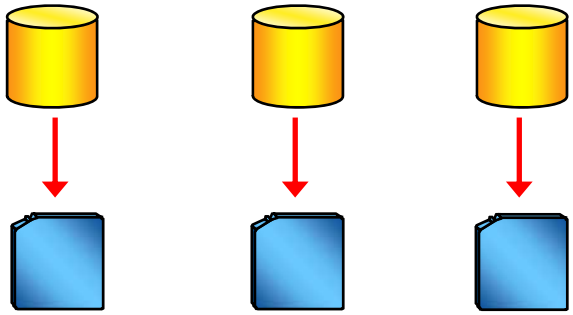
Storage Pool Backup

Random Access

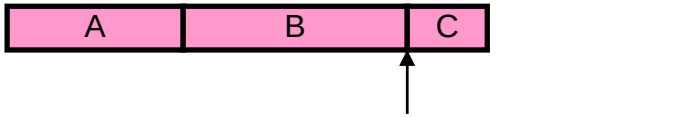


- If node is grouped and target pool is collocated by group, parallel backup processes each work on a different group
- Otherwise, parallel backup processes each work on a different node
- Each physical file (aggregate or non-aggregated file) must be checked during every storage pool backup

Sequential Access +



- Parallel backup processes each work on a different source volume
- Optimization: For each primary pool volume and copy pool, database stores offset of volume that has already been backed up (no need to recheck during each backup)
- Optimization can be especially important for long-term storage of data on disk



Space Recovery

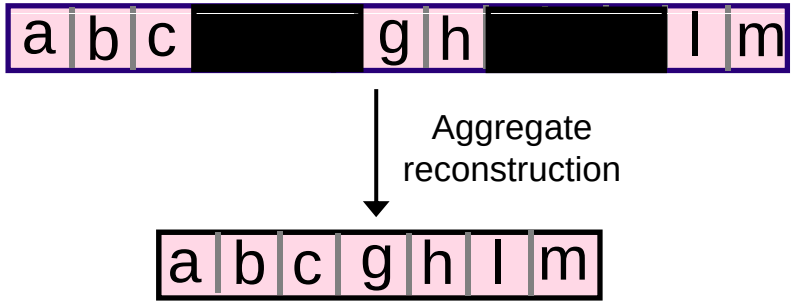
Random Access

- When physical file is moved to another pool (if caching not enabled)
- Space occupied by cached data is recovered as needed
- When physical file is deleted (for aggregates, all files in aggregate must be deleted)
- No reconstruction of empty space within aggregates, a disadvantage if aggregated files are stored for long periods of time

Empty space accumulates until entire aggregate is deleted

Sequential Access +

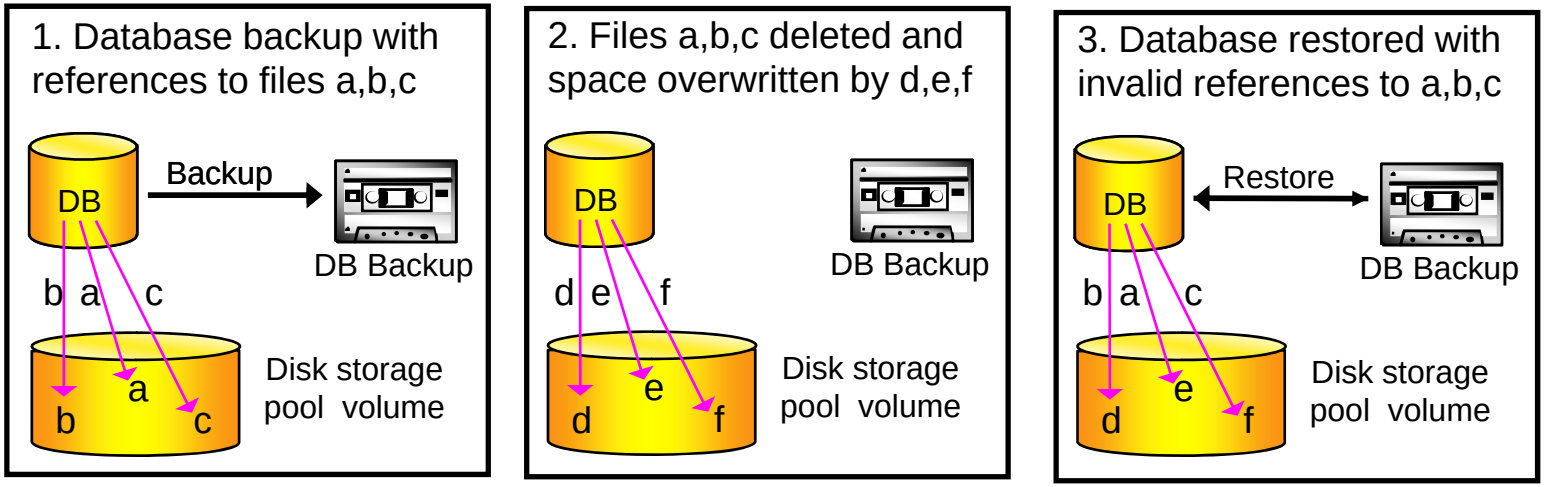
- Space is not immediately recovered after data movement or deletion, but is recovered via reclamation
- During reclamation processing, aggregates are reconstructed to recover space occupied by deleted files



Fragmentation

	Random-Access Disk	Sequential-Access Disk +
Aggregate fragmentation caused by expiration of files	Empty space accumulates in aggregates until all logical files in aggregate are deleted. May result in wasted space for long-term storage on disk.	Empty space is recovered by aggregate reconstruction during reclamation.
Fragmentation of space within TSM volumes caused by deletion of physical files	Volume fragmentation can occur due to allocation of multiple extents if client size estimate is too low. Fragmentation can degrade performance, but is relieved by migration if no TSM caching.	Deletion of physical files results in empty space within volumes, but this is recovered during reclamation.
File system fragmentation leading to fragmentation of files that constitute TSM volumes	Fragmentation is usually minimal because volumes are predefined or created by space trigger.	Use of scratch volumes causes fragmentation because volumes are extended as needed. Fragmentation can be avoided either by predefining volumes or using space trigger.

Database Regression



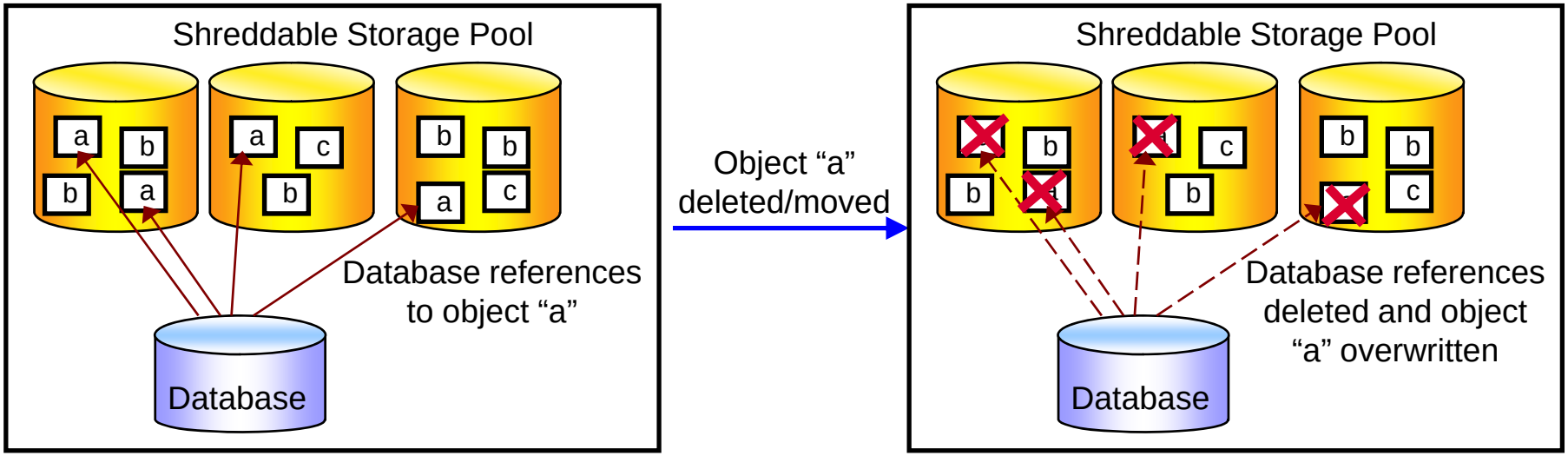
Random Access

- After database regression, all volumes must be audited
- This may be time-consuming for large DISK pools (for example, pools used for long-term data storage)

Sequential Access +

- After database regression, audit only volumes that were reused or deleted after database backup OR
- With REUSEDELAY set, volume audit can be avoided completely
- Time delays for volume audits during critical recovery operations can be minimized or eliminated

Shredding of Data Stored on Disk



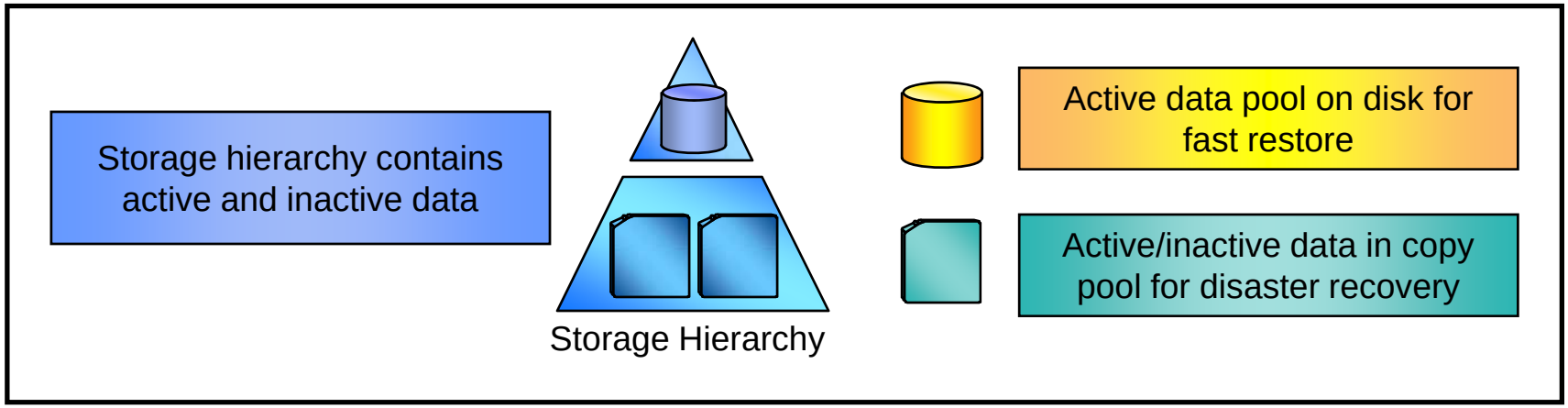
Random Access +

- Disk storage pools can be designated as “shreddable”
- When a data object is moved or deleted from a shreddable pool, TSM server overwrites the object
- Sensitive data objects are destroyed when deleted/moved, preventing undesirable data discovery

Sequential Access

- Not supported (future candidate)

Active Data Pools



Random Access

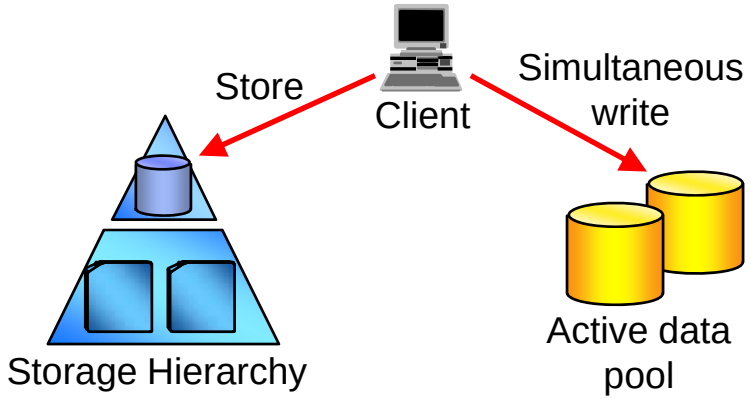
- Not supported

Sequential Access +

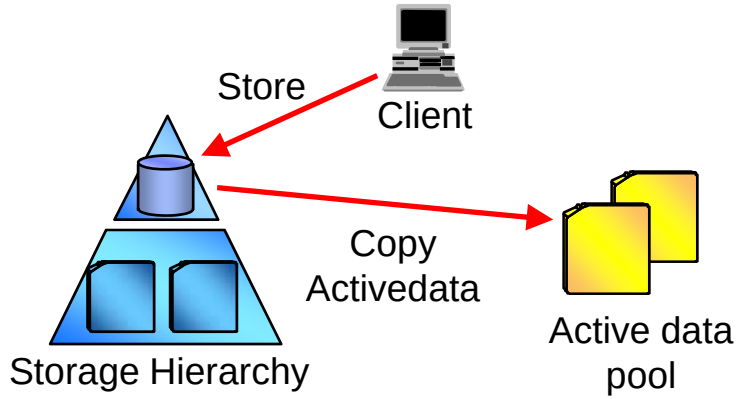
- Typical restores require active data only
- Benefits of active data storage pools
  - Optimized access to active versions for fast restore
  - Reduced size of disk pools if only active versions are stored
  - Avoids data movement to disk in preparation for restore of active data

More on Active Data Pools

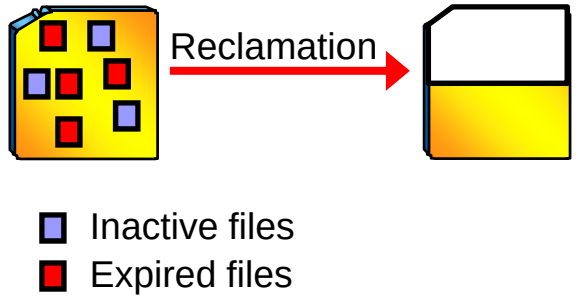
Data copied to active data pool using simultaneous write



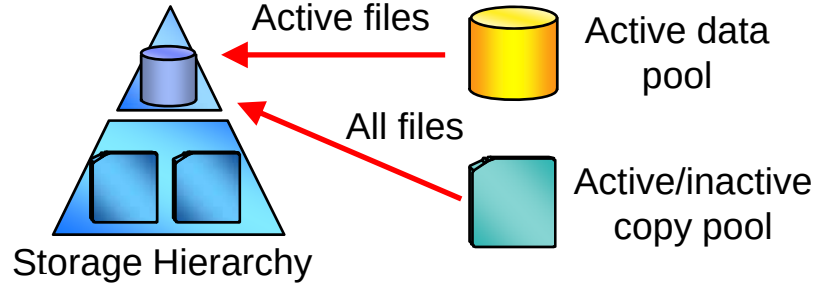
Data copied to active data pool using Copy Activedata command



Reclamation of active data pool recovers space used by inactive and deleted files

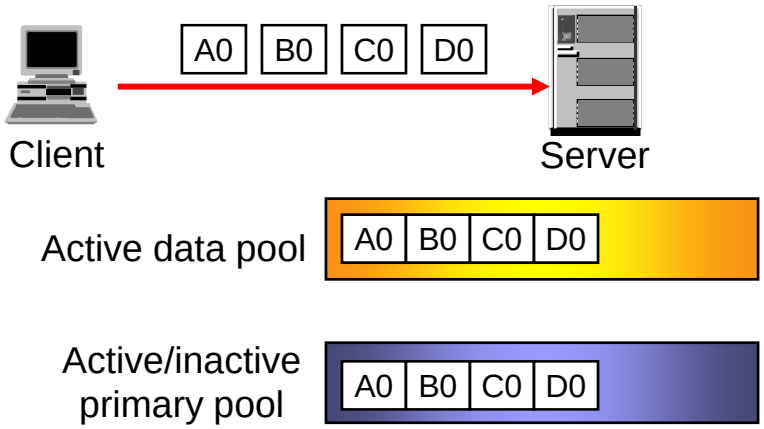


Storage pool restore from active data pool allows restore of active data only

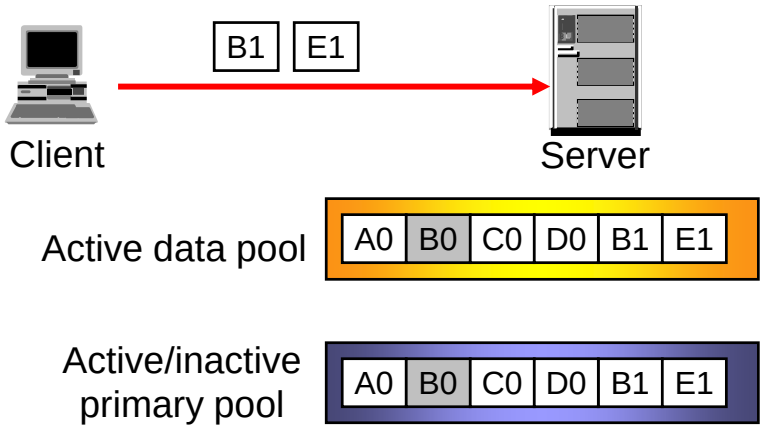


Active Data Pools: Example

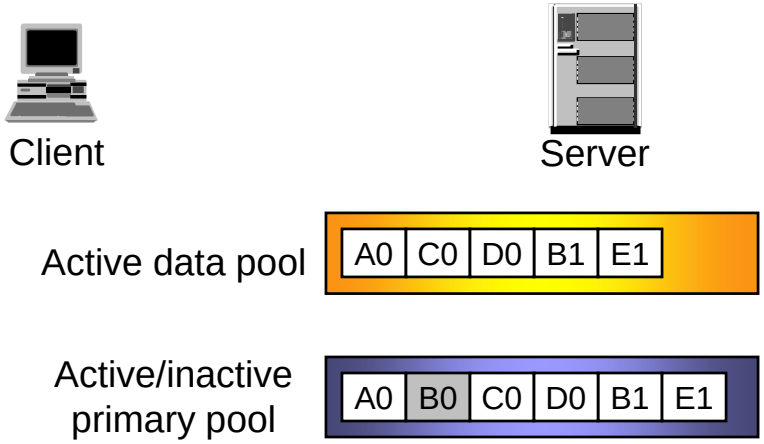
1. Client backs up A0, B0, C0, D0 to primary pool with simultaneous write to active data pool.



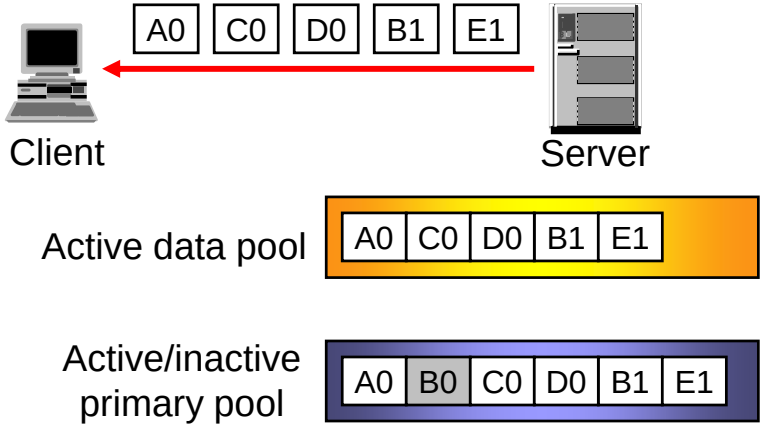
2. Client backs up B1, E1 with simultaneous write to active data pool. B0 deactivated.



3. Reclamation removes inactive B0 from active data pool.



4. Client restores active files A0, C0, D0, B1, and E1 from active data pool.



Random vs. Sequential Disk: Which is Best?

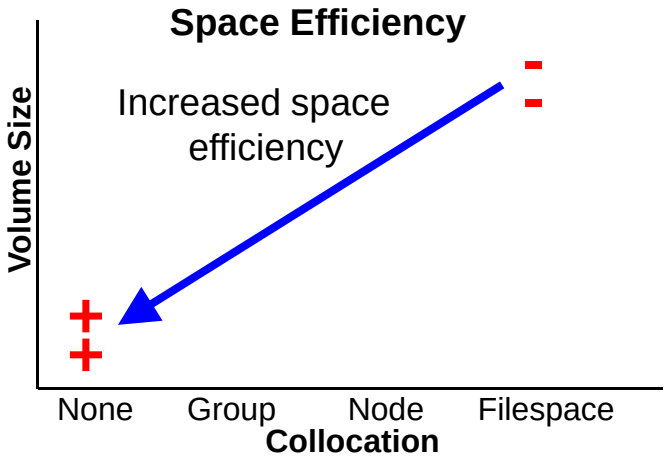
Disk Storage Usage	Recommendation
Traditional disk usage ▪LAN-based storage to disk ▪Daily migration from disk to tape	Either random or sequential, depending on requirements
Long-term storage of data on disk	Sequential offers significant advantages ▪Reconstruction recovers space in aggregates ▪Optimized storage pool backup ▪Reduced volume fragmentation ▪Multi-session restore ▪Avoidance of volume audit
Exploitation of new disk storage features	Sequential-access disk may be required
LAN-free data transfer between client and disk storage	Sequential or VTL
Data shredding	Random

Agenda

- Background
- Random- and sequential-access disk in TSM
- Special considerations for sequential-access disk
- Potential future enhancements for disk storage

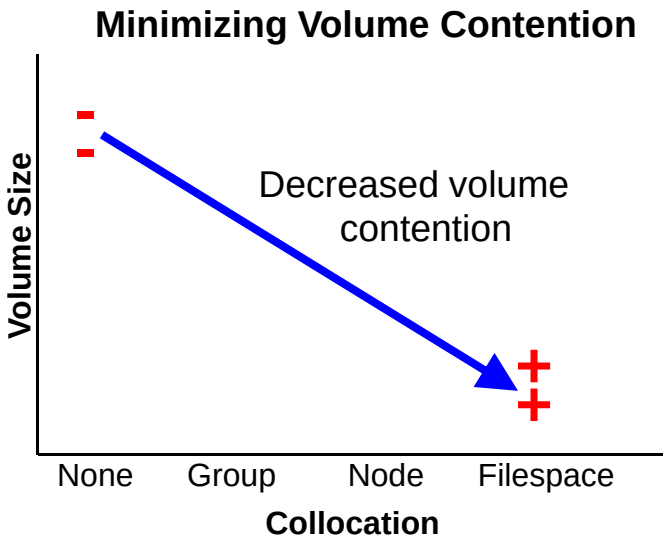
Optimizing Space Efficiency

- Scratch volumes
  - Volumes are created and extended only as needed
  - Space is conserved at the expense of file-system fragmentation
- Non-scratch volumes (created by Define Volume command or space trigger)
  - No collocation is more space-efficient
  - Smaller volumes are more space-efficient
- Reclamation should be performed regularly to recover space
  - Efficient because no mount/dismount
  - Many volumes can be reclaimed concurrently



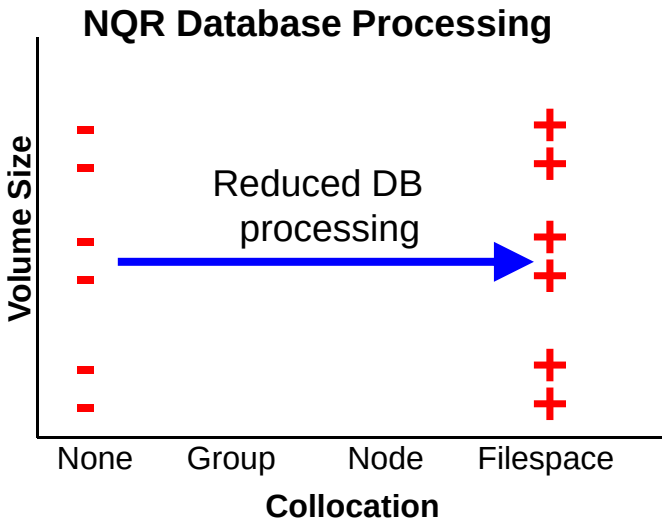
Reducing Volume Contention

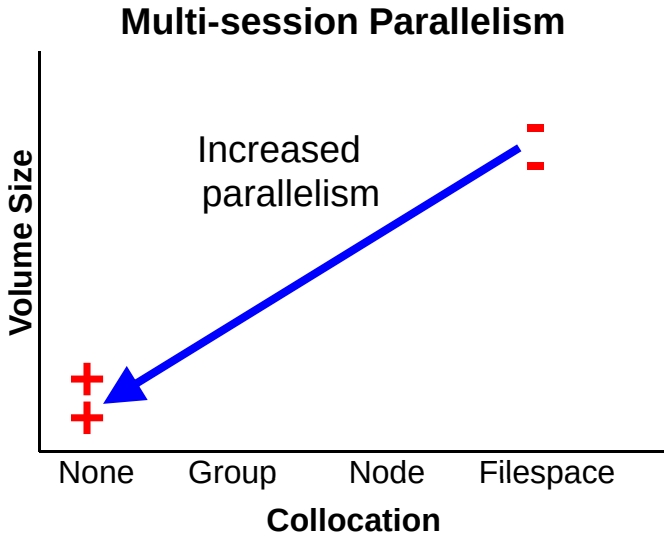
- High-granularity collocation (by node or filespace) reduces contention
- Smaller volume sizes reduce contention
- Volume contention greatly reduced with introduction of concurrent volume access in v5.5



Improving Client Restore Performance

- For no-query restore operations (used for most large restores of file data), database scanning is greatly reduced if data is well collocated
- Multi-session restore operations achieve greater parallelism if data is spread over multiple sequential-access volumes, indicating that parallelism may be increased by
  - Lower-granularity collocation
  - Smaller volumes





Avoiding Fragmentation

- Perform reclamation regularly
- Avoid use of scratch volumes
- Predefine volumes using Define Volume command
- Use space trigger to provision additional volumes as needed

Striking a Balance

- Configuration of sequential-access pools involves tradeoffs, but the following may be a reasonable starting point for most environments
- Define volumes and use space triggers for additional volume provisioning
- Collocate by node or group of nodes
- Use volume size scaled to the size of stored objects
  - For file systems, volume size of 2 GB to 10 GB
  - For databases and other large objects, volume size of 100 GB
- Set reclamation threshold at 20-60% and allow multiple reclamation processes
- Consider use of active data pools to achieve fast restore for active data while reducing disk storage requirements

Agenda

- Background
- Random- and sequential-access disk in TSM
- Special considerations for sequential-access disk
- Potential future enhancements for disk storage

Potential Future Enhancements for Disk Storage

- Enhancements specifically for sequential-access disk pools
  - Migration thresholds based on percentage occupancy rather than volumes containing data (planned 5.5)
  - Concurrent access for volumes (planned 5.5)
  - Performance improvements for sequential-access disk on z/OS server (next release candidate)
- LAN-free to sequential-access disk volumes in GPFS (next release candidate)
- Data deduplication (next release candidate)
- Data shredding for sequential-access disk (future candidate)
- Improvements to snapshot support (5.5, next release candidate, future candidate)
- Additional exploitation of continuous data protection (CDP) technology (future candidate)

Summary

- Trend toward increasing use of disk for long-term data storage in the TSM hierarchy
- TSM supports both random- and sequential-access disk, which differ in how disk is managed and operations supported
- Sequential-access disk is considered strategic and continues to be enhanced