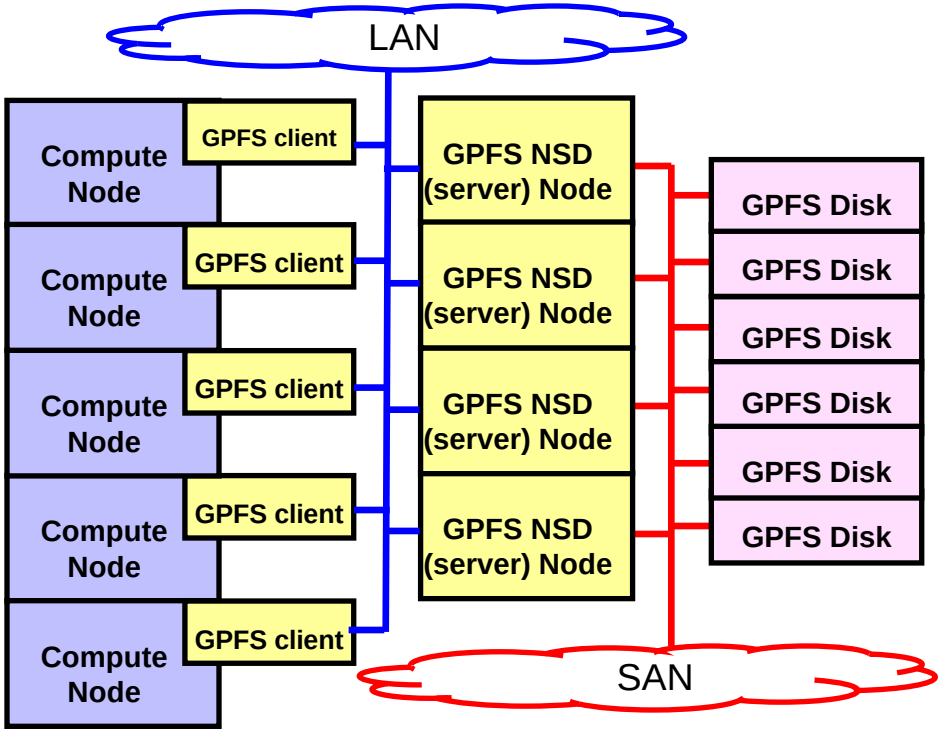
High Performance Storage System: GPFS + HPSS

* Please review disclosure statement on last slide

GPFS conceptual architecture

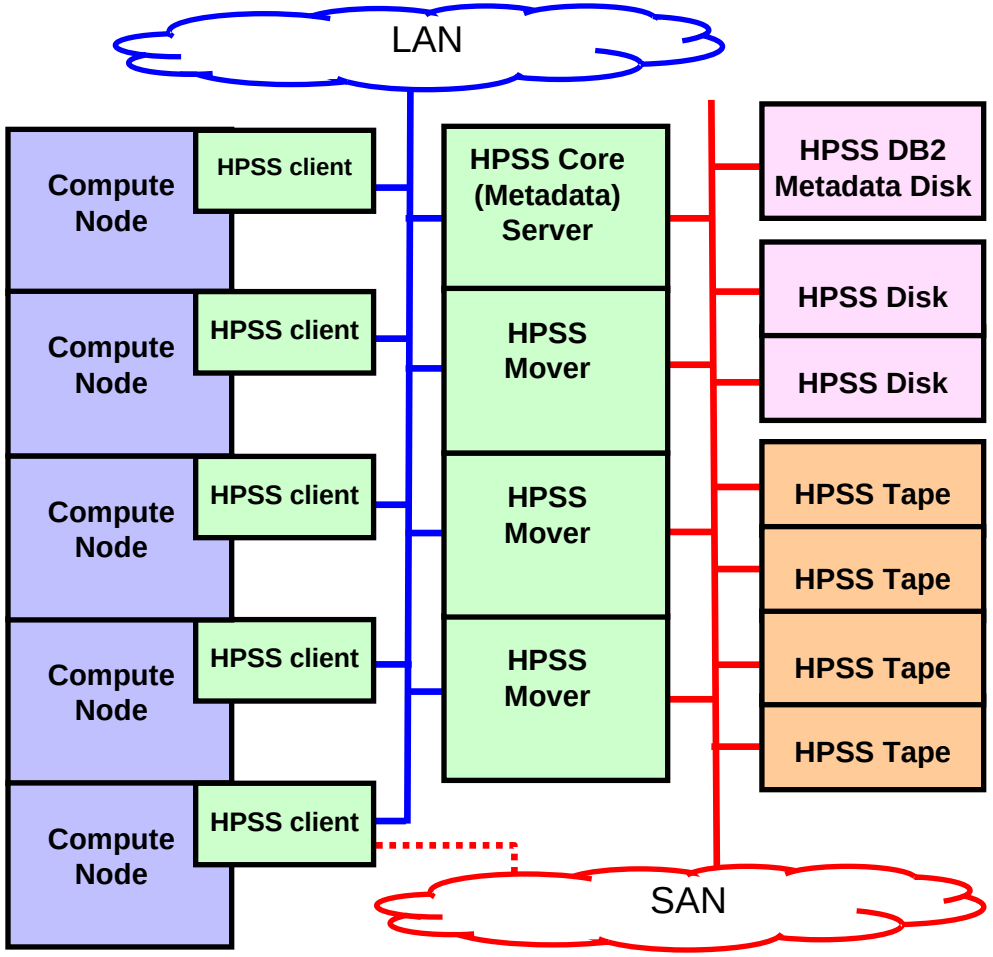
- General Parallel File System (GPFS) is IBM's highly scalable parallel clustered file system.
- Individual files are broken into multiple blocks and striped across the resources of multiple servers.
- Metadata i-nodes are also usually distributed over multiple servers.
- Information Lifecycle Management (ILM) policy scans are capable of scanning a billion files in just minutes.



- GPFS Client nodes and Network Storage Device (NSD) nodes differ only in whether they access disks directly over SAN or access disks only through other nodes.

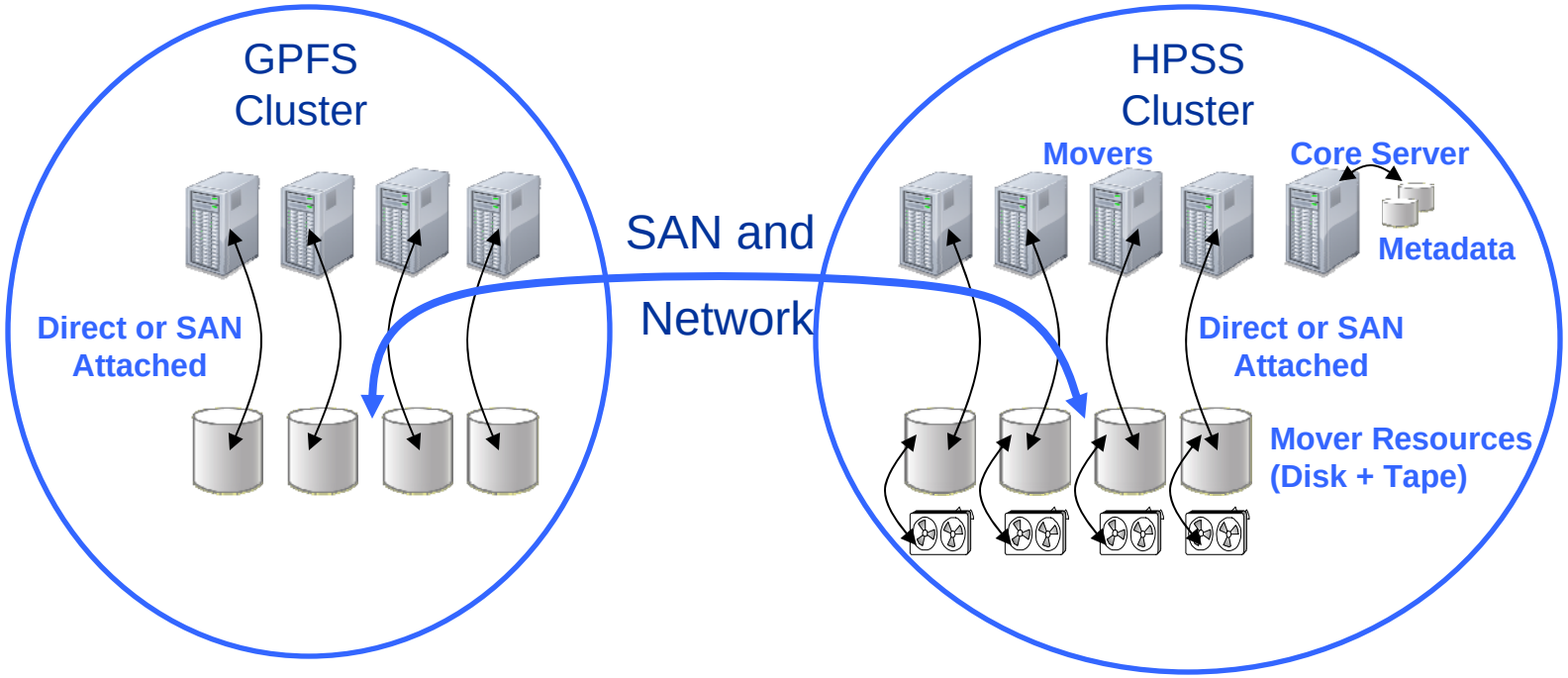
HPSS conceptual architecture

- HPSS cluster architecture is topologically similar to GPFS cluster architecture.
- HPSS has a DB2-based metadata architecture, while GPFS has i-node metadata as do most file systems.
- HPSS is primarily a tape-oriented system, whereas GPFS is a disk system.
- HPSS clients can send and receive data over a LAN or a SAN, as can GPFS. However, very large systems typically use LAN, shown here.

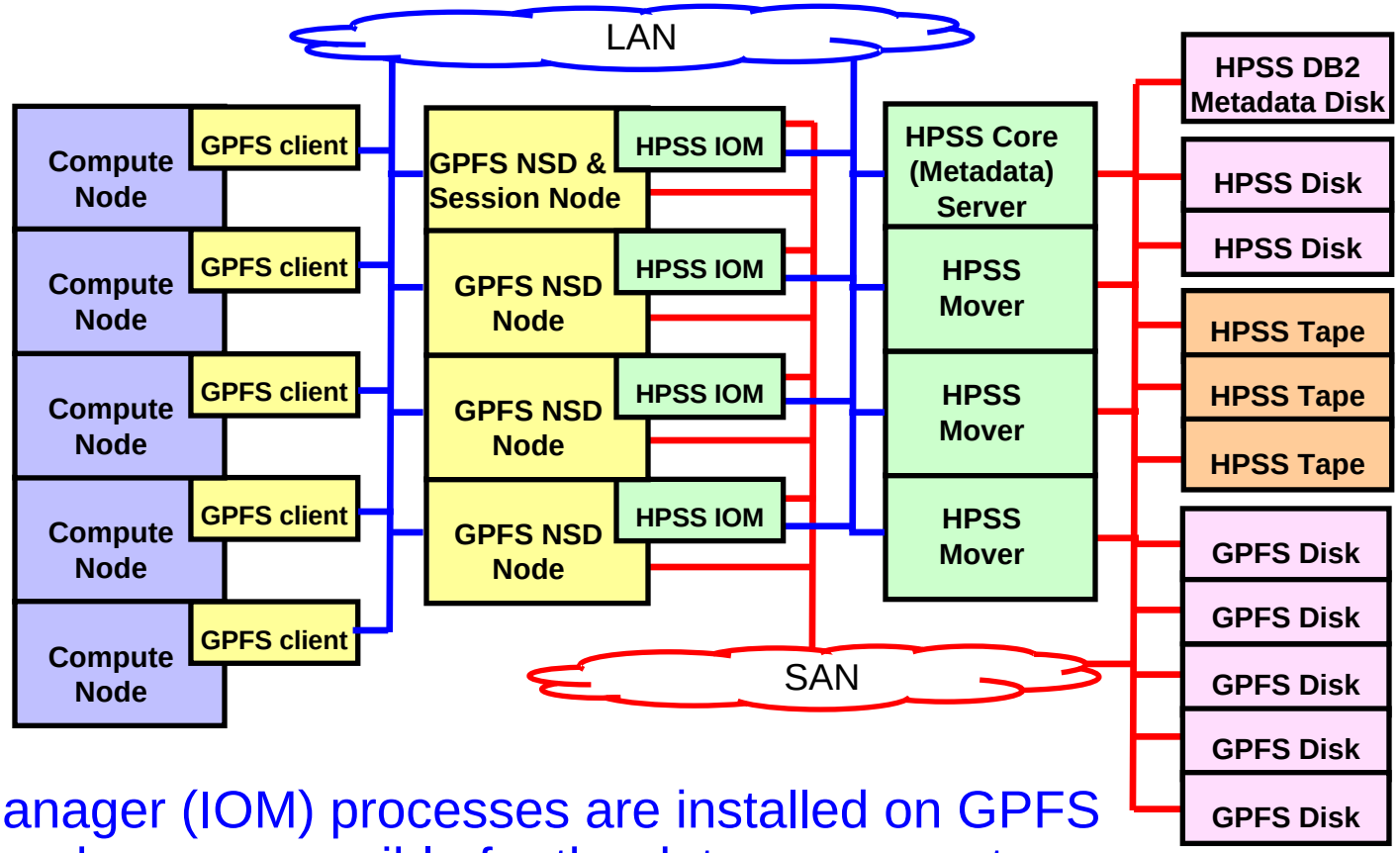


GPFS + HPSS

- HPSS can provide a highly scalable GPFS tape pool
- HPSS provides two tightly coupled services for GPFS:
  - Hierarchical space management of GPFS file space
  - Disaster recovery backup of GPFS file systems



GPFS + HPSS conceptual architecture



- HPSS I/O Manager (IOM) processes are installed on GPFS NSD nodes and are responsible for the data movement between GPFS and HPSS.
- This enables HPSS to provide hierarchical file migration and recall as well as disaster recovery backup for GPFS.

GPFS HPSS Interface

- GPFS + HPSS
  - Near-seamless file migration between GPFS and HPSS
  - Uses the GPFS rule-based ILM policies to simplify administration
  - Provide unequaled file system scalability to billions of files
  - Integrated disaster recovery for very large GPFS file systems
  - Support very high-performance data acquisition
- Optional feature
  - GPFS is sold separately
  - GPFS HPSS Interface Feature is an add-on feature of HPSS
- Joint project of
  - IBM HPSS - Houston
  - IBM GPFS Product Development
  - IBM Almaden Research Center
  - Our customers and collaboration members

- GPFS + HPSS = EXTREME SCALABILTY

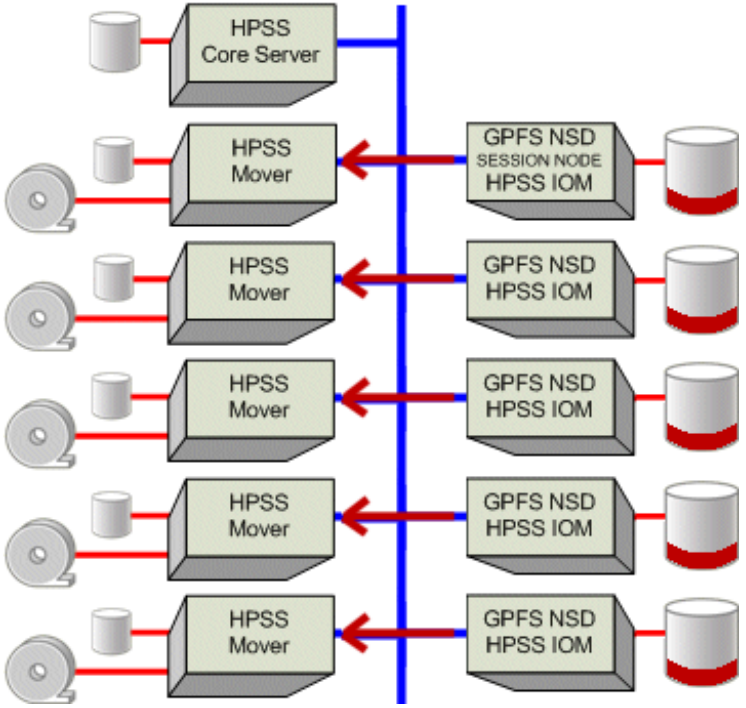
HPSS + GPFS

Why this combination is the best high performance storage solution

- Both GPFS and HPSS are cluster storage software solutions that support very large numbers of files and storage devices.
  - Both support file sharing among members of the cluster.
  - Both are used with some of the world's largest supercomputers.
- GPFS provides a disk file system suitable for home directories and working storage.
  - Its very low latency supports fine-grain parallel I/O.
  - GPFS is arguably the most scalable cluster file system anywhere.
- HPSS provides storage for long term retention and rapid staging.
  - Originally a stand-alone system, it now can serve as deep storage for GPFS.
  - Data is usually kept on tape, but HPSS provides up to five levels of hierarchy of disk and tape storage.
  - Its very rugged DB2 metadata architecture is suitable for medium- and coarse-grain file access.
  - HPSS is arguably the most scalable file retention and staging system anywhere.
- Together GPFS and HPSS provide extreme storage scalability and very high rate data transfers for large institutional repositories.

GPFS space management using HPSS

- GPFS files are continuously copied to HPSS.
- When GPFS space thresholds are reached, HPSS identifies candidates to be purged.
- All but the first block of the data are then purged to free disk resources.
- As files are needed by the user, HPSS will automatically stage files back to GPFS.
- Users can also schedule requests to recall groups of files, in bulk, from HPSS.



GPFS disaster recovery using HPSS

- Backup
  - File data are continuously copied to HPSS as part of space management process, so minimal data movement is necessary during a backup.
  - HPSS captures a snapshot of the GPFS namespace and stores it into HPSS.
  - One copy of a file, on tape, can serve both backup and space management.

- Restore
  - The disaster recovery process is simply a restore of the GPFS namespace.
  - Once the namespace has been fully recreated, the file system can be returned to the users.
  - Files can be recalled in bulk as needed, or on demand when accessed by the user.

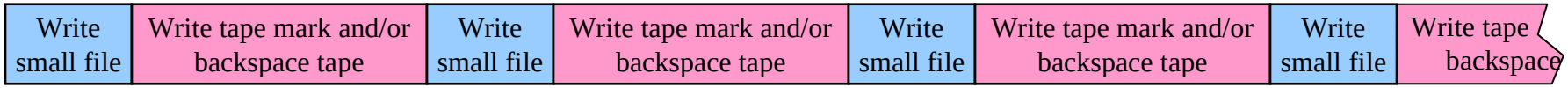
Copying GPFS files to HPSS with GHI

- Space management policy is governed by the GPFS Information Lifecycle Management (ILM) policy engine.
- HPSS initiates a GPFS ILM policy scan to identify files that need to be copied to HPSS.
- ILM policy results are sent to the HPSS Session Node processes.
- Using the policy results, HPSS I/O Managers (IOMs) copy data from GPFS to HPSS.
- Huge files can be striped to reduce migration time.
- Small files are aggregated to larger objects to improve performance.

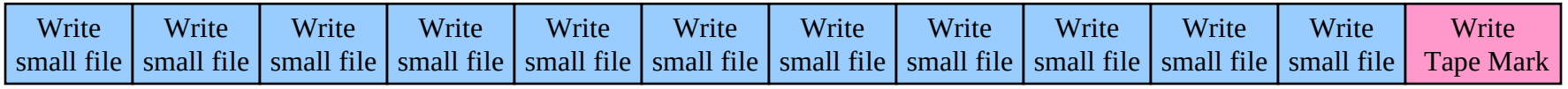
Small file aggregation with GPFS + HPSS

- GPFS says, 90% of the files use 10% of disk space, while 10% of the files use 90% of the resources.
- Writing one small file at a time to tape, diminishes tape drive performance – tape marks and/or backspacing tape.
- Large aggregates allow data to stream to the tape drive.
- GHI supports multi-threaded, double buffered aggregation of small GPFS files.
- HPSS v7.1 (when used alone without GPFS) also has small file aggregation, though the method is different.

Without aggregation of small files:



With aggregation of small files:



Sample ILM rules to migrate files

- Exclude all files in the scratch area (`/gpfs1/scratch`)

```
RULE ‘exclude' EXCLUDE FROM POOL 'system'
  WHERE path_name like '/gpfs1/scratch/%'
```

Rule Name

GPFS Pool

Qualifier

- Copy files larger than 10K to HPSS as non-aggregates

```
RULE 'tohsm_nonaggr' MIGRATE FROM POOL 'system'
  WEIGHT(CURRENT_TIMESTAMP - ACCESS_TIME)
  TO POOL 'hsm_nonaggr'
  WHERE FILE_SIZE > 10240
```

HPSS Pool

Order processing based on access time

Qualifier

- Copy files smaller than 10K to HPSS as aggregates

```
RULE 'tohsm_aggr' MIGRATE FROM POOL 'system'
  WEIGHT(CURRENT_TIMESTAMP - ACCESS_TIME)
  TO POOL 'hsm_aggr'
  SHOW ('-s' FILE_SIZE) WHERE FILE_SIZE <= 10240
```

Add more details to the scan results

Freeing GPFS disk space with GPFS + HPSS

- GPFS ILM policies are invoked when upper threshold limits are reached.
- An ILM policy is used to identify files to be freed.
- The session node will then remove all but the first block of these older GPFS files to free up disk resources.
- File metadata remains in the inodes of the GPFS file system (backed up by HPSS).
- From the user's point of view, the files look the same, but the data is no longer on GPFS resources.

Sample ILM rules for threshold limits

- Exclude files that have not been copied to HPSS

```
RULE 'exclude' EXCLUDE FROM POOL 'system'
  WHERE MISC_ATTRIBUTES NOT LIKE '%M%'
```

Qualifier that identifies if the file is in HPSS

- When the file system is at 90%, identify enough files to bring me down to 80%, but only give me the oldest of files that are larger than one block (256 KB).

```
RULE 'toHsm' MIGRATE FROM POOL 'system'
  THRESHOLD(90,80)
  WEIGHT(CURRENT_TIMESTAMP - ACCESS_TIME)
  TO POOL 'hsm_punch'
  WHERE FILE_SIZE > 256000
```

High and low watermarks for threshold processing

Staging files back to GPFS from HPSS

- Files are staged back to GPFS on demand or in bulk.
- Stage on demand
  - Triggered when a purged file is accessed.
  - GHI processes DMAPI events to copy files from HPSS back to GPFS.
- Bulk recall
  - ILM policies can be crafted to stage groups of files from HPSS back to GPFS.
  - Files are organized to minimize tape mounts and tape movement.
  - Can be scheduled to run after hours.

Tightly integrated HSM and backup

- GPFS + HPSS has tightly integrated the HSM and backup features.
  - HSM managed data is used by the backup and restore components.
  - Eliminates the need to manage separate tapes for space management and backup.
- Backup process includes:
  - HSM copy new file data to HPSS
  - Capture point-in-time snapshot of the entire GPFS directory structure
  - Capture attributes and other details of new files
- Restore process includes:
  - GPFS namespace is created using the point-in-time snapshot, and the attribute details of the files.
  - DMAPI events or ILM recall policies are used to stage file data back to GPFS based on the user's priorities.

Sizing GPFS for a GPFS + HPSS installation

- GPFS resources must be adequate to support the user's requirements AND the HSM/Backup activities of HPSS
- GPFS ***data*** resources must be provisioned to handle the additional bandwidth needed to migrate files to and from HPSS.
- GPFS ***metadata*** resources must be provisioned to handle:
  - Additional IOPS needed for ILM policy scans and increased extended attributes operations.
  - Additional space required for the extended attributes being stored with each file copied to HPSS.
- Network provisioning must take into consideration the additional network capacity required for file migration and recall and for file system backup.

GPFS HPSS Interface Failover support

- GPFS Session node supports HA by automatically failing over to a surviving GPFS quorum node.
- HPSS I/O Manager nodes support HA by redistributing the workload to surviving I/O Manager nodes.
- HPSS itself has optional high availability capability for core server and movers, available only in Red Hat Linux version

HPSS Contacts

Jim A. Gerry – jgerry@us.ibm.com

Harry Hulen – hulen@us.ibm.com

Patrick Schaefer – pschaef@us.ibm.com

Bob Coyne – coyne@us.ibm.com

Disclaimer

- Please obtain and read product documentation before deciding to acquire HPSS.
  - Documentation includes the HPSS License Agreement, the Statement of Work, and HPSS and other product manuals.
  - In case of conflict between information herein and product documentation, the documentation shall take precedence.
- Forward looking information including schedules and future product capabilities reflect current planning that may change and should not be taken as commitments by IBM.
- IBM may at its sole discretion discontinue, add, or change HPSS features and function without notice.