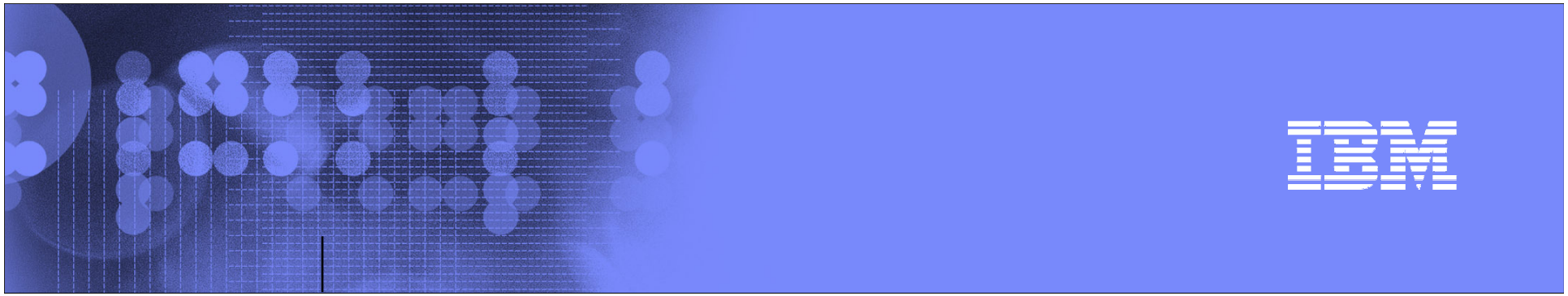


GPFS Update to SciComp 2007

Kalyan Gunda
kgunda@us.ibm.com
Scalable I/O Software Development

Outline1

- GPFS 3.2
  - ILM Integration with HSM
  - Multiple NSD servers per LUN
  - Heterogeneous platform support (Solaris & Windows)
  - Performance improvements
- Questions

GPFS 3.2

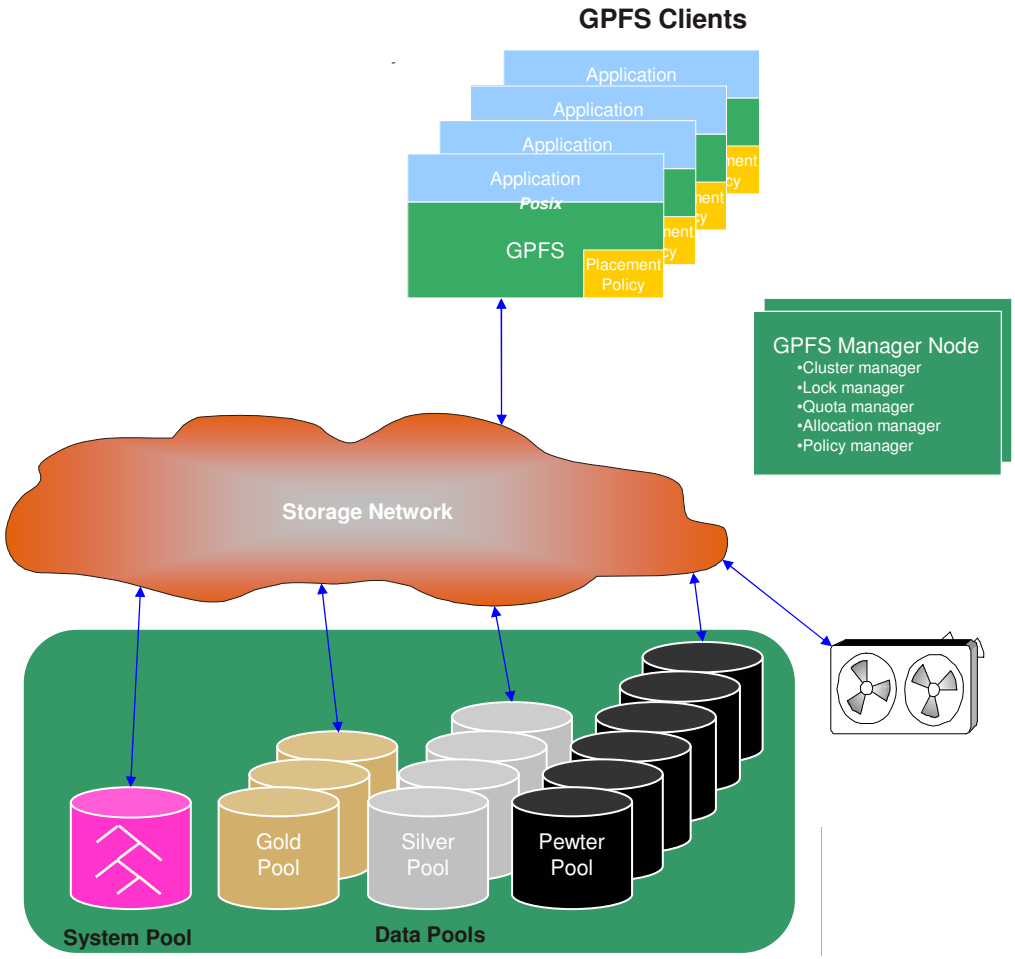
- Policy managed disk-tape migration
- Multiple active NSD servers per LUN
- Rolling migration/co-existence with 3.1
- More platforms (Solaris 9/SPARC, Windows)
- Infiniband – non-IP protocol support
- Admin/Manageability improvements
  - Tracing improvements (mmtracectl command)
  - SNMP support – GPFS monitoring
  - Improved failover times for database usage (persistent reserve)
  - Reconnecting broken sockets
- Support for NFS4 on Linux

GPFS 3.2

- Incorporation and support of HA-NFS
- RHEL5 support
- Rolling Migration from 3.1 & 3.2
- Scaling and performance improvements
  - Parallel file create in the same directory
  - Larger pagepool support (256 GB)
  - More mounted filesystems (256)
  - SMP scaling improvements
  - Parallel mmdefragfs
  - Mmfsck improvements
  - Control over placing cluster manager(mmchmgr –c)

GPFS Information Lifecycle Management

- GPFS supports Information Lifecycle Management (ILM) via three new abstractions: storage pools, filesets, and policies
  - –Storage pool – group of LUNs
  - –Fileset: subtree of a file system
  - –Policy – rule for assigning files to storage pools
- Types of policy rules
  - – **Placement**, e.g. place database files on RAID1 storage, place other data files on RAID6
  - – **Migration**, e.g. move project files to SATA storage after 30 days, to tape after 60 days
  - – **Deletion**, e..g delete scratch files after 7 days



GPFS - ILM Integration with HSM

- The idea: Integrate GPFS Policies with HSM
- The advantages:
  - Integration of disk-to-disk and disk-to-tape into fully-tiered storage
  - Finer control of data movement across storage tiers
  - More efficient scans and data movement using internal GPFS functions
  - Possibility of coalescing small disk files into large tape files
- GPFS/HPSS Integrated ILM Technology Demo at SC06
  - New list-oriented interface to HSM system (move list of files to/from disk)
  - HPSS prototype drives parallel data movement to tape over this interface

Policy language Enhancements

- Support for External Pools in policy language
  - Used to migrate/recall data from HSM
  - Trigger migrate using NO_SPACE/LOW_SPACE events
  - Eg:

    RULE EXTERNAL POOL 'hsm' LIB '/var/mmfs/etc'

    RULE ‘migx' MIGRATE FROM POOL ‘gold pool' THRESHOLD(90,85)
    WEIGHT(CURRENT_TIMESTAMP - ACCESS_TIME) TO POOL ‘hsm’

    WHERE FileSize > 1024KB

    This says migrate files from online pool goldpool to external pool hsm when goldpool is more than 90% full. Migrate only files greater than 1024KB and migrate files with the oldest access time first until the goldpool becomes less than 90% full. Then premigrate the files using same criteria an additional 5%.
  - **Limit Option for placement, migration and restore rules**.
    - This allows to say, “I like to use gold pool, but if that is already/nearly filled up, my second choice is silver, but if that is nearly full, my third choice is…”

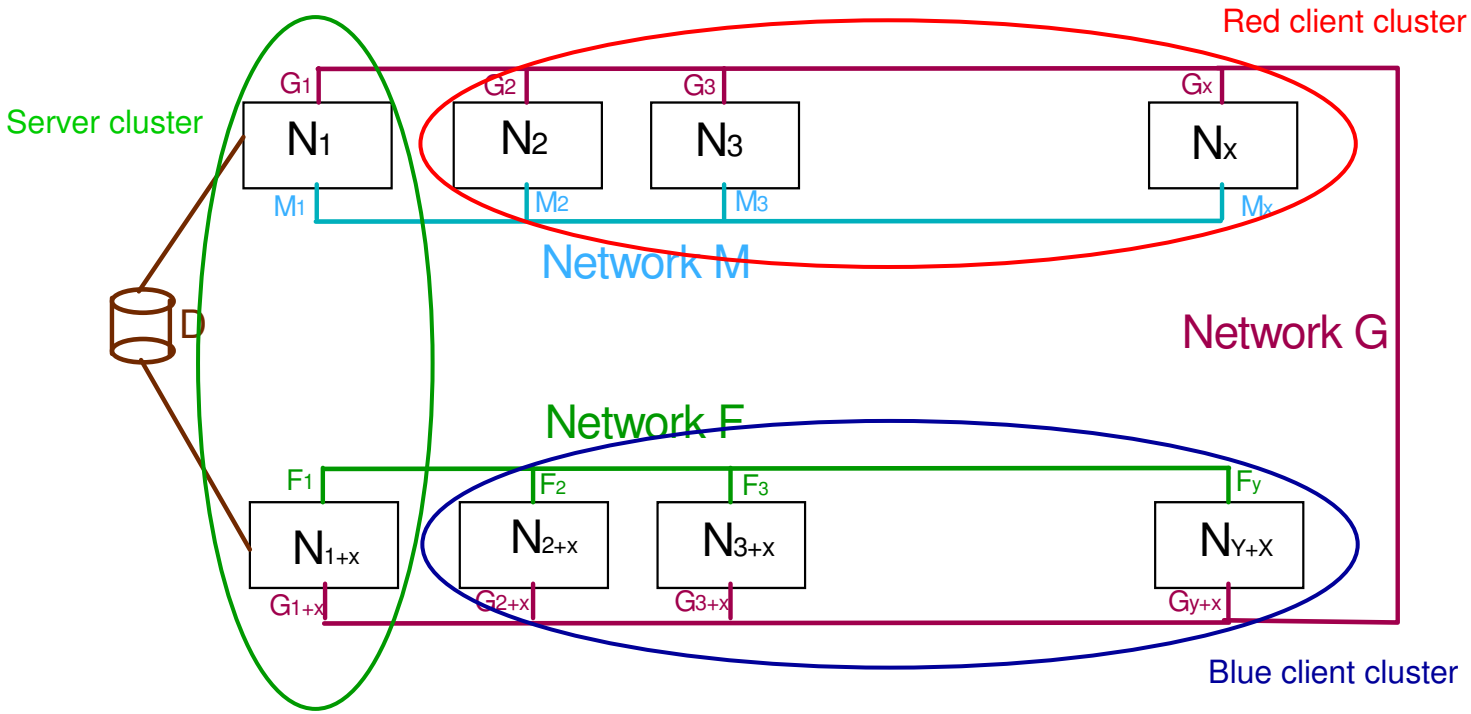
Details

- Mmapplypolicy generates a list of files to be migrated or backed up
- Invokes external scripts provided by TSM/HPSS
  - Eg: mmfsmigrate, mmfspremigrate, mmfsrecall, mmfspurge
  - Use LIB keyword to specify the location of these scripts
  - These scripts take the fileLists and other optional parameters and will invoke vendor specific cmds(eg: dsmmigrate)

Multiple NSD Servers per LUN

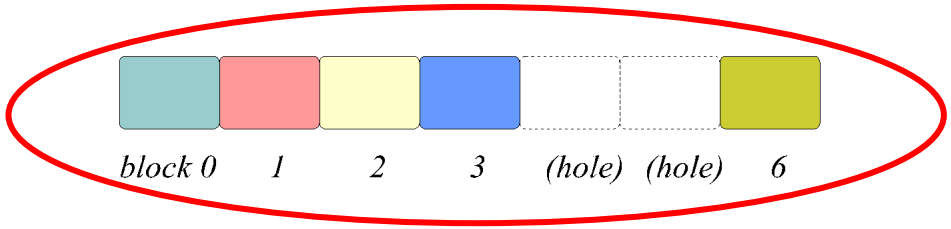
- What: the ability for different clients to go to different NSD servers to access a particular disk
- Why:
  - Access to storage across local cluster fabric
    - E.g. local Federation, Infiniband, etc.
    - Much better performance than across routed Ethernet

Multiple active NSD servers

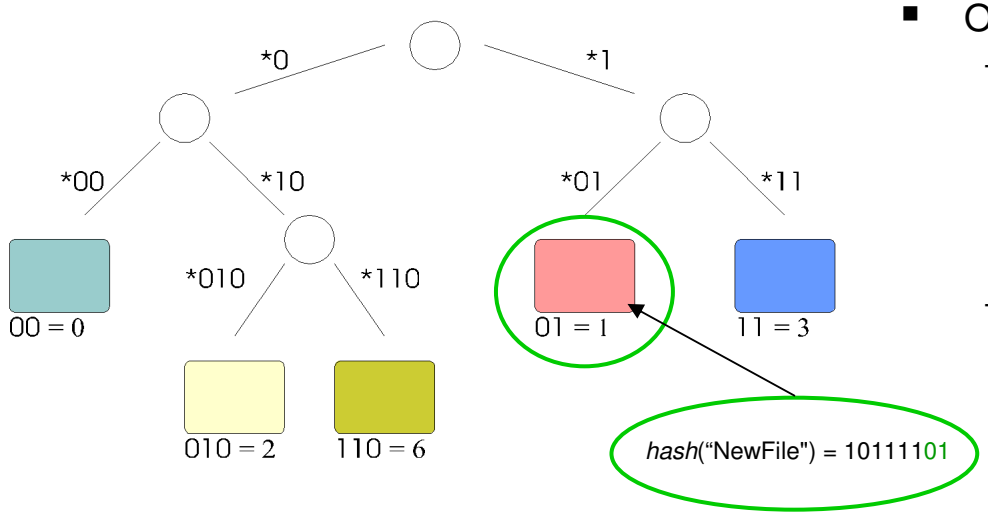


- **Client nodes favor NSD servers on the same subnet (presumably the same fabric)**
- **High availability - when the NSD server for a LUN is not reachable:**
  - Fall back to other NSD servers on the same subnet
  - Next, fall back to NSD servers on other routable subnets
- **Multi-cluster implications**
  - NSD servers for a file sysem must all be in the same cluster
  - If using multiple active NSD servers in multi-cluster environment, configure NSD node on client subnet as a member of one server cluster exporting the file system

Parallel file create



- File create previously locked the entire directory
  - Fine if files are in different directories
  - Not the best if they're in the same directory
  - Unfortunately, this is a natural way to organize files, and programmers are hard to re-educate
    - Checkpoints, output for a time-step



- Optimization to improve write sharing
  - GPFS directories use extendible hashing to map file names to directory blocks
    - Last n bits of hash value determine directory block number
    - Example: *hash*("NewFile") = 10111101 means the directory entry goes in block 5
  - File create now only locks the hash value of the file name
    - lock ensures against multiple nodes simultaneously creating the same file
    - actual update shipped to metanode
    - If create requires directory block split, lock upgraded to cover both old and new block
    - Parallel file create performance no longer depends upon whether or not the files are in the same directory
      - (except when splitting a block)

GPFS Futures(3.3)

- Petascale computing work - DARPA High Productivity Computing Systems program (HPCS) Phase III - which continues through 2010
  - Larger file systems (Petabyte)
  - More nodes and more powerful single nodes (larger scale SMPs)
  - 32K creates/sec
  - 30K node clusters
  - 10K metadata ops/sec
  - 32GB/sec to single node
  - 4TB/sec aggreate bandwdith
  - 1 Trillion files with potentially 1 billion online
  - HSM support with 300GB/sec required at peak
- pNFS
- Management of huge numbers of files
- Improved file search
- Improved robustness with increasing numbers of components
- React to the increases in network capabilities
- More Admin/Monitoring GUI/Usability work
- Software RAID
- More ILM work

Questions?