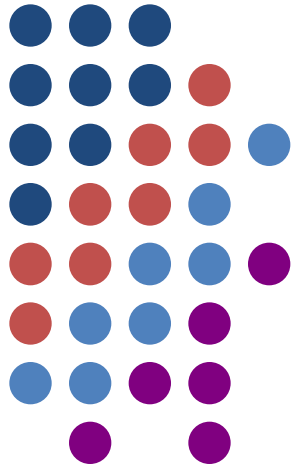
The GLADE Environment

GPFS User Forum SC13

18 November 2013

Pamela Gillman, NCAR

Manager, Data Analysis Services Group



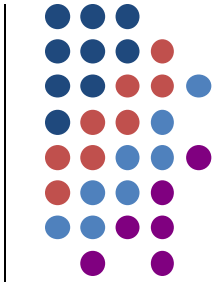
Data Analysis Services Group

NCAR / CISL / HSS / DASG

- Data Transfer and Storage Services
  - Pamela Gillman
  - Joey Mendoza
  - Craig Ruff
- High-Performance File Systems
- Data Transfer Protocols
- Visualization Services
  - John Clyne
  - Alan Norton
  - Scott Pearse
  - Miles Rufat-Latre (student)
- VAPOR development and support
- 3D visualization

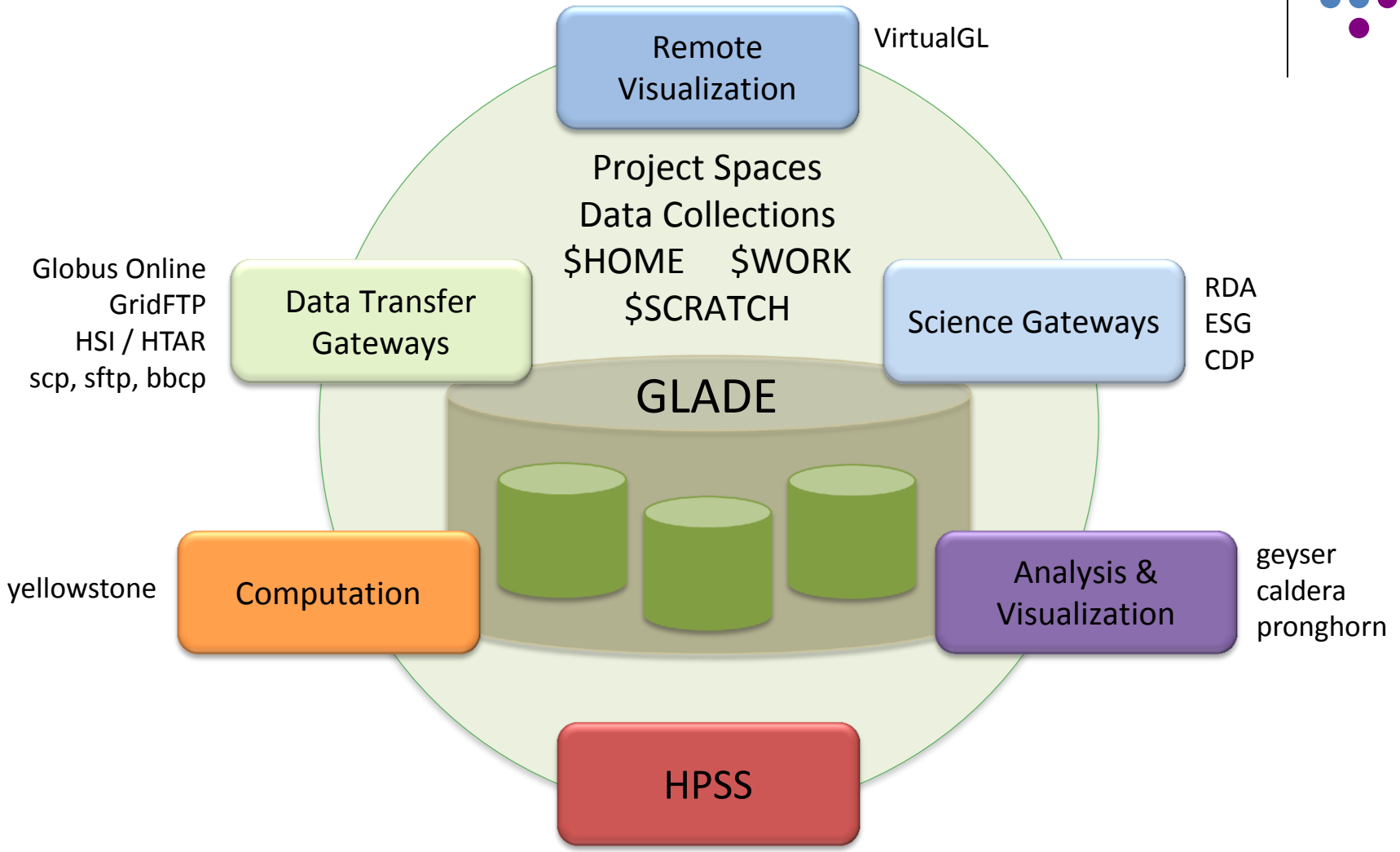
GLADE

GLobally Accessible Data Environment

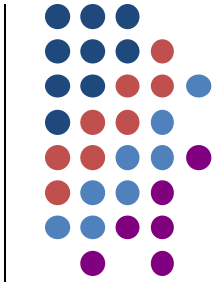


- Unified and consistent data environment for NCAR HPC
  - Supercomputers, Data Analysis and Visualization Clusters
  - Support for project work spaces
  - Support for shared data transfer interfaces
  - Support for Science Gateways and access to ESG & RDA data sets
- Data is available at high bandwidth to any server or supercomputer within the GLADE environment
- Resources outside the environment can manipulate data using common interfaces
- Choice of interfaces supports current projects; platform is flexible to support future projects

GLADE Environment

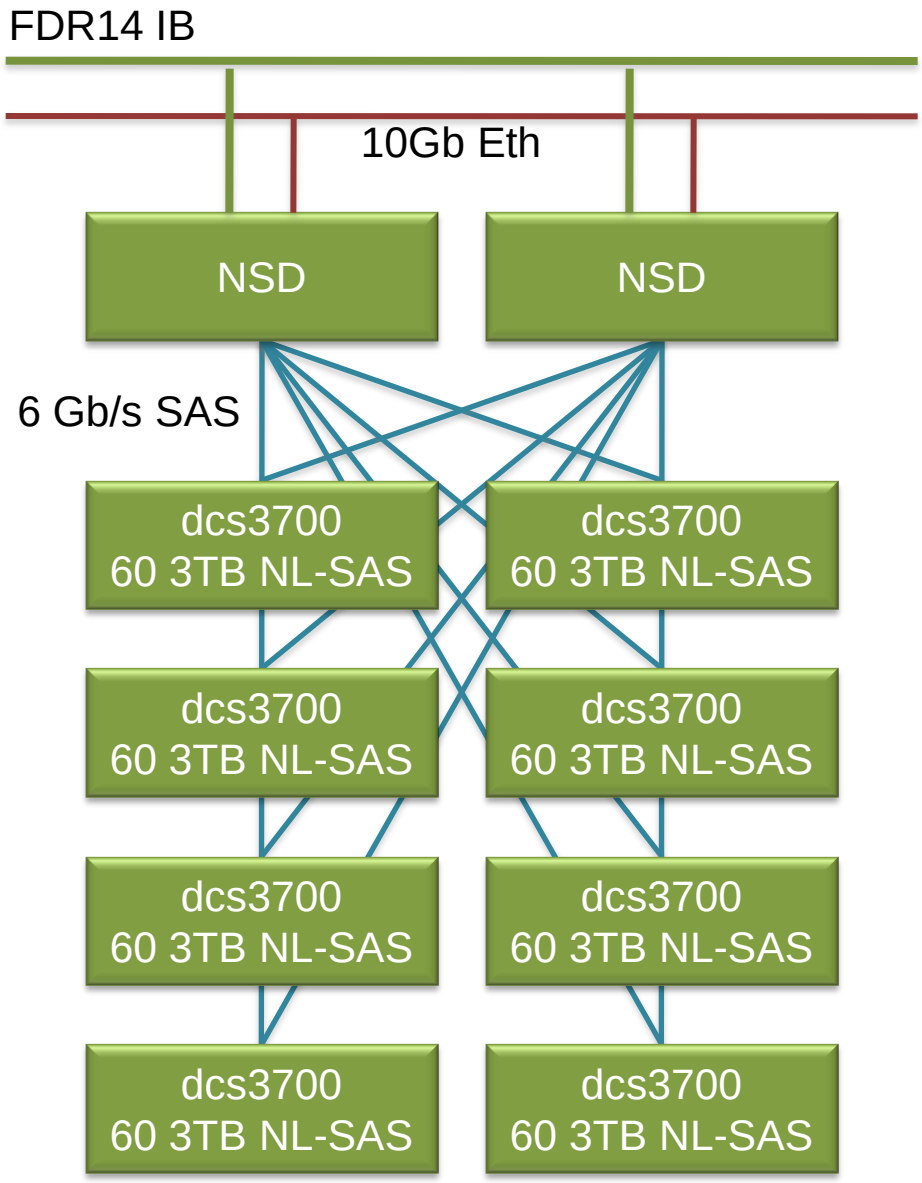


GLADE Overview



- 10.5 PB useable
  - + 6 PB useable, total 16.4 PB usable (2014)
- 76 DCS3700 systems
  - + 76 EXP3700 expansion drawers (2014)
- 4560 3TB drives
  - + 2280 3TB drives (2014)
- 20 NSD servers, 6 management nodes
- 2 InfiniBand management nodes
- 4 data mover nodes
- 1 108-port IB FDR 14 switch, 6 ethernet switches
- 21 racks

GPFS Building Block



Per Node Configuration

- 16 Cores, 64GB memory
  - 1 x IB FDR14 Card (6 GB/s)
  - 4 x 4 port SAS Card (6 Gb/s per lane)
    - 4 lanes per port
    - Up to 2.2 GB/s per port
  - 1 x 10Gb Card (1400 MB/s)

Per DCS3700 Configuration

- 4 x SAS (6 Gb/s), 4GB Cache

Streaming Rate per DCS3700

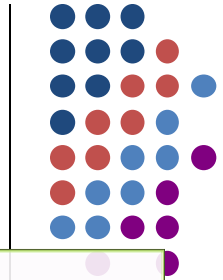
- write < 1.5 GB/s
- read < 2 GB/s

Streaming Rate Per Building Block *

- write < 12 GB/s
- read < 12 GB/s

* Building block aggregate performance is gated by the IB adapters.

GLADE Manager Nodes



glademgt1

- Primary xCAT cluster manager
- Secondary GPFS cluster manager, quorum node

glademgt2

- Secondary xCAT cluster manager
- Primary GPFS cluster manager, quorum node

glademgt3

- token manager, quorum node, file system manager

glademgt4

- token manger, quorum node, file system manager

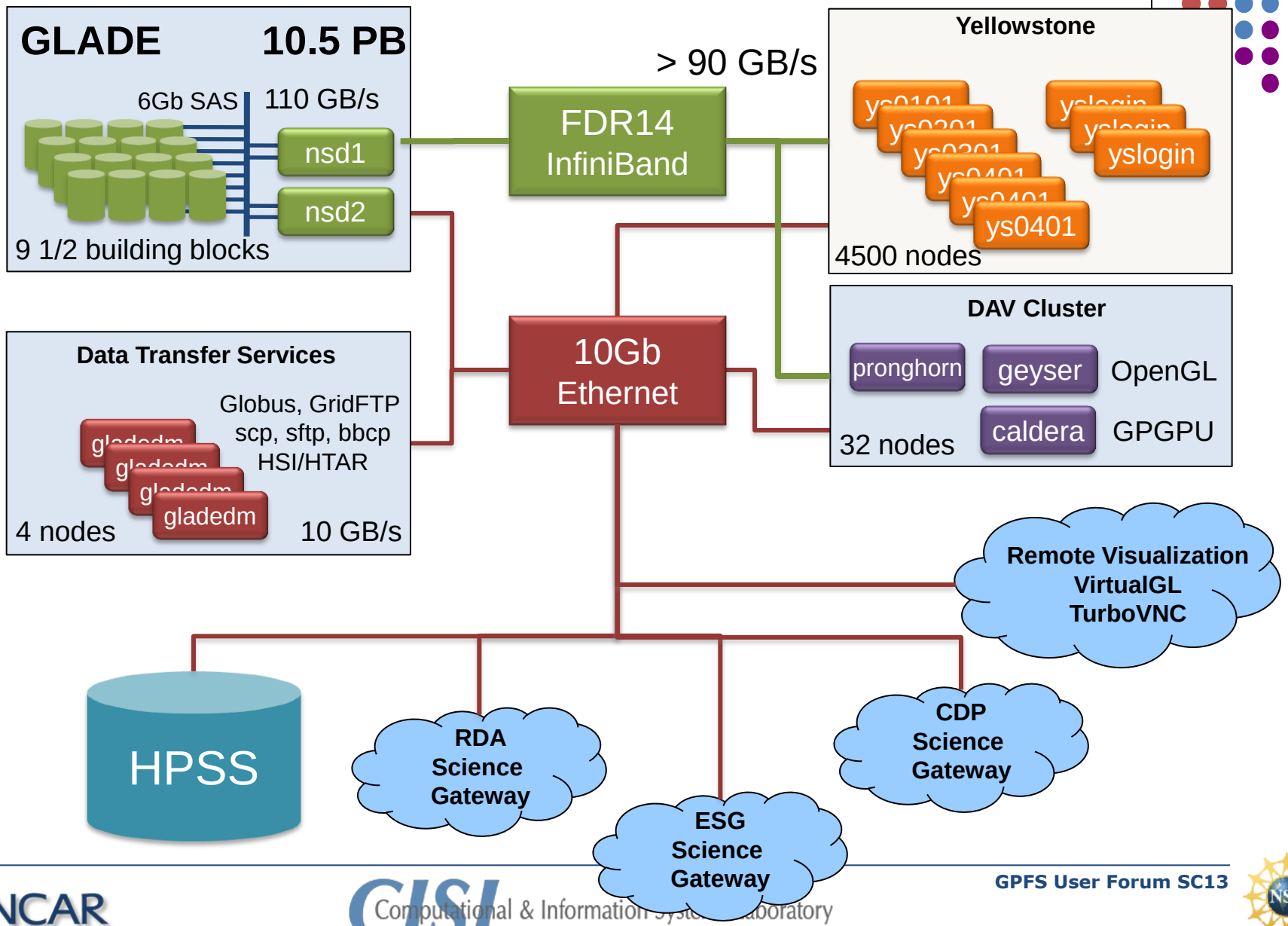
glademgt5

- Primary GPFS configuration manager
- token manager, quorum node, file system manager, multi-cluster contact node

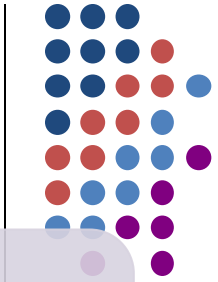
glademgt6

- Secondary GPFS configuration manager
- token manager, multi-cluster contact node

GLADE I/O Networks



GLADE File System Configurations



/glade/scratch

- 5 PB total space
- 4 MB block size
- 10 TB $SCRATCH per user
- 90 day purge policy

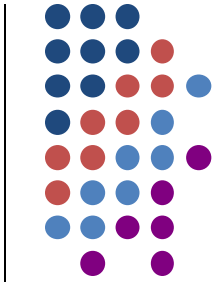
/glade/p

- 5 PB total space
- 4 MB block size
- 500 GB $WORK per user
- Allocated project spaces
- 2 PB allocated to data collections (RDA, ESG, CDP)

/glade/u

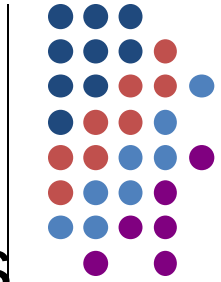
- 500 TB total space
- 512 KB block size
- 10 GB $HOME per user, 10 TB total, backed up
- Application software repository
- Special project allocations

Space Implementations



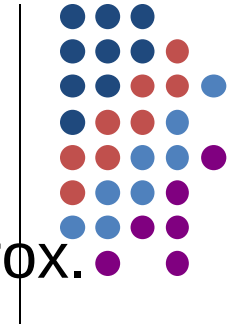
- $HOME implemented within a fileset
  - quotas enabled within the fileset per user
  - trying to use fileset level snapshots
- $WORK implemented within a fileset
  - quotas enabled within the fileset per user
- $SCRATCH implemented as a full file system
  - quotas enabled per user
  - ILM rule used to set 90 day purge policy
- Project spaces implemented within filesets

Space Allocation Rules



- $HOME, $WORK, $SCRATCH directories are created using the username
  - accounting provided per user per space
- Project spaces are created using the project name
  - access is controlled using a group matching the project name
  - accounting provided per user per project space

Storage Accounting Process



- Accounting process runs weekly and produces approx. 4,000 records
- Projects are charged for their allocation, accounting provides a record of how well they use their space
- Accounting Implementation
  - mmlsfileset
    - store the fileset name – which is also the group name
    - store the full path for project directories
  - mmrepquota – run per file system
    - pull all the USR records
    - sort per fileset
    - store the # of files and total space use for each user

Storage Accounting Record

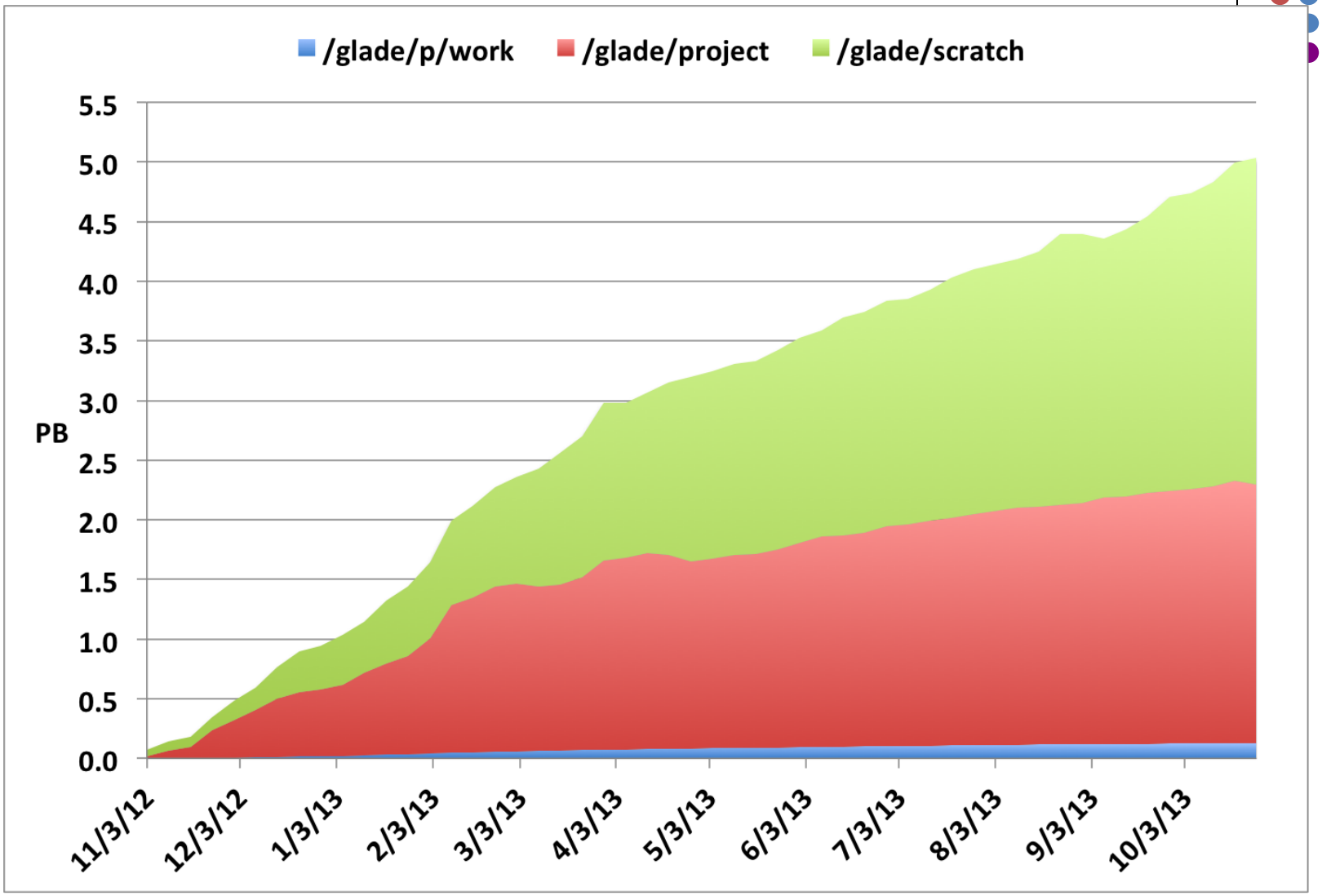
- Record Format

"eventtime", "projdir", "group", "user", "nofiles", "kbytesused", "period", "QOS”

"2013-10-25","/glade/p/ucol1413”,”ucol1413","pjg”,”162035", ”20670844288","7","0”

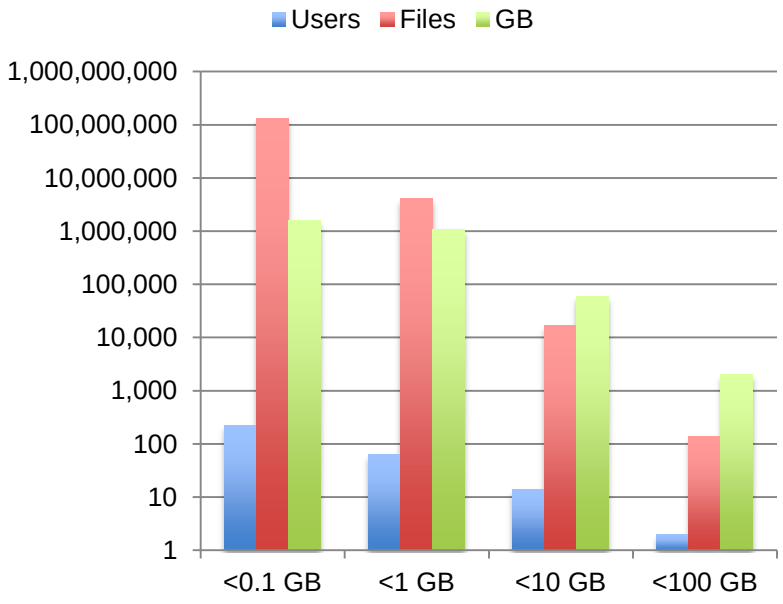
  - Event time – date record was collected
  - Project directory – generally the fileset name by policy
  - Group – Unix group, generally the fileset name by policy
  - Username
  - Number of files
  - kB used
  - Period – reporting interval, in days
  - QOS – quality of service (for future use)

GLADE Growth in 2013

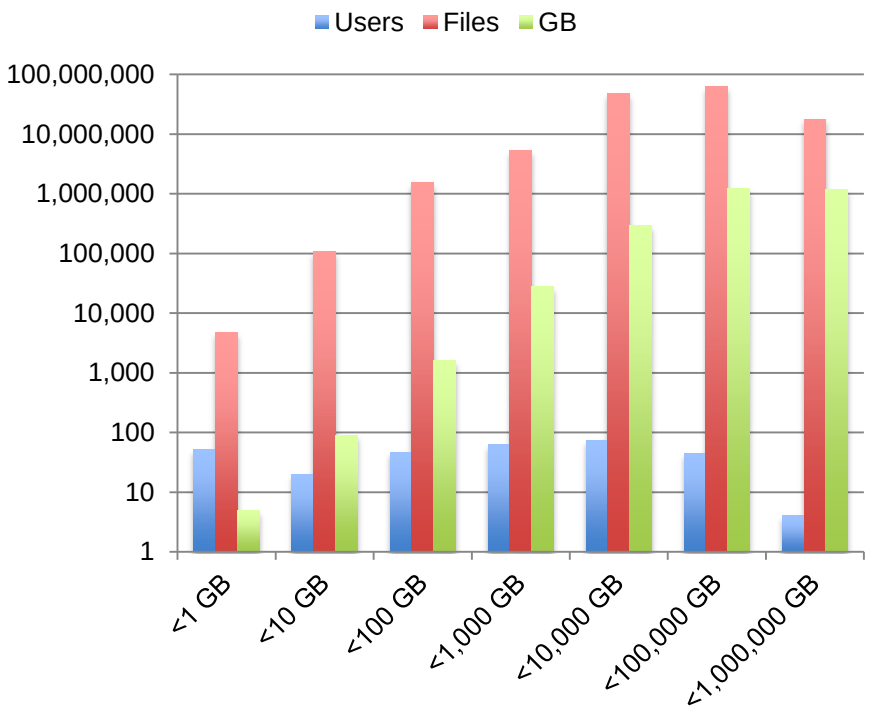


Profiling “Big Data”
GLADE Project Space Allocations

By average file size



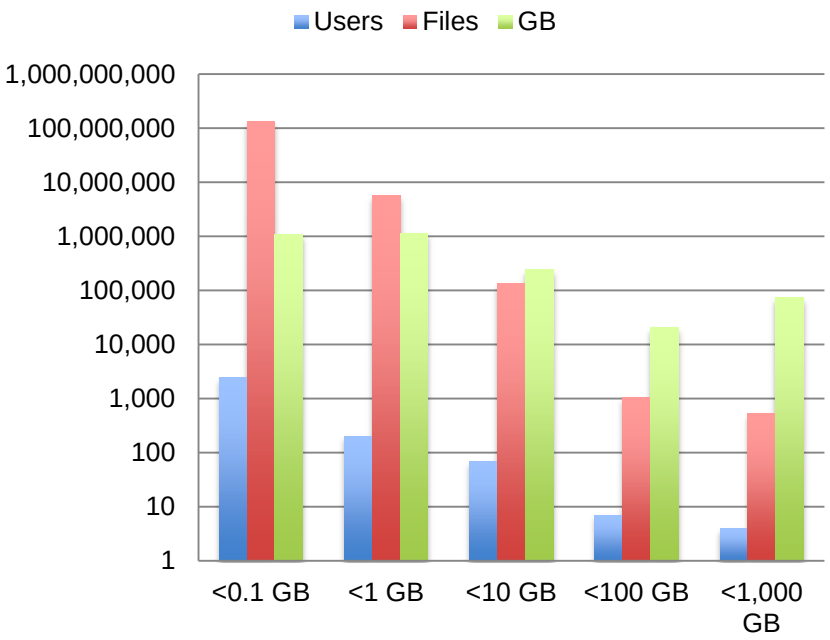
By GB stored



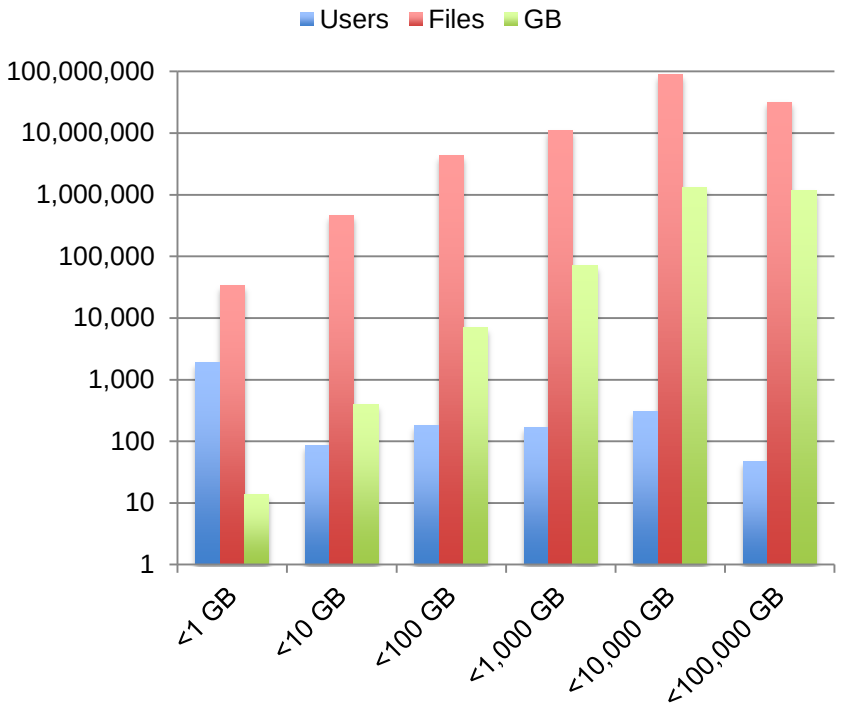
Average file size *vs.* Total data holdings

Profiling "Big Data"
GLADE Scratch Space Allocations

By average file size

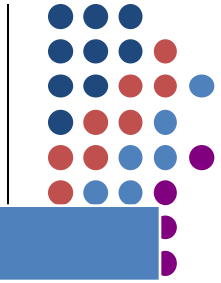


By GB stored



Average file size *vs.* Total data holdings

GPFS Monitoring Framework

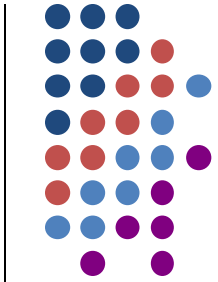


Component	Function
gpfsmond	Gathers GPFS status information from clients and forwards it to glademgt1.
gpfsmonitor	Collects the gpfsmond reports and stores them in the GPFS monitor database.
gprsmonproxy	Forwards gpfsmond reports from remote clusters to glademgt1.
gpfsreporter	Analyzes the recent entries in the monitor database and summarizes them to the web page and ganglia.
gpfsexpeller	Automates expelling and unexpelling nodes from GPFS.
Ganglia	Provides graphs of the summary reports, general I/O profiles and general system status.
Nagios	Provides operational monitoring and response procedures.
PostgreSQL	Provides storage for and analysis of the gpfsmond reports.

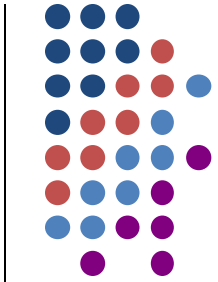
GPFS Monitoring Tools

- gpfsmond
  - runs on GPFS client nodes every 5 min and every 2 min on GPFS server nodes
  - checks the current GPFS status on the node
  - checks the VERBS status on the node
  - monitors the status of file systems mounts and will attempt a remount if necessary
  - GPFS attemps remounts on it's own in certain situations
- gpfsexpeller
  - staff can add a node to a list to be expelled or unexpelled
  - daemon runs every 5 mins to process the lists
  - keeps a list of currently expelled nodes

Report Examples



- GPFS Monitor Report
- Ganglia



pjg@ucar.edu

QUESTIONS?