NCAR Globally Accessible User Environment

Spectrum Scale User Group – UK Meeting
17 May 2016
Pamela Hill, Manager, Data Analysis Services

Data Analysis Services Group
NCAR / CISL / HSS / DASG

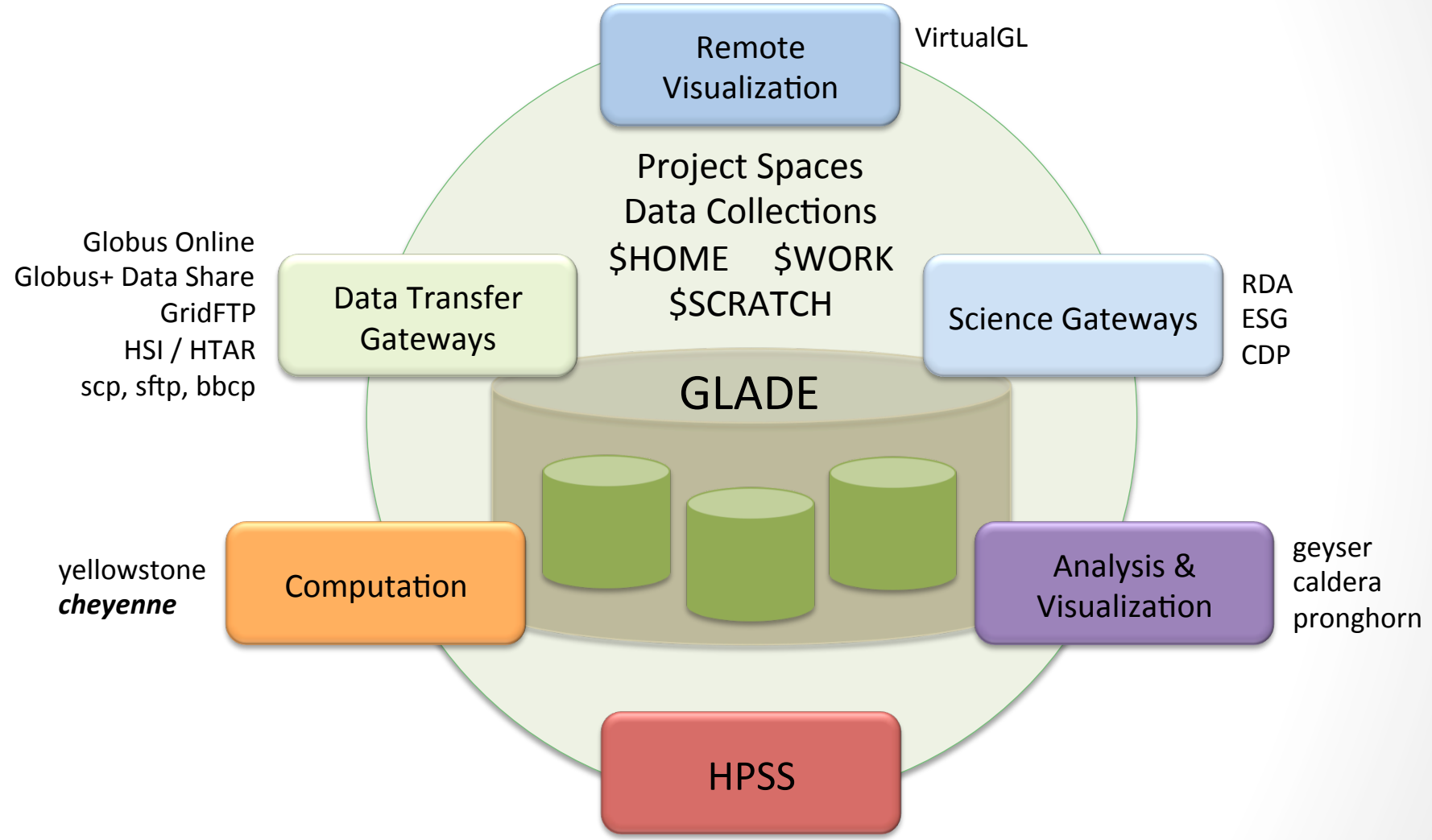
- Data Transfer and Storage Services
  - Pamela Hill
  - Joey Mendoza
  - Craig Ruff
- High-Performance File Systems
- Data Transfer Protocols
- Science Gateway Support
- Innovative I/O Solutions
- Visualization Services
  - John Clyne
  - Scott Pearse
  - Alan Norton
- VAPOR development and support
- 3D visualization
- Visualization User Support

GLADE Mission

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 RDA & ESG 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



IBM GPFS

DDN STORAGE

GLADE Today

- *90 GB/s bandwidth*
- *16 PB useable capacity*
- *76 IBM DCS3700*
- *6840 3TB drives*
  - *shared data + metadata*
- *20 GPFS NSD servers*
- *6 management nodes*
- *File Services*
  - *FDR*
  - *10 GbE*

Expansion Fall 2016

- *200 GB/s bandwidth*
- *~21.5 PB useable capacity*
- *4 DDN SFA14KX*
- *3360 8TB drives*
  - *data only*
- *48 800GB SSD*
  - *Metadata*
- *24 GPFS NSD servers*
- *File Services*
  - *EDR*
  - *40 GbE*

GLADE File System Structure

GLADE GPFS Cluster

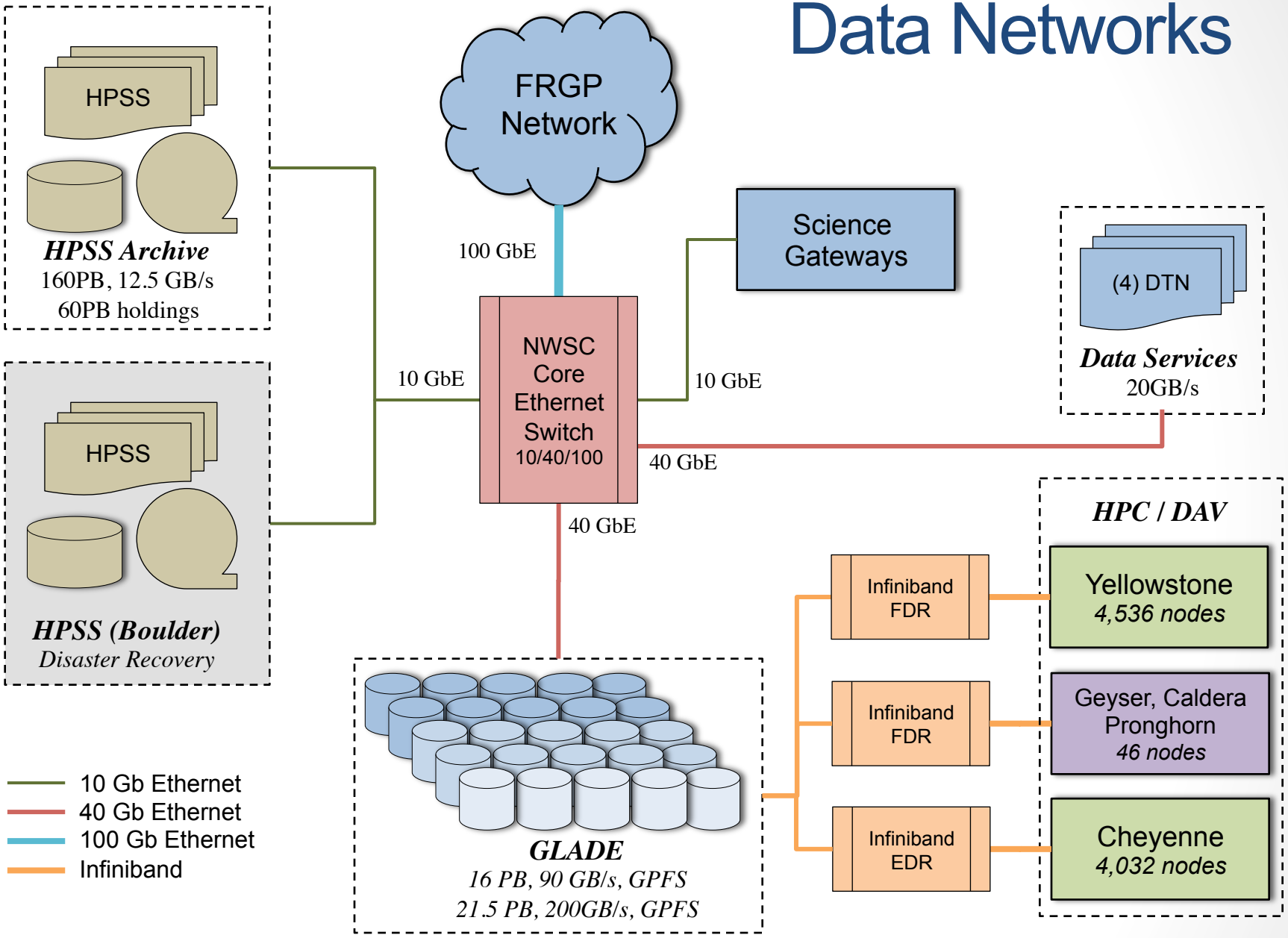
scratch	p	p2	home apps
SSD metadata	DCS3700 10 PB 4M block 512 byte inode	SSD metadata	SSD metadata
SFA14KX 15 PB 8M block, 4K inode		SFA14KX 5 PB 8M block, 4K inode	SSD applications
		DCS3700 5 PB	DDN 7700 100 TB 512K block, 4K inode

- ILM rules to migrate between pools
- Data collections pinned to DCS3700

GLADE I/O Network

- Network architecture providing global access to data storage from multiple HPC resources
- Flexibility provided by support of multiple connectivity options and multiple compute network topologies
  - 10GbE, 40GbE, FDR, EDR
  - Full Fat Tree, Quasi Fat Tree, Hypercube
- Scalability allows for addition of new HPC or storage resources
- Agnostic with respect to vendor and file system
- Can support multiple solutions simultaneously

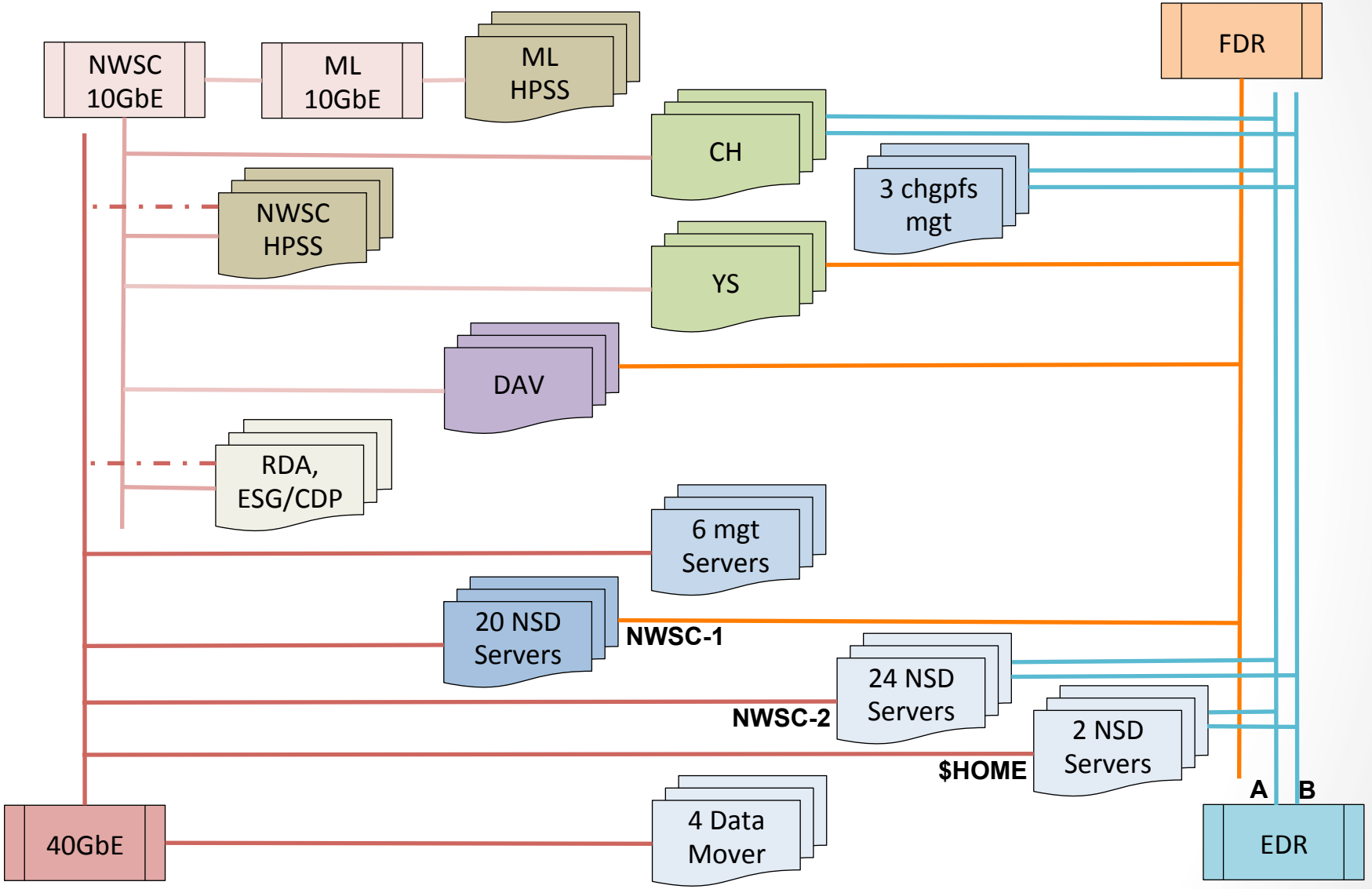
Data Networks



GLADE I/O Network Connections

- All NSD servers are connected to 40GbE network
- Some NSD servers are connected to the FDR IB network
  - Yellowstone is a full fat tree network
  - Geyser/Caldera/Pronghorn quasi fat tree, up/dwn routing
  - NSD servers are up/dwn routing
- Some NSD servers are connected to the EDR IB network
  - Cheyenne is an enhanced hypercube
  - NSD servers are nodes in the hypercube
    - Each NSD server is connected to two points in the hypercube
- Data transfer gateways, RDA, ESG and CDP science gateways are connected to 40GbE and 10GbE networks
- NSD servers will route traffic over the 40GbE network to serve data to both IB networks

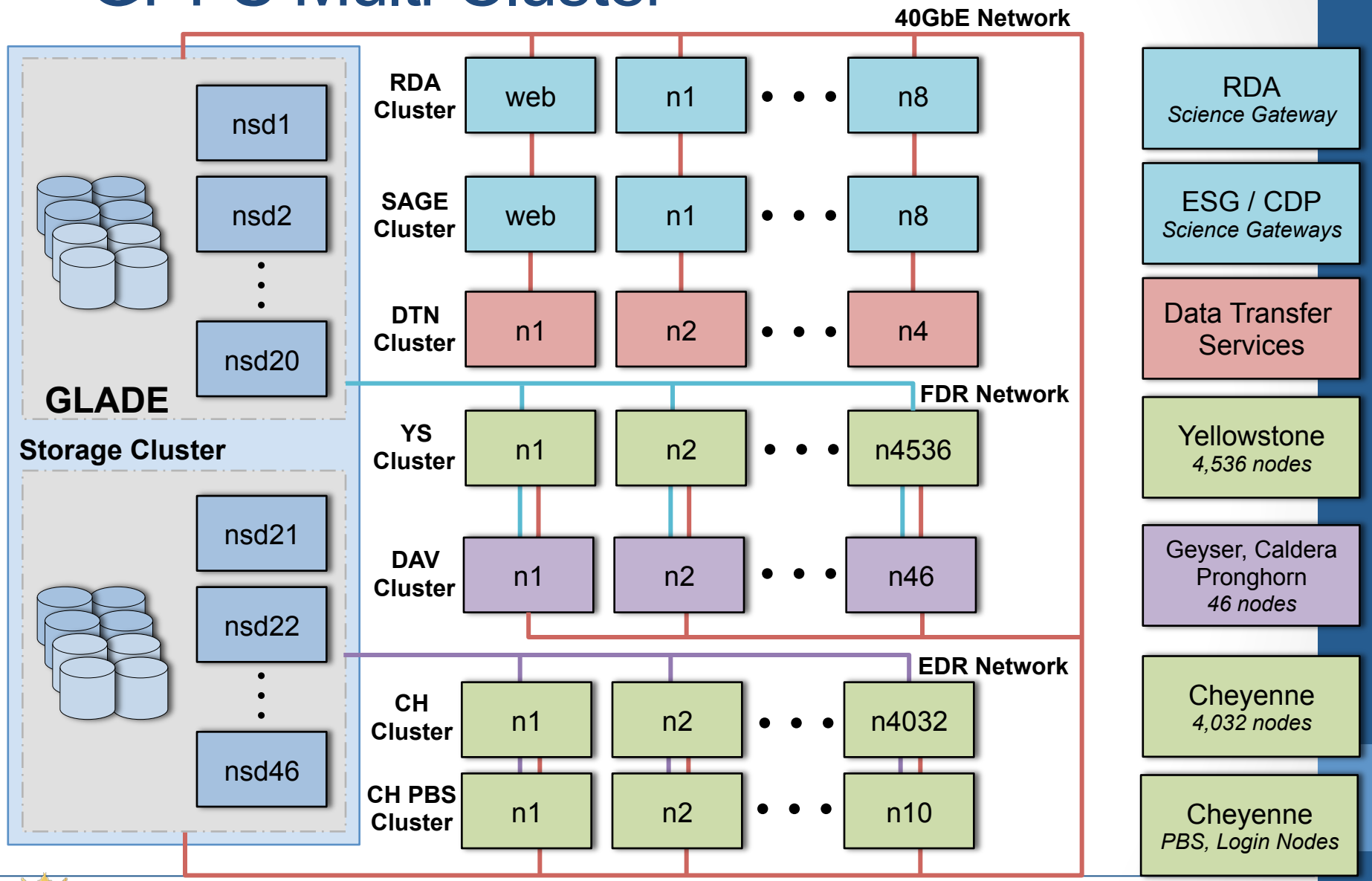
HPC Network Architecture



GPFS Multi-Cluster

- Clusters can serve file systems or be diskless
- Each cluster is managed separately
  - can have different administrative domains
- Each storage cluster controls who has access to which file systems
- Any cluster mounting a file system must have TCP/IP access from all nodes to the storage cluster
- Diskless clusters only need TCP/IP access to the storage cluster hosting the file system they wish to mount
  - No need for access to other diskless clusters
- Each cluster is configured with a GPFS subnet which allows NSD servers in the storage cluster to act as routers
  - Allows blocking of a subnet when necessary
- User names and group need to be sync'd across the multi-cluster domain

GPFS Multi-Cluster



NCARP – Requirements

The NCAR Common Address Redundancy Protocol

- Must provide IP routing between Infiniband-only GPFS clients and Ethernet-only GPFS servers.
  - Some GPFS protocol traffic is always carried over IP, even if RDMA is used for I/O
  - All GPFS protocol traffic can be carried over IP if necessary
- Need high availability of the virtual router addresses (VRAs) to provide the clients uninterrupted file system access
  - Routine maintenance
  - Unattended automatic VRA failure recovery
  - Use gratuitous ARP to reduce fail over time
  - Actively detect network interface failures
  - Use digital signatures for valid message detection and spoofing rejection

NCARP – Requirements

The NCAR Common Address Redundancy Protocol

- Make use of existing hardware where possible
  - Current need is for 1 year, then all NSD servers migrate to the EDR Hypercube network
- Be scalable if client cluster requirements change
  - Phase in/phase out of client clusters and cluster membership over time
  - There are thousands of client nodes in multiple clusters
  - Each router node has limited bandwidth available
    - IP over IB throughput is the limiting issue in testing
  - Partition nodes among the routers
    - Allows a measure of load balancing
    - Reduce the number of static rules in the Linux kernel routing table
    - Usable for both servers and clients

NCARP – Requirements

The NCAR Common Address Redundancy Protocol

- Must run on subnets managed by the organization's Networking Group as well as on the cluster private I/O networks.
  - Must not collide with Virtual Router Redundancy Protocol (VRRP) on the organization's subnets.
  - Must not appear as general purpose routers for non-GPFS traffic.
    - Will not support MPI job traffic
- Use a common configuration description for all participating networks and routers.

NCARP

The NCAR Common Address Redundancy Protocol

- NCAR Common Address Redundancy Protocol
- Based loosely on the BSD Common Address Redundancy Protocol (CARP)
- Reduce the number of NCARP packets traversing the networks to keep overhead lower
- Each node should handle a primary Virtual Router Address (VRA) and some number of secondary VRAs
- Use the node's load input when deciding to assume mastership of a secondary VRA.

NCARP – Deployment Strategy

The NCAR Common Address Redundancy Protocol

- Deploy NCARP on existing NSD server nodes.
  - CPUs usually mostly idle even under file I/O load
  - Spare memory and PCI bandwidth available
- Deploy dedicated inexpensive router nodes if bandwidth insufficient or impacts GPFS too much.
- Use 40G (or 100G) Ethernet as the interchange network between the router nodes and the GPFS servers.
  - Insulates us from IB dependencies and restrictions
  - Have to have this network anyway for non-IB attached GPFS clusters.

pjg@ucar.edu

QUESTIONS?