BLUE WATERS

SUSTAINED PETASCALE COMPUTING

2/28/13

PRAC Workshop 2/2013
Overall Storage Environments

BlueWaters

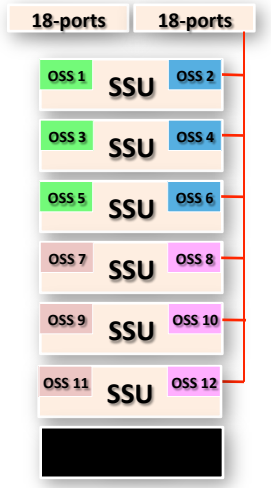

GREAT LAKES CONSORTIUM
FOR PETASCALE COMPUTATION

Overall Environment

- Lustre file system > 1TB/sec aggregately
  - Utilizing XIO nodes as Lnet routers to OSSs and OSTs
    - Direct connect and location dependent;
    - Kalyanna will cover more
  - Lustre 2.1+ on servers; clients 1.8.6
    - OSS Object Store Server – Lustre Server
    - OST Object Store Targe – disk devices on OSS
    - MDS Metadata server – separate metadata server
    - All data currently striped RAID 6 8+2

Cray BlueWaters data storage systems

- Lustre
  - Hardware
    - SonExion 1600 (OEM from Xyratex)
      - Self contained Lustre “appliance”
      - Xtremely small footprint
      - RAID6 protected; 2TB nearline SAS disks
      - Separate metadata environment
    - Direct connected through Cray XIO Lnets
    - Metadata through IB environment



SonExion 1600 rack

SonExion 1600

- scratch
  - >21 PBytes usable space
    - 180 Scalable Storage Units (SSU)
    - 14,400 2 TByte 7200 RPM NL-SAS disks = 28.8 PBytes raw storage
      - 360 global hot spares
  - 960 GBytes/sec. IOR performance
    - 480 LNET routers @ 2 GBytes/sec. for data
    - Two (2) additional LNET routers for metadata

SonExion

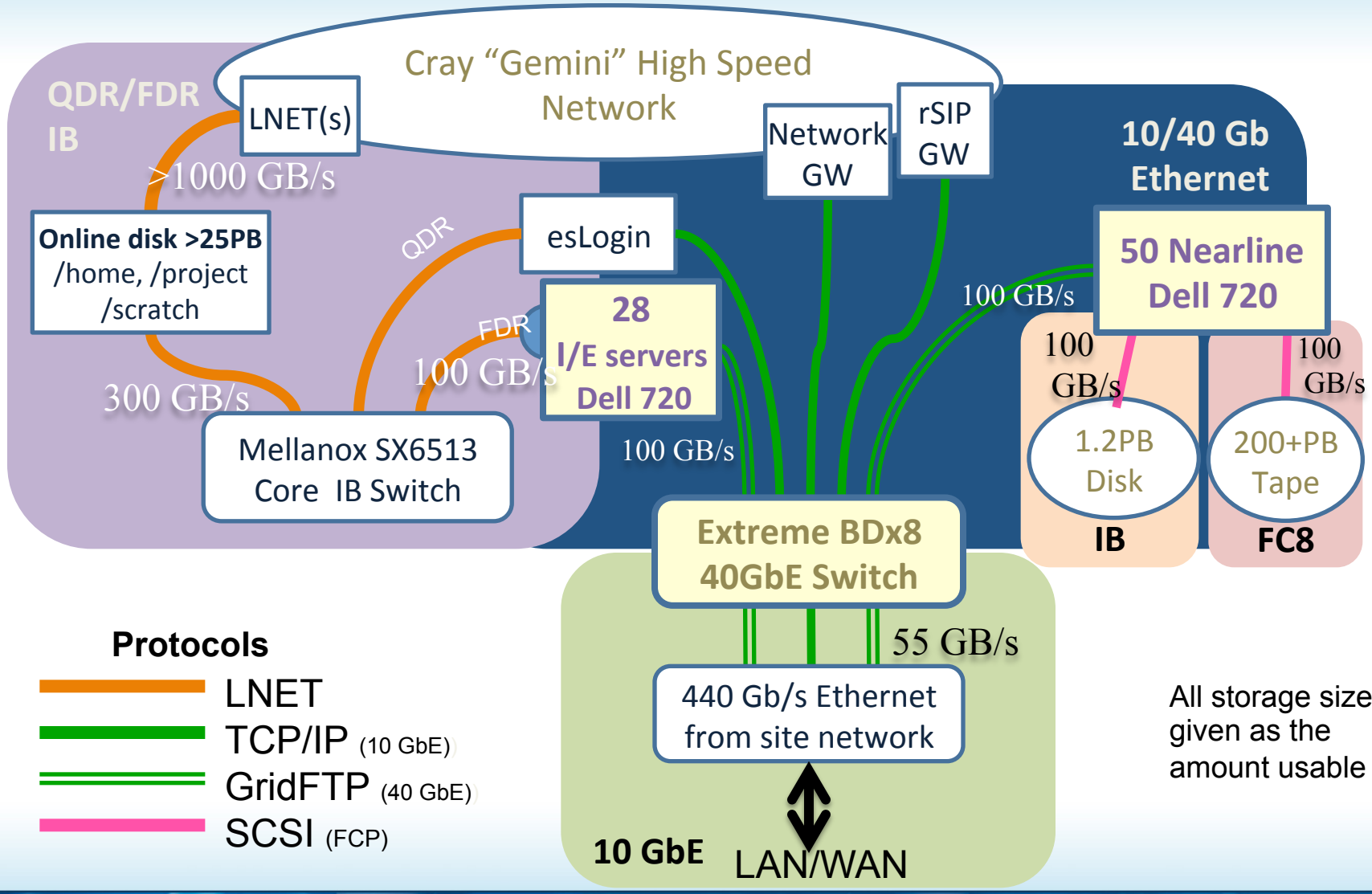
- /home & /project (Each file system)
  - 2 PBytes usable space
    - 18 Scalable Storage Units (SSU)
    - 1,440 2 TByte 7200 RPM NL-SAS disks = 2.9 PBytes raw storage
      - 36 global hot spares
  - 96 GBytes/sec. IOR performance
    - 48 LNET routers @ 2 GBytes/sec. for data
    - Two (2) additional LNET routers for metadata

Home & Project File System

- Blocksize 1M with default stripe depth 1 OST
- All user data backed up once every 24 hours
- Currently soft and hard quotas are being tested. Coming shortly
- home: directory hierarchy all user directories flat and same consistent quota level
- Project: directory hierarchy organized by project and PI owns “space and directory structure” and more quota can be requested.

Scratch File System

- Config 21.6PB
- Block 1M with default stripe 1 OST
- Not Backed up - traditional scratch for now
  - Managed file system on the horizon
- Users have soft and hard quotas
  - Coming shortly
- Directory hierarchy is flat
- Scratch purge will be in effect; won't purge files for 30 days



Protocols

- LNET
- TCP/IP (10 GbE)
- GridFTP (40 GbE)
- SCSI (FCP)

All storage sizes given as the amount usable

IE Functionality

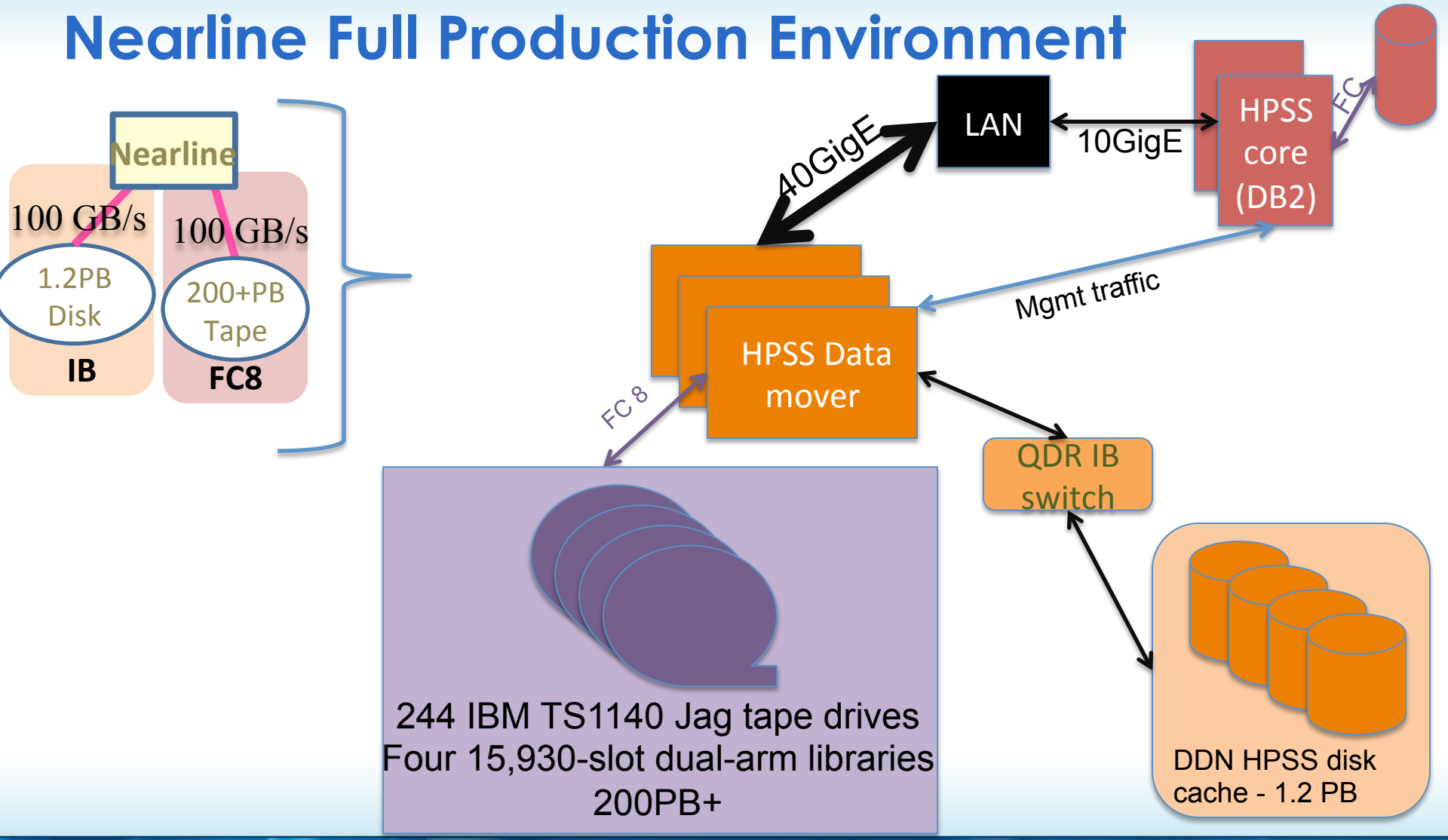
- Provide nodes with Lustre file systems mounted with good network connections to move data to and from Lustre to either HPSS or to outside world
- Using Globus-Online to provide service of managing the data movement.
  - Will be where we post all of our current changes for striping data into the Lustre and HPSS environment
  - GO will make the best decision for filesize for data striping and within the server and file system
- Provide at least 100GB/s out of the Lustre into HPSS. Other sites will not be able to transfer at that speed.

StopGap HPSS environment

- In Production today: ncsa#Nearline_gap

- HPSS system in place until end of 3/2013
  - 8 tape drives on 2 movers with 1 core server
  - When large system is deployed; NCSA will move data from gap to production environment

Nearline Full Production Environment



HPSS Enviornment

- HPSS 7.5 with RAIT on 50 movers
- 1.5PB(1.2PB useable) size disk cache connected over IB
- ~100GB/s aggregate throughput to tape drives
- HA core servers with 230TB of disk for databases
- 4 spectra-logic libraries with 240 TS1140 (Jag) drives
- 2013 2 additional libraries with 120 more tape drives
  - (6 libraries at 15930 slots *4TB = 383TB raw yeilding ~300PB useable)
- Nodes running HPSS, gridftp, RAIT, and disk and tape devices

Questions?