

Welcome to the PVFS BOF!

Rob Ross, Rob Latham, Neill  
Miller

Argonne National Laboratory

Walt Ligon, Phil Carns  
Clemson University



# An interesting year for PFSs

- At least three Linux parallel file systems out there and in use now
  - PVFS
  - GPFS
  - Lustre
- Many “cluster file systems” also available
- Verdict is still out on what of these are useful for what applications
- Now we’re going to complicate things by adding another option :)

# Goals for PVFS and PVFS2

- Focusing on parallel, scientific applications
- Expect use of MPI-IO and high-level libraries
- Providing our solution to a wide audience
  - Ease of install, configuration, administration
- Handling and surviving faults is a new area of effort for us

Application

High-level I/O Library

MPI-IO Library

Parallel File System

I/O Hardware

# Outline

- Shorter discussion of PVFS (PVFS1)
  - Status
  - Future
- Longer discussion of PVFS2
  - Goals
  - Architecture
  - Status
  - Future
- Leave a lot of time for questions!



# PVFS1 Status

- Version 1.6.1 released in the last week
  - Experimental support for symlinks (finally!)
  - Bug fixes (of course)
  - Performance optimizations in stat path
    - Faster ls
    - Thanks to Axiom guys for this and other patches

# PVFS1 Future

- Community support has been amazing
- We will continue to support PVFS1
  - Bug fixes
  - Integrating patches
  - Dealing with RedHat's changes
- We won't be making any more radical changes to PVFS1
  - It is already stable
  - Maintain it as a viable option for production



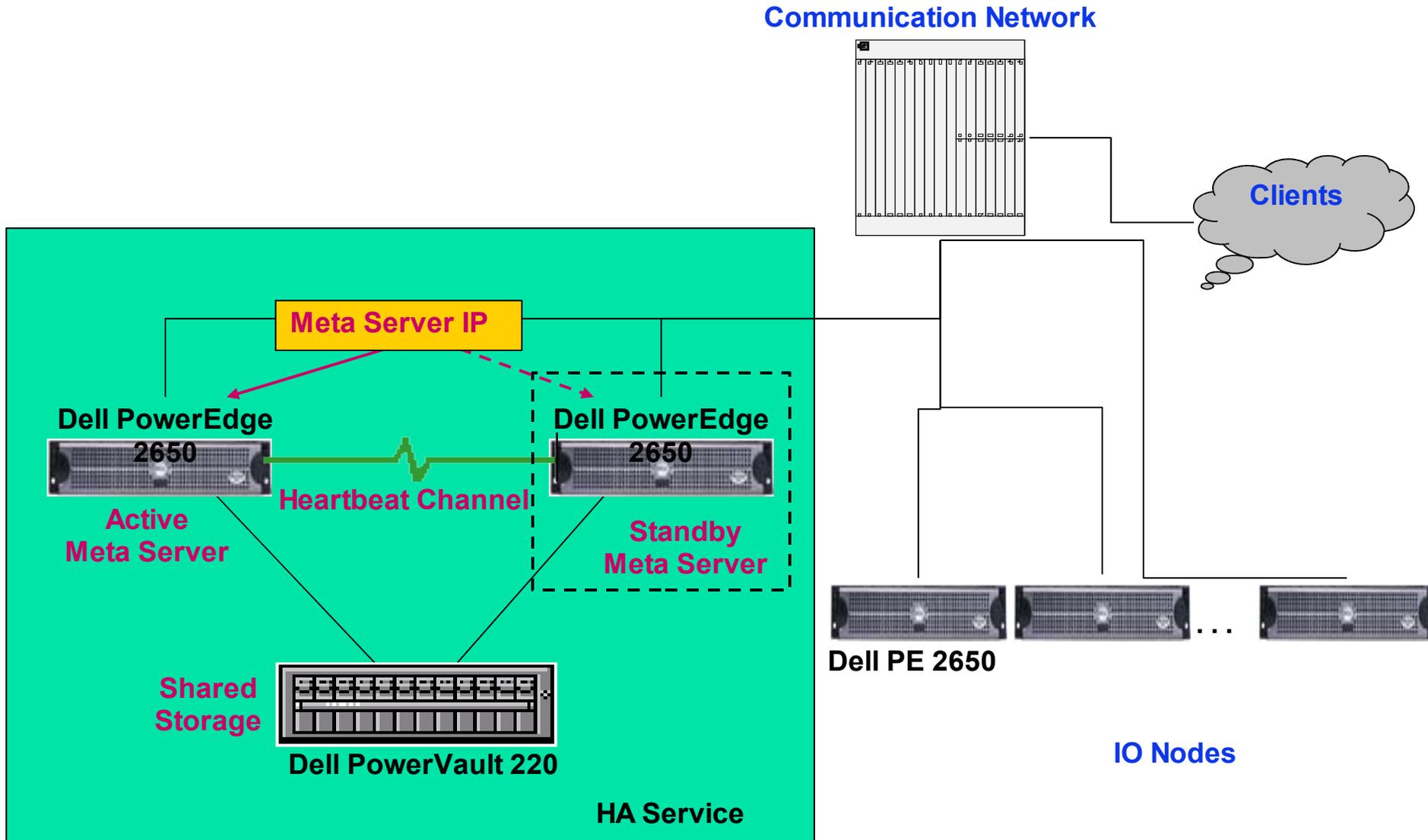
# Fault Tolerance!

- Yes, we do actually care about this
- No, it's not easy to just do RAID between servers
  - If you want to talk about this, please ask me offline
- Instead we've been working with Dell to implement failover solutions
  - Requires more expensive hardware
  - Maintains high performance

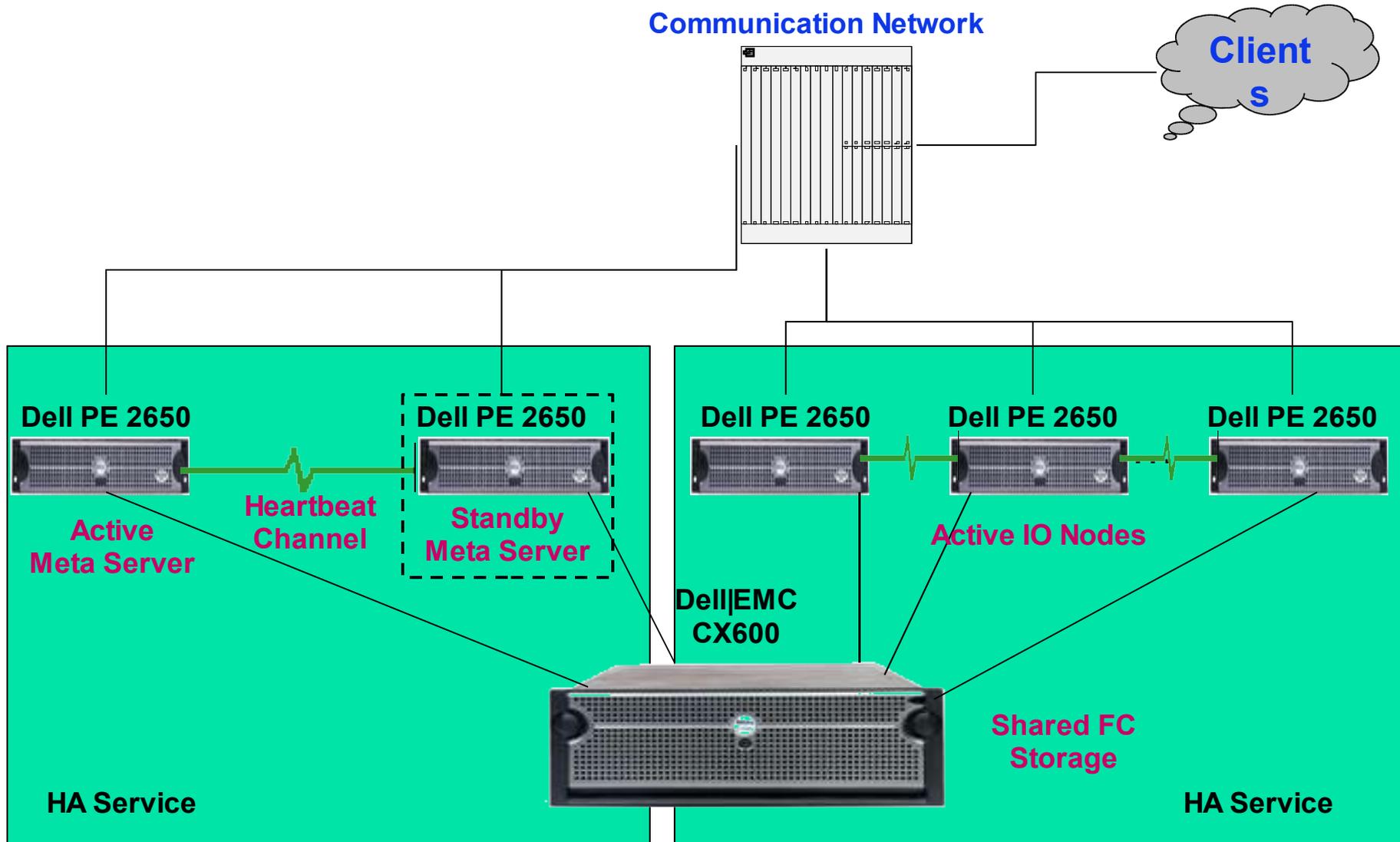
# Dell Approach

- Failover protection of PVFS meta server, and/or
- Failover protection of I/O nodes
- Using RH AS 2.1 Cluster Manager framework
  - Floating IP address
  - Mechanism for failure detection (heartbeats)
  - Shared storage and quorum
  - Mechanism for I/O fencing (power switches, watchdog timers)
  - Hooks for custom service restart scripts

# Metadata Server Protection



# Metadata and I/O Failover





# The Last Failover Slide

- These solutions are more expensive than local disks
- Many users have backup solutions that allow the PFS to be a scratch space
- We'll leave it to users (or procurers?) to decide what is necessary at a site
- We'll be continuing to work with Dell on this to document the process and configuration



# PVFS in the Field

- Many sites have PVFS in deployment
  - Largest known deployment is the CalTech Teragrid installation
    - 70+ Terabyte file system
  - Also used in industry (e.g. Axiom)
- Lots of papers lately too!
  - CCGrid, Cluster2003
  - Researchers modifying and augmenting PVFS, comparing to PVFS
- You can buy PVFS systems from vendors
  - Dell, HP, Linux Networx, others

# End of PVFS1 Discussion

---



Questions on PVFS1?