Argonne
NATIONAL
LABORATORY

... for a brighter future

U.S. Department
of Energy

UChicago ▶
Argonne LLC

A U.S. Department of Energy laboratory
managed by UChicago Argonne, LLC

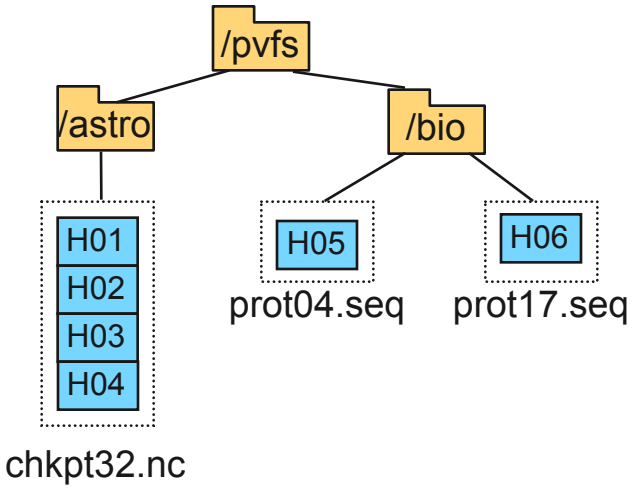
PVFS at 100 Teraflops

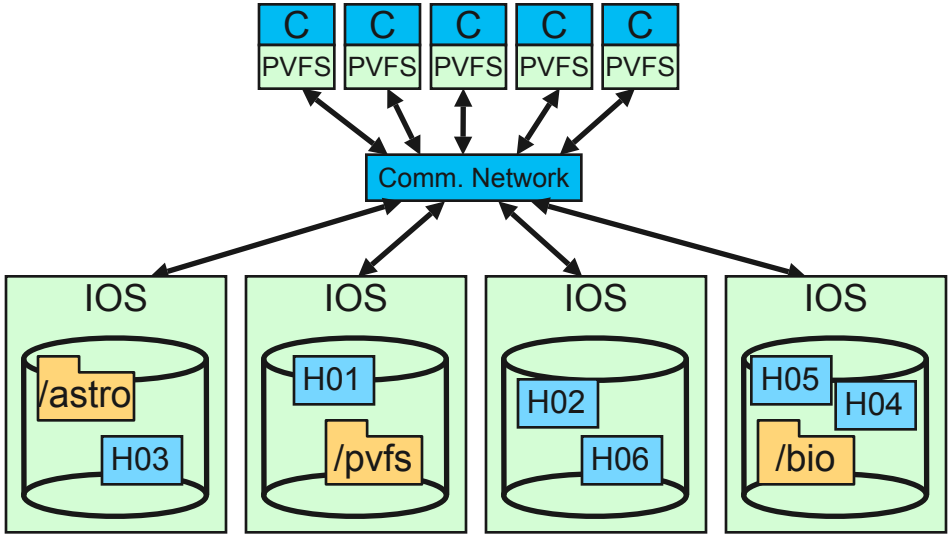
Talk Outline

- BGP System Overview
- PVFS Deployment on BGP
- BGP Unique I/O Challenges
  - I/O Forwarding
  - High-bandwidth networking
  - Fault tolerance
- Long term challenges and research

PVFS Overview

- Parallel File System
  - Asymmetric Client-Server Architecture
  - High-performance concurrent I/O from many clients
  - Servers manage the filesystem on local disk or shared storage
  - Tight integration with MPI-IO





The Growing PVFS Community

- R. Ross, S. Lang, R. Latham, P. Beckman, W. Gropp, R. Thakur, S. Coghlan, K. Yoshii, K. Iskra, and K. Harms
  Argonne National Laboratory
- P. Wyckoff and T. Baer
  Ohio Supercomputer Center
- W. Ligon and B. Settlemyer
  Clemson University
- M. Vilayannur
  Ex-ANL PVFS guru
- P. Carns and D. Metheny
  Acxiom Corporation
- B. Bode
  Ames Laboratory
- G. Gibson, M. Polte, and S. Patil
  Carnegie Mellon University
- T. Ludwig, J. Kunkel, and D. Buettner
  University of Heidelberg
- X.-H. Sun and S. Byna
  Illinois Institute of Technology
- G. Grider and J. Bent
  Los Alamos National Laboratory
- P. Honeyman
  University of Michigan
- P. Geoffray and S. Atchley
  Myricom Corporation
- A. Choudhary and A. Ching
  Northwestern University
- D.K. Panda
  Ohio State University
- A. Malony and A. Nataraj
  University of Oregon
- J. Nieplocha, J. Piernas-Canovas, and E. Felix
  Pacific Northwest National Laboratory
- L. Ward, R. Klundt, and J. Schutt
  Sandia National Laboratories

OSC Innovations in computing, networking, and education

CLEMSON UNIVERSITY

ACXIOM REDISCOVER YOUR POTENTIAL

center for information technology integration

THE OHIO STATE UNIVERSITY

Los Alamos NATIONAL LABORATORY EST. 1943

Sandia National Laboratories

NORTHWESTERN UNIVERSITY

Carnegie Mellon

Myricom

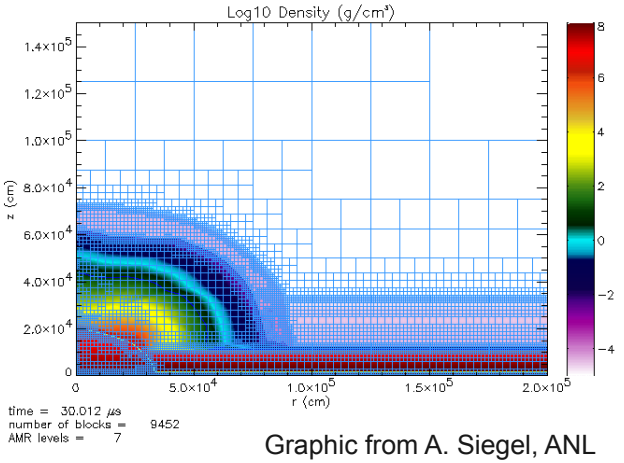
AMES LABORATORY

ILLINOIS INSTITUTE OF TECHNOLOGY
Transforming Lives. Inventing the Future. www.iit.edu

UNIVERSITY OF OREGON

Pacific Northwest National Laboratory
Operated by Battelle for the U.S. Department of Energy

BGP Example Apps



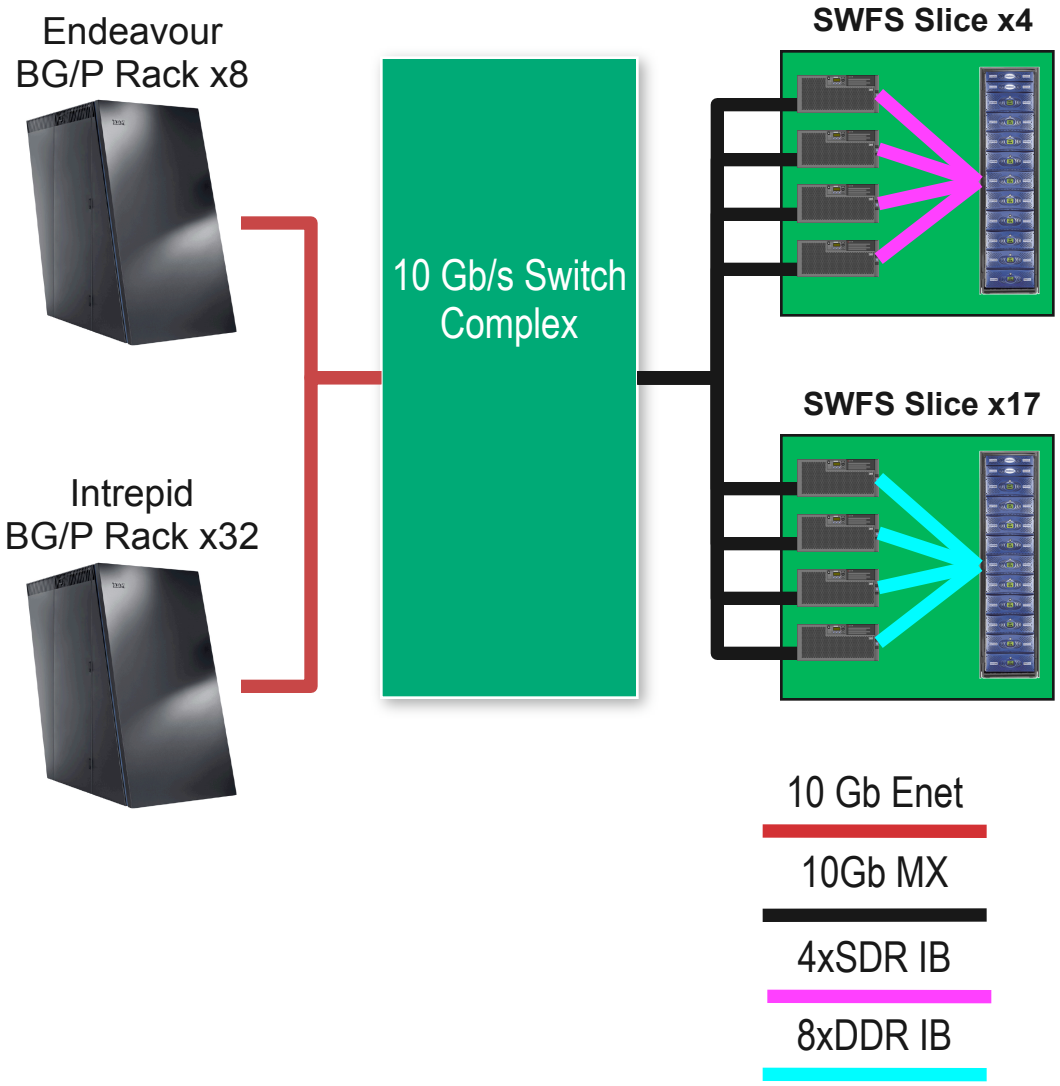
Graphic from A. Siegel, ANL

- FLASH: Astrophysics simulations
  - Writes large blocks of type variables from all nodes
  - Checkpoint 15% of RAM around every 30 minutes
  - Bursty writes of ~ 6TB, with expected bandwidth at ~ 50-100 GB/s
  - I/O using pNetCDF and MPI-IO

- QCD: Computations on particle physics results
  - Large contiguous and non-contiguous reads and writes of data
    - *~ 10 TB per run*
    - *100s of files*

BGP System Overview

- Endeavour (100T)
  - 8K CN = 16TB RAM
  - 4 SANs (8.8 GB/s)
  - 16 File Servers:
    - *192 GB RAM*
    - *40 GB/s Myrinet*
  - 512 TB raw storage
- Intrepid (500T)
  - 32K CN = 64TB RAM
  - 17 SANs (78 GB/s)
  - 68 File Servers:
    - *816 GB RAM*
    - *170 GB/s Myrinet*
  - 4.3 PB raw storage



BGP Single Rack

- 1024 Compute Nodes
  - 4 core 850MHz
  - 2GB memory
  - running IBM CNK
- I/O forwarded to I/O nodes
  - 1 I/O node per 64 CNs
  - Collective Tree Network 1.7GB/s
  - 10GigE NIC to storage
  - running Linux

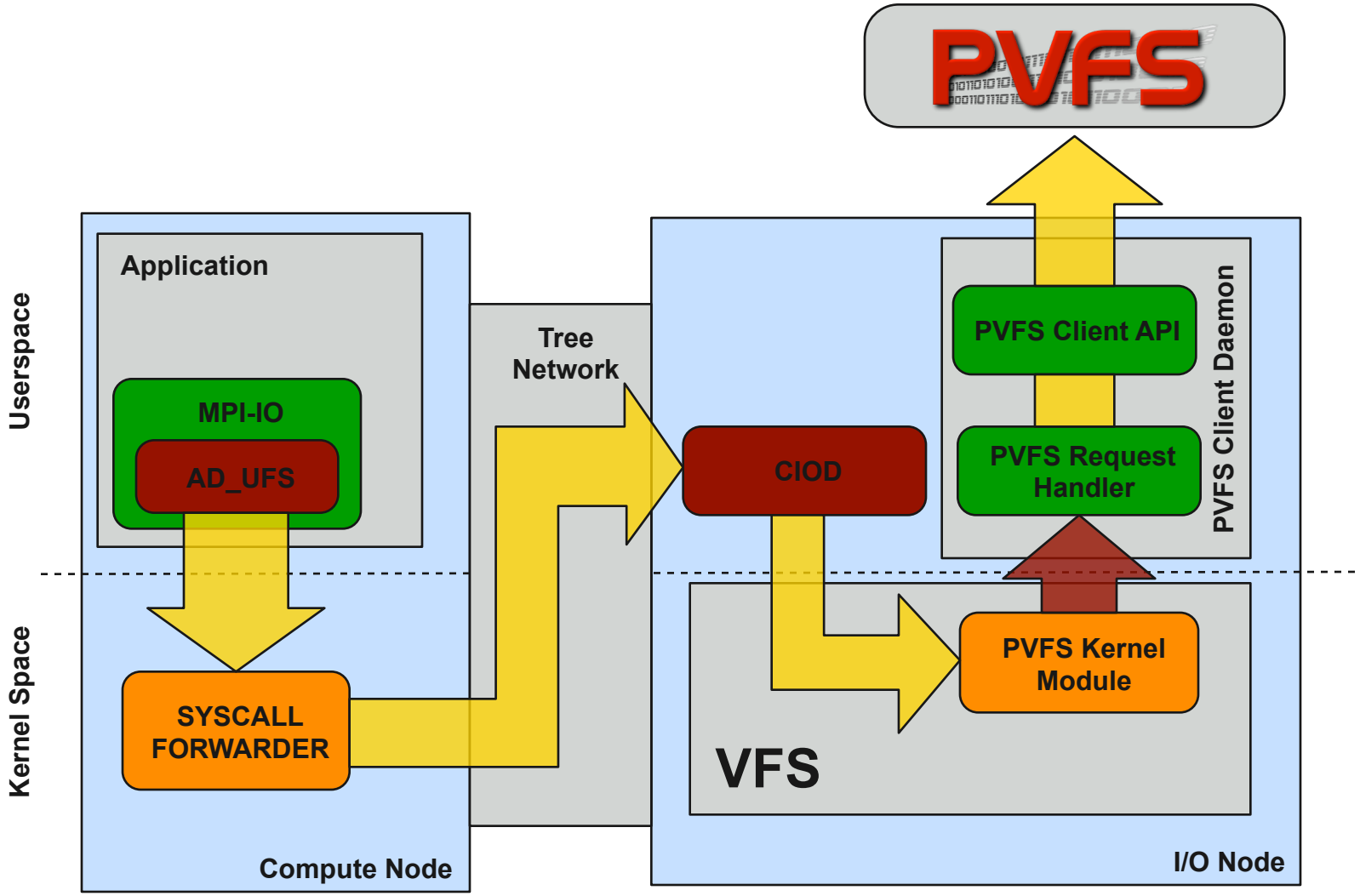
BGP PVFS Deployment

- Storage Nodes
  - SAN LUNs exported over IB to storage nodes
  - XFS filesystem mounts LUNs on storage nodes
- PVFS Servers
  - 2 Servers running on each x3655 storage node
  - Servers access shared storage over local XFS volume
- PVFS Clients
  - One client running on each I/O node, handles requests forwarded from compute nodes
  - IBM I/O forwarding daemon requires mounting PVFS
  - Communicate to PVFS servers over MX and Myrinet

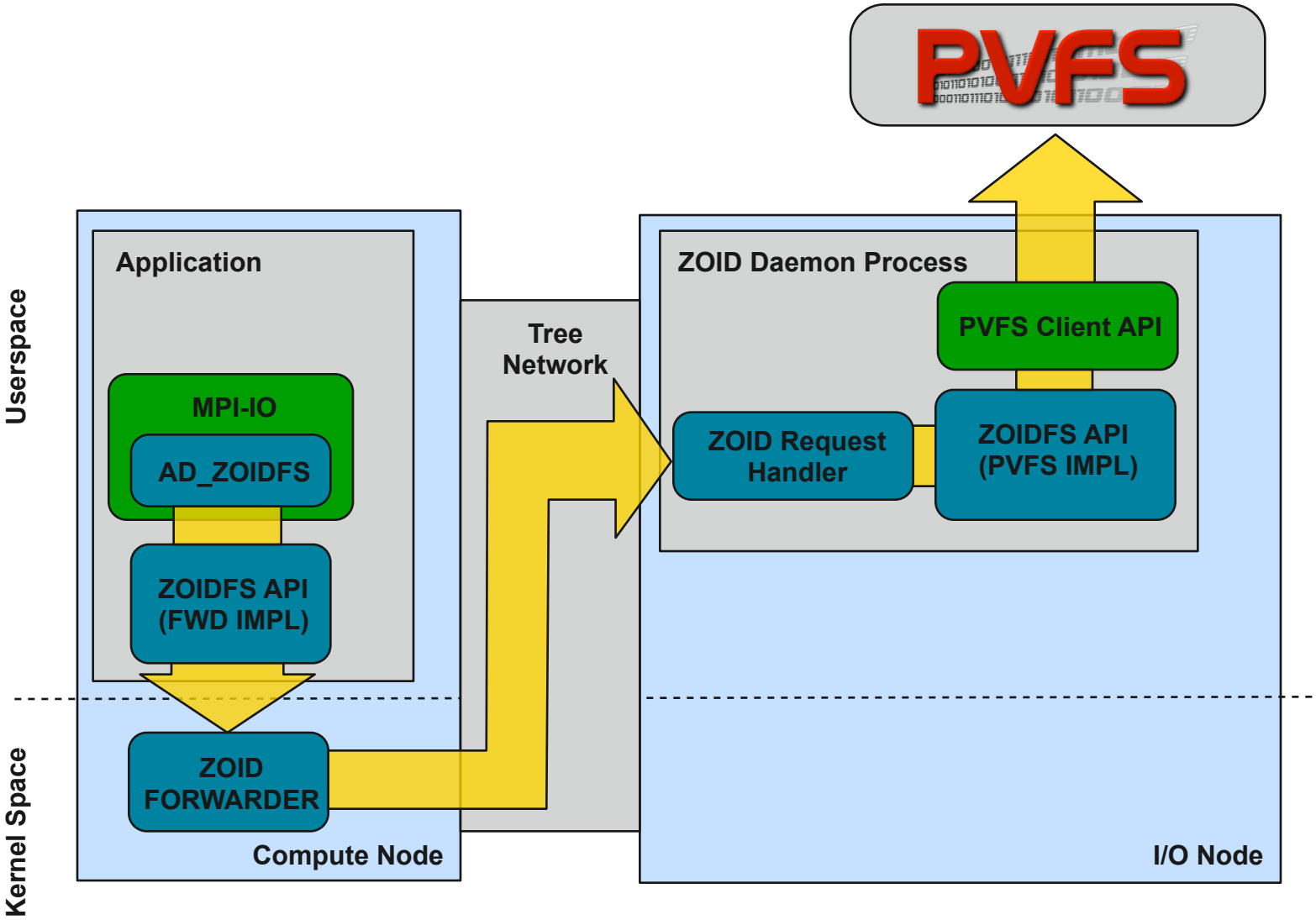
BGP Unique Challenges

- I/O is forwarded from CNs to IONs
  - Usual PVFS setup is PVFS clients on CNs
  - BGP architecture forwards I/O over the Collective Network
- High-speed networking
  - Large 10G switch complex to storage
  - 500 I/O aggregators
- Fault-tolerance
  - Source routed network

I/O Forwarding (without ZOID)



I/O Forwarding (with ZOID)



ZOID: Efficient I/O Forwarding

- Smarter forwarding daemon
  - Don't serialize requests
  - Plug into I/O calls directly
- ZOIDFS: Standardize I/O forwarding interface
  - Can be used on both compute and I/O side
  - Provide vectored I/O interface
  - Allow for hints, stateless file references
- Components:
  - MPI-IO driver for ZOIDFS
  - ZOIDFS to PVFS driver for ZOID daemon
  - ZOIDFS marshaling driver
  - Requires implementing a ZOID tree driver
    - *Linux implementation for Linux compute node kernel*
    - *IBM CNK source available*

PVFS over High Speed Networks

- TCP has known problems at scale
- Especially true for BGP I/O node
- Myricom's OpenMX allows
  MX over Broadcom 10GigE NIC
- BMI-MX module for PVFS
  - written by Myricom's Scott Atchley
- Cray Portals
  - written by Pete Wyckoff

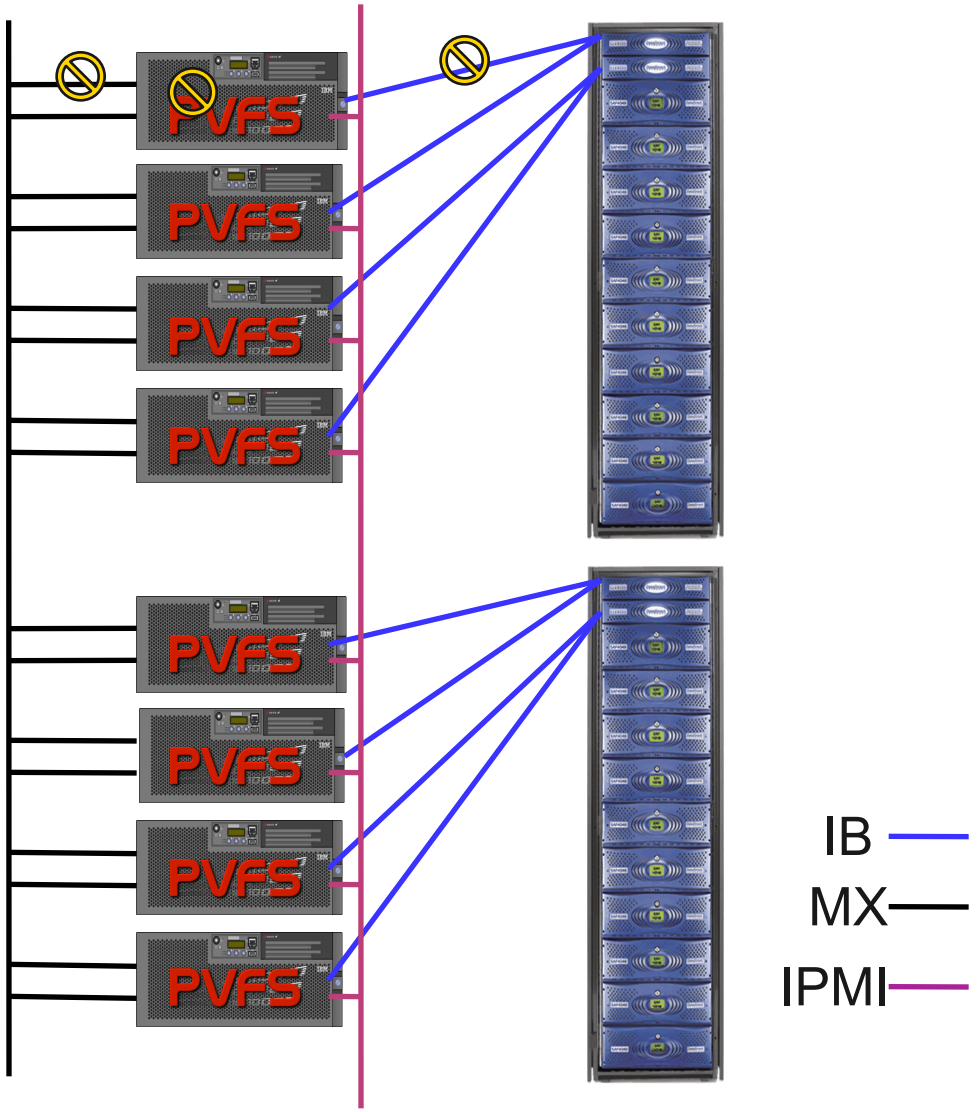
PVFS Fault Tolerance on BGP

- Hardware Support
  - DataDirect SANs provide:
    - *Disk-level redundancy (RAID6)*
    - *Online disk rebuilds*
    - *Shared access from multiple storage nodes*

- PVFS Support
  - Object-based storage allows painless fail-over from one storage node to another
  - Multiple PVFS servers per storage node can be managing different storage areas on the same SAN
  - Address failover support in PVFS clients
  - Clients are stateless, integrate well with standard HA deployments

PVFS Fault Tolerance on BGP

- Failover group of 4 servers
- Failure points:
  - Myri10G NIC
  - x3655 server node
  - IB NIC
  - DDN controller



Client Failover in PVFS

- Standard IP failover: Use IP address aliasing
  - requires ARP update at the switch
- Myrinet is source routed
  - PVFS clients need to provide address failover
- PVFS needs Multi-address support in PVFS
  - Modified config file endpoint format
    - *old:*

```
Alias hosta mx://hosta:0:3
```

    - *new:*

```
Alias hosta mx://hosta:0:3 mx://hostb:0:4 mx://hostc:0:5 mx://hostd:0:6
```

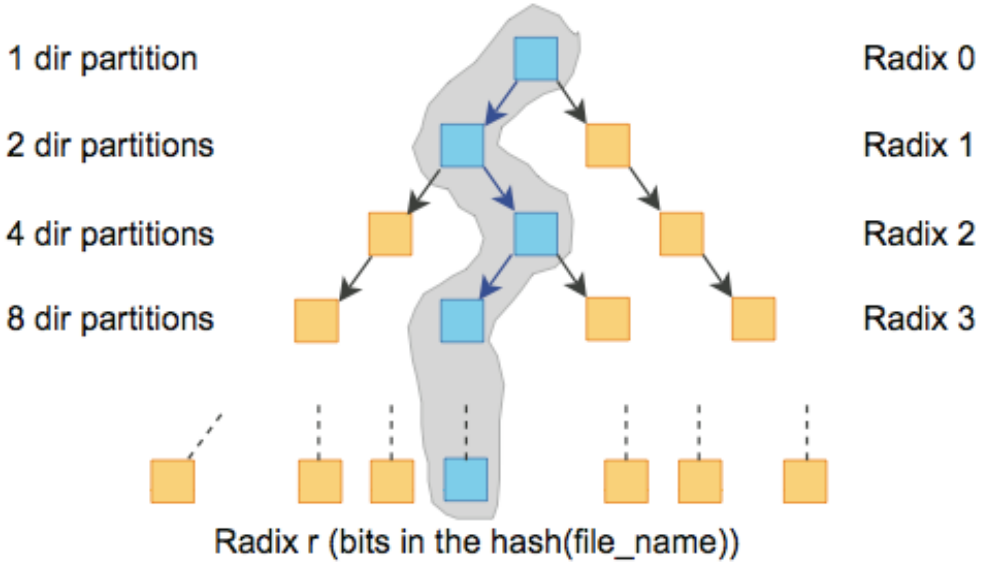
  - Clients timeout on address, try the next

PVFS Long-term Challenges

- Storage architectures have greater complexity
  - Solid-state disks
  - Huge memory caches
  - Multi-core SMP systems
- Applications want to create billions (even trillions) of files
  - Scaling issues with directories with millions of entries
- Small files
  - Creating many at once
  - Striping isn't efficient
- Searching and finding data in multi-petabyte storage
  - POSIX interfaces don't work so well
- Monitoring and optimization challenges with larger systems

PVFS Research

- Event driven PVFS servers: coalescing, lock-free queuing
- Small file support: Inode stuffing, Pre-allocation, storage levels
- Giga+ Directories
  - CMU: Garth Gibson, Swapnil Patil, Milo Polte
  - billions of entries in a directory
  - 100K entries/sec
- Multi-dimensional extensions
  - Milo Polte, John Bent
  - Allow SQL-style searching within PVFS
- Complete path tracing and visualization



Conclusion and Info

- PVFS: Deploys easily to increasingly larger scaled systems
- PVFS: Integrates well with I/O forwarding system
- PVFS: Good platform for Parallel I/O research
- Web site: http://www.pvfs.org/