# GridNFS: Scaling to Petabyte Grid File Systems

Andy Adamson
Center For Information Technology Integration
University of Michigan

# What is GridNFS?

- GridNFS is a collection of NFS version 4 features and minor versions with supporting daemons, data bases, and tools that enables NFSv4 to integrate with emerging Grid technologies to provide petabyte scale file service for the Grid.
- NFSv4 is the IETF standard for distributed file systems that is designed for security, extensibility, and high performance.

# Outline

- NFSv4 Feature Overview
- Global Name Space
- Parallel NFS (pNFS)
- Scaling Issues
- The Big Picture
- Status

# NFSv4 Features

- The only IETF Distributed File System
  - Minor versions
- Joins Windows and UNIX semantics
- State
  - OPEN/CLOSE: share and deny access
  - Byte range locking
  - Delegations

# NFSv4 Features

- Security is a required feature
- RPCSEC_GSS: adds the GSSAPI to RPC
  - Kerberos V
  - SPKM3 (PKI; similar to SSL)
  - LIPKEY (like TLS)
- ACLs: modeled after Windows ACLs
  - Superset of POSIX ACLs

# NFSv4 Features

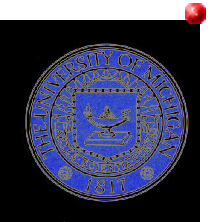
- NFSv4 Administrative Domain
  - Unique UID/GID mapping space
  - Can contain multiple DNS, NIS, Kerberos, PKI
  - v4domain; one of the DNS domains
- Names on the wire, not UID/GID
  - RPCSEC_GSS principals;  Kerberos principal@REALM
  - ACL names (UNIX owner/group); user@v4domain
  - Enables ACLs for foreign users on local files
  - Secure name to ID mapping required on server

# NFSv4 Features

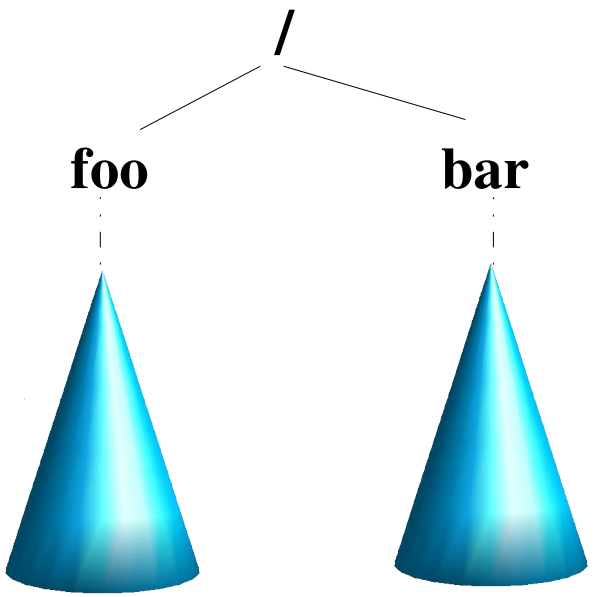
- Global name space can be constructed
  - Details currently being addressed
- /nfs/<nfsv4domain>/[local nfsv4 namespace]
  - <nfsv4domain> : DNS SVR record to give IP address(es) for the root of the nfsv4domain namespace server(s)
  - [local nfsv4 namespace] : Constructed using server Pseudo File Systems and the File System Locations attribute

# NFSv4 Server Pseudo File System

- NFSv2/3 clients explicitly mount server's exported file systems
- NFSv4 pseudo file system: read only file system to join exports on the server
- NFSv4 server mounts exports on pseudo file system leafs
- NFSv4 client can mount pseudo file system root.
- User traversal from pseudo file system into exported file system causes client 'under the cover' automatic mount.
- Access to exports via user credentials



# NFSv4 Server Pseudo File System




pseudo fs root

...

pseudo fs leafs

exported file systems mounted on pseudo fs leafs

# File System Locations Attribute

- fs_locations: GETATTR attribute that provides a list of servers where a file system resides
- Intended for replication, migration, and referral
  - Replication: data is on current server as well as other servers in list
  - Migration: data is no longer on current server, try another server on the list
  - Referral: data was never on current server, go to a server on the list

# File System Locations Attribute

- User traverses server pseudo file system into a referral node
- Server returns NFS4ERR_MOVED to client
- Client obtains fs_locations list, automatically redirects NFSv4 requests to new server

# Linux NFSv4 Server: fs_locations

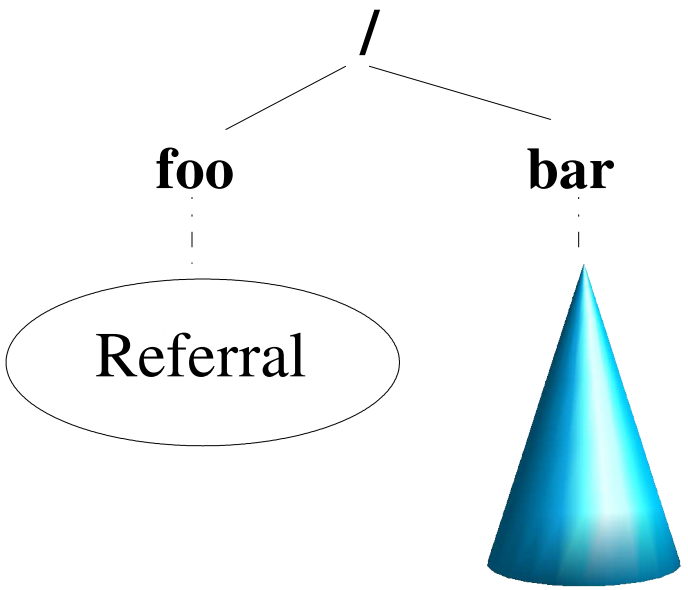
- New exports option:
  - fs_loc=”file locations list method"
- Methods include:
  - LDAP lookup
  - DNS TXT record
  - Flat file on server
- Graphic tool for name space construction under development

# NFSv4 Server Pseudo File System

pseudo fs root

...

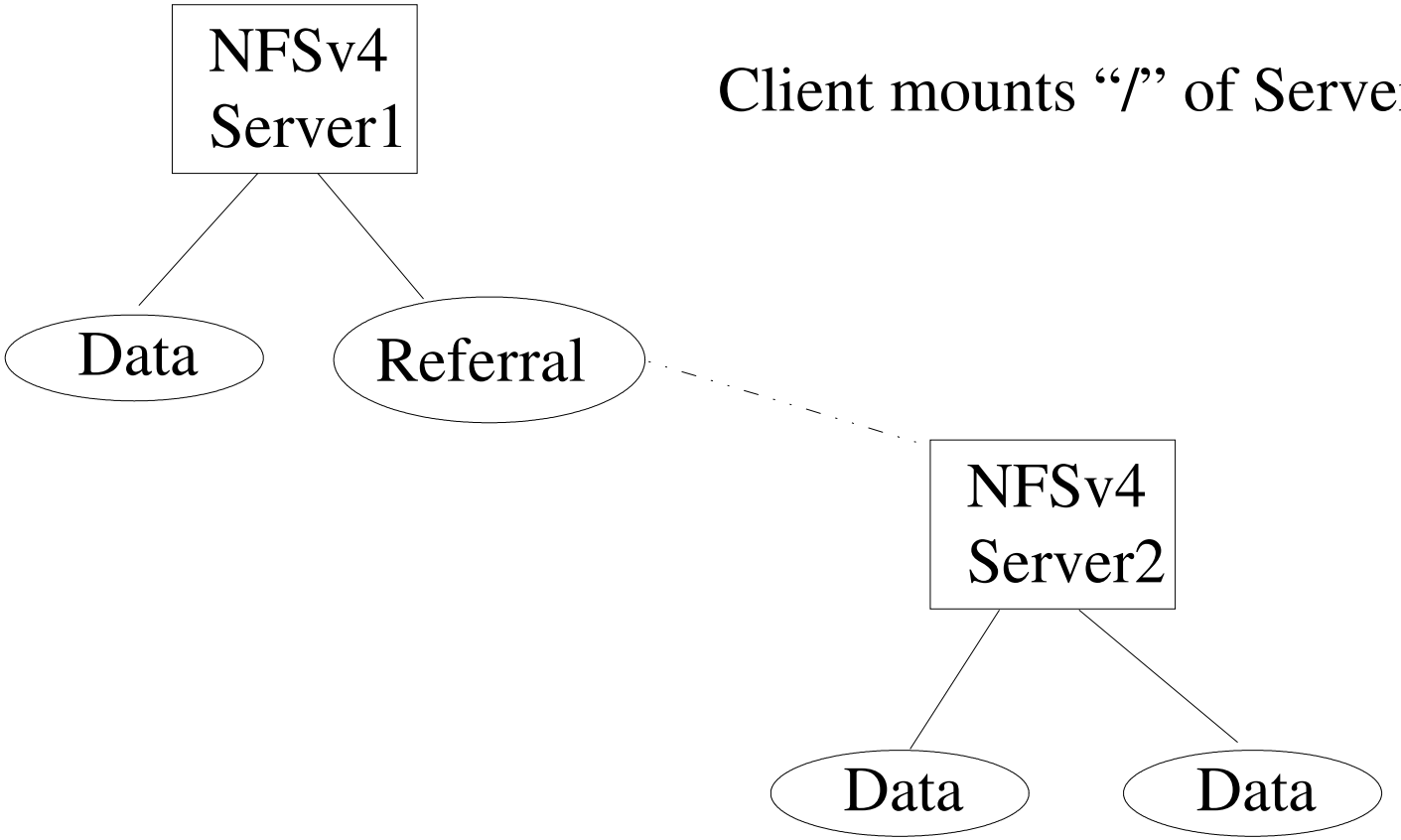
pseudo fs leafs




referral node, a single directory, mounted on pseudo fs leaf

exported file system mounted on pseudo fs leaf

# NFSv4 Name Space Construction



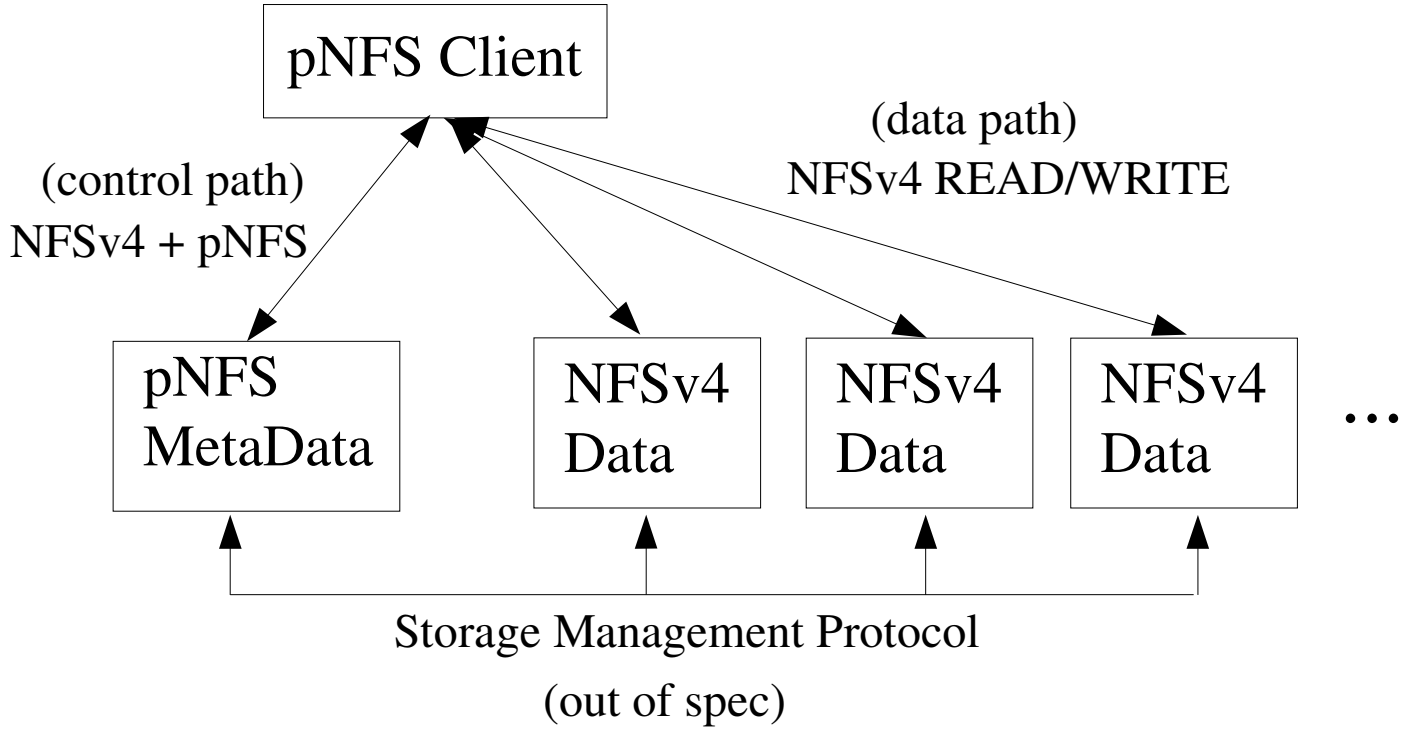
# Parallel NFSv4 (pNFS) Overview

- NFSv4 IETF minor version candidate
- Separate Data (READ/WRITE) from Control path
- NFSv4 server logically divided in two:
  - Metadata server:  also a fully functional NFSv4 server
  - Data server(s): available for pNFS clients
- Data path can (eventually) include different storage protocols; file (NFS), or OSD, or block, ...
- GridNFS is currently only interested in the file storage protocol for the Data path.

# Parallel NFSv4 (pNFS) Overview

- pNFS client negotiates storage protocol details with pNFS Metadata server
- pNFS client requests LAYOUT, which describes where the data is on the Data servers
  - File LAYOUT list of: server, filehandle, stateid, stride, etc
- LAYOUT allows striping of files across Data servers
  - LAYOUT with single server == file location redirection, can be used for load balancing

# Parallel NFSv4: File LAYOUT




# Parallel NFSv4 (pNFS) Overview

- pNFS MetaData server required to service the full NFSv4 protocol including READ/WRITE
  - Service non-pNFS clients
- Storage Protocol is not described; Data servers need to follow NFSv4 with respect to MetaData server receiving:
  - SETATTR with set ACL
  - RENAME, UNLINK, etc
  - RPCSEC_GSS security on exported file system

# NFSv4 Scaling: Enterprise Service

- NFSv4 is a protocol, not a file system
- Scaling characteristics of a single NFSv4 server is dependent upon the file system being exported
  - Maximum file size, delivery of data through file system
  - Server hardware configuration, disk configuration, etc
- Name space joins mulitple servers; more servers equals more file space.....

# NFSv4 Scaling: Enterprise Service

- Desirable to separate name space from storage for load balancing and back-end management
- Currently, a single NFSv4 server can 'fill the pipe' for a single client with a Gigabit Ethernet NIC.
- Servers with multiple NICs, large amount of RAM, high speed PCI-X buses, etc can help but...
- A single NFS client getting data from a single NFS server is a bottleneck.

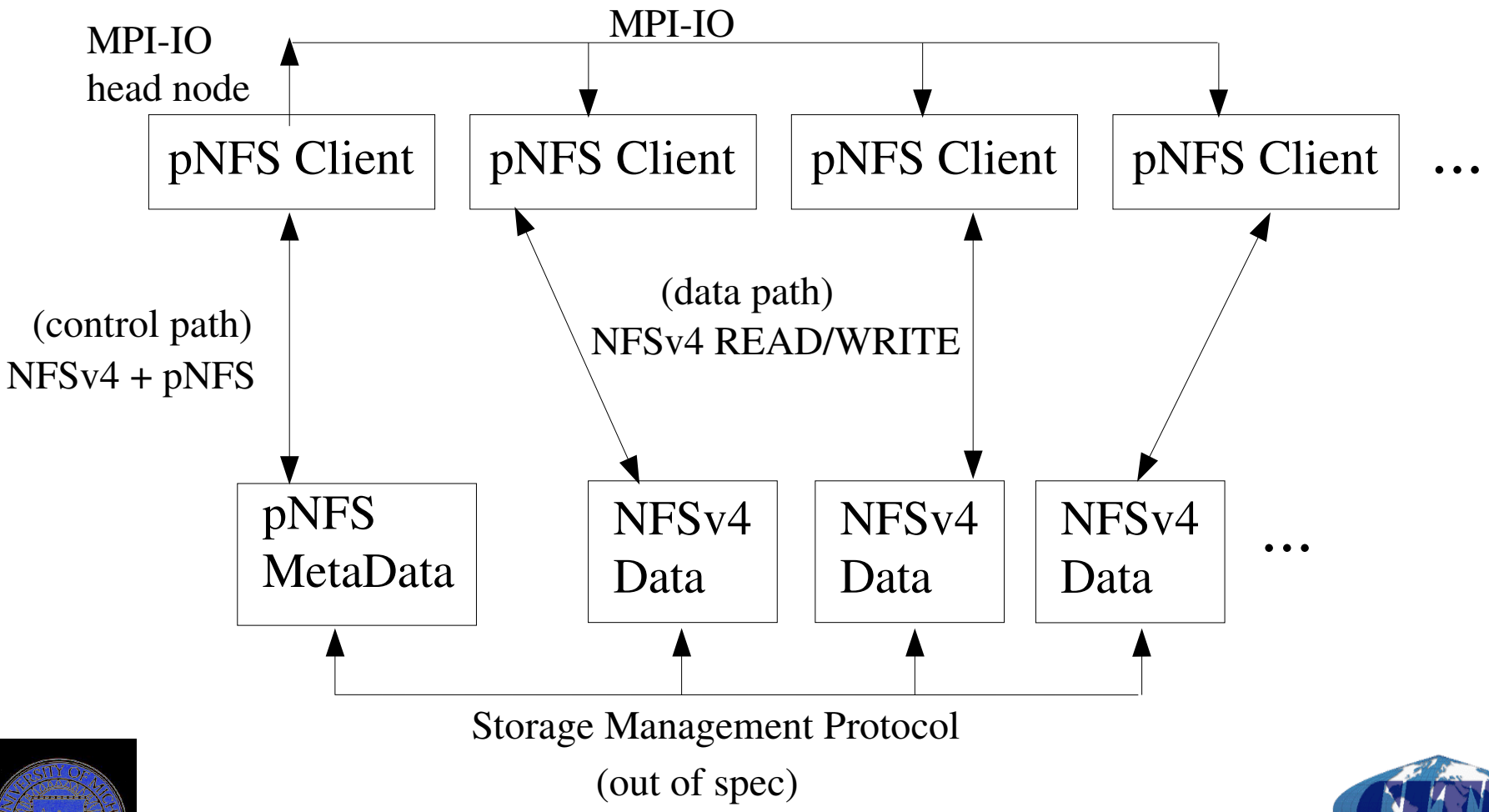
# NFSv4 Scaling: pNFS Service

- The NFSv4 pNFS minor version with file layouts can help
- A file layout with a single server enables relocation of a file, separating the file storage from the namespace
- A file layout with multiple servers provides
  - File sizes beyond the exported file systems maximum file size (max file size * number of data servers)
  - Performance beyond the single client to single server bottleneck (single client to multiple data servers)

# NFSv4 Scaling: pNFS Service

- The pNFS working group is also looking at MPI-IO systems where MPI-IO nodes are also pNFS clients
- MPI-IO head node uses pNFS to obtain a multiple server file layout
- MPI-IO head node distributes layouts via the MPI-IO layer to nodes in it's group, one server layout per node
- Multiple pNFS clients can then access data to multiple pNFS Data servers; High speed inter-cluster transfer

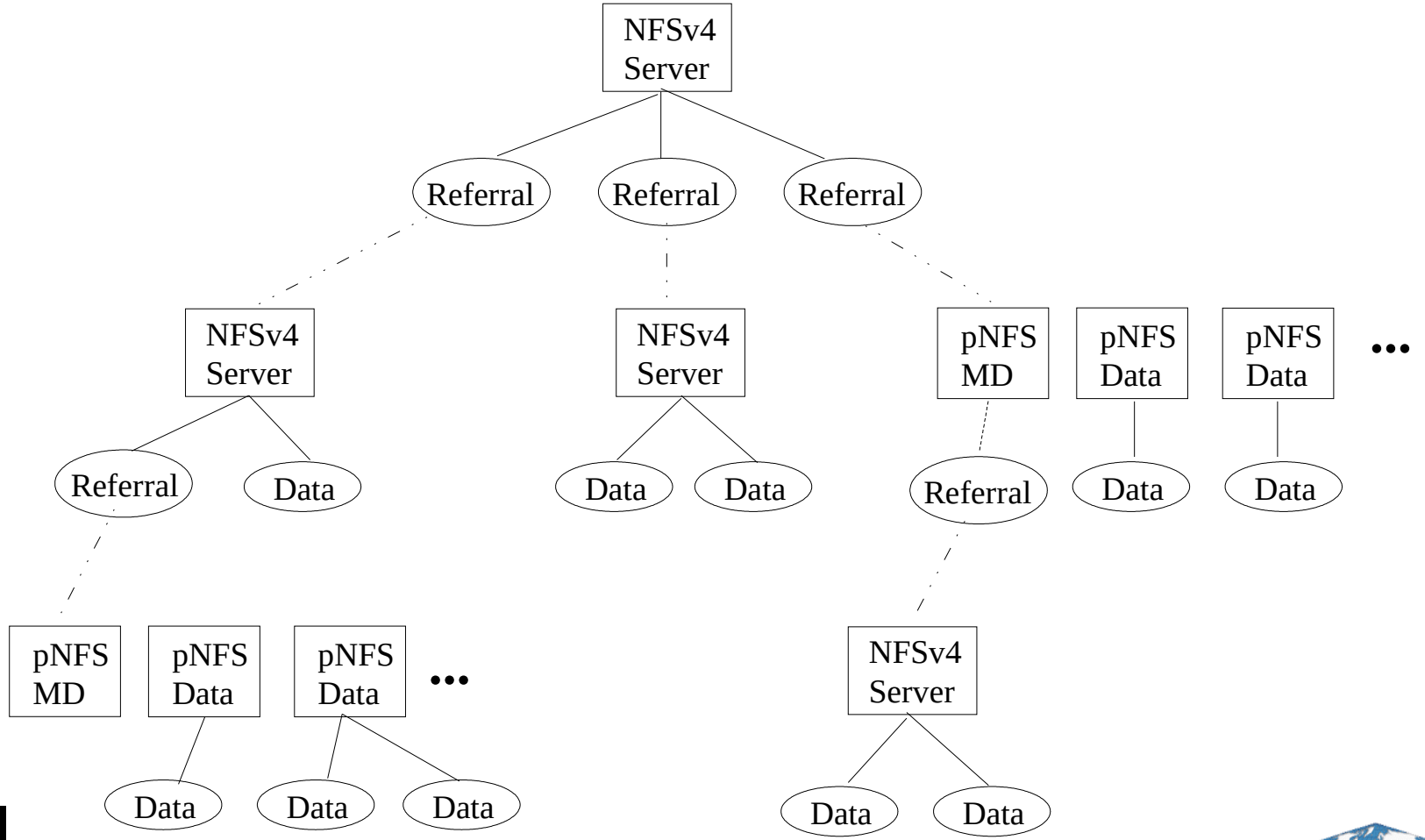
# Parallel NFSv4 and MPI-IO



# NFSv4 Scaling to Petabyte Service

- The NFSv4 pNFS minor version with file layouts in combination with enterprise NFSv4.0 service has the potential to deliver a global distributed file service scaling to petabytes of data.
- The IETF NFSv4 minor version feature enables the protocol to evolve to address issues as they arise.
  - e.g. we can fix what is broken

# The Big Picture



# Status

- NFSv4.0 product
  - Solaris, Linux 2.6.X, AIX, Network Appliance, Hummingbird (Windows)
  - Coming soon: EMC, HP, full support in Linux distros
- NFSv4.0 interoperability with Kerberos V, byte-range locking, delegations
- Name space, replication, migration being developed
- WAN performance yet to be tuned

# Status

- Minor versions
  - Some protocol 'fixes'
  - NFS over RDMA
  - Directory delegation
- pNFS
  - IETF working group formed, operations RFC
  - Initial implementations underway

# Any Questions?

http://www.citi.umich.edu/projects