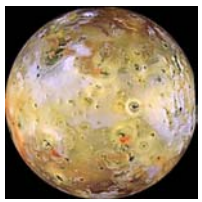
The ASCI Search for the Intergalactic File System



04/07/2003

Gary Grider

Los Alamos National Lab

LA-UR 03-1818

“We Saw it Coming”

- For Sandia, LLNL, LANL and DOD, the need for a global parallel file system was there from the beginning of clustered based parallel computing,
  - few solutions existed,
  - none were heterogeneous,
  - none were open source,
  - none were based on standards, and
  - none were secure on a public net.
- This is primarily for our giant clusters, secondarily for our enterprise, and lastly across multiple enterprises/sites
- We saw Linux clusters coming in the future which made the problem very real and very evident

FS Requirements Summary

- **From Tri-Lab File System Path Forward RFQ (which came from the Tri-labs file systems requirements document) ftp://ftp.lanl.gov/public/ggrider/ASCIFSRFP.DOC**
  - **POSIX-like Interface**
  - **Works well with MPI-IO**
  - **Open Protocols, Open Source (parts or all)**
  - **No Single Point Of Failure**
  - **Global Access**
    - *Global name space, …*
  - **Scalable Infrastructure for Clusters and the Enterprise**
    - *Scalable bandwidth, metadata, …*
  - **Integrated Infrastructure for WAN Access**
    - *WAN Access, Global Identities, Wan Security, …*
  - **Scalable Management & Operational Facilities**
    - *Manage, tune, diagnose, statistics, RAS, build, document, snapshot, …*
  - **Security**
    - *Authentication, Authorization, Logging, …*

Link to more RQMTS

It Has to Scale with Our Machine Appetite

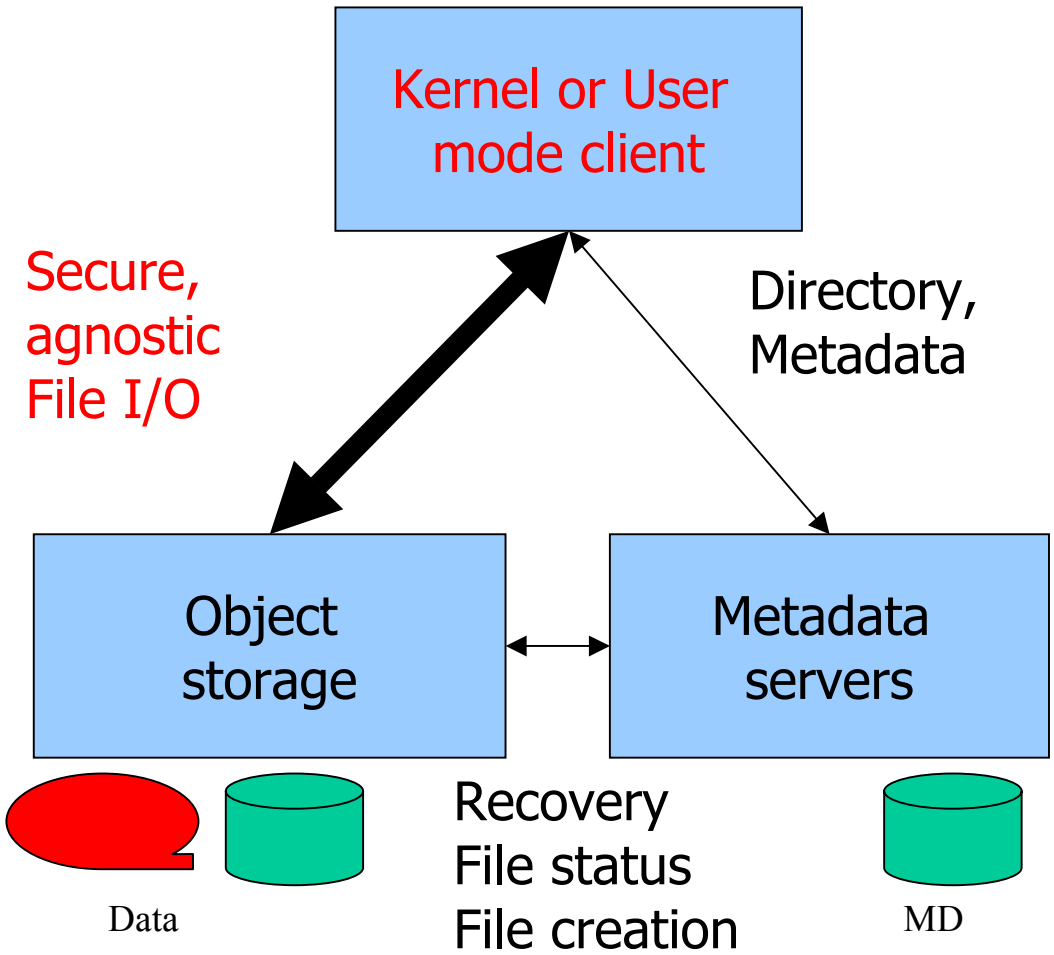
Aggregate Bandwidth Rates for One Parallel Job
Simulation & Physics Model Aggregate FS Requirements

	1999	2003	2005	2008
Teraflops/Clients	**3.9 / 6K**	**30 / 12k**	**100 / 50K**	**400 / 100k**
Memory Size (TB)	**2.6**	**13-20**	**32-67**	**44-167**
I/O Rates (GB/s) N to N and N to 1	**4 – 8**	**20-60**	**50-200**	**80-500**

OBFS's Most Worthy

- NAS file systems don't scale to our levels
  - lack of parallelized metadata operations like allocation (especially for a single file or directory)
- SAN file systems don't scale to our levels
  - no network security and SAN cost, bypassing sending data through a file server is great **if** its secure
- Need a model that makes possible secure scalable networking and scaling of important metadata operations
- OBSD
  - has good network security model allowing for data path scaling, and
  - securely allows for metadata offload functions to the storage devices (needed to enable massively parallel writes)
  - has the promise of pushing even more of the I/O workload to smarter and smarter devices

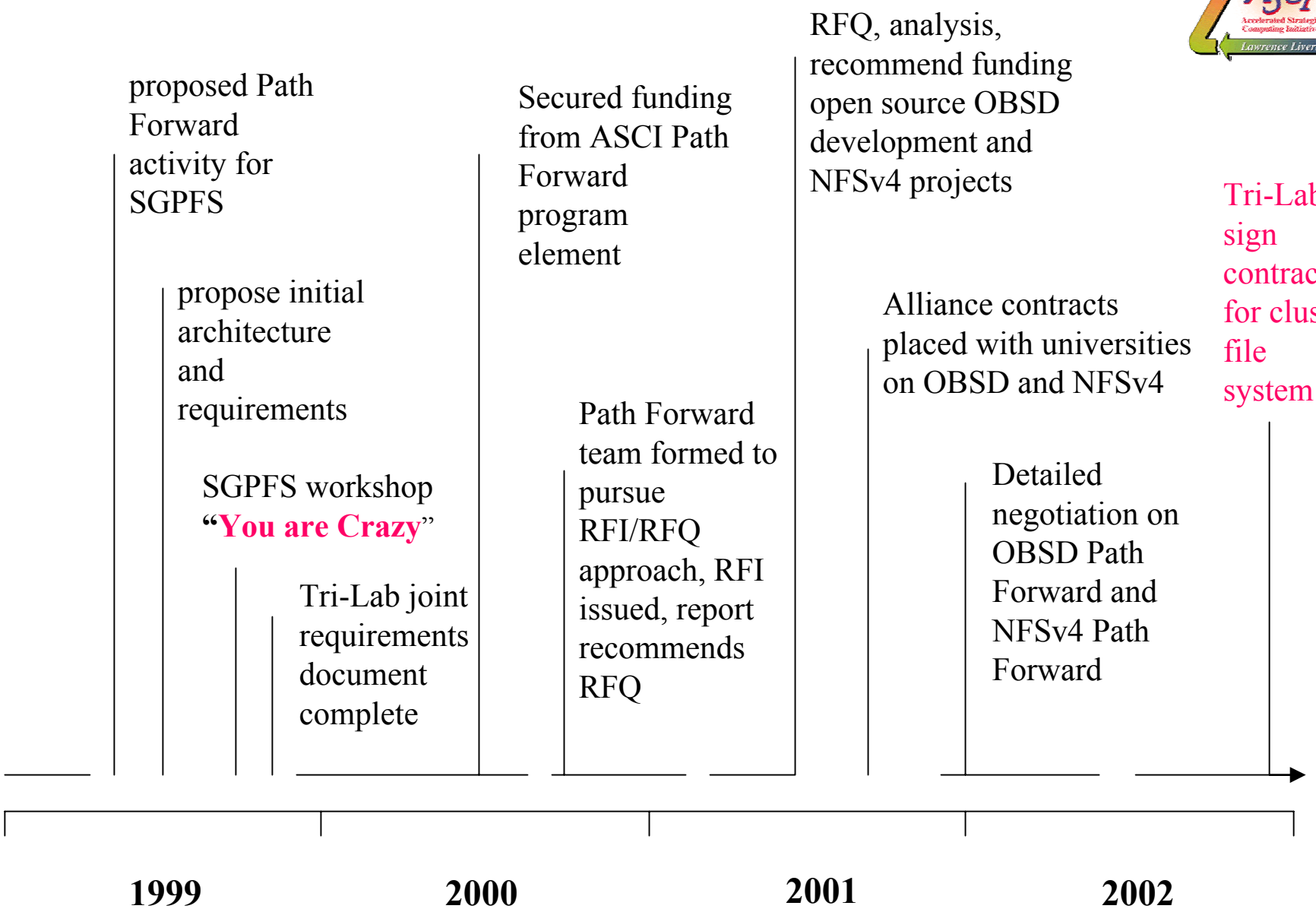
OBFS Approach



Keys to Getting a Galactic OBSD File System Solution that Will Endure

- **Client needs to be in OS Kernel typically, we need supportable penetration**
  - Open source client for Linux (required for our funded efforts, but that may not be enough to ensure long run support)
  - A way for non Linux OS's to be supported (NFSv4 seemed most likely given DAFS, NFS on RDMA, etc.) (required for our funded efforts)
  - We decided to get involved with NFSv4 via U of Michigan alliance to help
- **We need open secure standard for devices, but device market or standards won't materialize, without useful software solution(s)**
  - We are encouraging through funding of cluster file systems and scalable NAS solutions
  - We made a part of our product development efforts to push standardization
- **We are prepared to encourage follow on "smarter storage" if standard secure infrastructure for this becomes widely available**

Historical Time Line



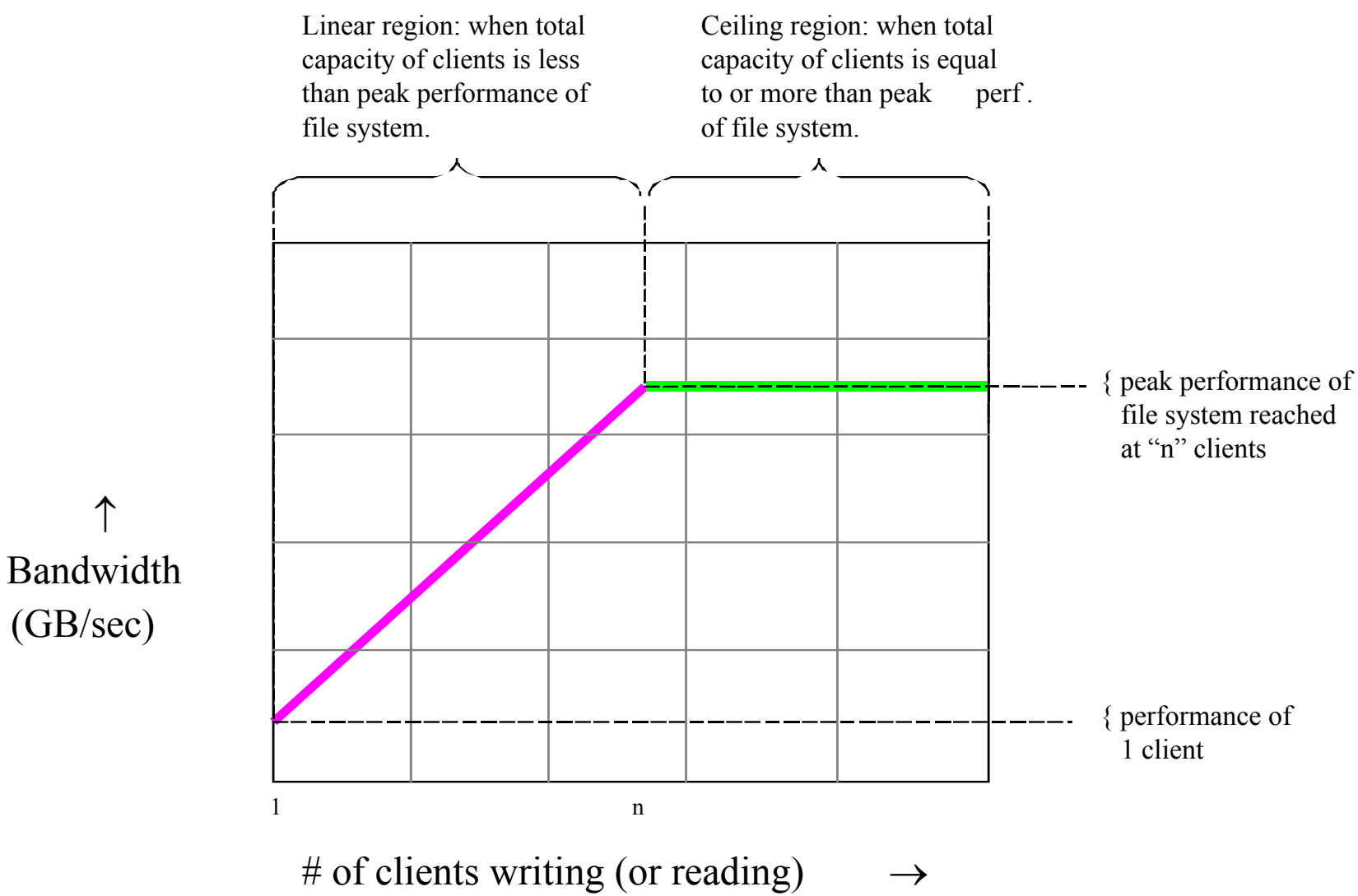
Current State

- Maybe we were not so "Crazy" after all
  - Clusters being deployed by the thousands, even large clusters are popping up everywhere
  - File System is still the most important missing piece for clusters
- Funding/working with OBSD vendors or "Vendors to Be" for cluster file systems and scalable OBSD scalable NAS
- Funding and working with Universities and Vendors on NFSv4 with parallel extensions and protocol agnostic capabilities so OBFS can be extended heterogeneously
- Hoping for some limited deployment in FY04
- NEED to begin to see progress on standards efforts soon!
- What more can we do?

Backup slides

Backup slides

It Has to Scale with Number of Processes



Capacity Has to Scale Too

File System Capacities				
	1999	2002	2005	2008
Teraflops	3.9	30	100	400
Memory size (TB)	2.6	13-20	32-67	44-167
File system size (TB)	75	200 - 600	500 -2,000	20,000
Number of Client Tasks	8192	16384	32768	65536
Number of Users	1,000	3,000	3,500	3,500
Number of Directories	5.0*10^6	1.5*10^7	1.8*10^7	1.8*10^7
Number of devices/subsystem	5000 (18GB drives)	10000 (72GB drives)	8375 (300GB drives)	8750 (1200 GB drives)
Number of Files	7.5*10^7 to 1.0*10^9	3.75*10^8 to 4.0*10^9	4.5*10^8 to 1.0*10^10	4.5*10^8 to 1.0*10^10

Even Meta-Data Operations have to Scale

File Create Performance –versus- Number of Nodes One parallel program creating multiple files (one per node) into a single directory. *N* = total number of processors in machine *R*=File create rate for one processor				
	1/4th machine	1/2th machine	3/4th machine	Full machine
Aggregate File Create Rate	.20*N*R	.40*N*R	.60*N*R	.75*N*R

* - Please note that multiple metadata servers with a reasonable decomposition of the operations is likely required

Other Requirements Besides Scalability

- Security more like AFS/DFS but better
  - Content based security, born on marks, etc.
- Global, Heterogeneous, Protocol Agnostic, open source, open protocols
- POSIX behavior with switches to defeat parts
  - Lazy attributes, byte range locks, etc.
- WAN behavior like AFS/DFS but better
  - Including ACL's, GSS, multi domain, etc.
- Scalable management (sorry, scalability keeps coming up)
- A product, supported by a market larger than the Tri-Labs

FS Requirements Detail 1

- **3.1 POSIX-like Interface**
- **3.2 No Single Point Of Failure**
- **4.1 Global Access**
  - *4.1.1 Global Scalable Name Space*
  - *4.1.2 Client software*
  - *4.1.3 Exportable interfaces and protocols*
  - *4.1.4 Coexistence with other file systems*
  - *4.1.5 Transparent global capabilities*
  - *4.1.6 Integration in a SAN environment*
- **4.2 Scalable Infrastructure for Clusters and the Enterprise**
  - *4.2.1 Parallel I/O Bandwidth*
  - *4.2.2 Support for very large file systems*
  - *4.2.3 Scalable file creation & Metadata Operations*
  - *4.2.4 Archive Driven Performance*
  - *4.2.5 Adaptive Prefetching*
- **4.3 Integrated Infrastructure for WAN Access**
  - *4.3.1 WAN Access To Files*
  - *4.3.2 Global Identities*
  - *4.3.3 WAN Security Integration*

FS Requirements Detail 2

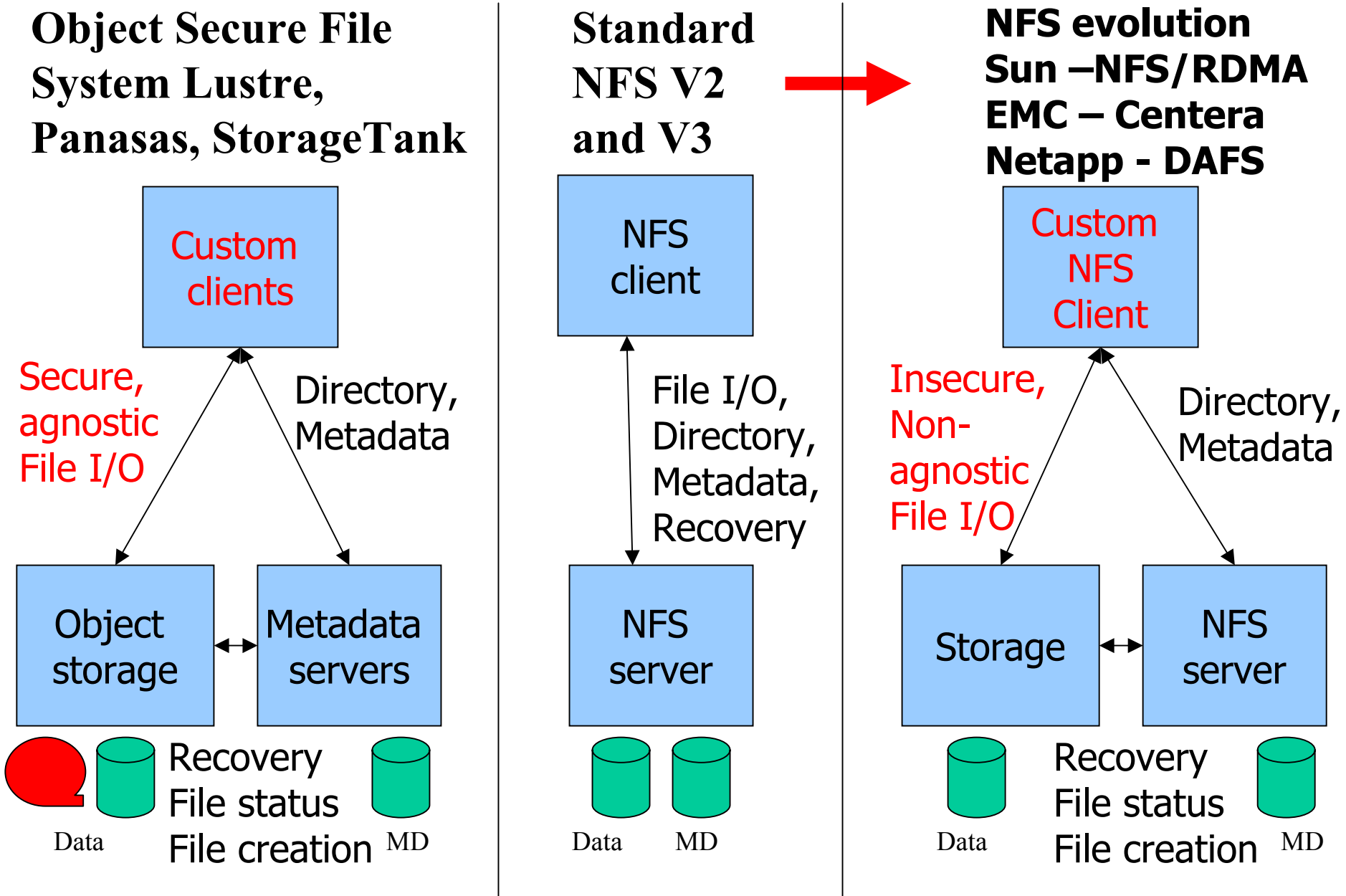
- **4.4 Scalable Management & Operational Facilities**
  - *4.4.1 Need to minimize human management effort*
  - *4.4.2 Integration with other Management Tools*
  - *4.4.3 Dynamic tuning & reconfiguration*
  - *4.4.4 Diagnostic reporting*
  - *4.4.5 Support for configuration management*
  - *4.4.6 Problem determination GUI*
  - *4.4.7 User statistics reporting*
  - *4.4.8 Security management*
  - *4.4.9 Improved Characterization and Retrieval of Files*
  - *4.4.10 Full documentation*
  - *4.4.11 Fault Tolerance, Reliability, Availability, Serviceability (RAS)*
  - *4.4.12 Integration with Tertiary Storage*
  - *4.4.13 Standard POSIX and MPI-IO 4.4.14 Special API semantics for increased performance*
  - *4.4.15 Time to build a file system*
  - *4.4.16 Backup/Recovery*
  - *4.4.17 Snapshot Capability*
  - *4.4.18 Flow Control & Quality of I/O Service*
  - *4.4.19 Benchmarks*

FS Requirements Detail 3

- **4.5 Security**
  - *4.5.1 Authentication*
  - *4.5.2 Authorization*
  - *4.5.3 Content-based Authorization*
  - *4.5.4 Logging and auditing*
  - *4.5.5 Encryption*
  - *4. 5.6 Deciding what can be trusted*

BACK

Some File System Approaches



Lets leverage NFSv4's existing metadata capabilities, our NFS level 3 alliance, NFS's huge market force, our OBFS PF, and other efforts to reduce risk in this overall area?

Combine the efforts: NFSv4 and OBFS
Scalable NAS, regular NFS, NFS with secure data channel, etc.