Overview of I/O systems, technology, and directions

Gary Grider, LANL

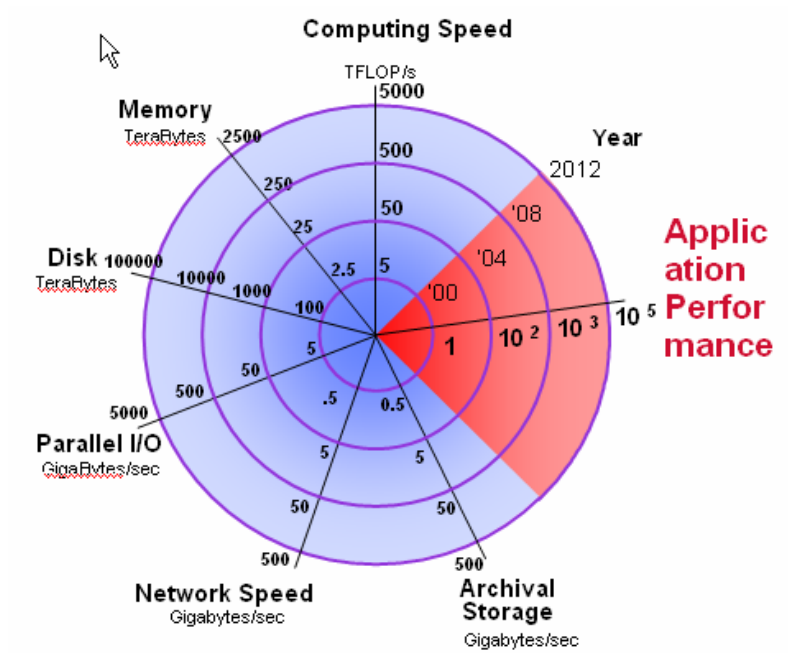
LA-UR 06-0397

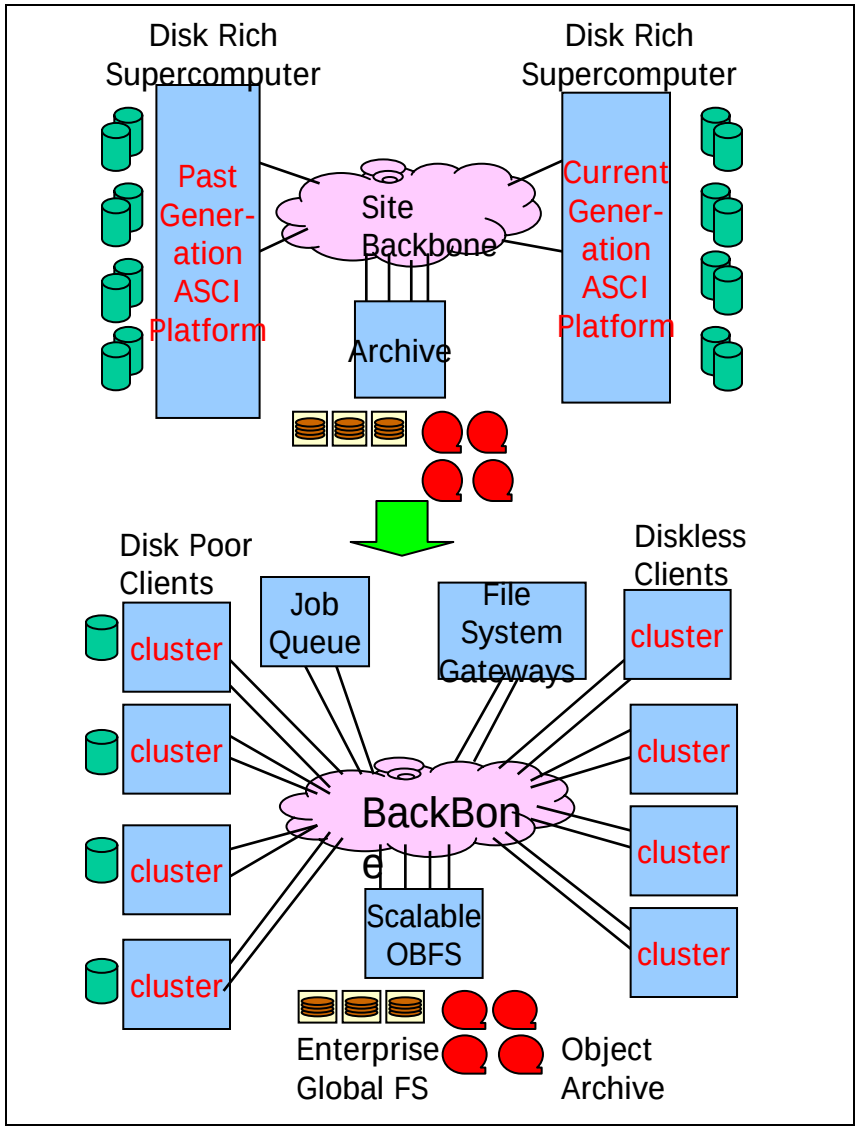
01/2006

What drives us?

- **Provide reliable, easy-to-use, high-performance, scalable, and secure, I/O**
- **Via standard and other interfaces**
  - MPI-IO, POSIX, etc.

Balanced System Approach





Requirements Summary

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 bandwidth, metadata, management, security ...*
  - *WAN Access, Global Identities, Wan Security, ...*
  - *Manage, tune, diagnose, statistics, RAS, build, document, snapshot, ...*
  - *Authentication, Authorization, Logging, ...*

FS Requirements Detail

- 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*
- 4.4 Scalable Management & Operational Facilities
  - *4.4.1 Need to minimize human management effort*
  - *4.4.2 Integration with other management tools*
  - *4.4.2 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*
- 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*

Lots of things have to scale

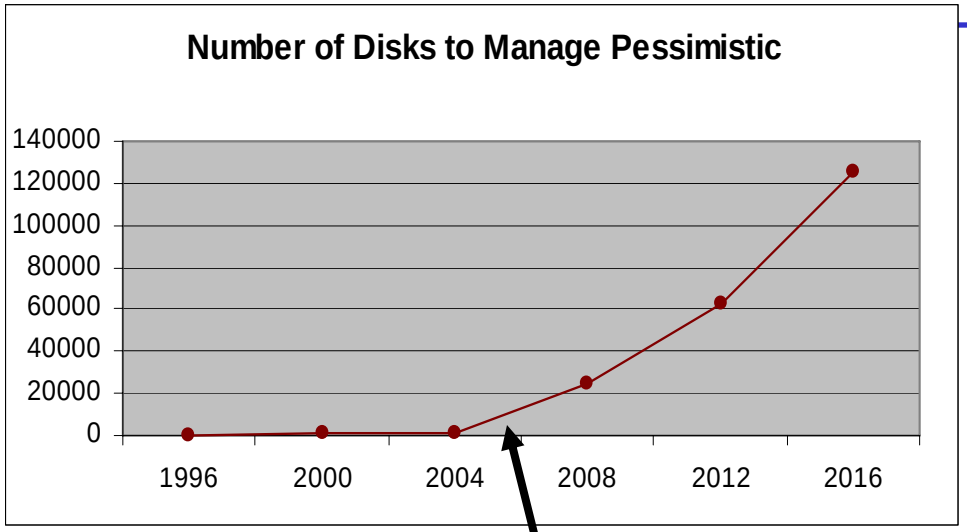
File System Attributes				
	1999	2002	2005	2008
Teraflops	3.9	30	100	400
Memory size (TB)	2.6	13-20	32-67	44-167
File system size (TB)	75	200 - 600	500 -2,000	20,000
Number of Client Tasks	8192	16384	32768	65536
Number of Users	1,000	4,000	6,000	10,00
Number of Directories	5.0*10^6	1.5*10^7	1.8*10^7	1.8*10^7
Metadata Rates Data Rate	500/sec 1 mds 3 GB/sec	2000/sec 1 mds 30 GB/sec	20,000/sec n mds 100 GB/sec	50,000/sec n mds 400 GB/sec
Number of Files	1.0*10^9	4.0*10^9	1.0*10^10	1.0*10^10

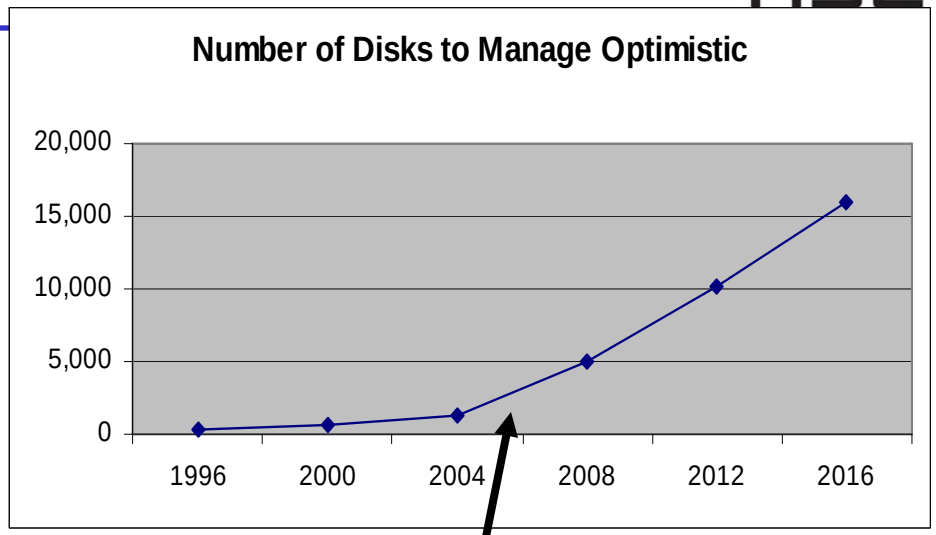
Other Requirements

- **Based on Standards**
- **Security**
  - Content based security, born on marks, hooks for end to end encryption, extensible attributes, etc.
  - Real transactional security on the SAN, not simple zoning and other poor attempts (ANSI T10)
- **Global, Heterogeneous, Protocol Agnostic, open source, open protocols**
- **POSIX behavior with switches to defeat portions**
  - Lazy attributes, byte range locks, etc.
- **WAN behavior like AFS/DFS but better**
  - Including ACL's, GSS, multi domain, directory delegation, etc.
- **Scalable management (sorry, scalability keeps coming up)**
- **A product, supported by a market larger than the Tri-Labs**

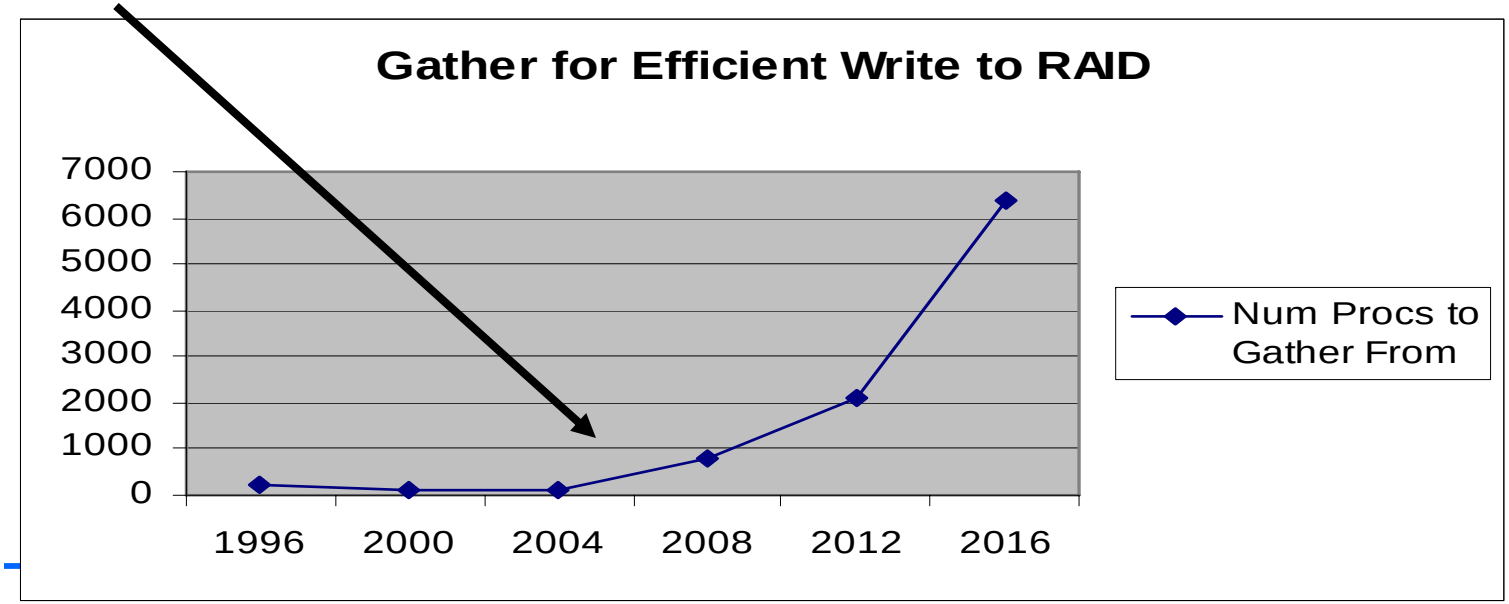
Trends and Emerging Issues

Some Disturbing Trends



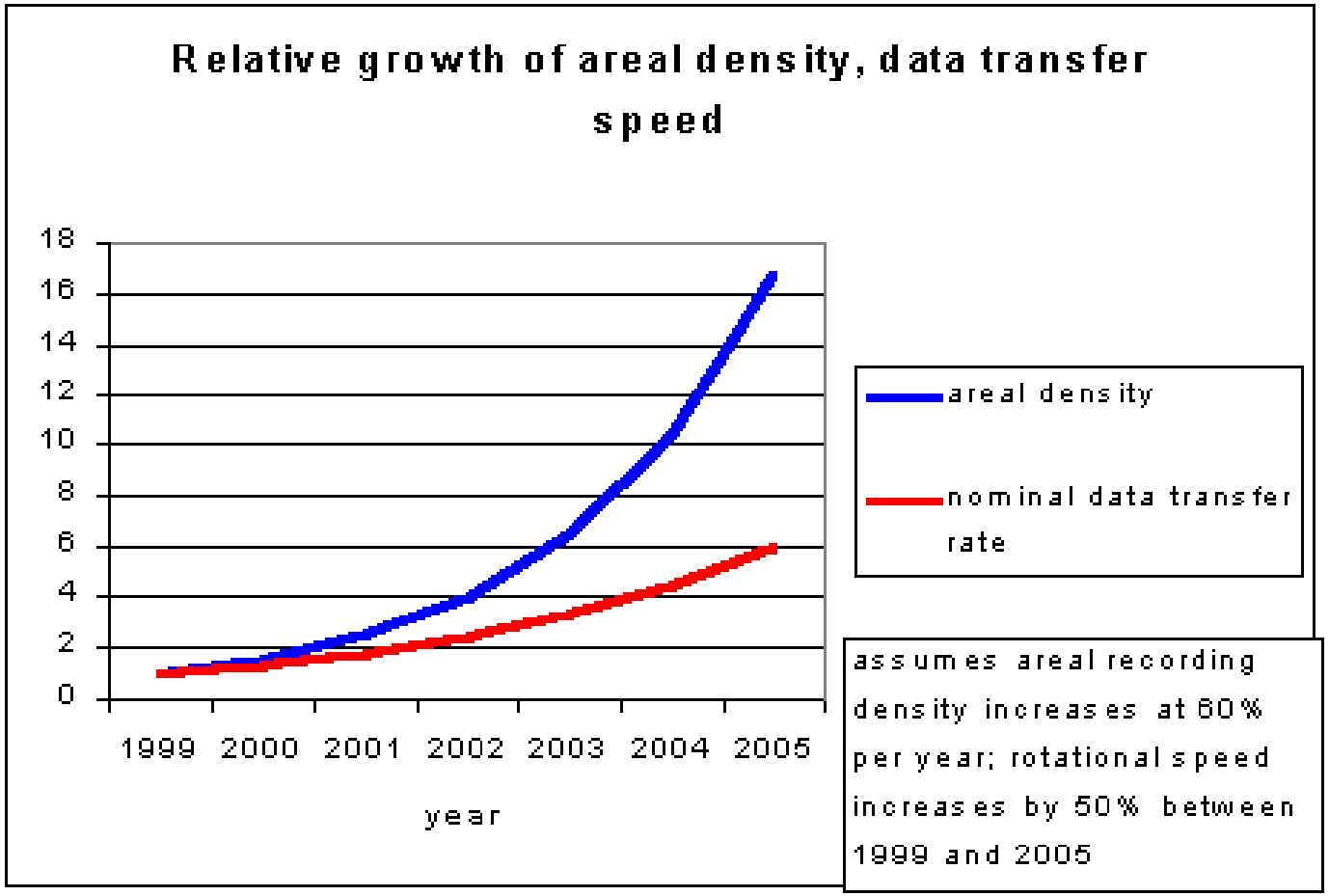


We are at the beginning of some pretty disturbing trends



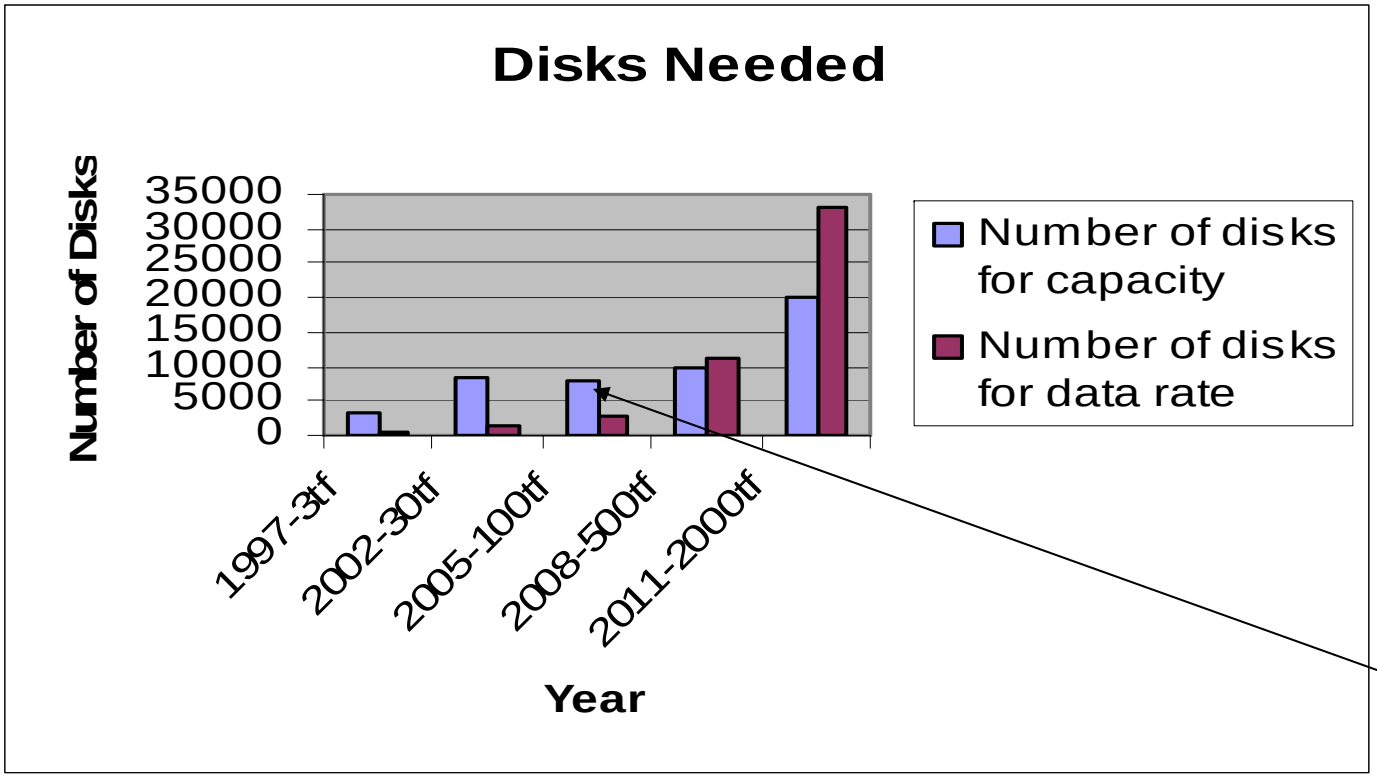
Emerging Issues with RAID

- Disks are getting much denser but not appreciably faster (bandwidth read/write)



RAID Oriented Implications of Capacity vs BW Trend

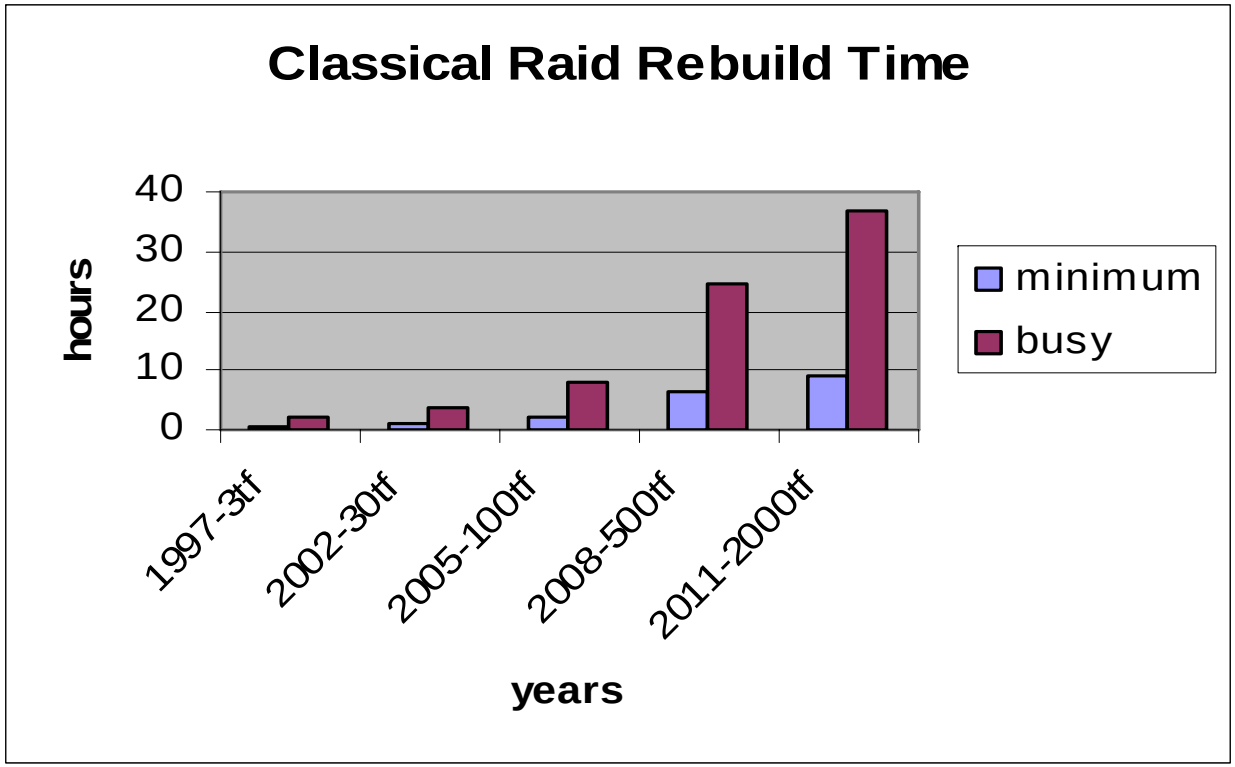
- We will be buying disks for BW rather than for Capacity
- Write size for single disk sweet spot keeps rising and thus, for full stride RAID continues to rise
- Files will be striped over a larger percentage of the disks on the floor on average to get desired data rate
- Reliability at scale becomes more and more important



SATA

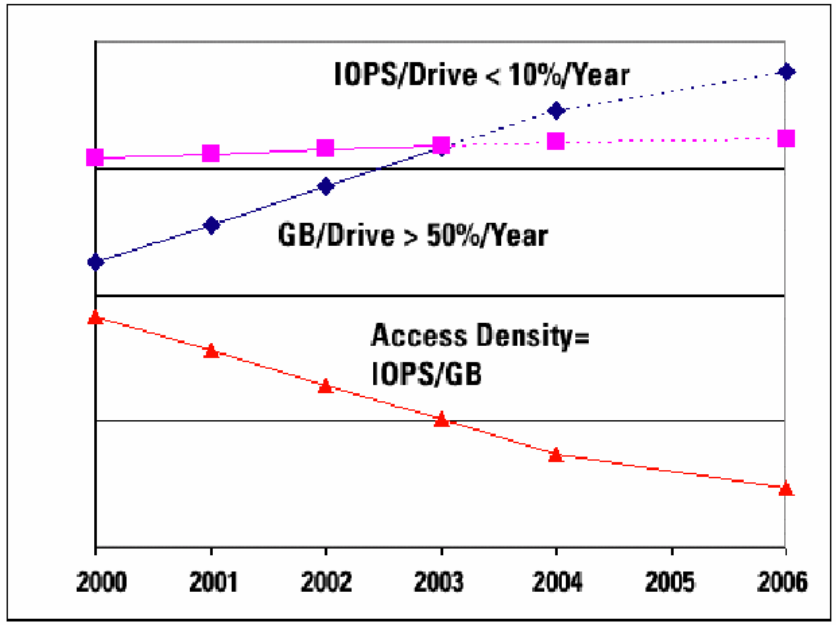
Classical RAID Rebuild Time

- **Rebuild times get worse and worse, from minutes, to hours, to days – raising chances of 2-3 disk failure more and more**



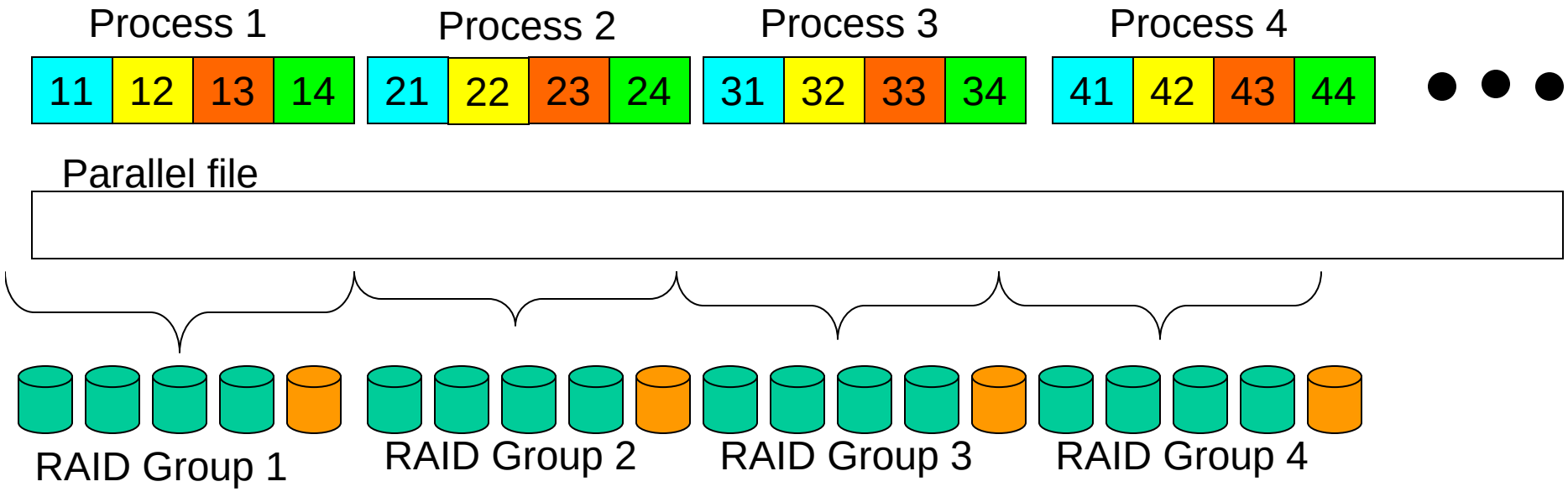
Scalable Metadata

- Disks not getting more agile, Metadata services must scale
- Due to growth in Global use from many clusters and due to usage patterns, N to N, N to 1 small ops, etc. metadata scaling issues are upon us. We saw this coming and started research efforts a while ago but we have some tactical things we are doing.



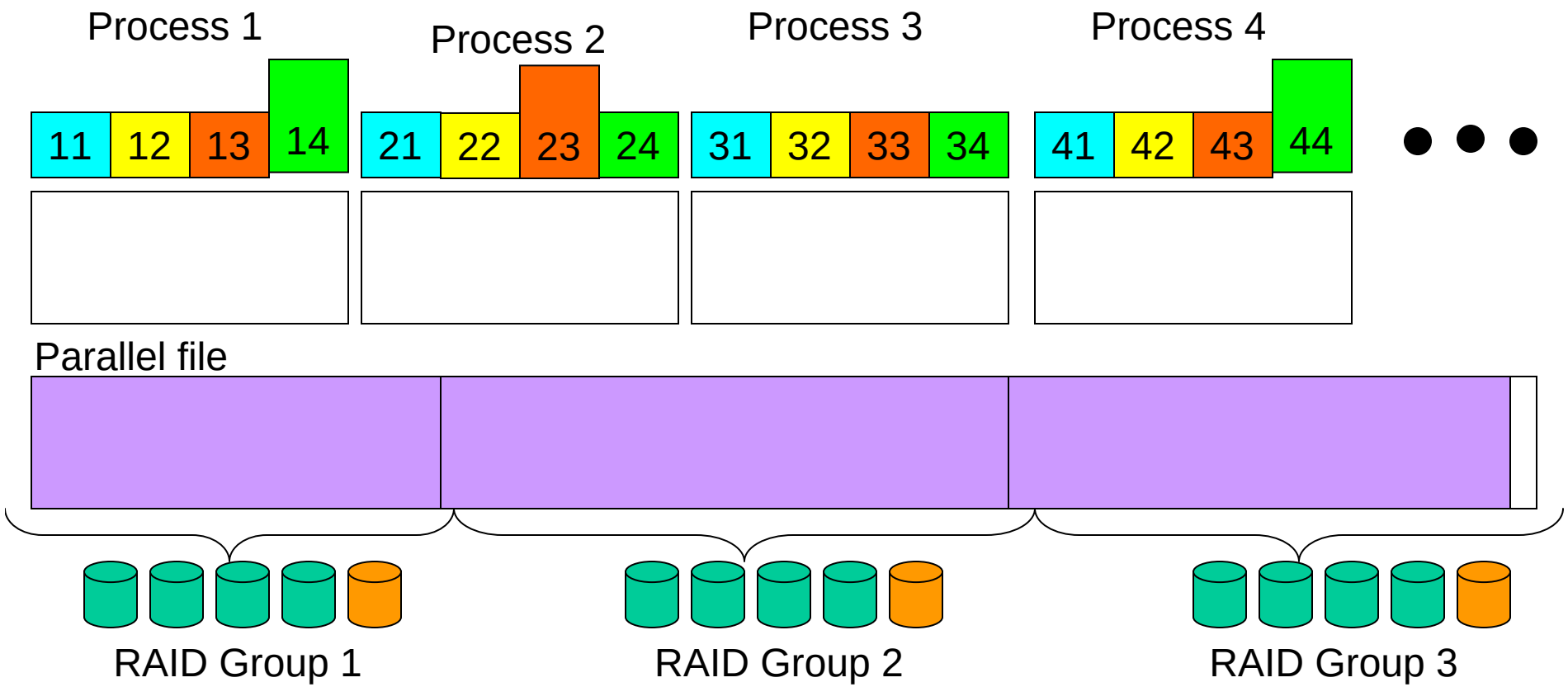
What do apps do?

Example of well aligned I/O



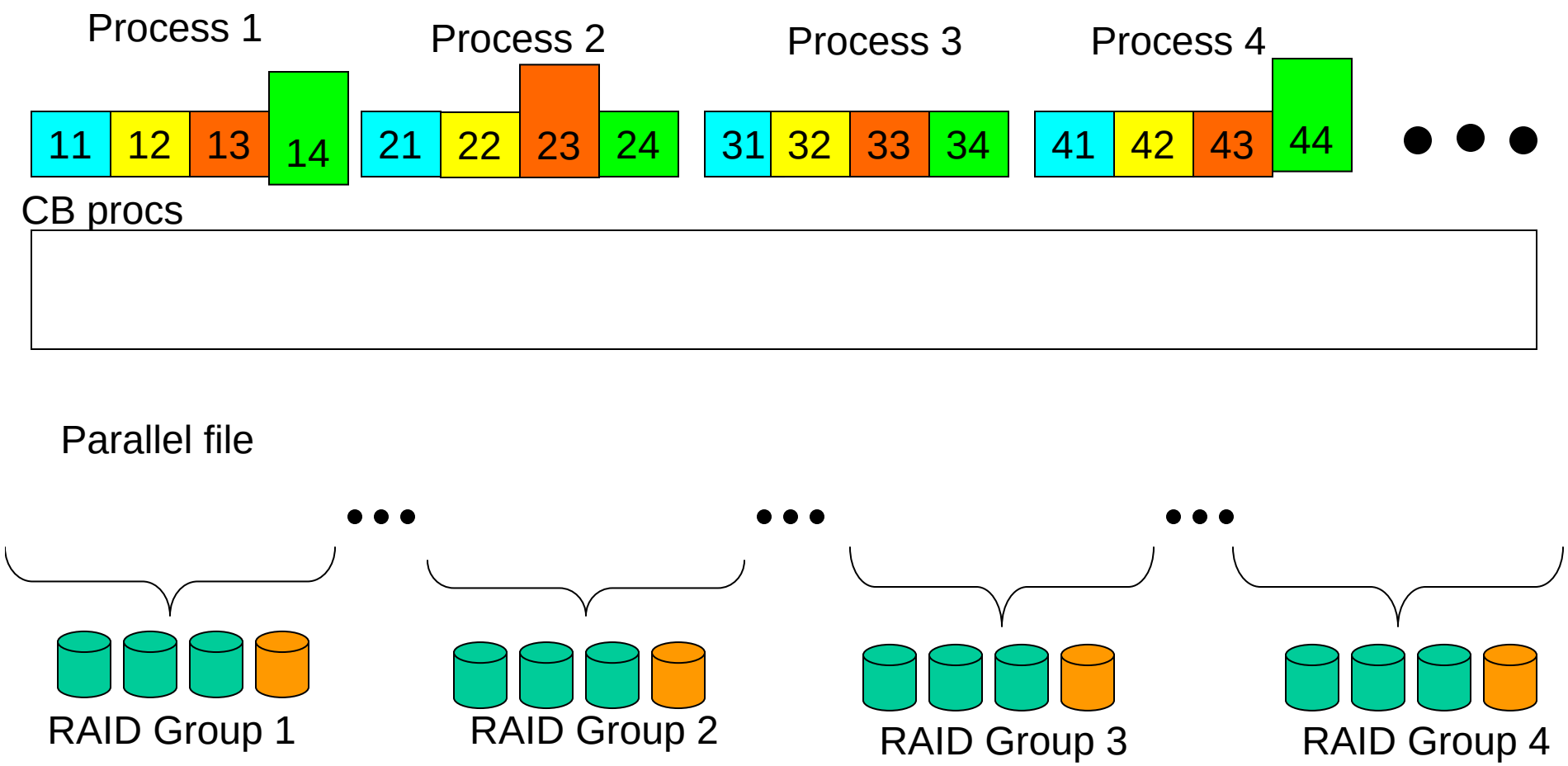
Oh, if applications really did I/O like this!

Real applications do small, unbalanced, and unaligned I/O



Notice every write is possibly a read/update/write since each write is a partial parity update. Notice that processes are serializing on their writes as well.

Middleware can help but more work is needed



Recent and Current I/O R&D Focus

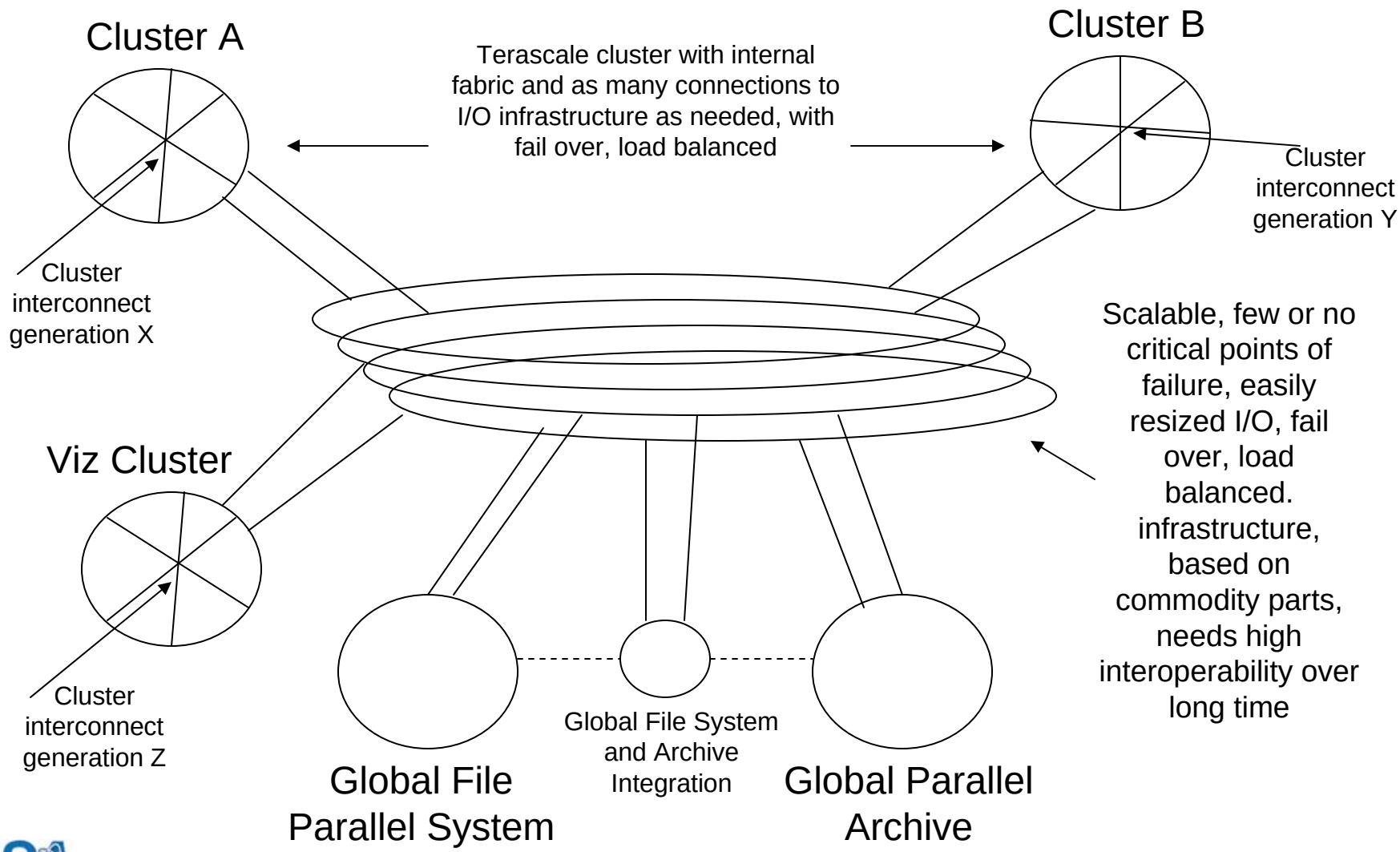
- ***Cluster File Systems that scale for aligned data operations and Symmetric NFS***
  - *Lustre*
  - *Panasas*
  - *PVFS2 (user space file system)*
  - *Light Weight file Systems*
- ***Beginning engineering and early R&D for other non BW scalability dimensions***
  - *Metadata (Lustre, Panasas and UCSC research)*
  - *Security (Lustre, Panasas and UCSC research)*
  - *NFSv4 security, WAN, Performance, and pNFS*
- ***Middleware to address usage, small unaligned I/O, and efficient connection to High Level I/O libraries***
- ***High Level I/O libraries***
- ***Early utilization of intelligence near disk drive and new storage devices (mems/mram)***
- ***Enterprise Sharing of Global Parallel File System – access from many clusters***
- ***Autonomics***
- ***Object Archive***
- ***Standards work***
  - *HPC POSIX I/O*
  - *NFSv4 and pNFS*
  - *ANSI T10 OBSD*
  - *ISER*

Is any of this stuff paying off?

- Lustre in use, in programmatic role at LLNL and Sandia
- Panasas in use, in programmatic and supports roles LANL, Sandia and LLNL
- Academic and other partnerships paying off
  - ANL and Northwestern work going into ROMIO, aggregation, two phase I/O, sieving, better caching coming
  - NFSv4 getting into Linux
    - » We showed multi-realm kerberized controlled access to a file system via NFSv4
  - UCSC ideas beginning to become appealing to our file system partners
- Implementing first generation enterprise class common parallel file systems, shared between multiple clusters (Linux natively, others via NFSv3)
- pNFS entering IETF to allow for non Linux clients to access object file systems natively, to be part of NFSv4.1
- MPI-IO beginning to be used in more apps
- Future works spinning up, POSIX api work, object archive, multi-agency I/O R&D investments strategy (HEC-IWG I/O emphasis resulting in $12M NSF HEC/URA call for I/O and FS research)

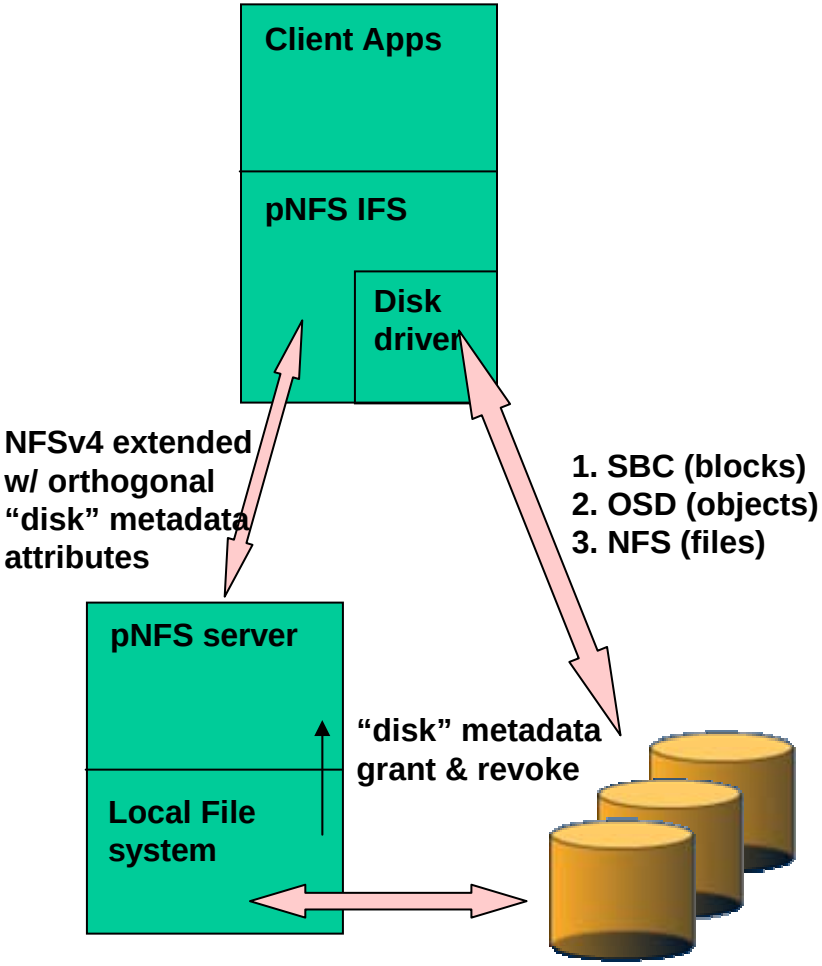
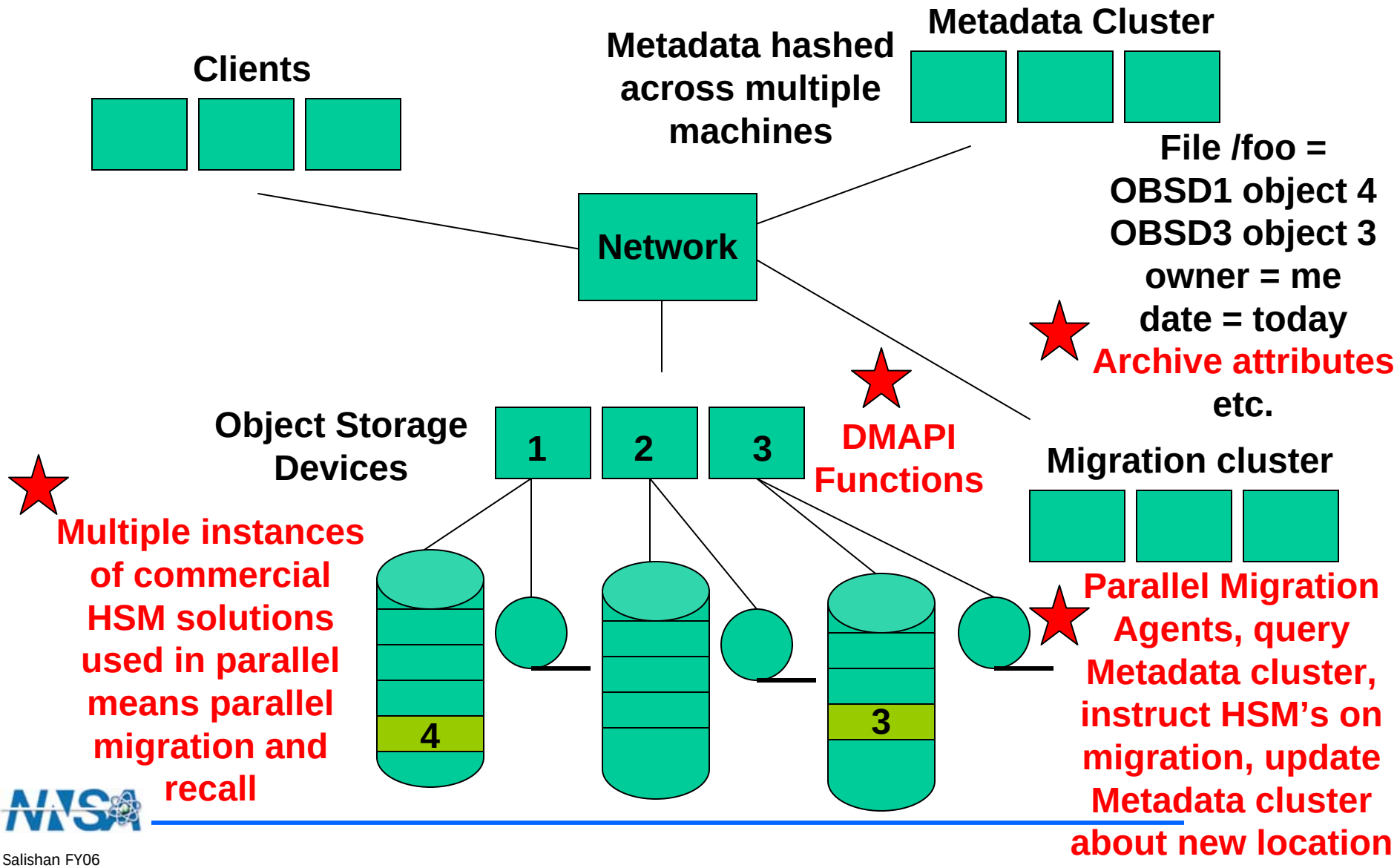
Backup Slides

Where do we want to go, the bigger picture?



pNFS Multiple Data Server Protocols

- **Inclusiveness favors success**
- **Three (or more) flavors of out-of-band metadata attributes:**
  - BLOCKS: SBC/FCP/FC or SBC/iSCSI… for files built on blocks
  - OBJECTS: OSD/iSCSI/TCP/IP/GE for files built on objects
  - FILES: NFS/ONCRPC/TCP/IP/GE for files built on subfiles
- **Inode-level encapsulation in server and client code**



Object Archives: What if we move the migration function from the clients to the OSD's?