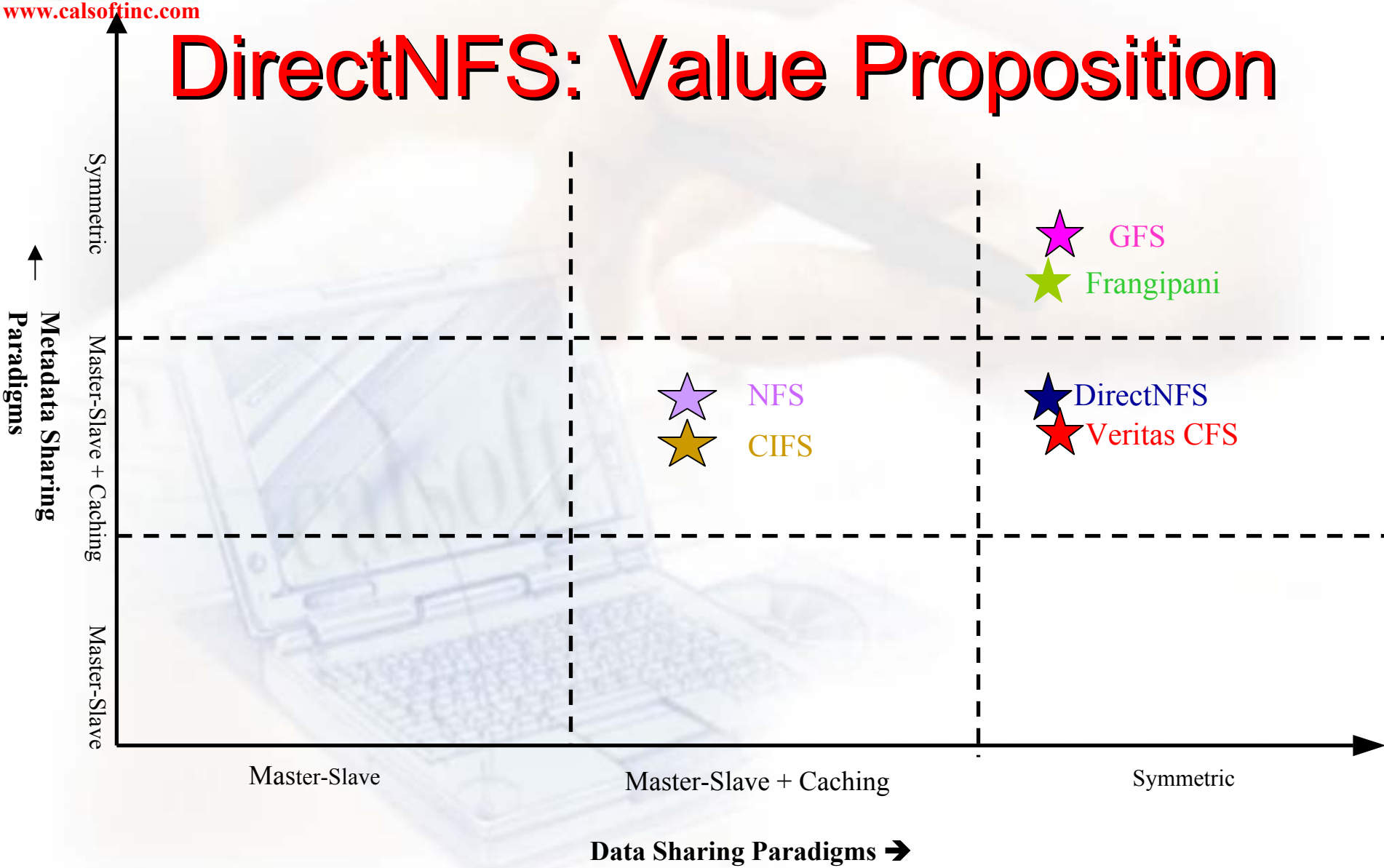
calsoft
www.calsoftinc.com

DirectNFS

Anupam Bhide
CalSoft Private Limited

calsoft
www.calsoftinc.com

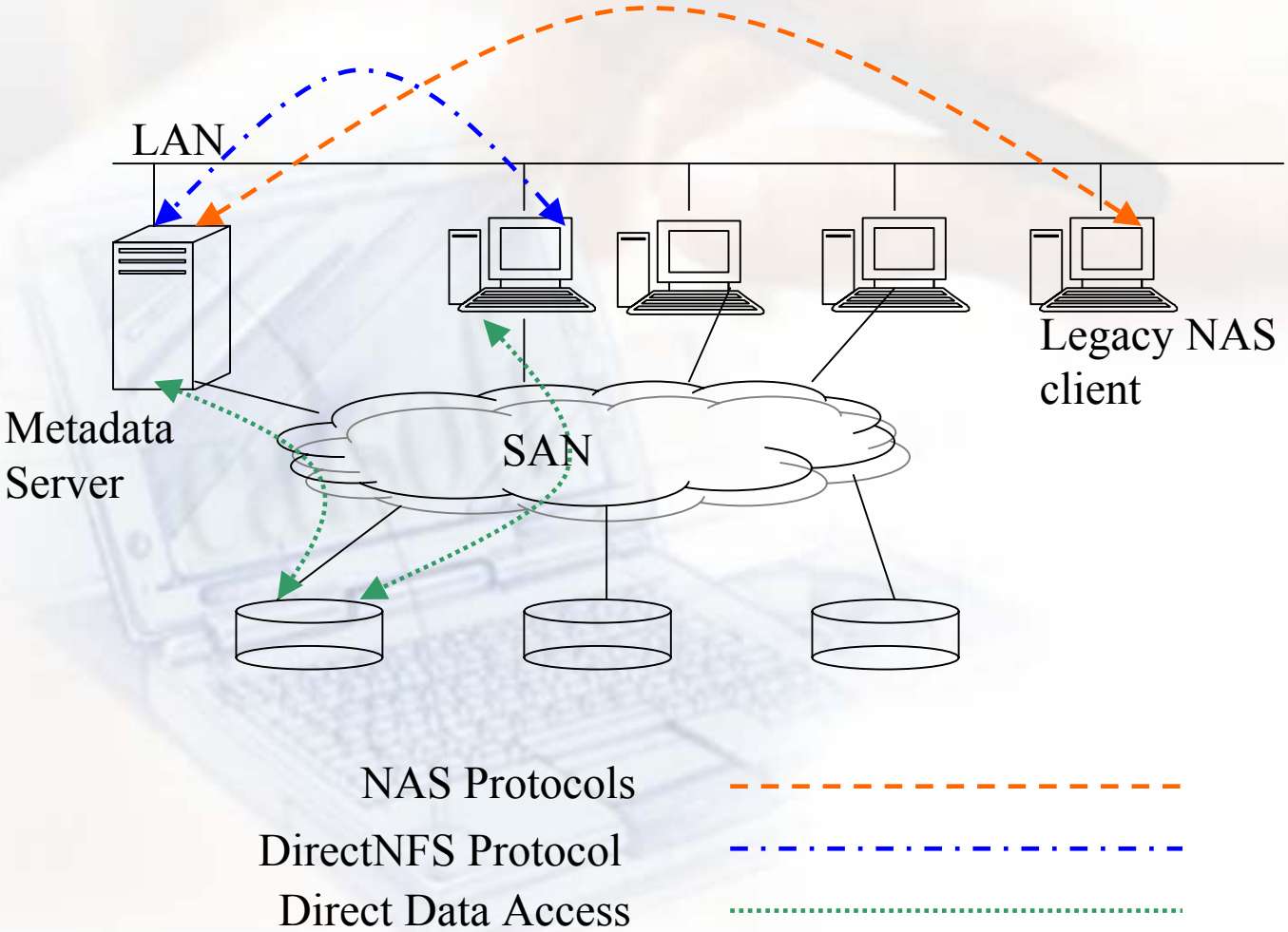
DirectNFS: Value Proposition



DirectNFS NAS to SAN

- NAS is Well established Standard
- SAN is slowly penetrating the Storage Industry
- Cost of Migration
- DirectNFS offers a seamless upgrade path, which No other cluster File System offers

DirectNFS Architecture



calsoft
www.calsoftinc.com

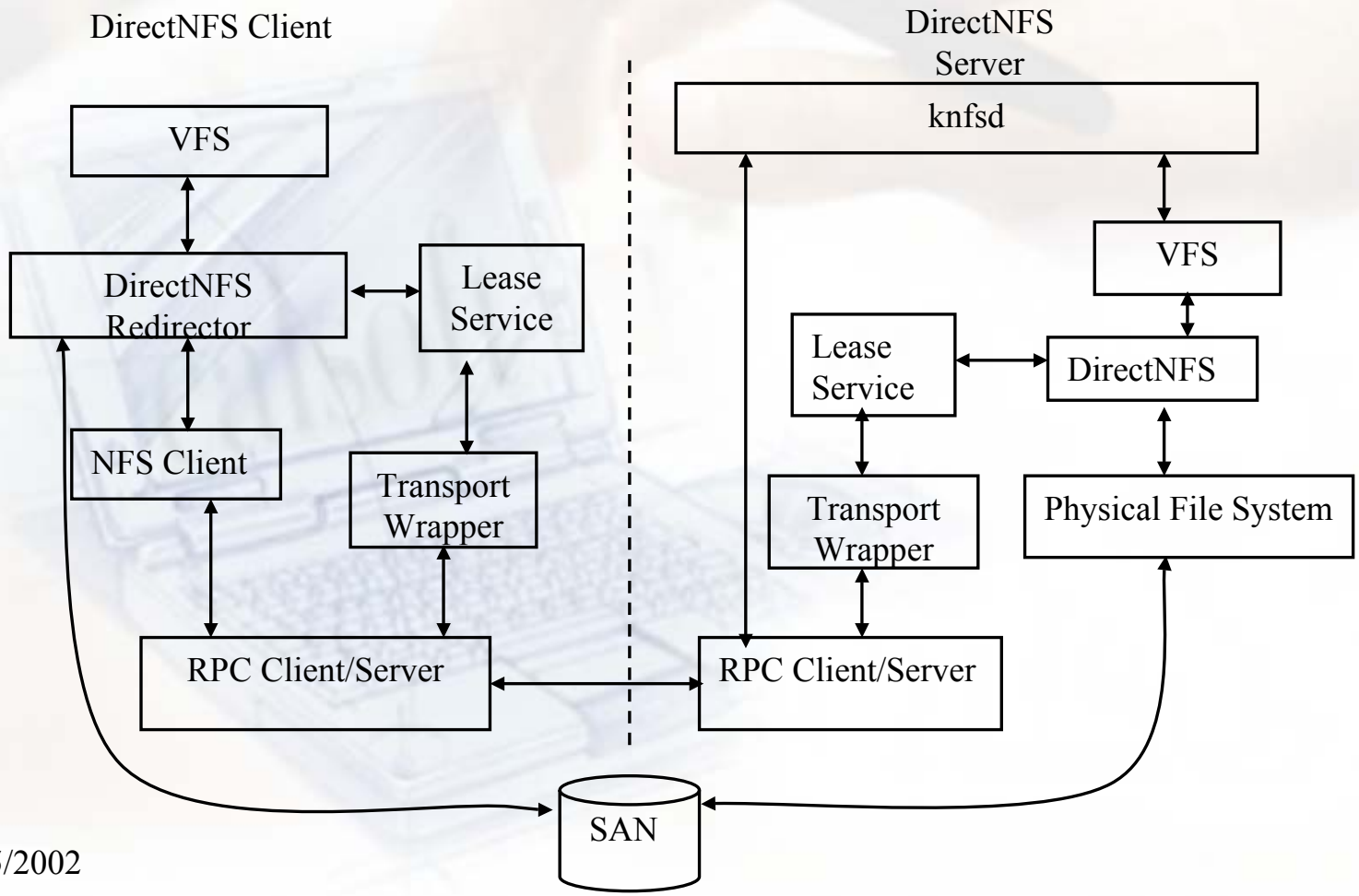
DirectNFS vs Cluster File Systems

- Simplicity
- Ease of Implementation
- De facto Standard
- Portability
  - Portability wrt. To Operating Systems
  - Portability wrt. To Physical File Systems

- Complex to Implement
- Suffer from Scalability problems
- No Standards
- Lack of Portability

calsoft
www.calsoftinc.com

DirectNFS Software Architecture



Overall architecture

- Extensions to NFS
- Metadata caching
- Cache Coherency through leases
- Write Gathering
- Security considerations

Extensions to NFS

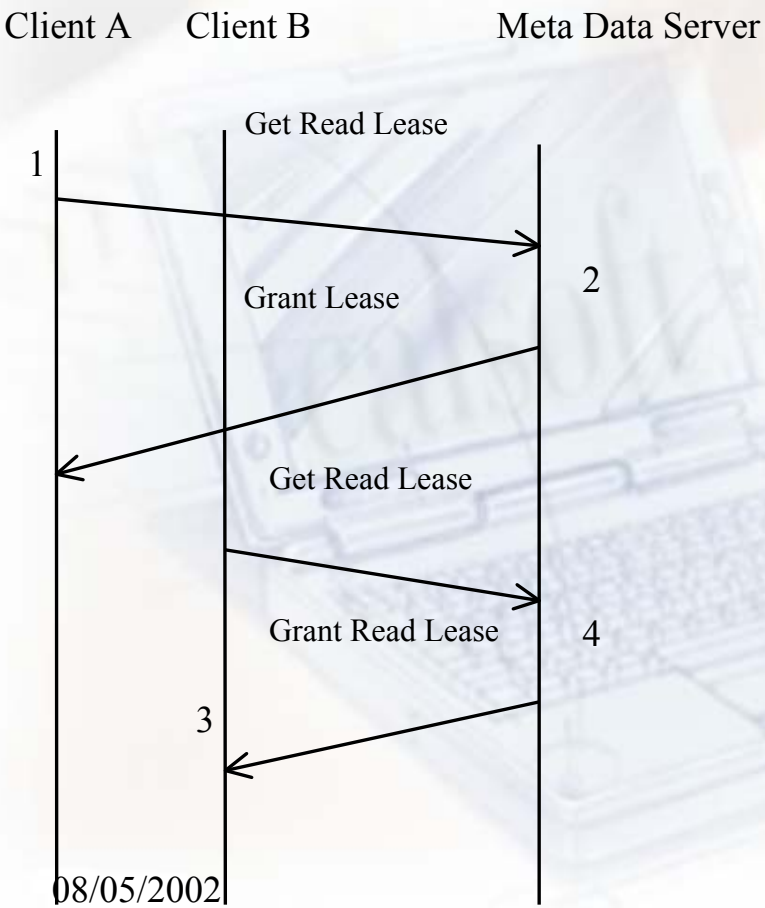
- GETBLKLIST
- Lease RPCs
  - GETLEASE
  - VACATELEASE
  - VACATEDLEASE

Metadata caching & coherency

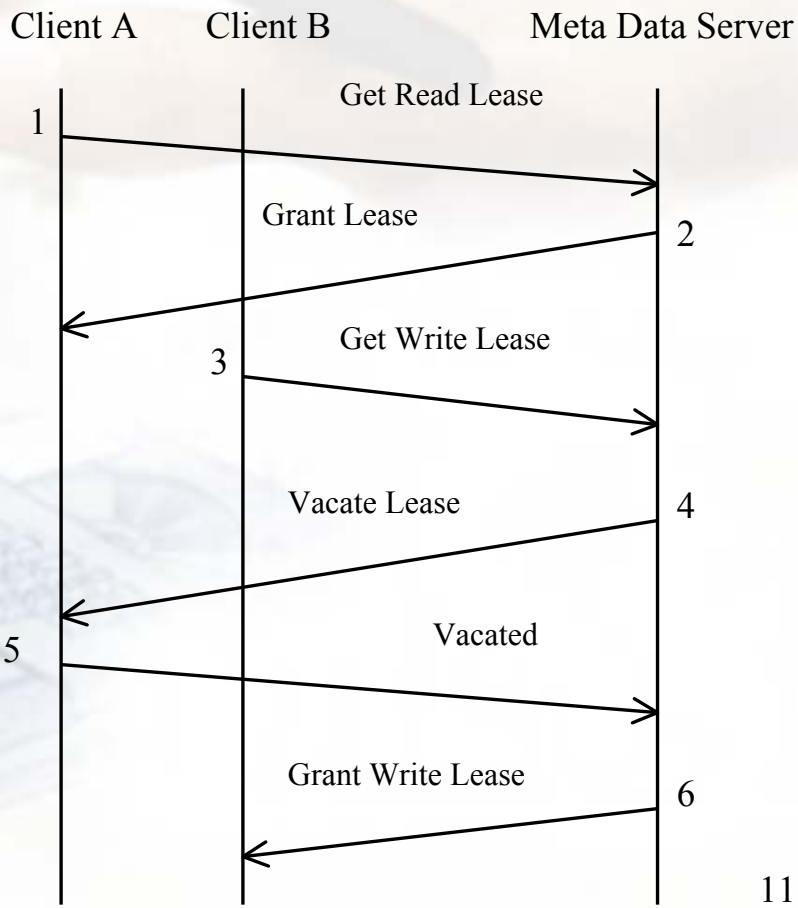
- Block lists are cached on client
- LRU style cache
- Need coherency: done through leases
- Leases: functionally similar to DLM
- Lease = Time bound lock
- Leases need periodic renewal
- Lease expiry implies:
  - Clean data needs to be discarded
  - Dirty data needs to be written out

Lease Protocol Interactions

- Read Lease



- Write Lease



Linux Prototype

- Used FiST
  - Stackable file system generator
  - Helps portability across OSes
  - DSL language for describing file-system filters
  - Generates code skeleton for pass-thru stackable file system
  - Used FiST to write DirectNFS Redirector module

Linux Prototype (contd)

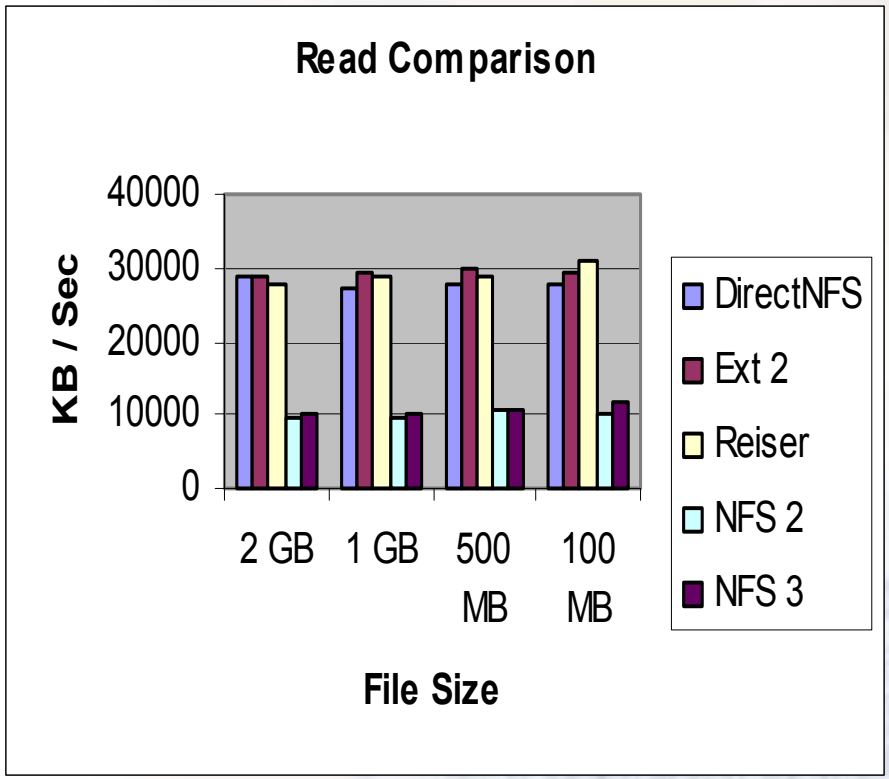
- Redirector module traps open, close, sync, unlink, read & write calls
- Lease obtained
- Read & Write: Block Number Cache is looked up
- Cache miss triggers request to server
- Reads & Writes go directly to disk

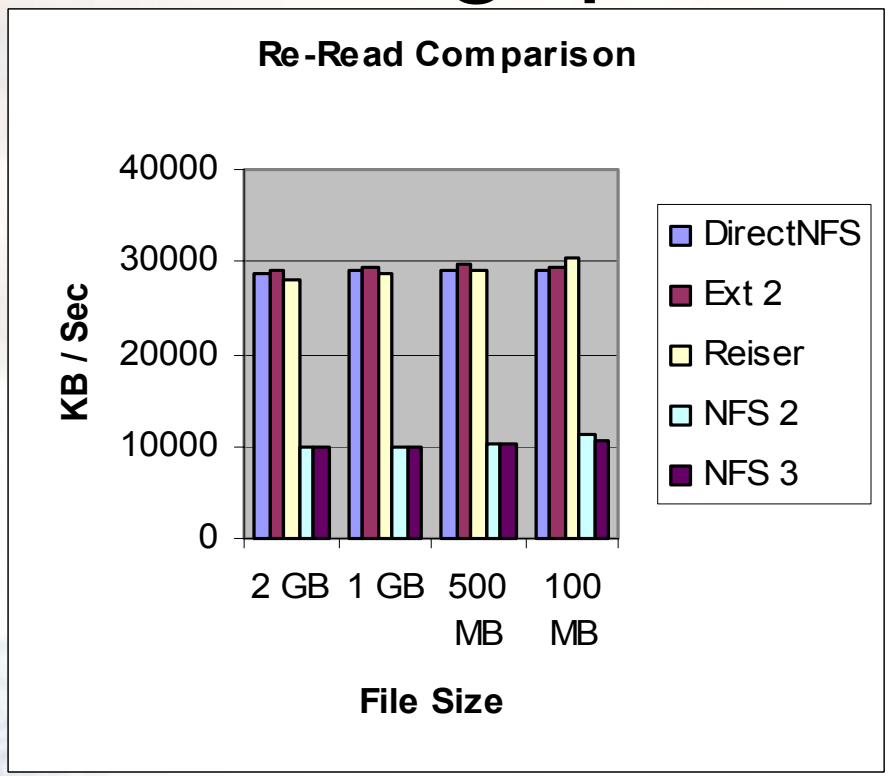
Performance Measurements

- Measured against Ext2, ReiserFS, NFSv2 & NFSv3
- Shared SCSI used to emulate a SAN
- Iozone used to study performance
- Varying file sizes and record sizes made no difference
  - Used 2GB file in all the experiments

calsoft
www.calsoftinc.com

DirectNFS Read Throughput

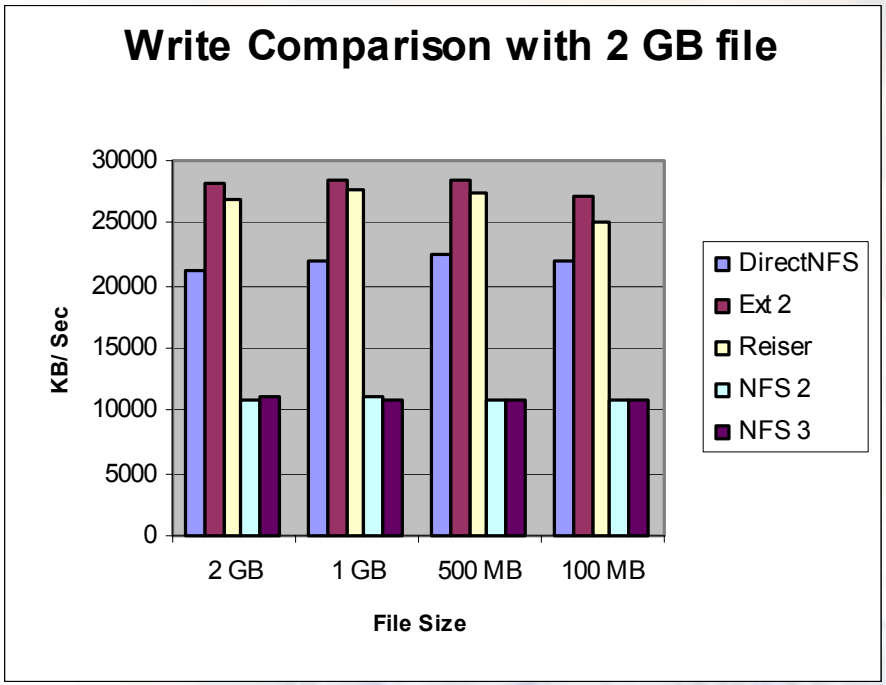


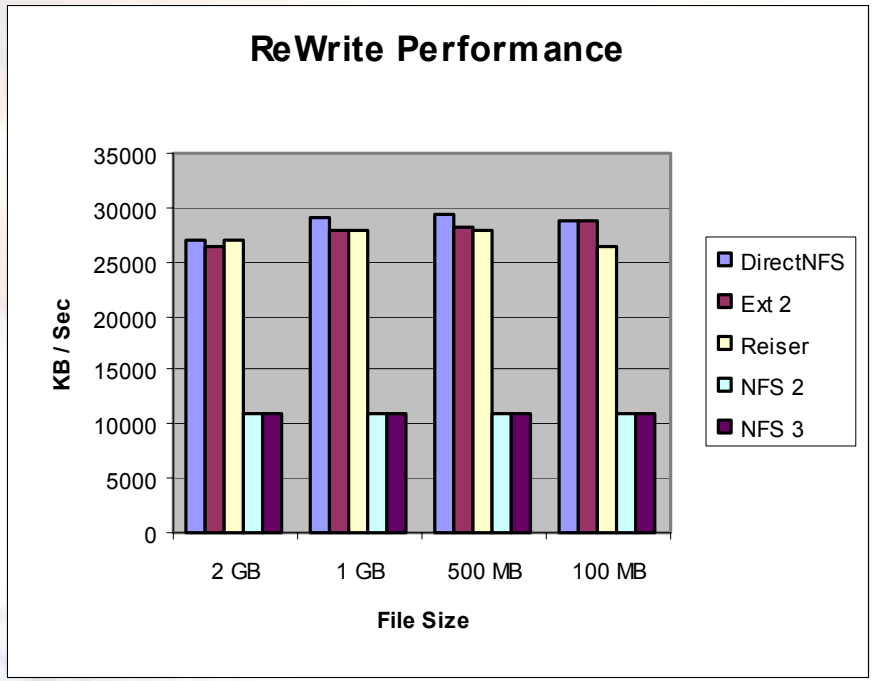


- Avg. Throughput = Local FS throughput
- Approx. 3x NFS v2/v3 throughput

calsoft
www.calsoftinc.com

DirectNFS Write Throughput



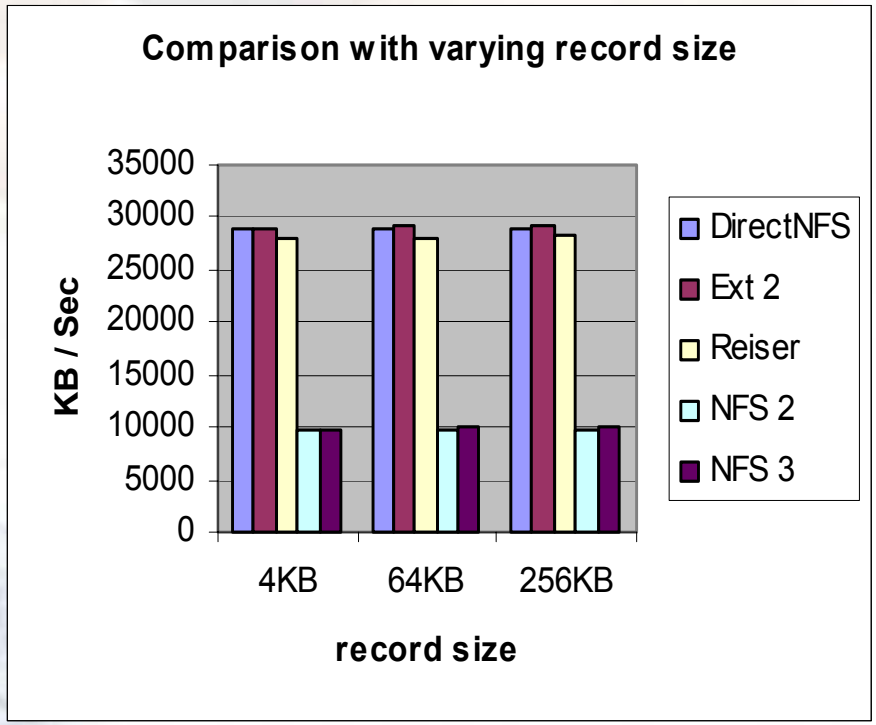


- Avg. Throughput > 60% of local FS
- Twice as fast as NFS v2/v3

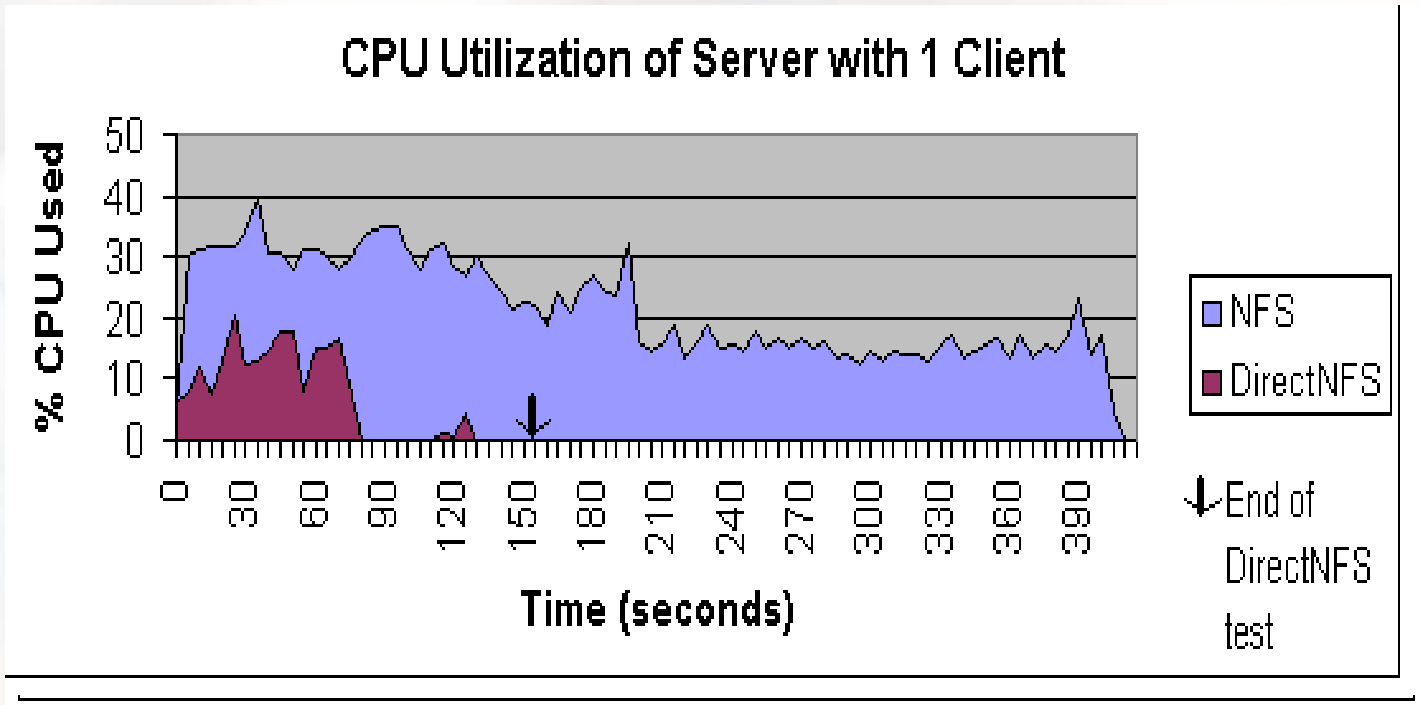
calsoft
www.calsoftinc.com

Effect of Record Size on Throughput

- ???



Server CPU Utilization



- Direct I/O Path Brings down Server CPU Utilization