Comparing Linux File Systems

Cary Whitney

NERSC

clwhitney@lbl.gov

Outline

- About Parallel Distributed System Facility (PDSF)
- Problems/Requirements
- First Attempt (Increase file system size)
- Second Attempt (NAS)
- Third Attempt (File system testing)
- Results
- Conclusion
- What next?

About PDSF

- PDSF is a Linux cluster of 200 dual Pentium class machines
- 30 storage nodes of .5 to 1 TB in size for a total of about 35 TB
- Fast Ethernet and Copper GigE interconnects
- Primarily serves the High Energy Physics community

Problems

- Problems
  - Required 20 TB of storage which then increased to ~50 TB for the year
  - Increased performance demands +80 compute nodes
- Constraint
  - Provide this storage for $300k
- Environment
  - HEP problems are data intensive
    - More reading than writing
    - Sequential in nature

Requirements/The Test

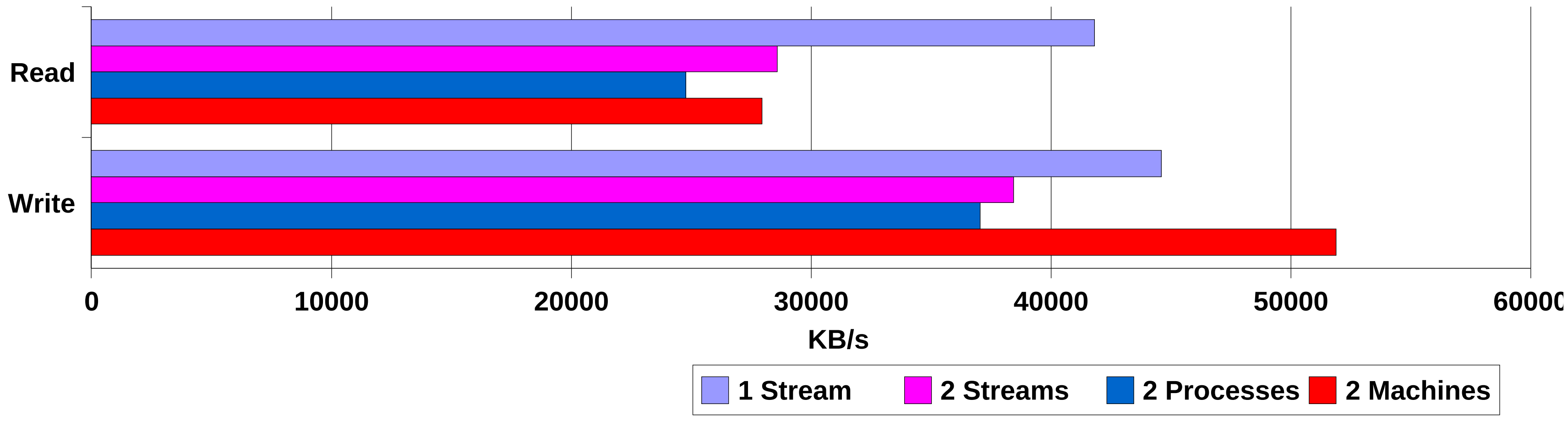
- Requirement order
  - Scales to large number of connections
  - Capacity that can be grown
  - Performance
  - Cost effective
- About the test
  - Upper level test. NFS
  - Any caching effects is on the server not client

Test Environment (PDSF)

- Server Dual 1.6GHz Athlon, 1GB memory
  - Raid Inc Fibre channel IDE Raid 5 box. (6 drives Raid + 1 Hot spare)
  - LSI 2 Gb fibre channel card
  - SysKonnect GigE card
  - 2.4.19-pre10 kernel with NFS_ALL patches
- Clients PDSF compute nodes
  - GigE machines for under 40 client tests
  - All systems above 40 clients

Benchmark

- Iozone was used as the benchmark
  - Sequential reads/writes with 1 GB files
  - Cycled through temporary files on the server
  - Limited to 1 process per client



First Attempt - Increase file system size

- Increase HD size. File systems now 1 TB
  - Up side
    - More data per system
    - Offered more storage for the budget
  - Down side
    - Increase demand per system
    - Overall system performance did not increase
  - Result
    - User disapproval because of performance

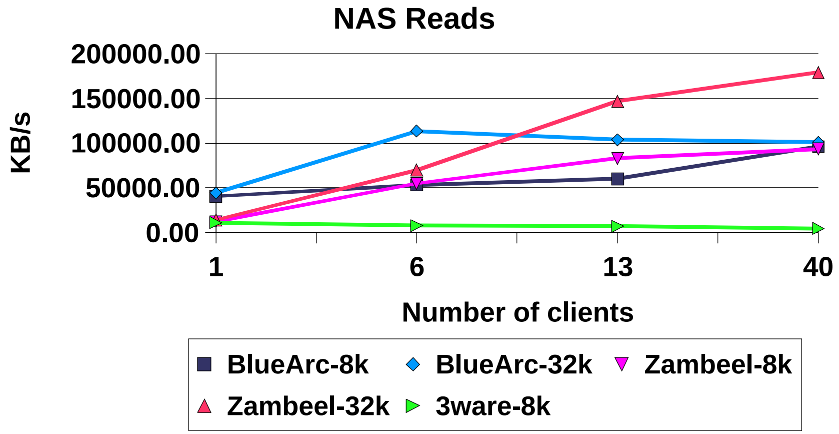
NAS - BlueArc

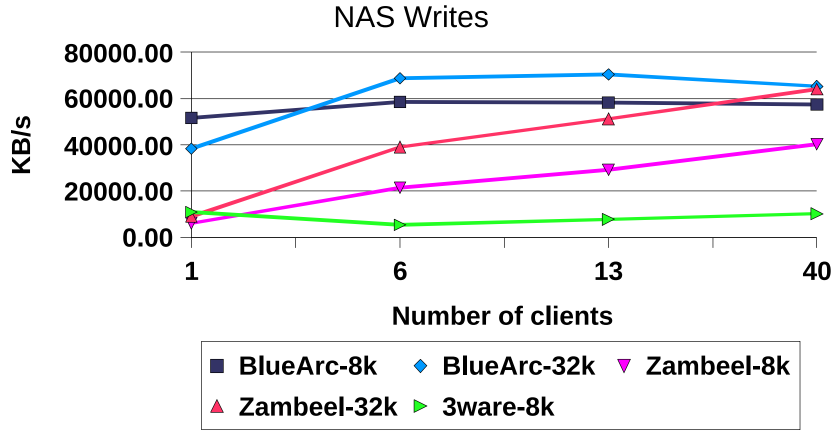
- A demo BlueArc
  - Good single stream performance
  - Limited to single GigE connection
  - Get maximum performance needed to use multiple volumes

NAS - Zambeel

- A beta Zambeel
  - Poor single stream performance
  - Expandable up to 22 GigE connections
  - No need for multiple volumes

Round 2 - NAS

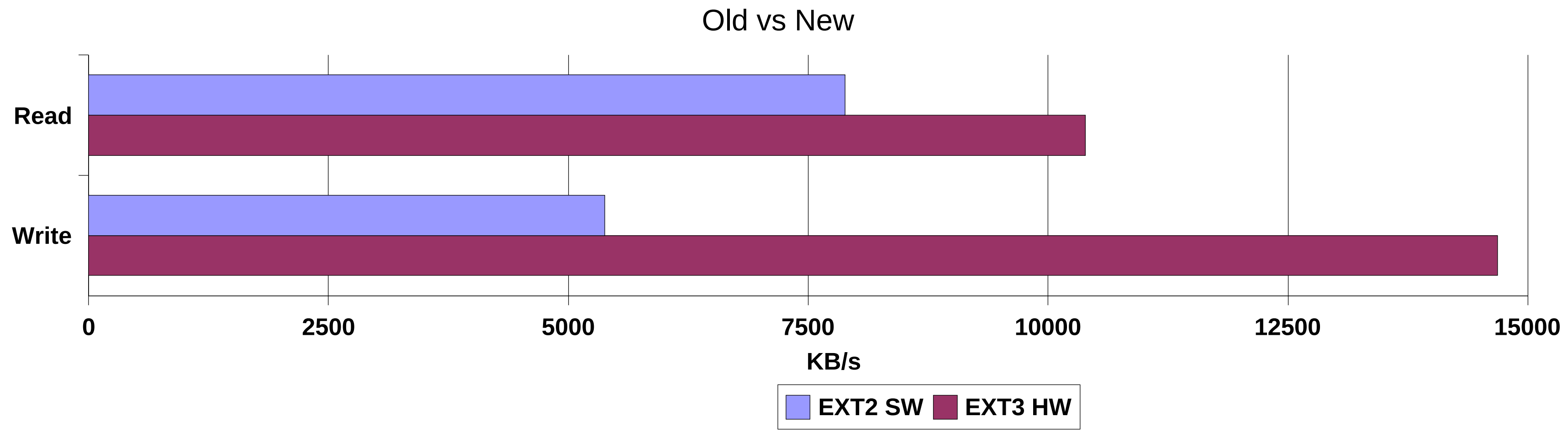




- NAS overview
  - Good scaling performance
  - Very reliable
  - Costly but could be used in certain areas
  - Change in storage requirements placed them out of range.

Round 3

- File system and configuration testing
  - SW vs HW raid
  - Move from a 2.2 to 2.4 kernel
  - EXT3 vs JFS vs ReiserFS vs XFS
  - NFS 8k vs 32k block size



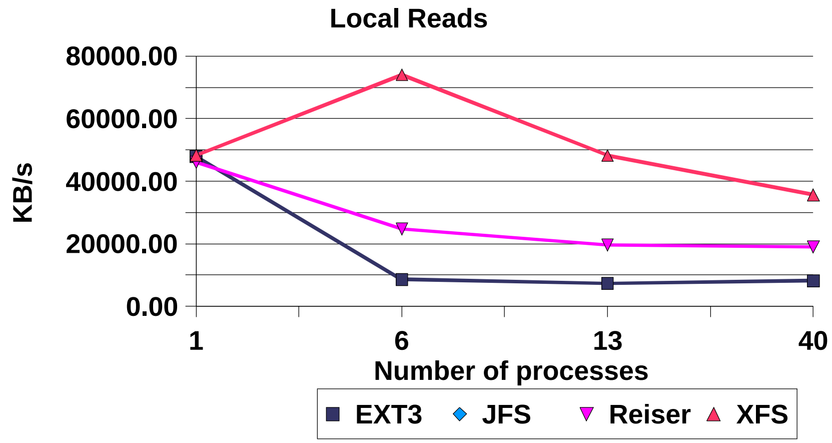
GPFS

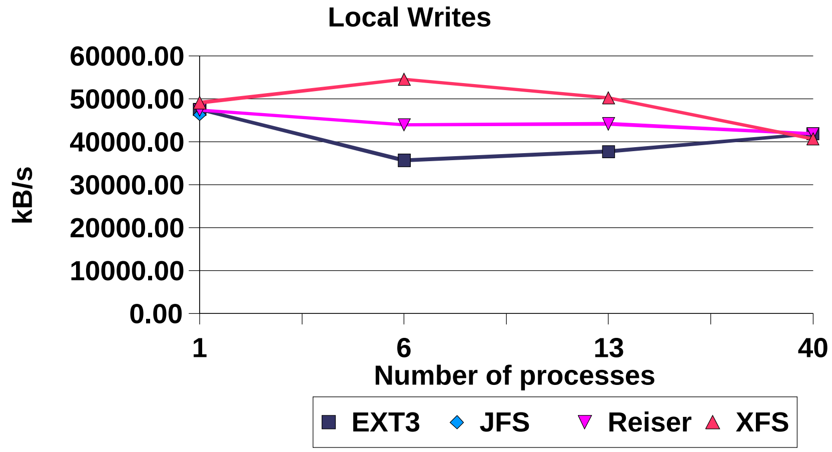
- Alvarez Linux cluster
  - 120 dual processor PIII 866 machines
  - 2 I/O Node GPFS servers with 2 GB memory each
  - Myrinet 2000 interconnects between computer and I/O nodes
  - The same Iozone setup

GPFS Limitations

- To be usable on PDSF we would need to do:
  - Run GPFS across FastE
  - Install some interconnect network for the file system
  - Down grade our kernel
  - Possible hardware changes
- Or treat GPFS as a NAS solution thus loosing the benefit of a cluster file system
  - Using Linux GPFS as the back end (Not tested yet)
  - Using Seaborg as the back end (Not sure if we can)

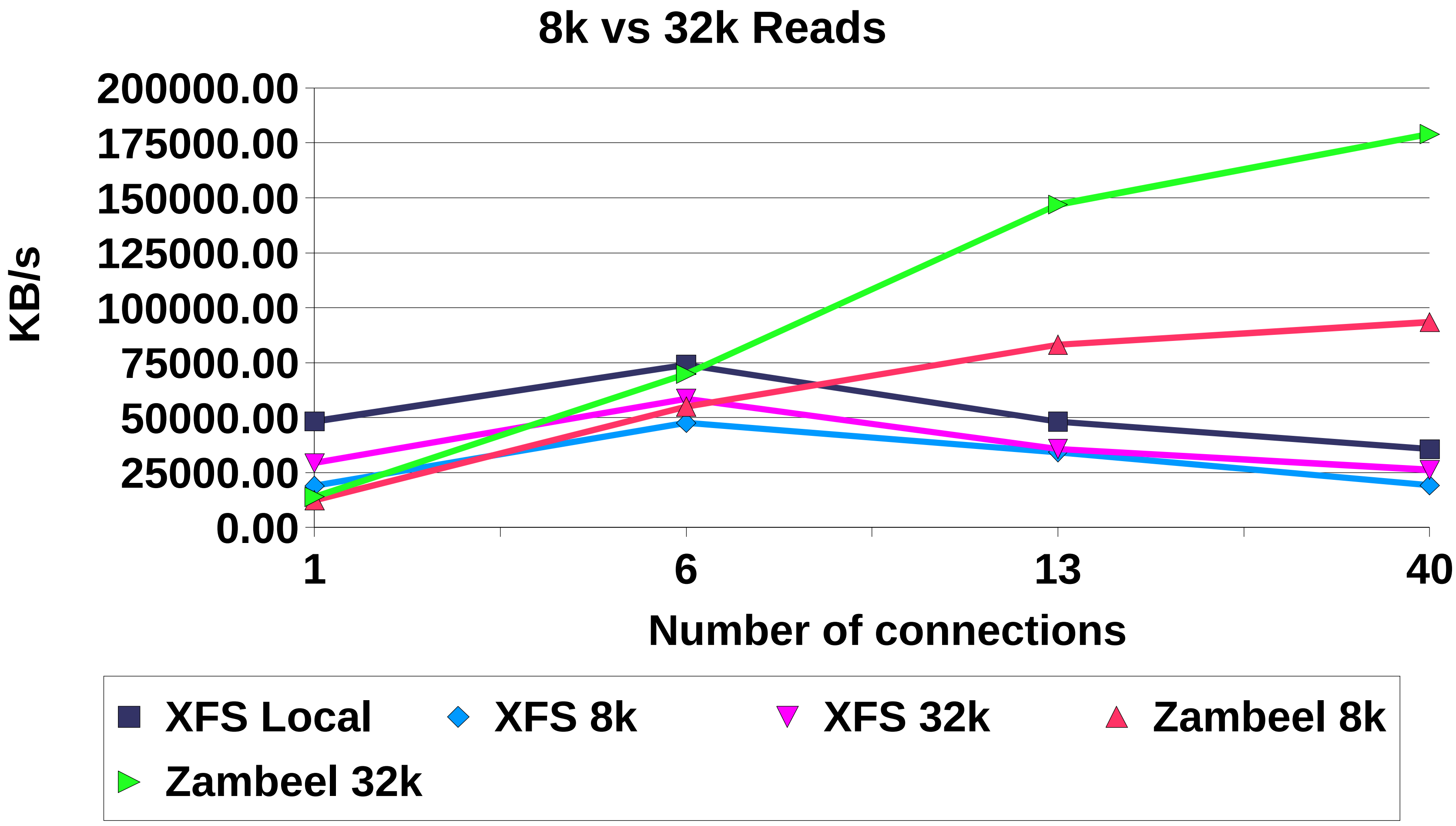
Local Test



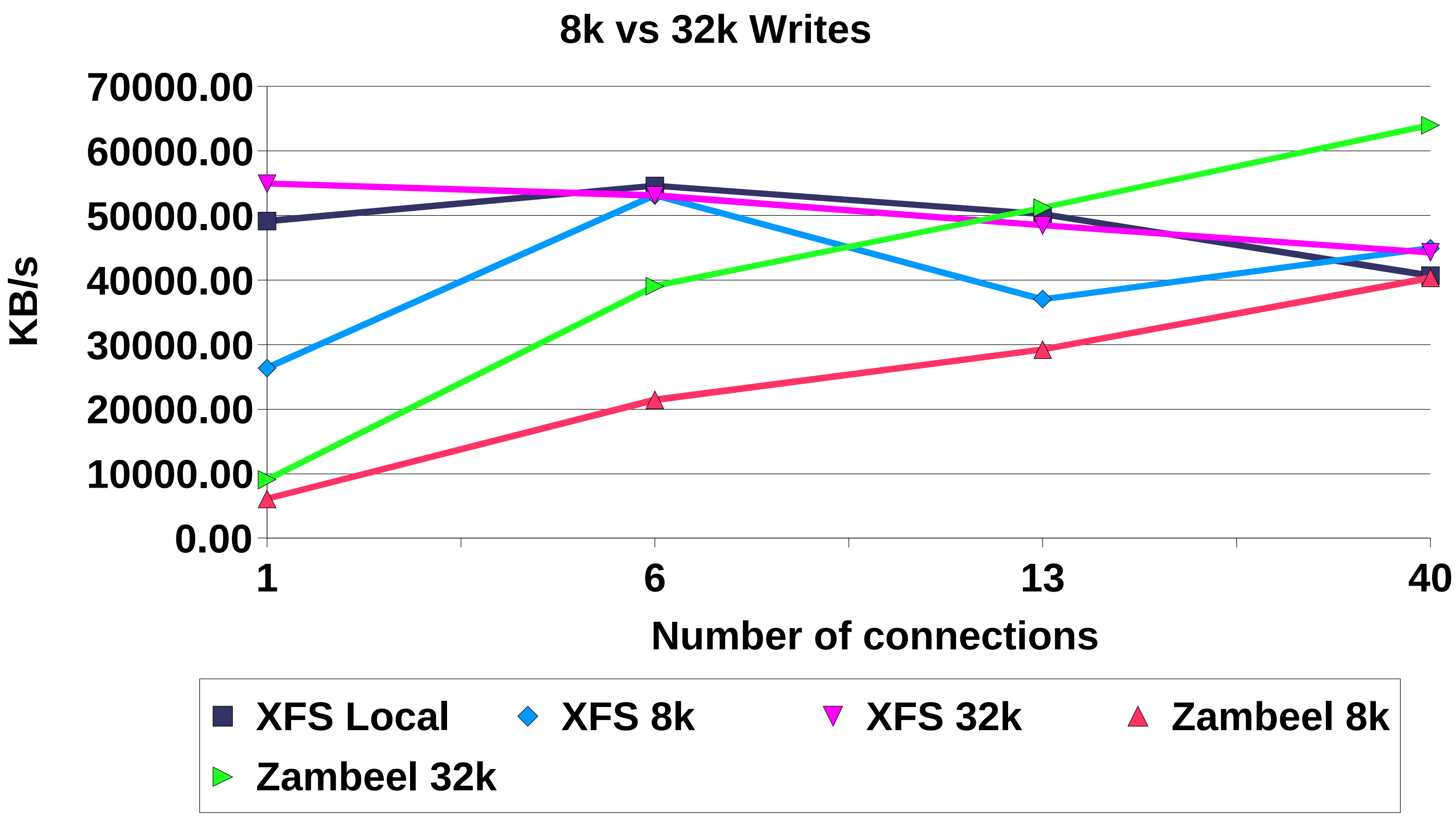


- Base numbers for NFS to reach
  - The best local should be the best for NFS?
  - Can the file system scale locally? Nfsd

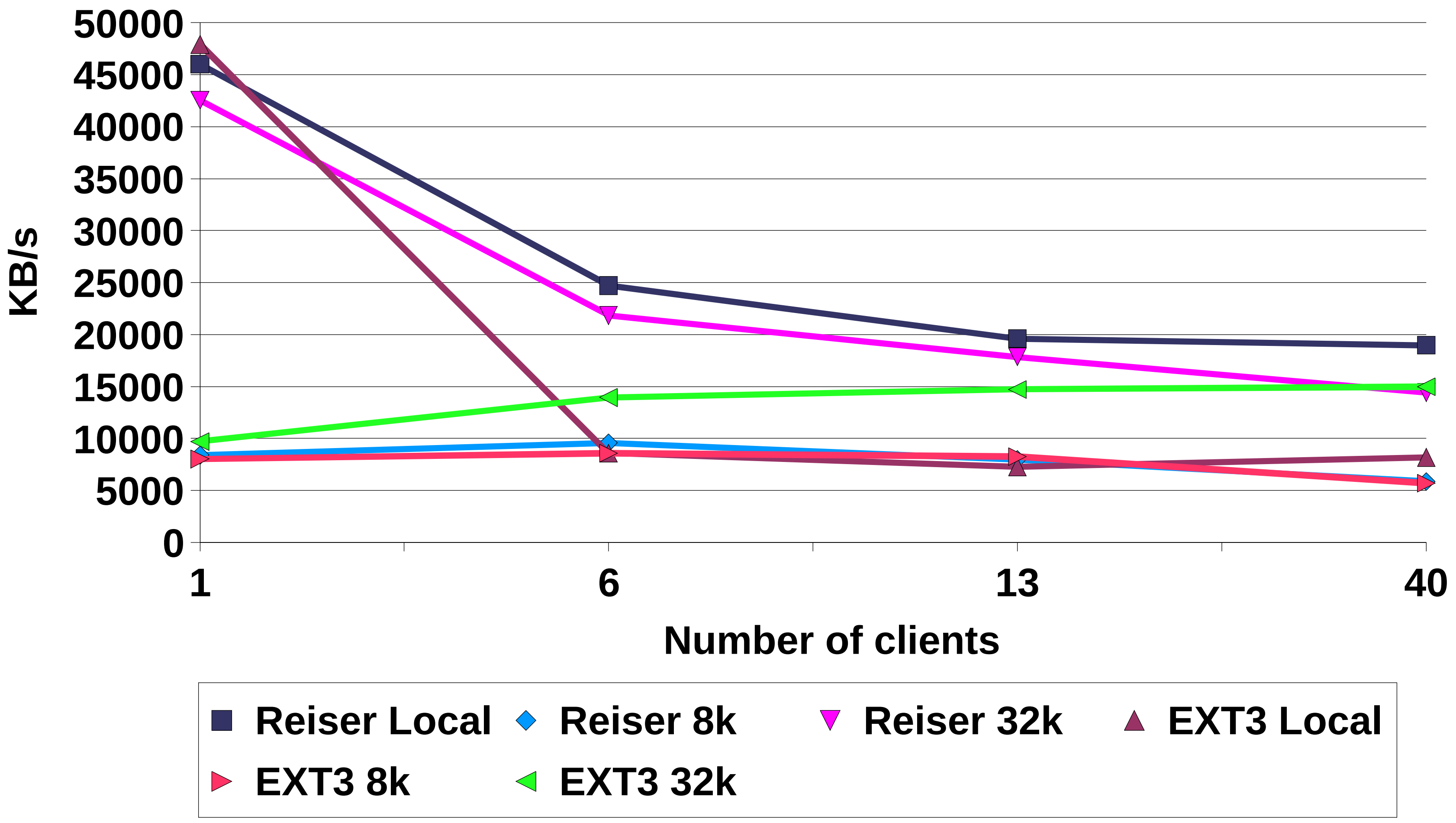
8k vs 32k Read



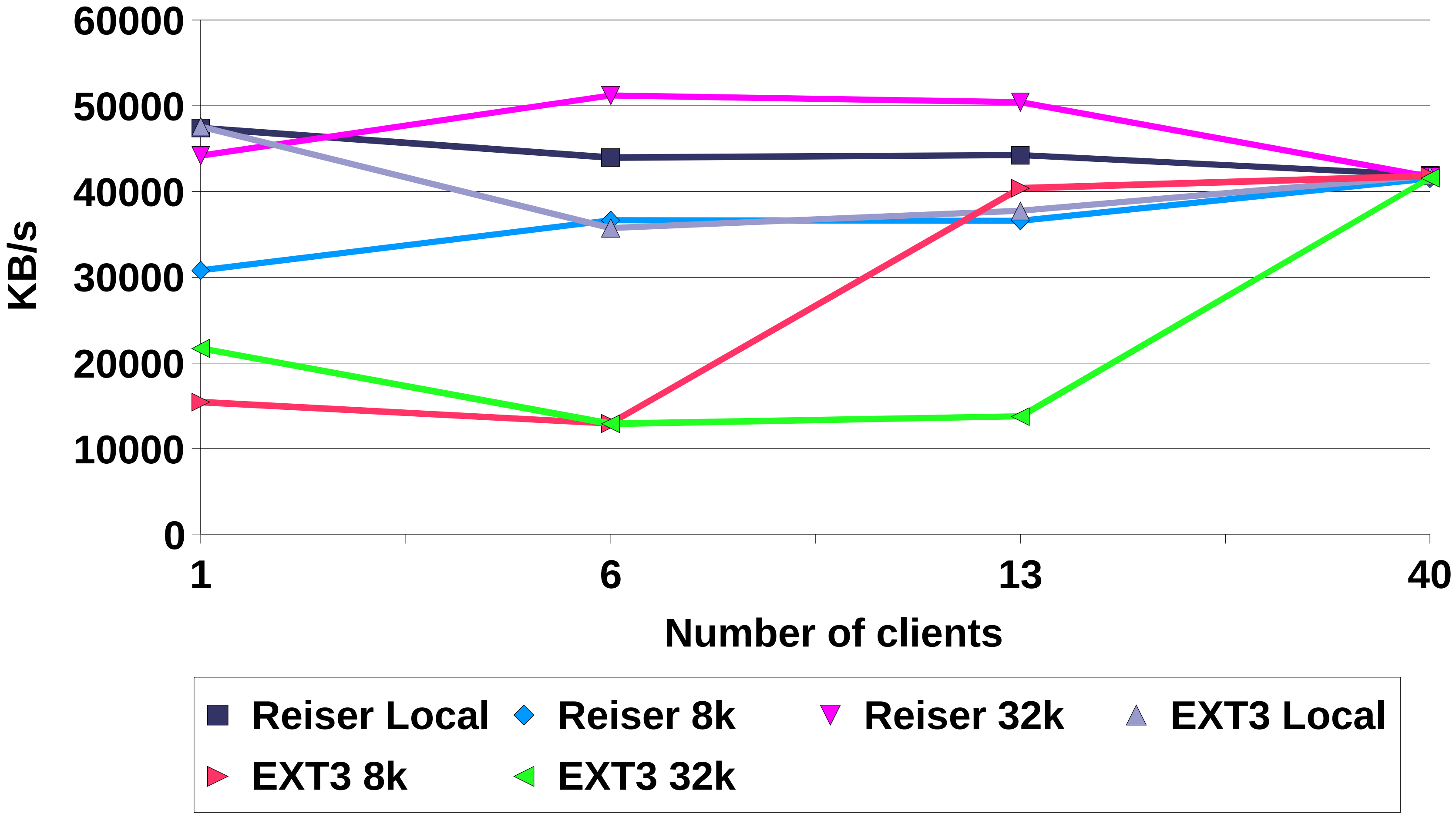
8k vs 32k Write



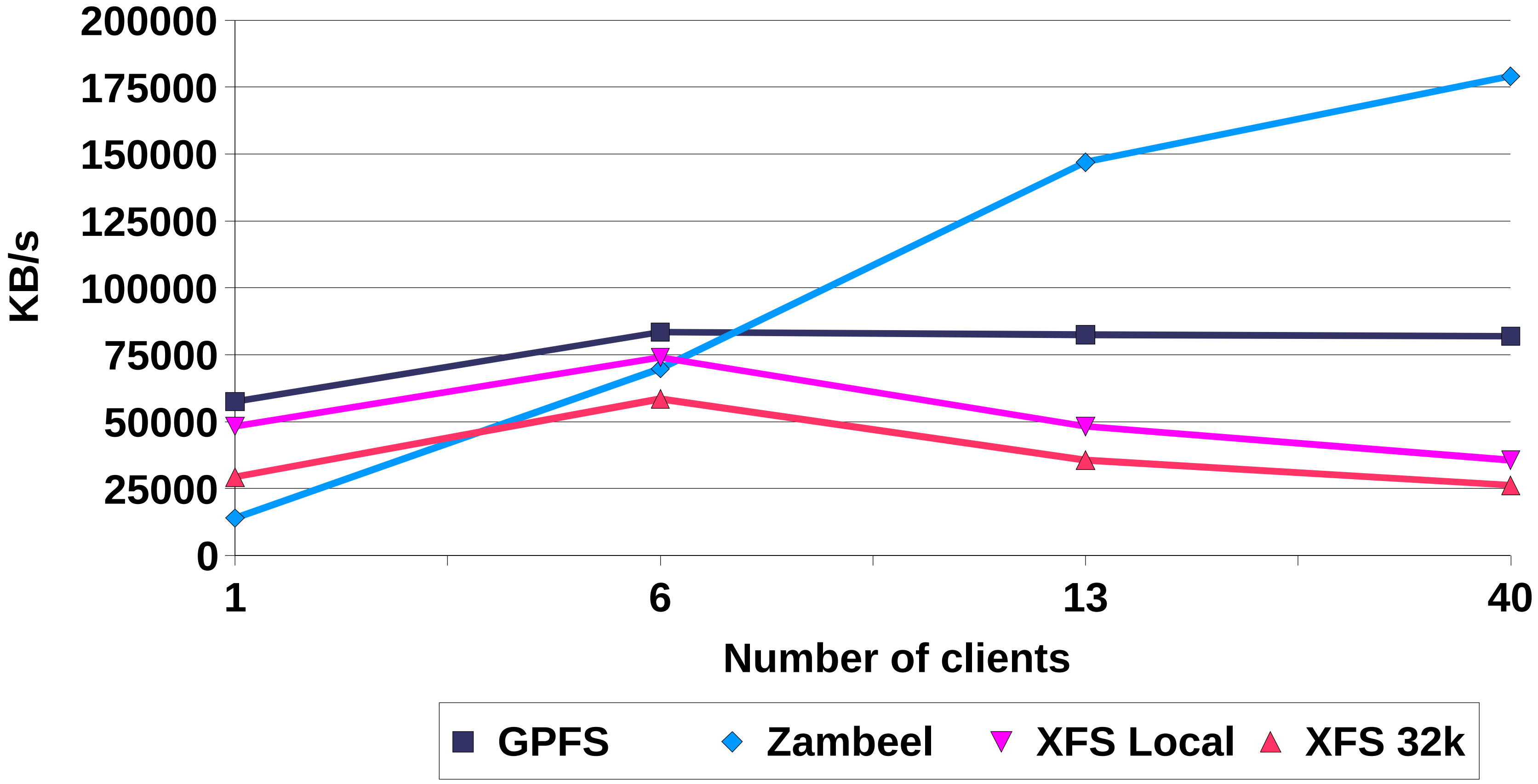
8k vs 32k Reads



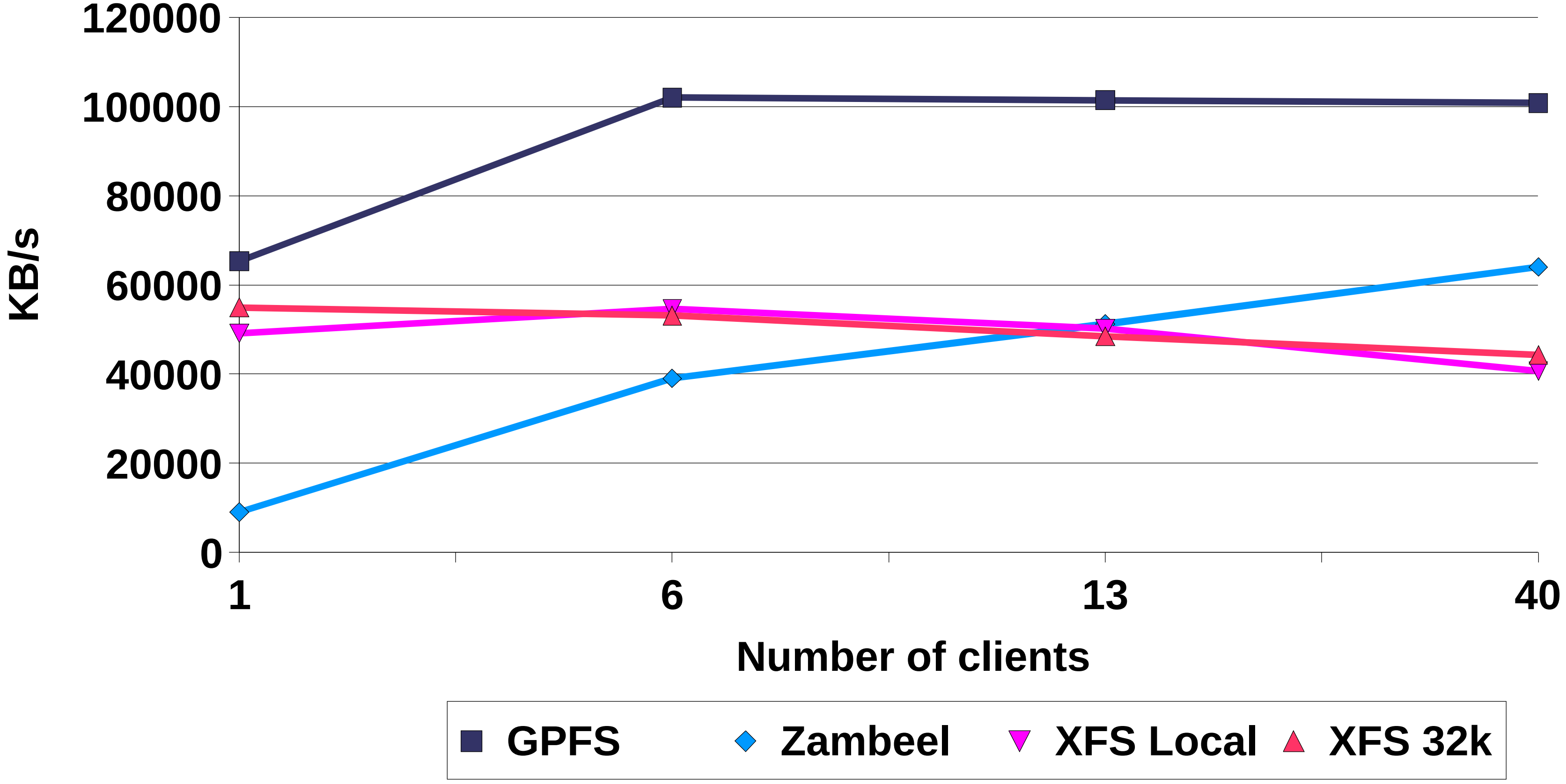
8k vs 32k Writes



The Best Reads



The Best Writes



Problems/Limitations

- JFS
  - Oops with NFS 40 client test and Qlogic card
  - System hang with 6 process local test
- GPFS
  - I/O servers failover
  - Limited to certain hardware and 2.4.9 RedHat kernel

Problems Encountered 2

- XFS or all the other native file systems
  - VFS modification by XFS
  - In general non-xfs kernel was better for 8k an xfs kernel was better for 32k
- Sun T3 Storage Edge really poor read performance no matter what configuration. Not repeatable with a different fibre channel disk setup.

Conclusion

- XFS looks like a match for us
  - ReiserFS also performed well
- We tested only one aspect of a file system.
  - We did not test meta data access.
  - We did not test random reads/writes.
- Selected Zambeel for home file system
- More information at http://pdsf.nersc.gov/

What Next?

- Server side changes
  - Big Kernel Locks (BKL)
  - Bounce Buffers
  - 2.5 kernel
  - 2 Gb fibre
  - PCI-X
- Network/Transport changes
  - NFS over TCP
  - Jumbo Frames