AIX NFS Client Performance Improvements for Databases on NAS

October 20, 2005

Sanjay Gulabani
Sr. Performance Engineer
Network Appliance, Inc.
gulabani@netapp.com

Diane Flemming
Advisory Software Engineer
IBM Corporation
dianefl@us.ibm.com

Agenda

- Why Databases on NAS?
- AIX NFS improvements - CIO
- Simple IO Performance Tests
- OLTP Performance Comparison
- Future work
- Conclusion
- Q&A

Databases on SAN

- Databases do block based I/O and prefer raw blocks
- But, most admins still put a volume manager and a file system: JFS2, VxFS, UFS over HW RAID storage

  Reason: Simplicity -- Easier backups and provisioning

NAS Myths

- **Myth**: NFS consumes more CPU for Databases

  Reality: Most NFS client performance problems with DB apps are due to kernel locking and are fixed in good NFS clients

NAS Myths

- **Myth:** Ethernet is slower than SAN
  Reality: Ethernet is 1Gb, FC is 2Gb. 4Gb FC is almost here but so is 10GbE.
  - Ethernet is likely to takeover bandwidth of FC soon; cost effectively
  - Its easy to setup multiple wires to match FC bandwidth with just 1GbE, today
  - Storage latencies at database layer are measured in msecs, differences between SAN and NAS latencies are in usecs
    78usec more for 8K blocks < 0.1msec difference!

Why NAS?

- Networking is simpler
- Networking promotes sharing
- Sharing = Higher utilization (~grid storage)
- NAS solutions are cheaper than SAN
- No one has won against Ethernet!
- Seriously, even blocks based storage is moving to Ethernet (iSCSI)

Why DB on NAS?

- Complex storage management is offloaded from the database servers to storage appliances
- Performance vs. manageability
  - Database requirements different than traditional NFS workloads: home directories, distributed source code
  - single writer lock can be a problem

Industry Support for NAS

- Oracle On-Demand hosts ERP+CRM 1000+ Linux servers on NetApp NFS
- Yahoo runs Database on NetApp
- Sun 2003 presentation McDougall + Colaco promoted NFS
- NetApp #1 Leader of NAS $1.59b(FY05)
- IBM AIX 5.3 ML-03 improvements now
- No one got fired for buying from IBM! (IBM resells NetApp products)

Oracle's Austin Datacenter

- More than 15,000 x86 servers
- 3.1 Petabytes of NetApp storage
- 100s of mission-critical hosted Oracle apps

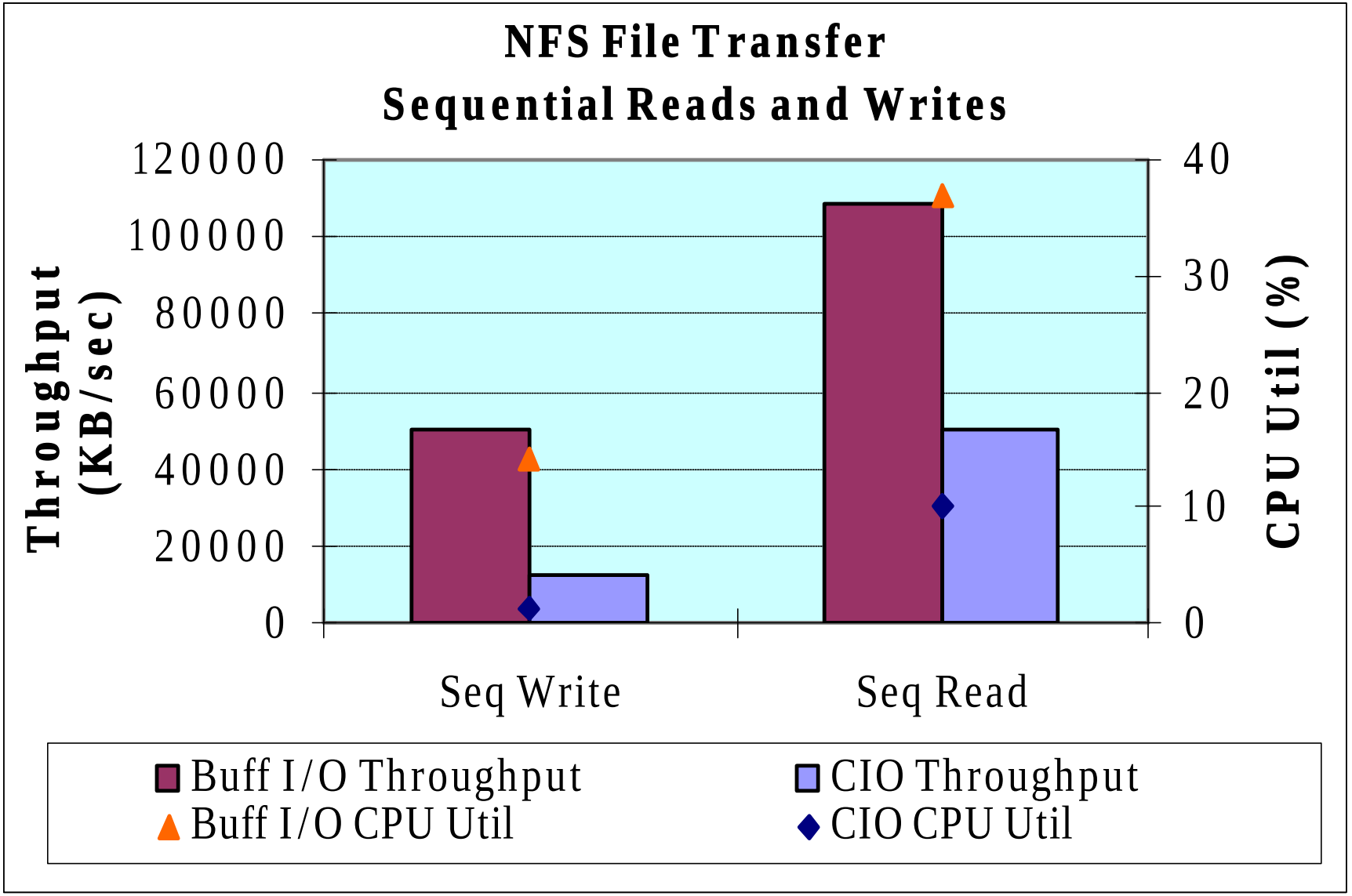
Source: Oracle Magazine Mar/Apr 2005

I/O, I/O… It's off to disk I go

- Raw logical volumes
  - No memory management overhead
- Buffered I/O
  - Variety of kernel services
- Direct I/O
  - No memory management overhead
  - defaults to synchronous accesses over NFS
- Concurrent I/O
  - DIO plus no inode/rnode contention

NFS File Transfer

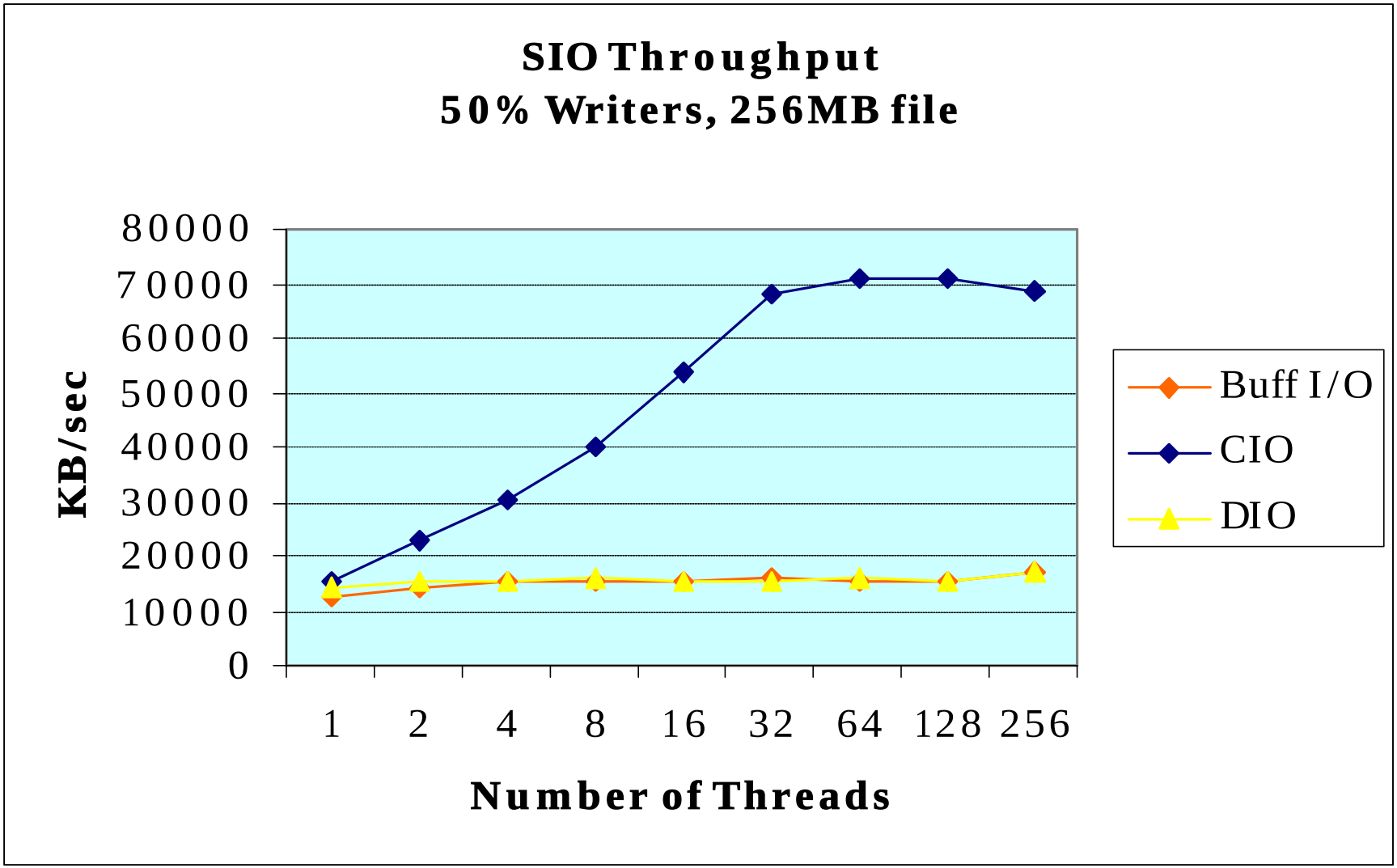
- Metrics:
  - CPU utilization and throughput in KB/sec
- Options:
  - Default (buffered I/O) and CIO
- AIX Client and Server
  - p630 1.45GHz 4-way
  - GbE 1500-byte MTU
  - AIX 5.3 ML-03
- Demonstrates algorithm behavior
  - Atomic, Single-Threaded



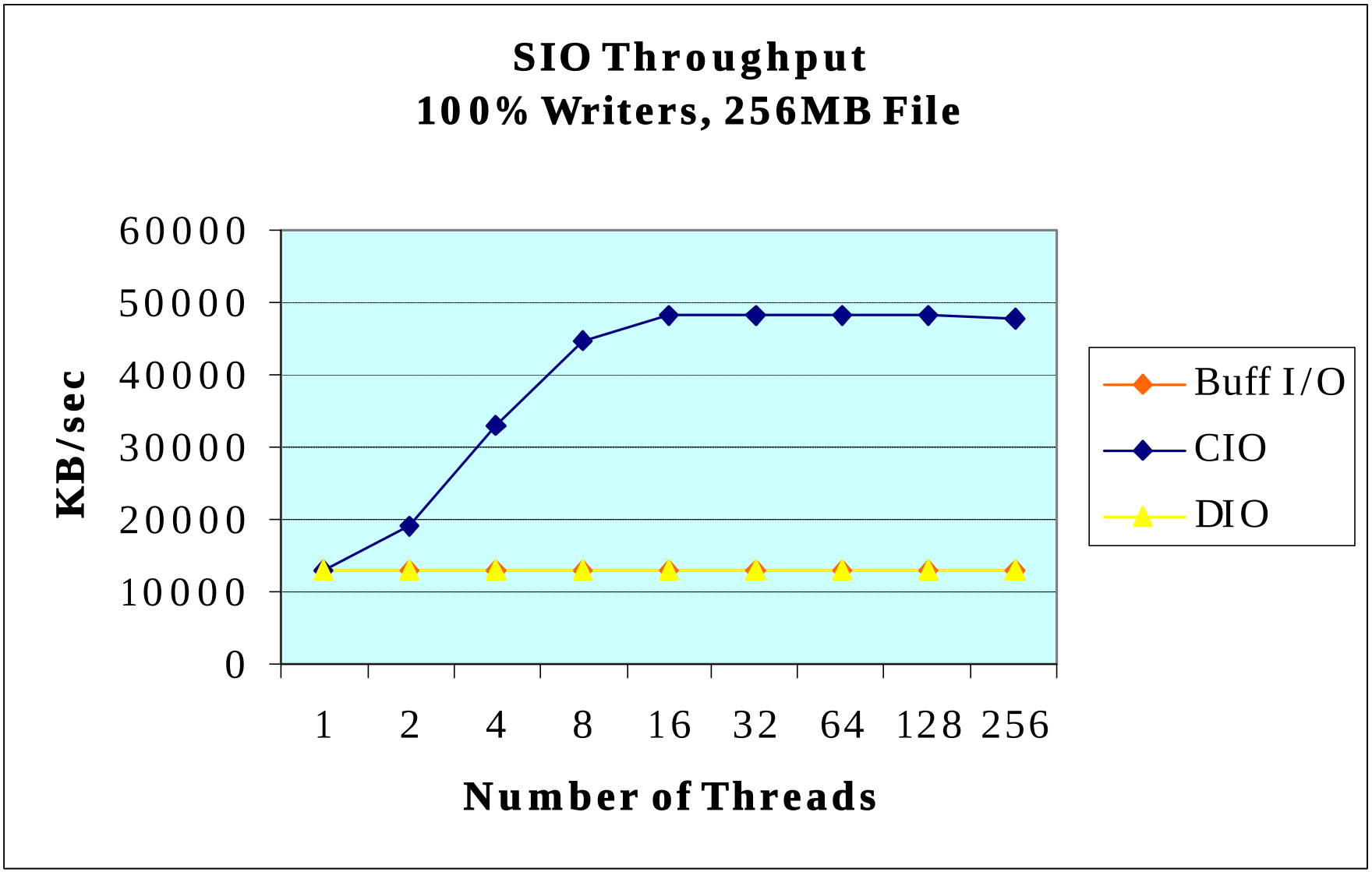
256MB File, NFS V3, TCP, auth=sys, 32KB RPCs

Simple I/O Load Generator

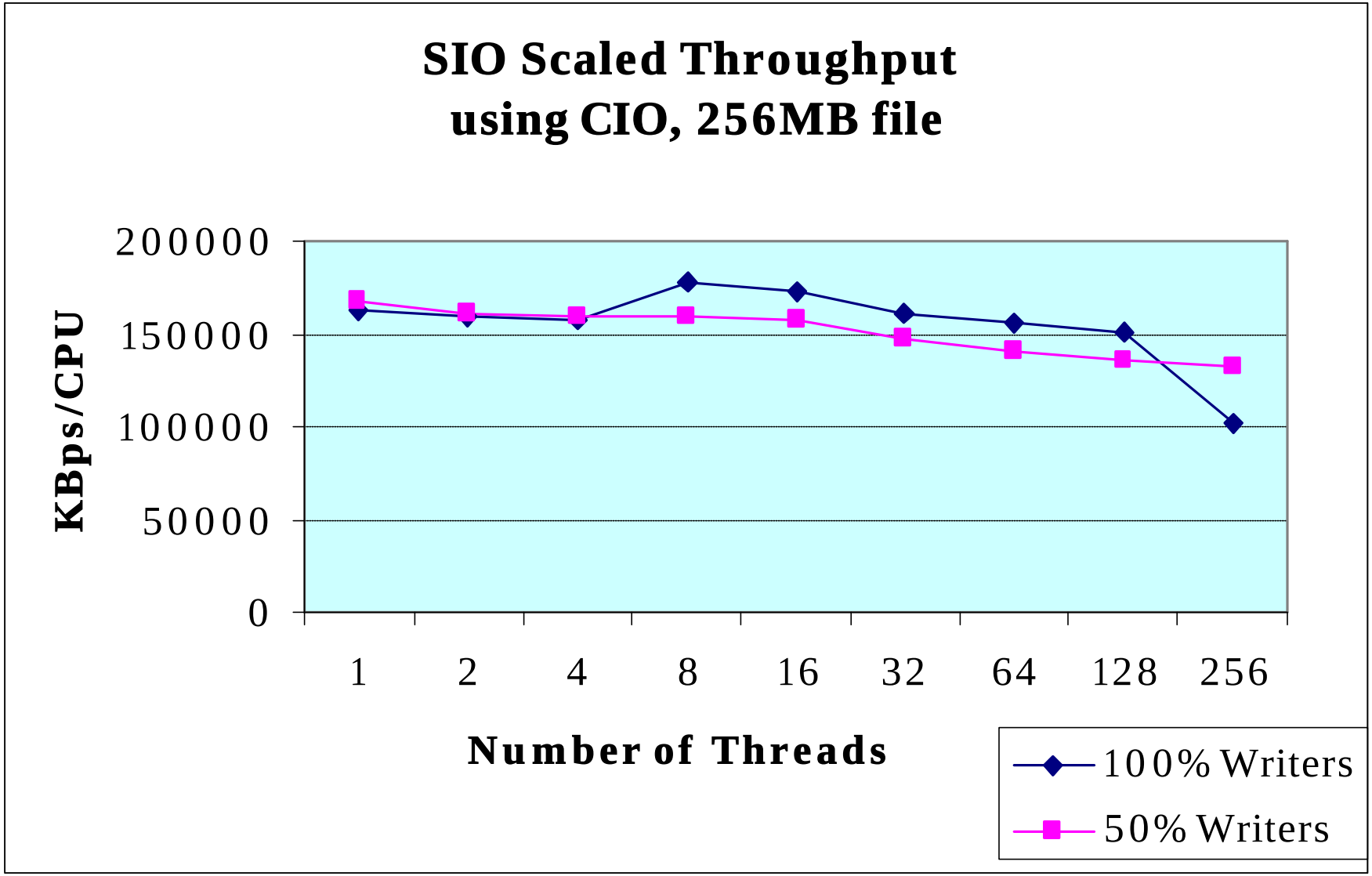
- Metrics
  - CPU utilization, IOPS, throughput, etc.
- Options
  - NFS client mount options to specify I/O mode: Default (buffered I/O), DIO, CIO
- NetApp Filer FAS880
  - ONTAP 6.5.1
- AIX Client Model p630 1.45 GHz 4-way
  - AIX 5.3 ML-03
  - GbE 1500-byte MTU



256MB File, NFS V3, TCP, auth=sys, AIX v5.3 ML-3 client



256MB File, NFS V3, TCP, auth=sys, AIX v5.3 ML-3 client



256MB File, NFS V3, TCP, auth=sys, AIX v5.3 ML-3 client

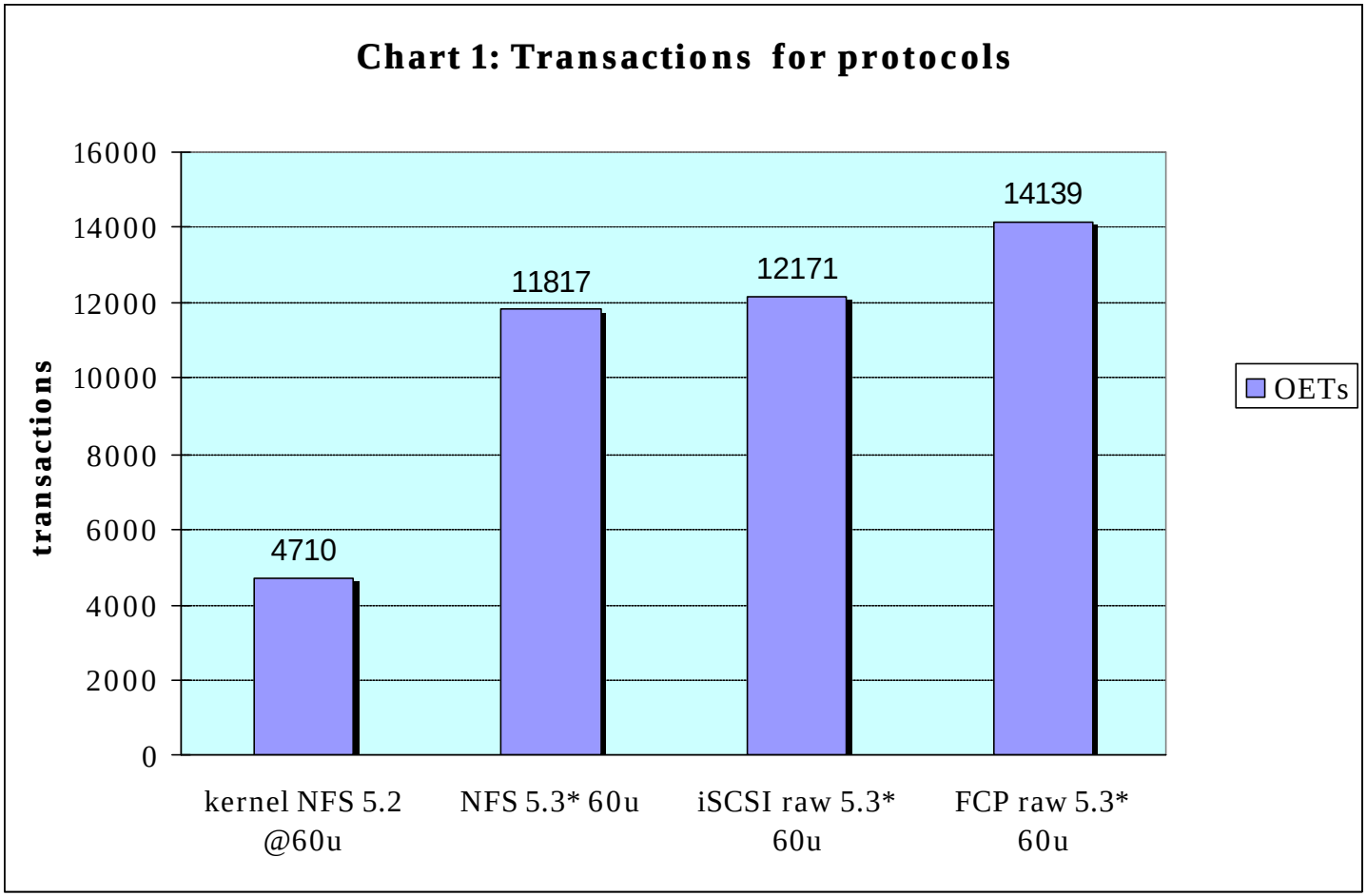
NFS FT and SIO Summary

- CIO not appropriate for single threaded sequential read/write workloads.
- CIO aimed at workloads with higher concurrency such as databases.
- Significant gain in raw throughput performance using CIO vs. Buffered I/O for workloads with higher concurrency.
- CPU utilization illustrates issue with TCP socket lock contention with increased concurrency.

OLTP Performance

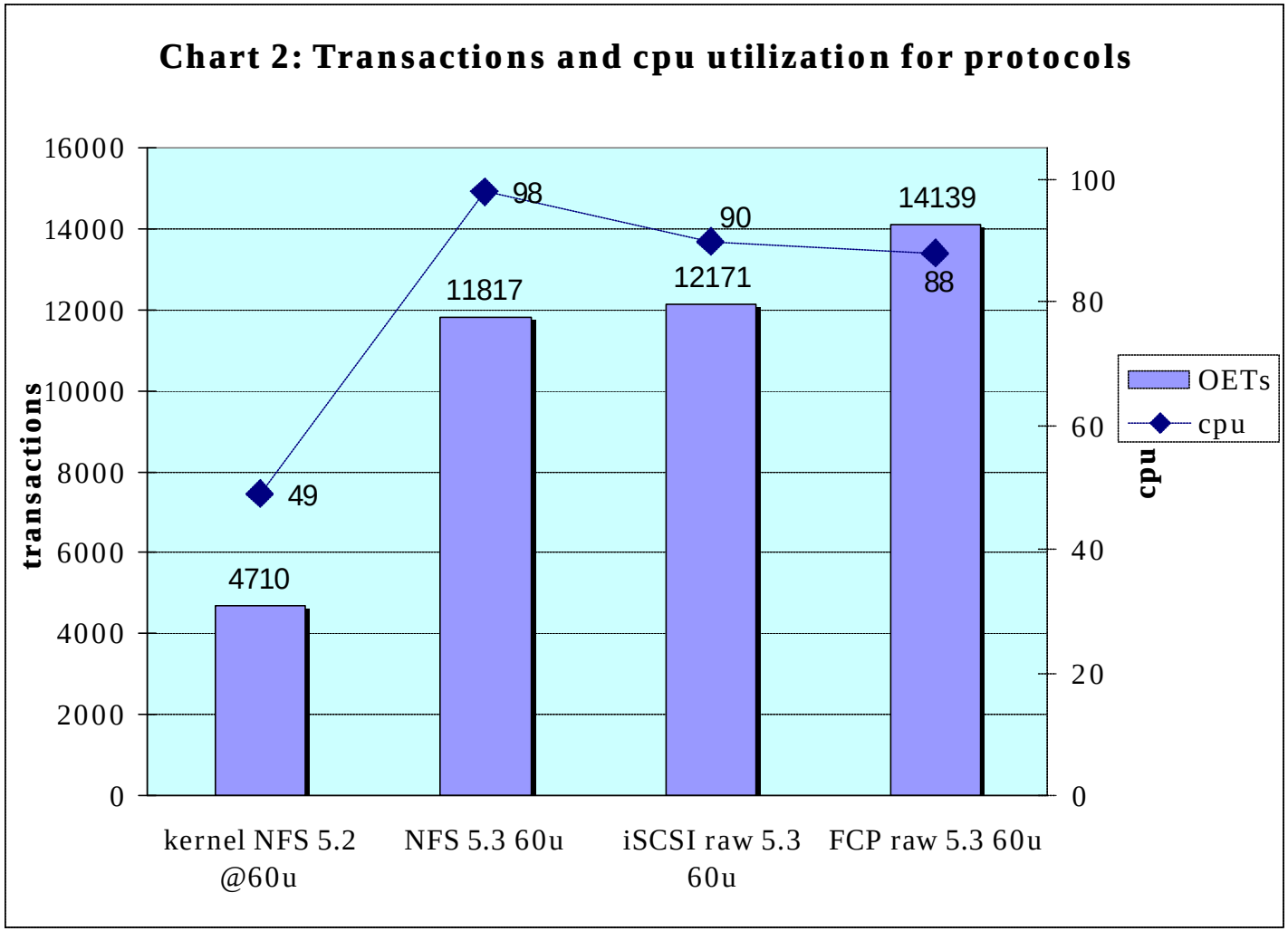
- AIX 5.2 ML-5 and AIX 5.3 ML-03 (beta)
- PowerPC pSeries 650 (2 * 1.2 GHz, 4 GB RAM)
- Oracle 10.1
- 1GbE for NFS or iSCSI or 2Gb FCP card
- FAS 940c
- 48 Disk Spindles (4 x DS14) 144GB 10K RPM
- ONTAP 6.5.1

Transactions on AIX 5.2, 5.3, iSCSI, FCP



OETs = Order Entry Transactions (an Oracle OLTP benchmark, approx 2:1 read to write ratio)

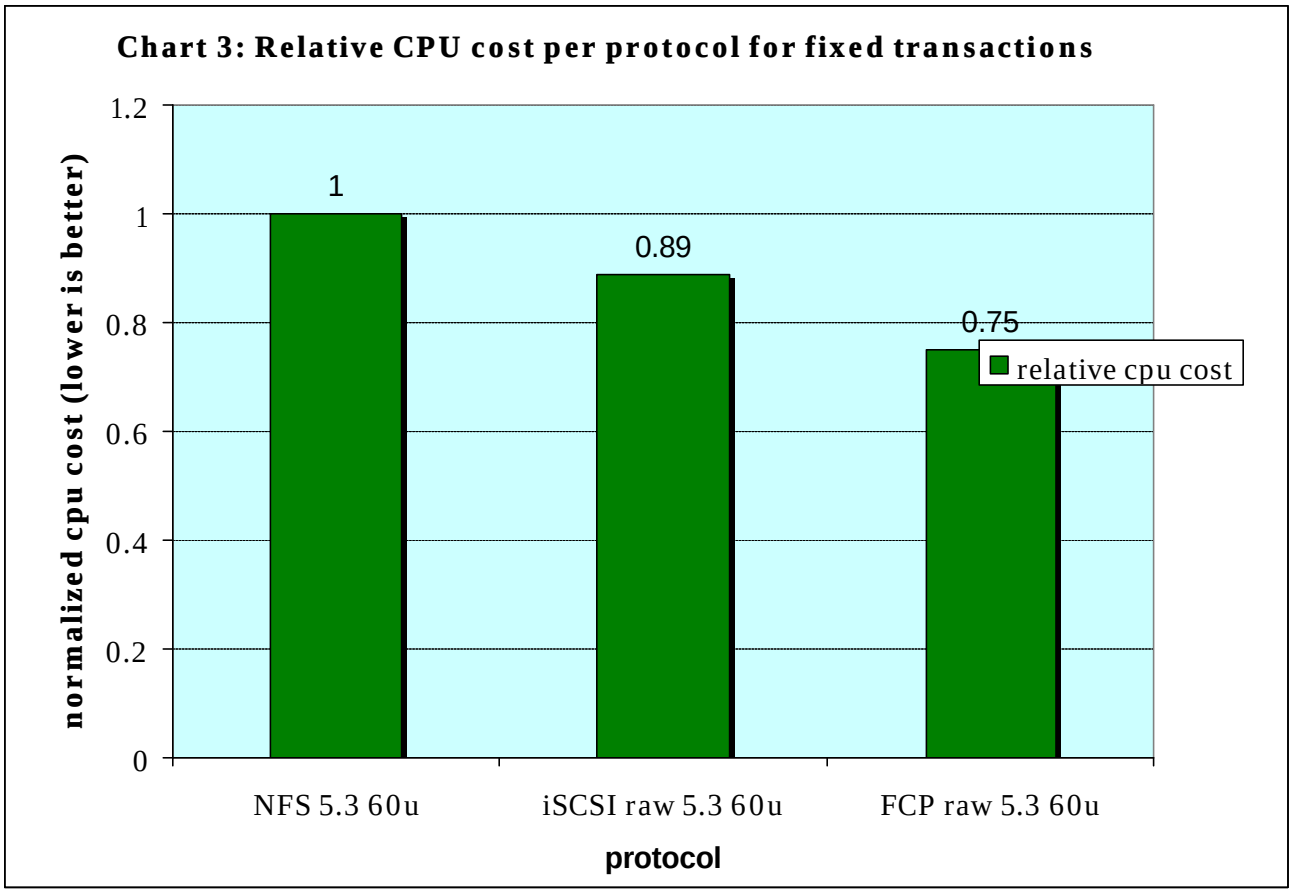
Host CPU Utilization



Protocol Efficiencies

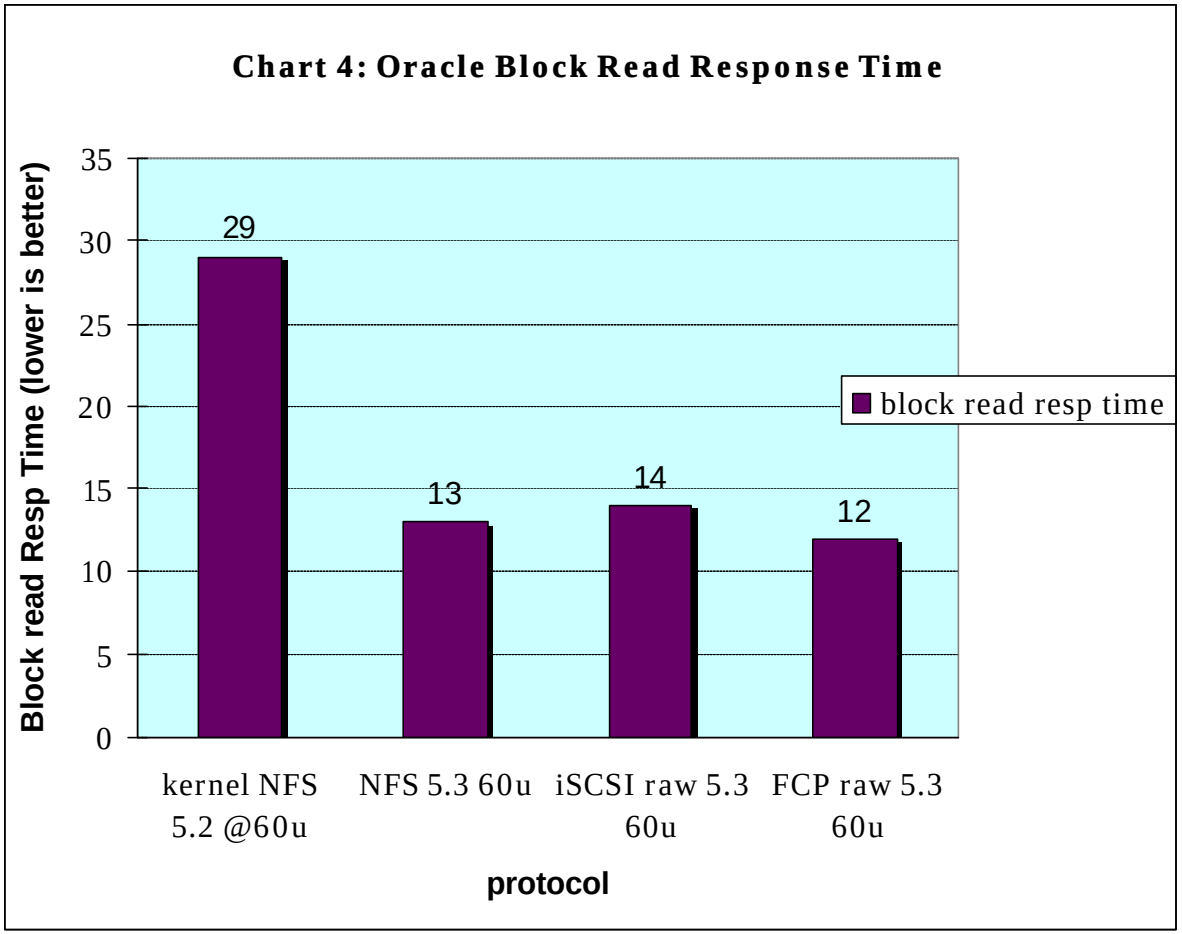
Computed using (%CPU used/ OETs) * K:

where K is a constant such that NFS CPU cost per K transactions is 1.0 and relative cost for iSCSI and FCP is computed using the same constant K transactions.



Oracle Block Read Response Time

Not typical response time. Random workload used. Filer with more cache improves block response time or more sequential reads improve avg. block response time.



Future Work

- Testing and analysis on systems with higher numbers of CPUs.
- Further investigation on socket lock scaling issue on AIX

Conclusions

- AIX 5.3 ML-03 with new Concurrent I/O ‘CIO’ mount option delivers an OLTP performance on NFS comparable to that of block-based protocols iSCSI and FCP
- Don’t be afraid to deploy databases on AIX NFS, we will support you.
- Proof of AIX NFS performance up to 4 CPU completed.

References

- AIX Performance with NFS, iSCSI and FCP Using an Oracle Database on NetApp White Paper @ http://www.netapp.com/library/tr/3408.pdf
- NetApp Best Practices Paper located at: http://www.ibm.com/servers/storage/support/nas/
- Download SIO tool
  http://now.netapp.com/NOW/download/tools/sio_ntap/index.shtml
- Improving Database Performance with AIX Concurrent I/O:

  http://www-03.ibm.com/servers/aix/whitepapers/db_perf_aix.pdf

Questions/Answers