

NFS-RDMA for Internet Search

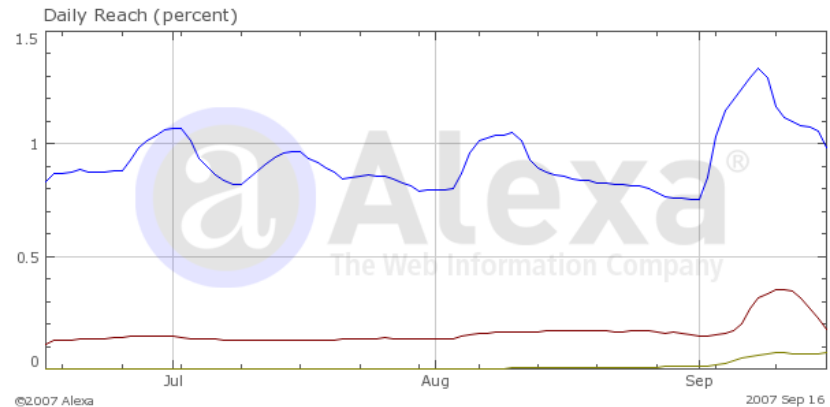
🔥 Hot searches: [vanessa hudgens](#), [pavarotti](#), [britney spears](#), [ipod touch](#), [steve fossett](#)

Speaker:
Dr. Ekechi Nwokah
Alexa Internet

Alexa Internet: Who we are

- Wholly owned Amazon.com subsidiary

- Best known for “traffic rankings”



- Search: our focus

- Alexa Search rapidly improving
 - Competitive with leading search providers
 - Alexa Search increasingly used on Amazon.com and other sites
 - Higher traffic, more exposure
 - Need faster, cheaper, more scalable storage infrastructure

Alexa Internet: A storage history



Tape



**External SCSI
with 100Mb
Ethernet**



**DAS with 1Gb
Ethernet**

Where to go from here?

Challenge

- 240TB of crawled web data
 - Homegrown text database
- Existing infrastructure cumbersome
 - Consolidate Hardware
 - Reduce CapEx/OpEx
 - Achieve Better ROI
- Large data mining apps
 - Data is write-once, read many
 - Strictly bi-modal access payloads
 - Data VERY cache-unfriendly
 - Highly Parallel



Storage Decision

- Solicited proposals from leading vendors
 - Expensive
 - Don't need large feature set
 - Price/Performance is key metric for Alexa

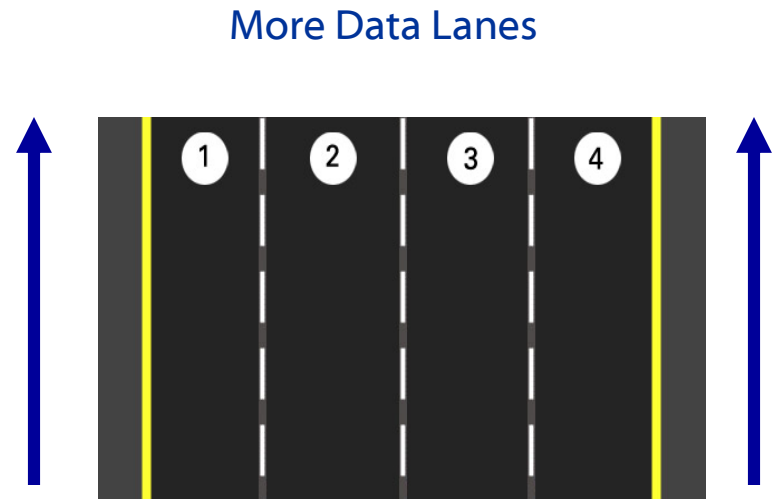


- Decided on home-grown/open-source solution
 - Most cost effective solution
 - Most scalable
 - Had dedicated storage/kernel people

Motivations

- Motivations for Infiniband

- Performance
- Hardware consolidation needs big pipes
- Line rate scalability
- Low Latency
- Lowest cost at scale
- OFED achieving stability



- Motivations for NFS-RDMA

- Performance: 630MB/s/SDR, 1.05GB/s/DDR
- Mostly internal apps (5 9's not required)
- Our apps know what to do with NFS mounts
- Mostly read only (low risk)
- Cost (free)

Lower Cost



Bottom Line: What's the ROI?

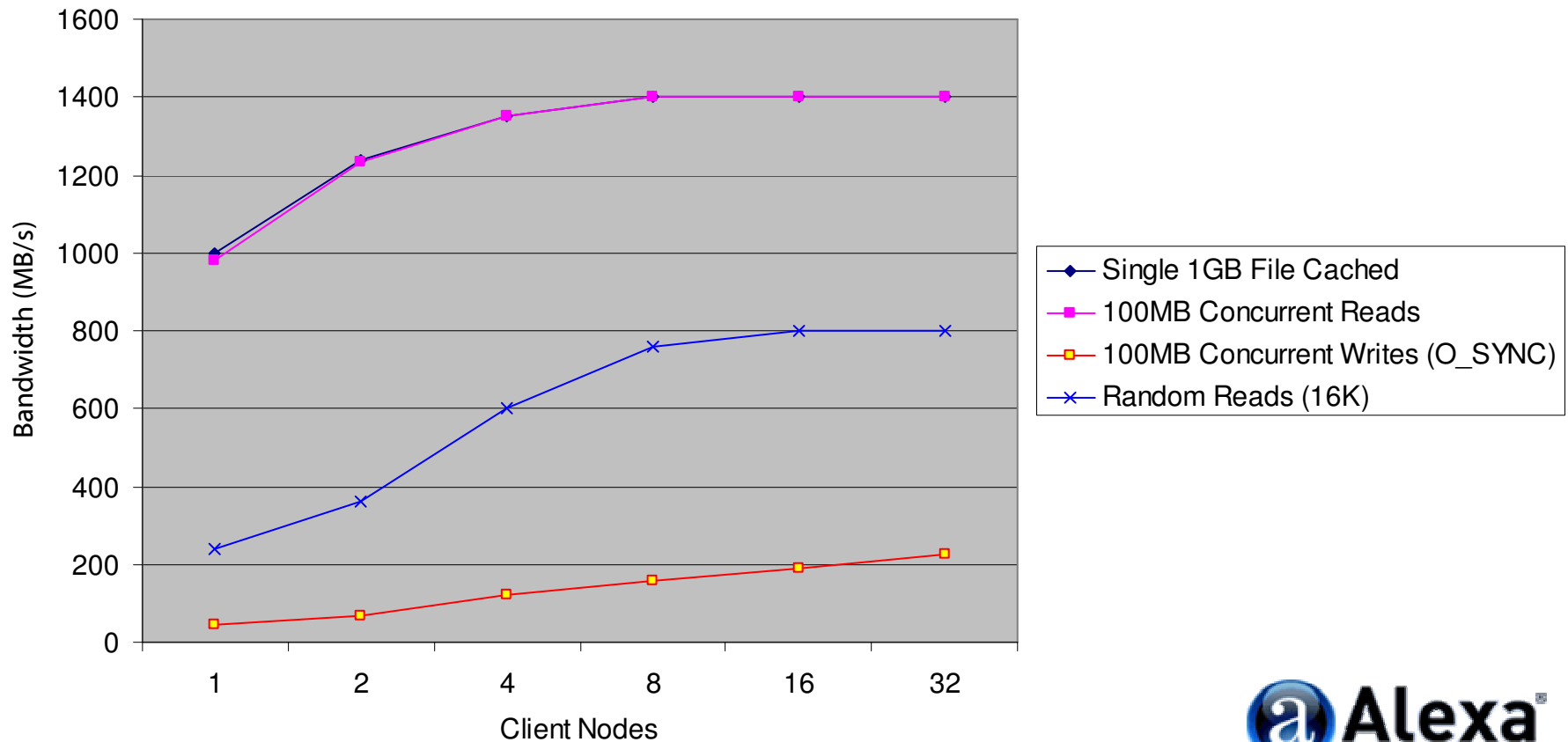
(240 TB Configuration using 500GB SATA)

NFS-RDMA/IB (24 10TB RAID6 nodes) NFS/GigE (120 2TB JBOD nodes) NFS-RDMA/10GigE (24 10TB RAID6 nodes) FC SAN (Mid-range)

	NFS-RDMA/IB (24 10TB RAID6 nodes)	NFS/GigE (120 2TB JBOD nodes)	NFS-RDMA/10GigE (24 10TB RAID6 nodes)	FC SAN (Mid-range)
STORAGE	\$280,000	\$189,000	\$280,000	\$1,500,000
SWITCHES	\$80,000	\$700,000	\$2,800,000	\$60,000
CABLING/NICS	\$51,000	\$4,800	\$108,000	\$800
RACK	\$3,000	\$4,000	\$3,000	\$4,000
FACILITY/COOLING	\$115,000	\$144,000	\$130,000	\$140,000
POWER	\$145,000	\$210,000	\$160,000	\$170,000
ENGINEERING/SUPPORT	\$70,000	\$250,000	\$70,000	\$60,000
DRIVE BREAK / FIX	\$55,000	\$55,000	\$55,000	\$0
3-yr TCO	\$799,000	\$1,556,800	\$3,606,000	\$1,934,800
Cables (Server/Client):	48/48	240/240	48/48	16/16
Bandwidth:	28.8 GB/s	19.2 GB/s	26.4 GB/s	3.4 GB/s

Early Testing

- Formed partnership with vendors as “alpha” customer
 - DDR Infiniband links
 - 32 NFS clients running NFS-RDMA (NFSv3) client version 7 on Linux 2.6.18.5
 - 1 NFS server running NFS-RDMA server version 6 on Linux 2.6.17:
 - Mellanox MTD2000 with 16 15K 146GB SAS drives



Production Server

- “NFS-RDMA” ready nodes
 - Enough disks to fill dual “NFS-RDMA over IB” pipes (2 x 600MB/s)
 - Dual SDR over DDR (cheaper)
- Consolidation
 - Can reasonably put 24 drives in chassis
 - 24 SATA drives: 1100MB/s
 - 2 RAID6 cards: 1400MB/s
- Motherboard considerations
 - 2 PCI-e x4 to accommodate RAID6 (2GB/s)
 - PCI-e x8 to accommodate dual port HCA (1.4 GB/s)
- RAID card/Drive firmware testing
 - Tested various drive and RAID firmware versions for performance
 - Sustained 520MB/s with RAID6, XFS, 500GB SATA drives



Early Experiences

- 2.6.17 Performance issues/Stability issues
 - Weird hangs on client (“ls”, etc.)
 - Mount would timeout due to inactivity, remount would hang
- 2.6.18 Mellanox SDK released
 - Testing and Performance Measurements
 - Worked with Mellanox to test and debug
 - Alexa IT Dept. moved to CentOS5
 - We modified SDK build scripts to support CentOS
 - Found major bug (Server Transport Lock)
 - Error from client caused server to permanently lock RDMA transport
 - Worked with developers to test and fix bug



Early Challenges

- NFS has problems with readahead
 - RAID cards don't do well without readahead
 - Linux readahead code needs surgery
 - Patched kernel
- IB cabling was a problem for hardware staff
 - Cables are heavy and prone to dislodge
 - New optical cables should solve this problem
- NFS-RDMA means custom OS
 - Separate management policies and procedures for IB cluster
 - Adds cost and complexity for IT staff
- Wrote thin middleware layer for Alexa apps
 - Parallelization, load balancing and namespace management



Current Status and Performance

- Running in production
 - Used by data mining applications
 - Also serving as backend for Alexa Search
 - Mounted read-only
 - Data written over GigE on back end
 - 4 server nodes, 30 clients
 - Will start scaling out soon
- Current Read Performance: 2.5GB/s uncached
 - Compare with NFS over IPoIB: 1.25GB/s
- Tuning necessary for good performance
 - Linux scheduler choice
 - NFS mount options: rsize, wsize
 - Block size, access patterns, drive options, etc.



The New Data Center

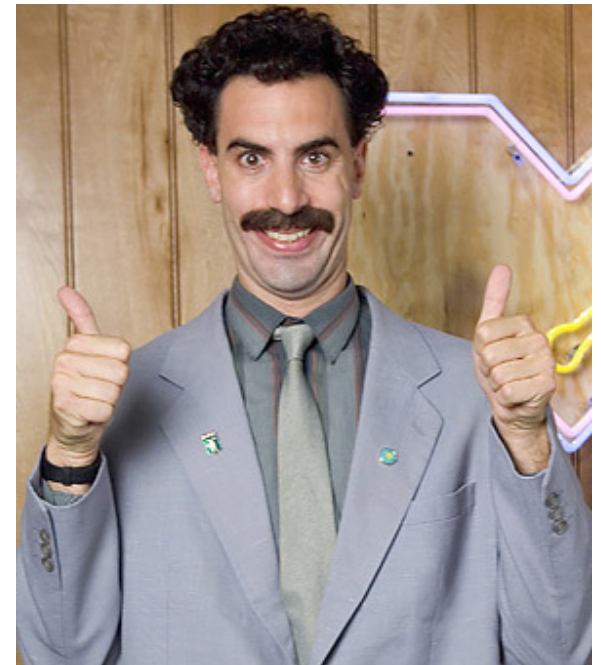
- Multi-core CPUs, Large RAM footprints
- Large and Fast Remote Storage
- Infiniband or 10GigE as interconnect
- Blades and Diskless Nodes
- Virtual/Restartable Machines
- Power!!



- We want our search to be better, cheaper, and faster
 - Must continue to be innovative in architecting our infrastructure
 - IB is best choice right now for storage network
 - NFS-RDMA is conceptually simple...it's just NFS

Conclusions and Next Steps

- Conclusion
 - NFS-RDMA with IB was well worth the effort!!
 - Best choice for price/performance, ease-of-use, TCO, etc.
 - 2x improvement over NFS with IPoIB
 - Mellanox SDK a good first step to production environment
 - Linux kernel still needs work
 - Will run custom kernel for the foreseeable future
- Next Steps
 - R/W NFS-RDMA servers in next few weeks
 - More Linux surgery
 - Scaling out (Global Namespace)
 - May augment Alexa middleware
 - Looking at pNFS
 - Considering Lustre and others
 - Integrate NFS-RDMA with Xen



Contributors

- We would like to thank the following people for their efforts:
 - Tom Tucker (Open Grid Computing)
 - Thad Omura, Fred Dickely, Vu Pham, Todd Wilde, Gilad Shainer, Eyal Waldman, Mehran Entazari, Graham Smith (Mellanox Technologies)
 - Jon Lewis, Chris Watson, Matt Jay (Silicon Mechanics)
 - Matt Dinola, Eric Dube (Voltaire)
 - Tom Talpey (Network Appliance)
 - Fengguang Wu (University of Science & Technology of China)