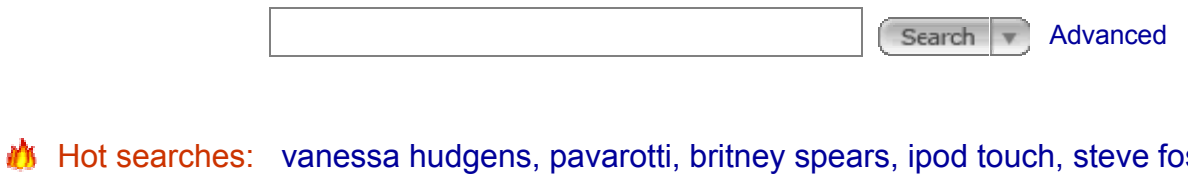
NFS-RDMA for Internet Search



Speaker:

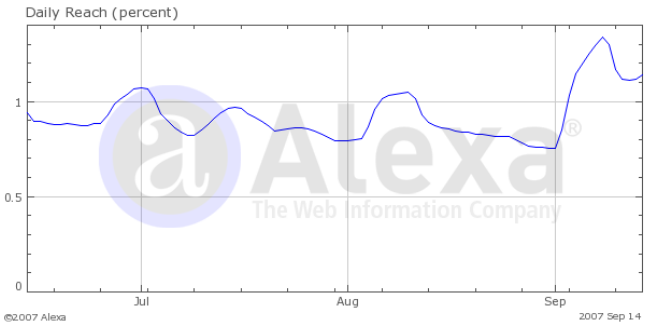
Dr. Ekechi Nwokah

Alexa Internet

Alexa®
The Web Information Company

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® The Web Information Company

Alexa Internet: A storage history

Tape

External SCSI with 100Mb Ethernet

DAS with 1Gb Ethernet

Where to go from here?

Alexa®
The Web Information Company

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

Alexa®
The Web Information Company

Storage Decision

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

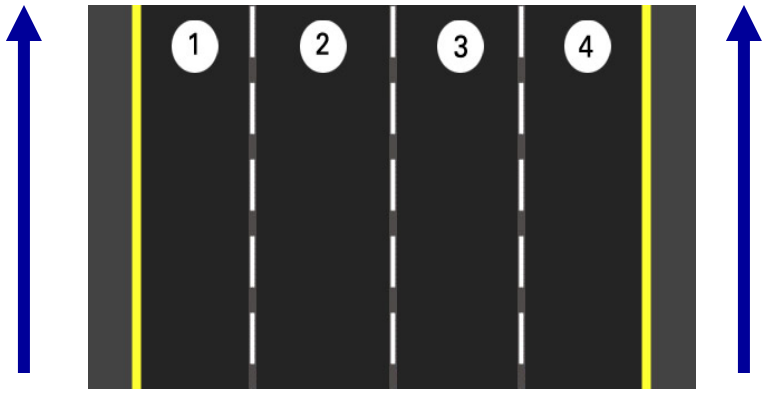
- Decided on home-grown/open-source solution
  - Most cost effective solution
  - Most scalable
  - Had dedicated storage/kernel people

Alexa®
The Web Information Company

Motivations

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

More Data Lanes



- 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

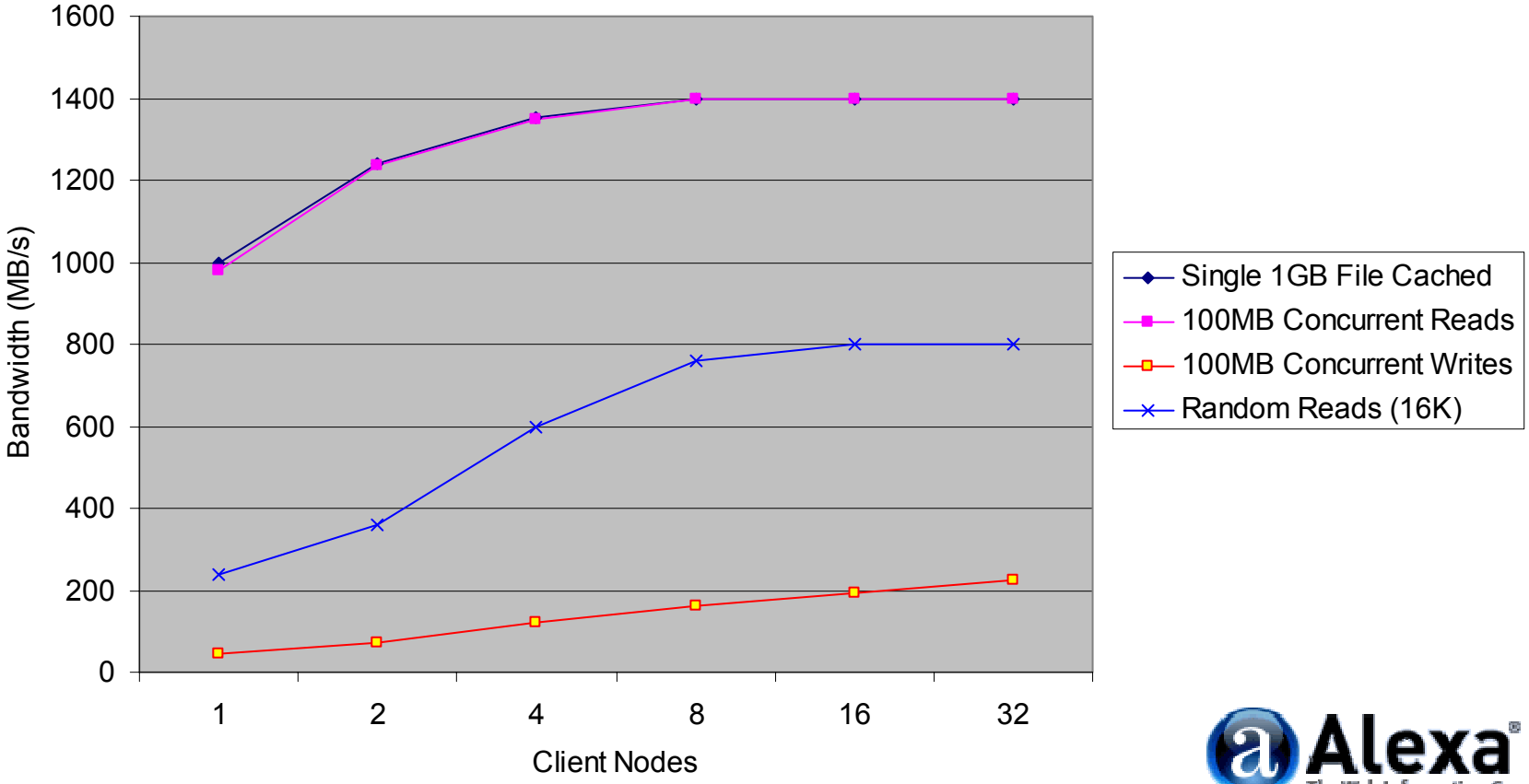
Bottom Line: What's the ROI?

(240 TB Configuration)

	NFS-RDMA/IB (24 10TB RAID6 nodes)	NFS/GigE (240 1TB nodes)	NFS-RDMA/10GigE (24 10TB RAID6 nodes)	FC SAN (Leading Vendor)
STORAGE	$277,909	$367,440	$277,909	$2,610,961
SWITCHES	$250,724	$678,812	$2,560,105	$447,712
CABLING/NICS	$19,200	$240	$60,048	$256
RACK	$2,931	$4,031	$3,297	$4,029
FACILITY/COOLING	$115,200	$158,401	$129,600	$158,400
POWER	$143,571	$218,023	$161,985	$213,602
ENGINEERING/SUPPORT	$68,475	$327,676	$68,475	$63,150
DRIVE BREAK / FIX	$52,891	$40,686	$52,891	$0
3-yr TCO	$930,901	$1,795,309	$3,314,310	$3,498,110
Peak Bandwidth:	28.8GB/s	28.8GB/s	26.4GB/s	3.2 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 630MB/s)
  - Dual SDR over DDR (cheaper and better performance from server)
- Consolidation
  - Can reasonably put 24 drives in chassis
  - 24 SATA drives: 1100MB/s peak
  - 2 3ware RAID6 cards: 1400MB/s peak

- Motherboard considerations
  - 2 PCI-e x4 to accommodate RAIDs (2GB/s peak)
  - PCI-e x8 to accommodate dual port HCA (1.4 GB/s peak)
- RAID card/Drive firmware testing
  - Tested various drive and RAID firmware versions for performance
  - 3ware 9650 maxed out at 520MB/s with RAID6, XFS, 500GB SATA drives

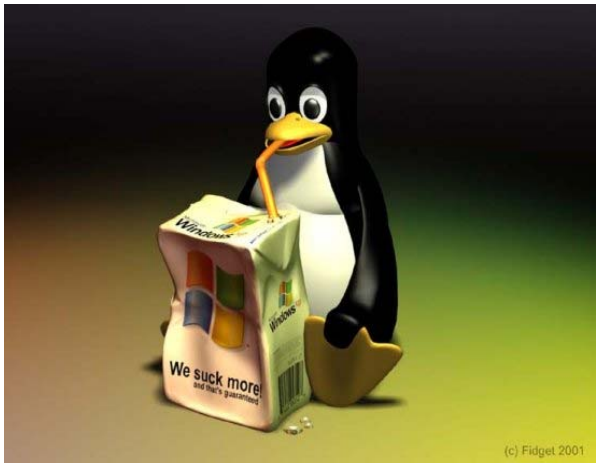
Early Experiences

- 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

Alexa®
The Web Information Company

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 kernel
  - Separate management policies and procedures for IB cluster
  - Adds cost and complexity for IT staff
- Wrote thin middleware layer
  - 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

- Read Performance: 2.5GB/s uncached
  - Compare with NFS over IPoIB: 1.25GB/s uncached

- Tuning necessary for good performance
  - Linux scheduler choice
  - NFS mount options: rsize, wsize
  - Block size, access patterns, drive options, etc.

Alexa®
The Web Information Company

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 networking technology
  - NFS-RDMA is conceptually simple…it's just NFS

Conclusion and Next Steps

- Conclusion
  - NFS-RDMA with IB was well worth the effort!!
    - 2.5GB/s for under $60K
    - 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 forseeable 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
  - Integrate NFS-RDMA with Xen

Alexa®
The Web Information Company

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)