Storage Evaluations at BNL

HEPiX at DESY
Spring 2007

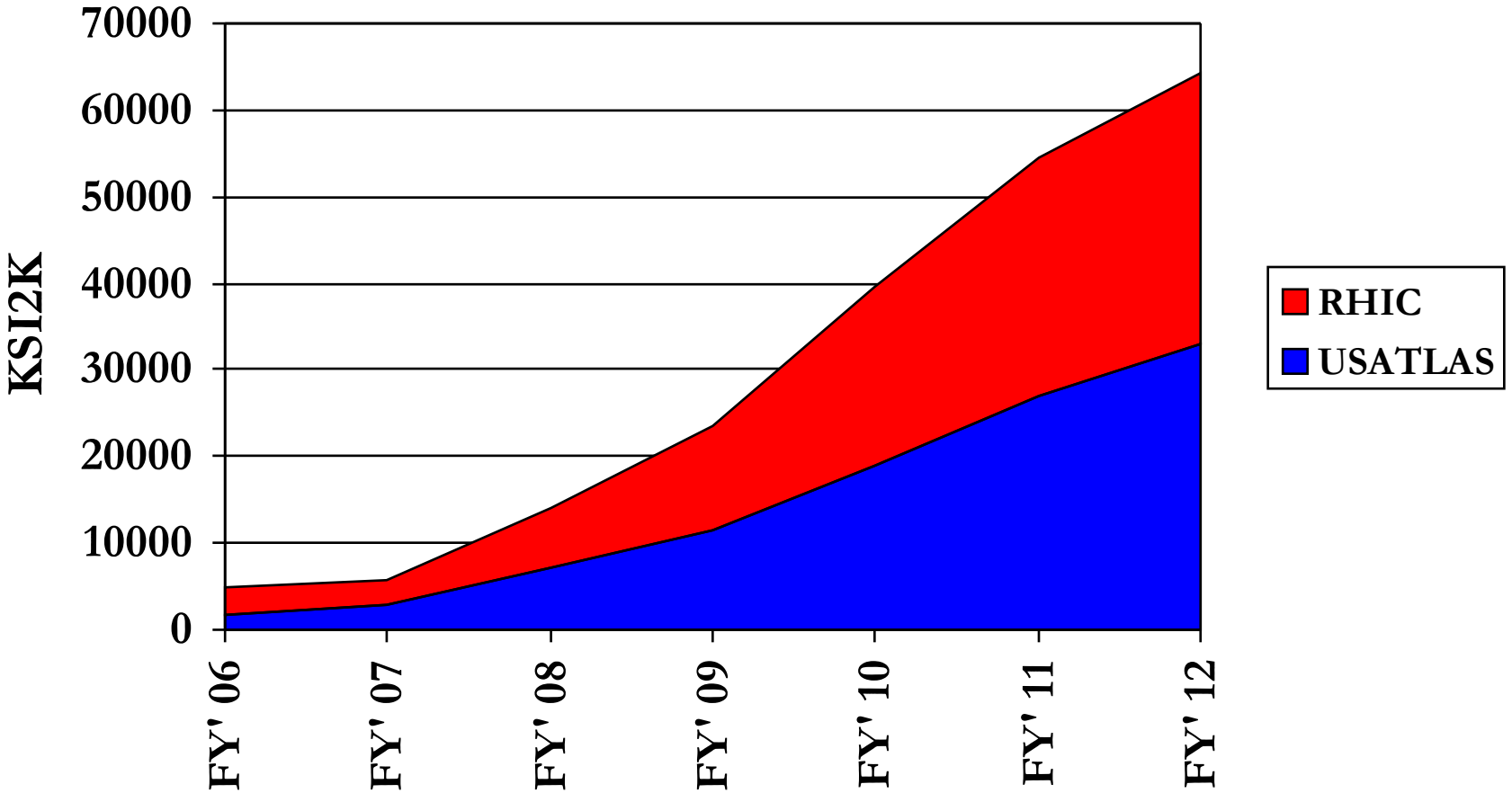
Robert Petkus

RHIC/USATLAS Computing Facility
Brookhaven National Laboratory

State of Affairs

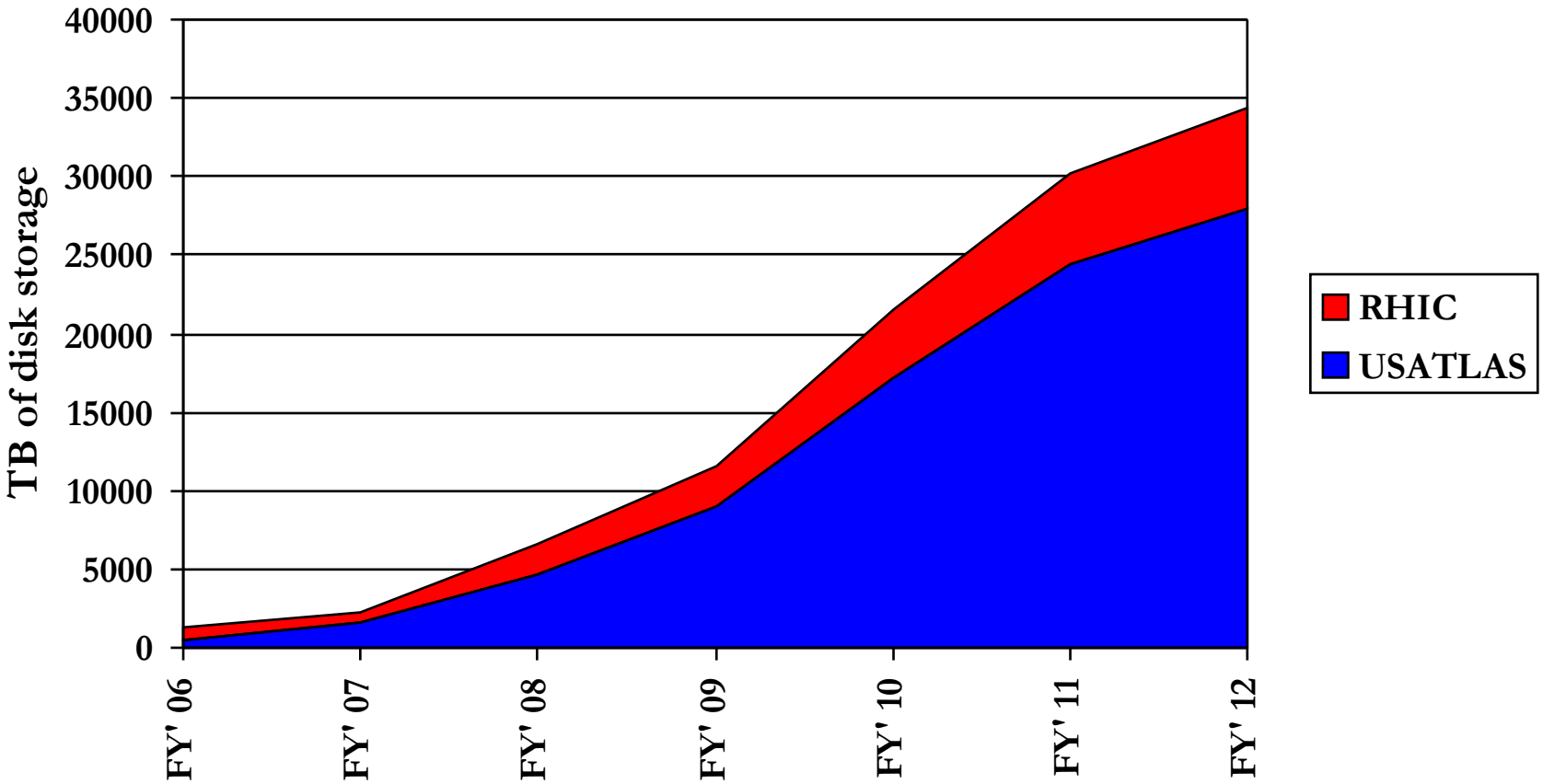
- Explosive disk storage growth trajectory over the next 5 years (>30 PB).
  - Projected storage requirements for distributed fileservers (dCache, xrootd) may require **more** farm nodes than necessary for computation alone.
  - Management and scalability concerns
    - Model using distributed dCache managed disk space on compute nodes may not prove viable or cost effective
    - Separate and consolidate the distributed storage component (dCache, xrootd) of the farm onto dedicated storage servers ?
  - Differentiate between two tiers of distributed storage – read-only vs. write (HA).
- Increasing demand for data center real estate, power, and cooling.
- Aging centralized storage infrastructure (Panasas, NFS).

Expected Computing Capacity Evolution



A. Chan

Expected Storage Capacity Evolution



A. Chan

Test Methodology

- Create a standard test/production configuration
  - Local I/O profiles using IOzone and FileOP
  - dCache and xrootd read bandwidth
  - dCache write pool stress testing (GridFTP in / PFTP out)
  - Mock production using pre-packaged applications in alliance with various experimental groups
- Obtain maximum I/O capacity for each system
- Identify the highest performing and most versatile storage systems
- Focus is on high-density disk arrays and NAS systems (e.g., SunFire x4500, Scalable Informatics, Nexsan SataBeast, Terrazilla, Xtore, 3Par)
- Compare and contrast SATA vs. SAS, Solaris/ZFS vs. Linux/ext3/XFS, HW vs. SW RAID, separate disk array vs. local disk

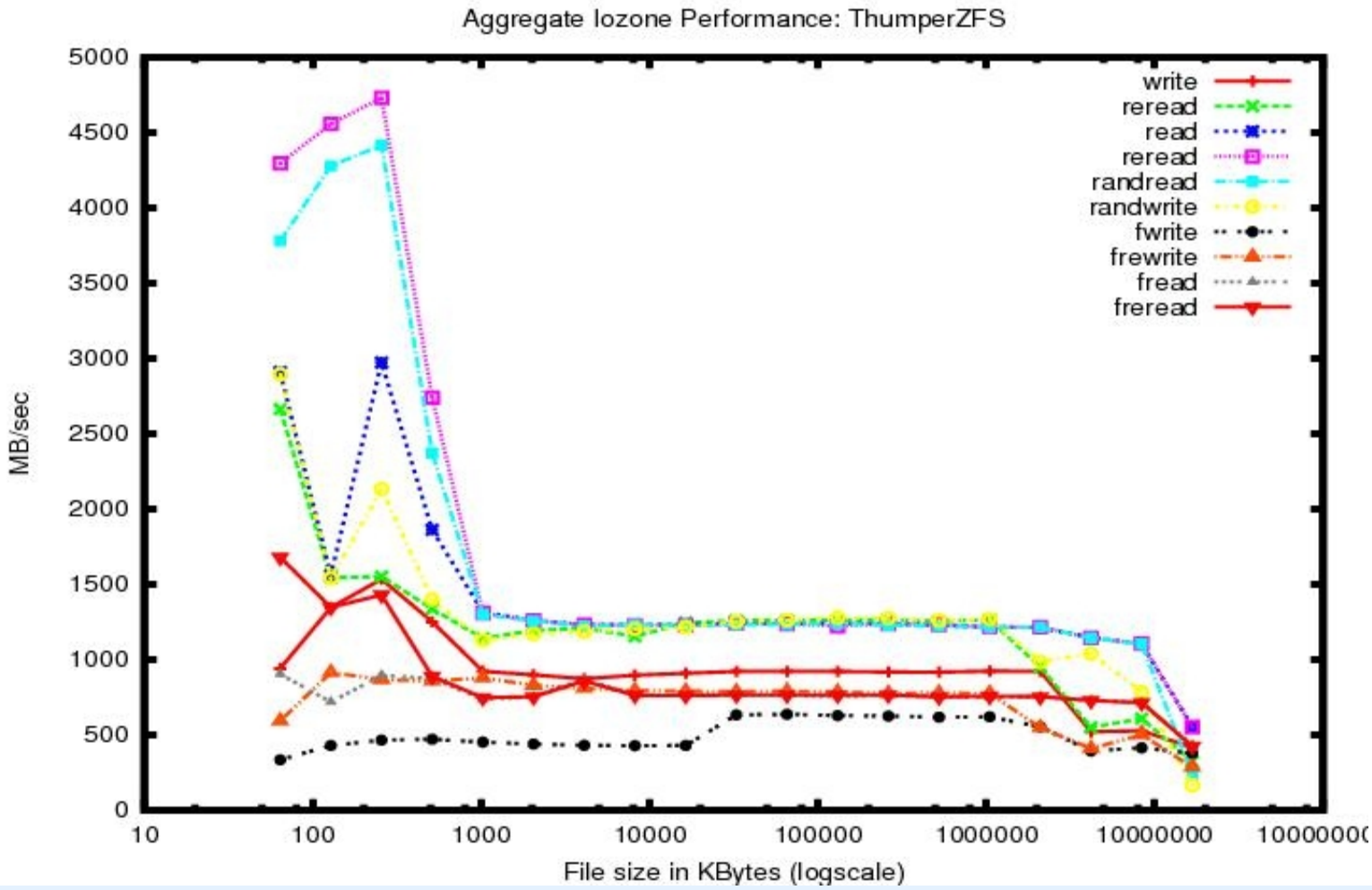
SunFire x4500 "Thumper"

- The SunFire x4500 is a promising potential dCache write pool node, NFS server, and/or iSCSI target.
- NAS solution consisting of
  - (2) dual-core AMD Opteron 285 processors (2.6Ghz) and (48) 500 GB SATA II drives yielding 24/20.5 TB RAW/usable capacity
  - (6) paired SATA controllers connected to HyperTransport PCI-X 2.0 tunnels
  - (2) dual-Gb ethernet controllers
  - 16 GB RAM

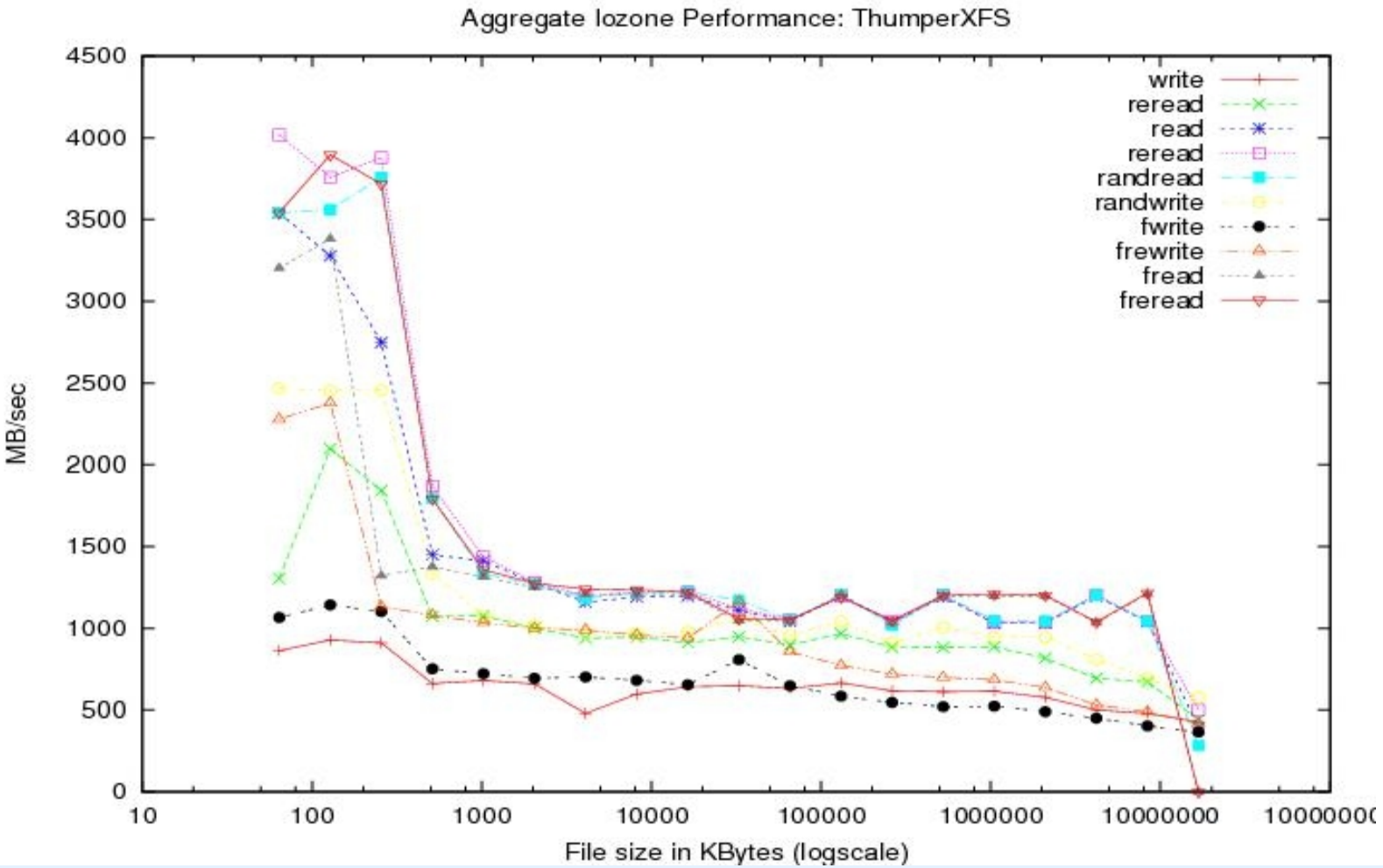
2 Thumper Test Configurations

- Solaris 10 update 2 (6/06) kernel 118855-19, ZFS filesystem
  - A single ZFS storage pool and file system was created from (8) raidz (RAID 5) sets in (5 + 1) and (6 + 1) configurations
  - Each RAID member disk resides on a different controller
  - 4 channel bonded interfaces
- Fedora Core 6, x86_64 2.6.19 kernel
  - (8) RAID-5 sets created using mdadm. Again, each member RAID disk on a different controller. 64k chunk.
  - Single RAID-0 (stride) volume using LVM2
  - EXT3 (not 64-bit mode ext4) and XFS (2.8.11)file systems created on the volume
  - 4 channel bonded interfaces

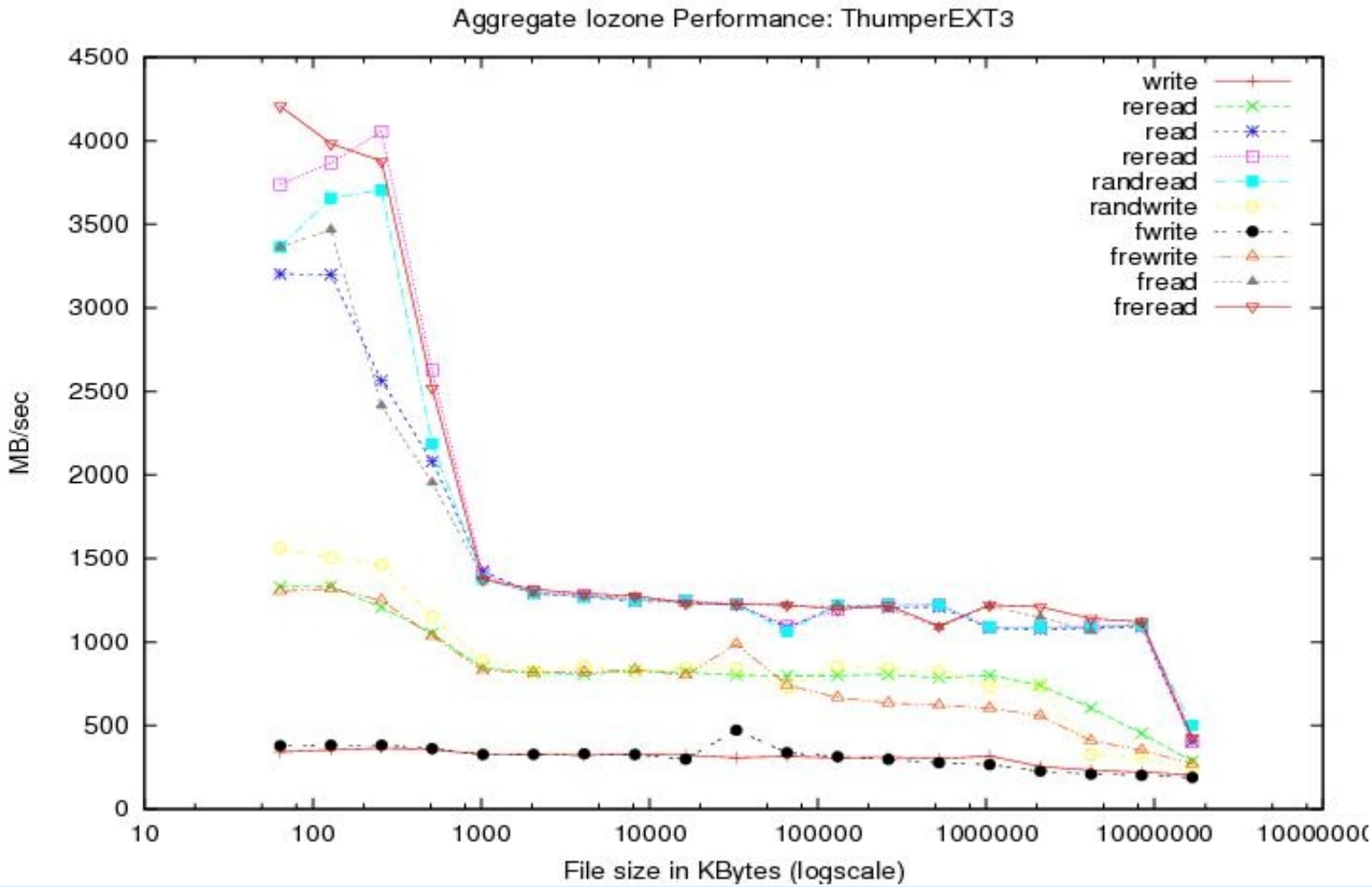
Thumper IOzone Results: Aggregate ZFS



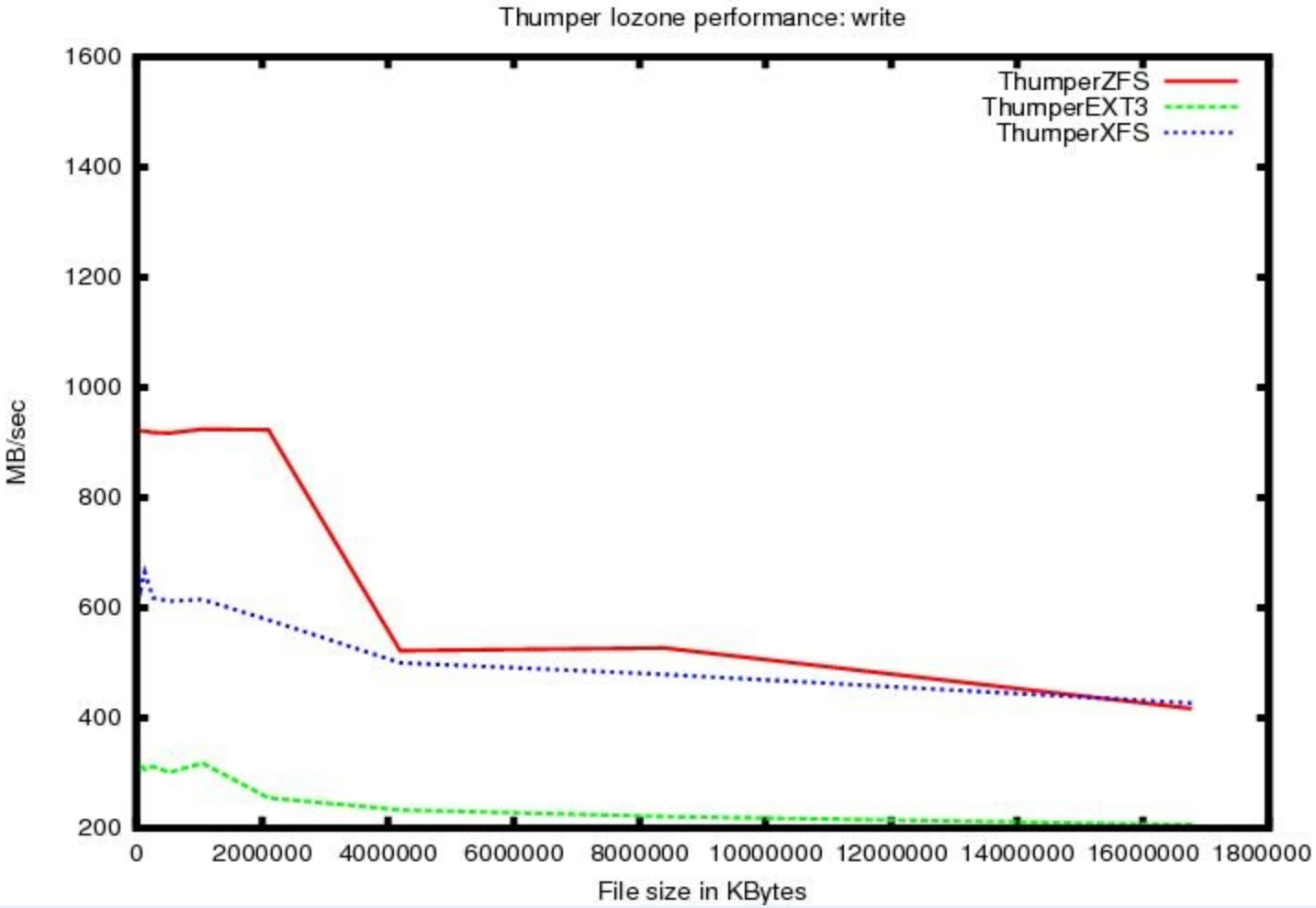
Thumper IOzone Results: Aggregate XFS



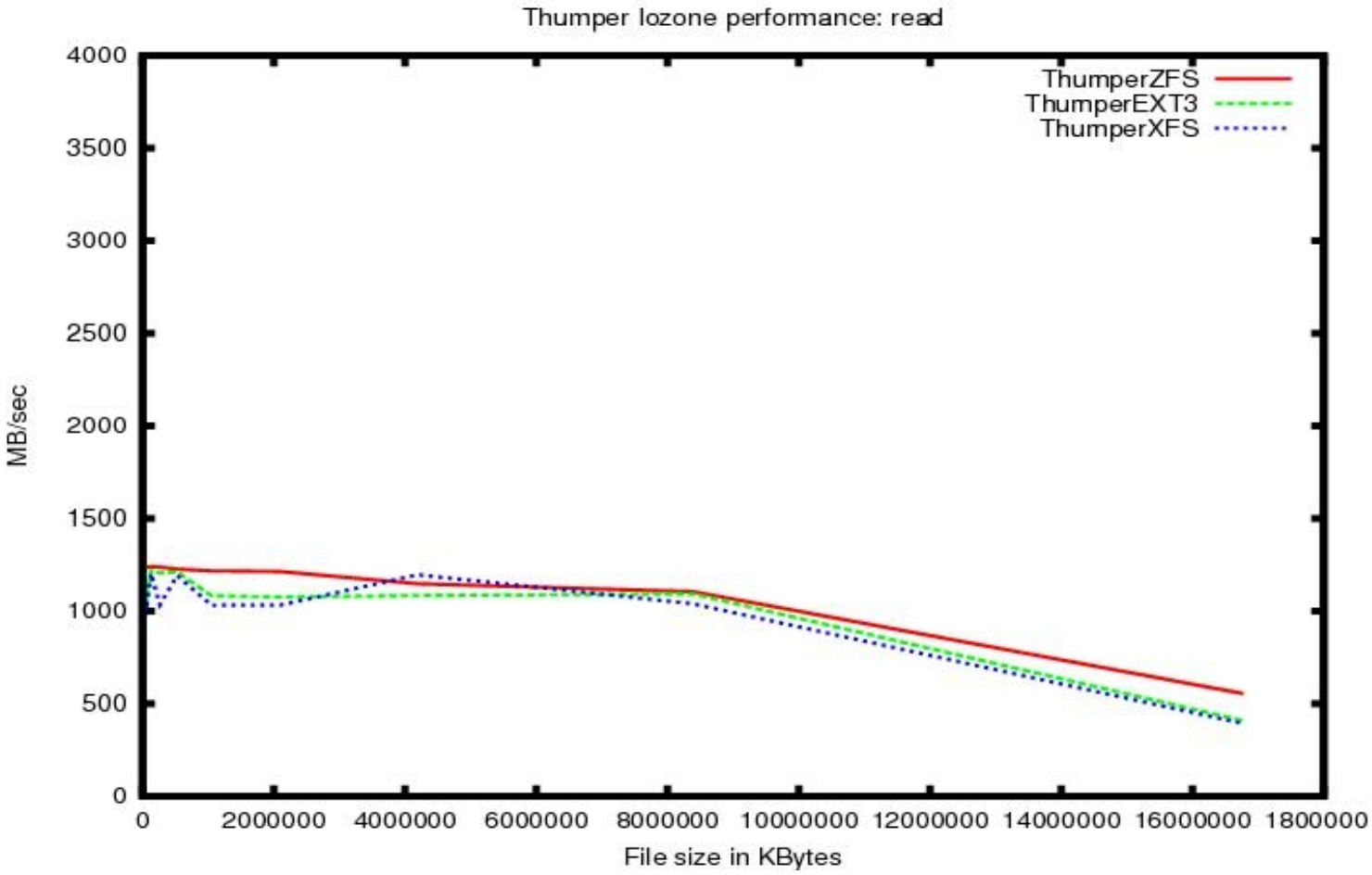
Thumper IOzone Results: Aggregate EXT3



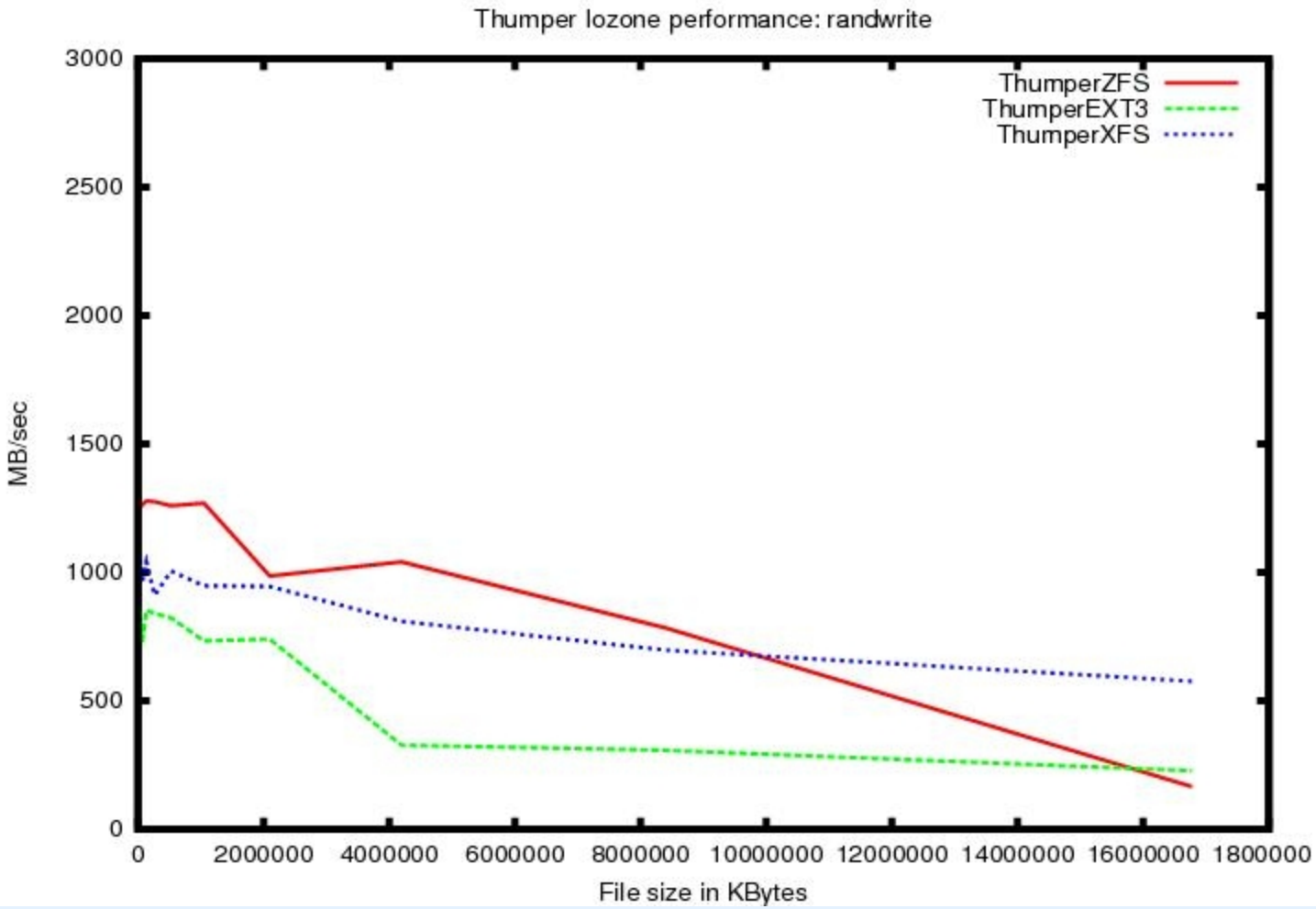
Thumper IOzone Comparative Results



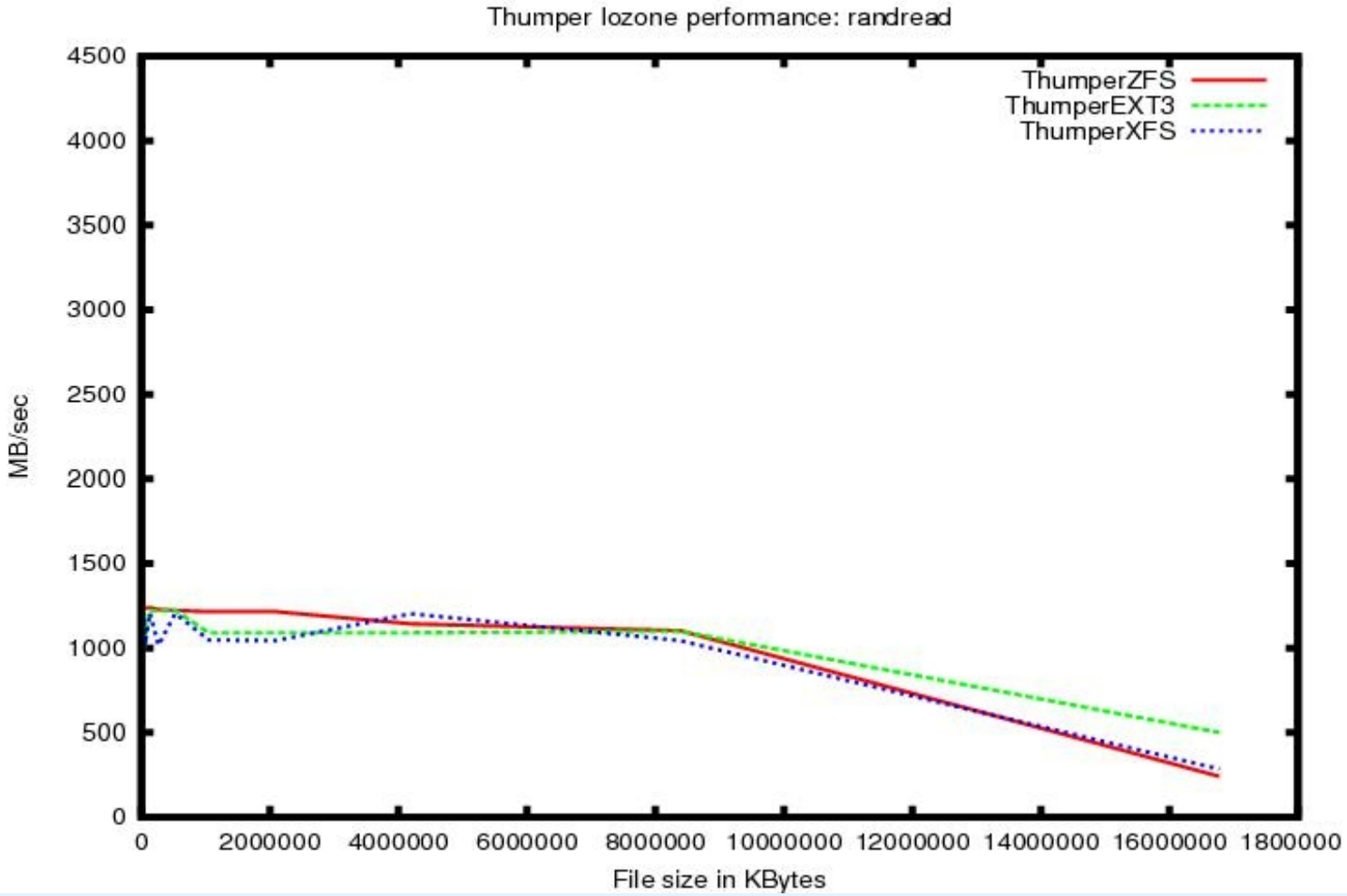
Thumper IOzone Comparative Results



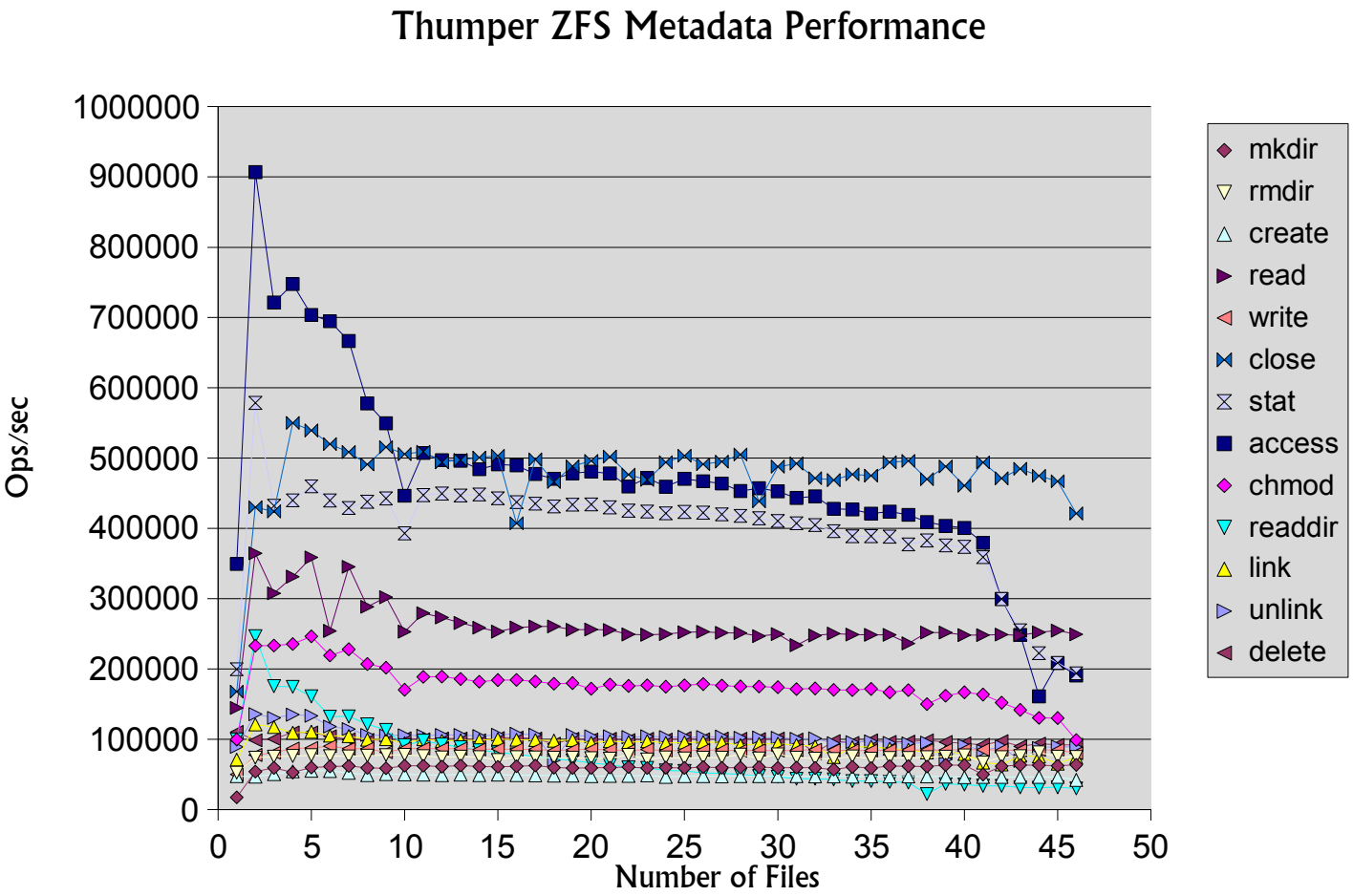
Thumper IOzone Comparative Results



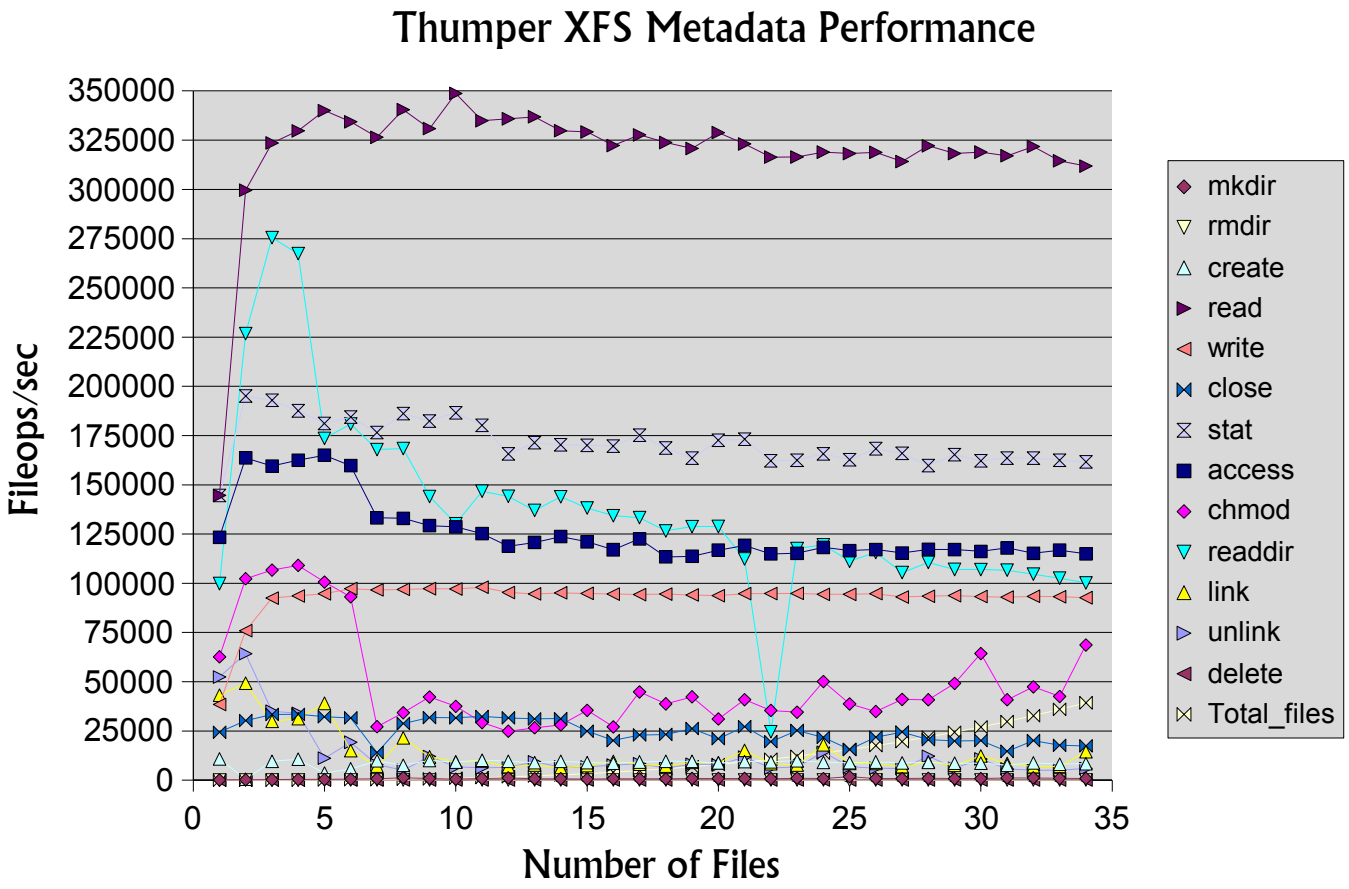
Thumper IOzone Comparative Results



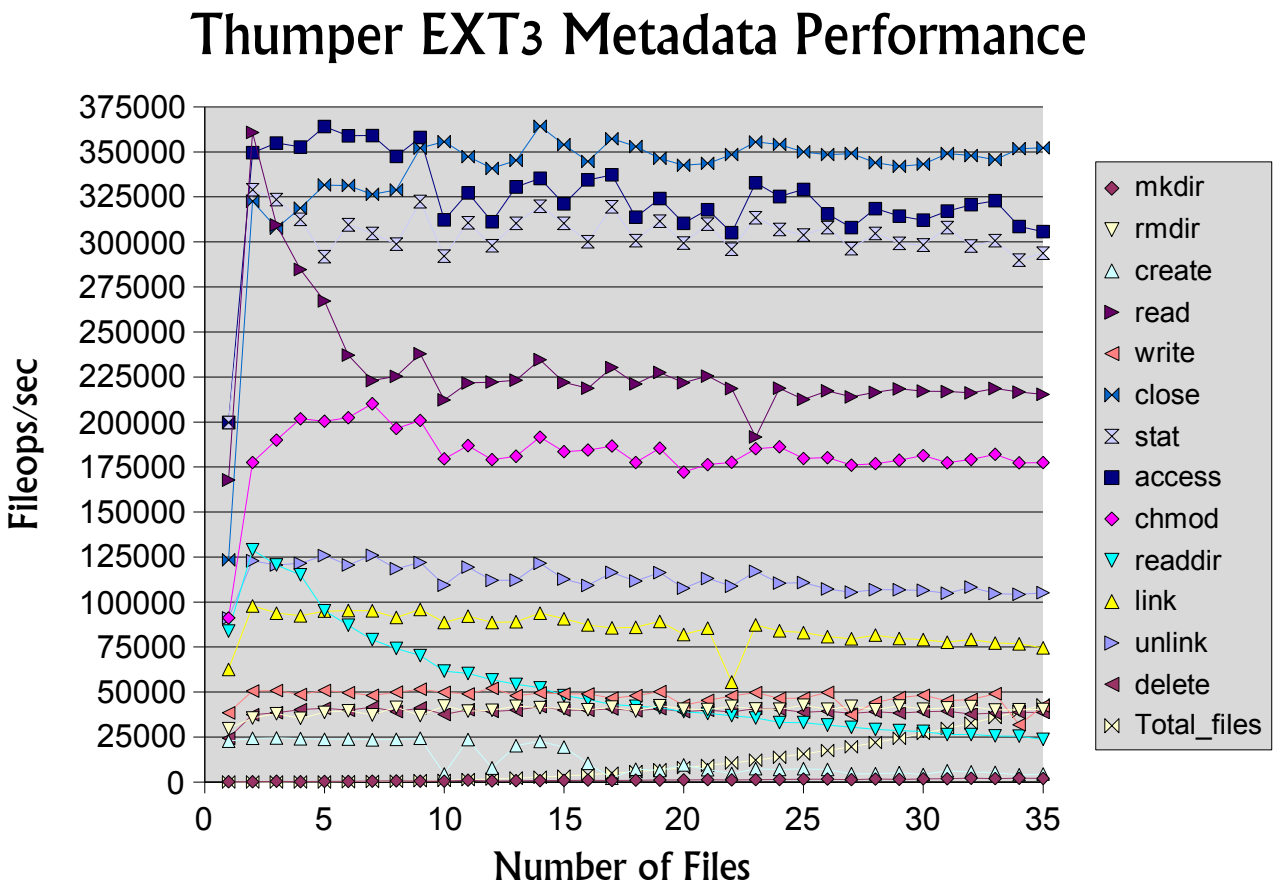
Thumper ZFS Fileop Tests



Thumper XFS Fileop Tests



Thumper EXT3 Fileop Tests



Thumper dCache Tests (ZFS - dccp)

RHIC Computing Facility Grid > Phenix SUN NFS Servers > thumper01.rcf.bnl.gov

Overview

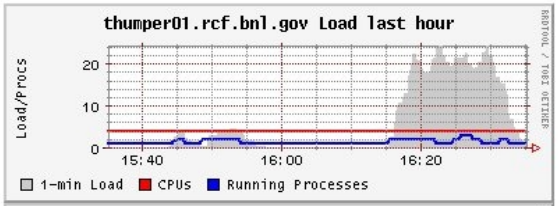
This host is up and running.

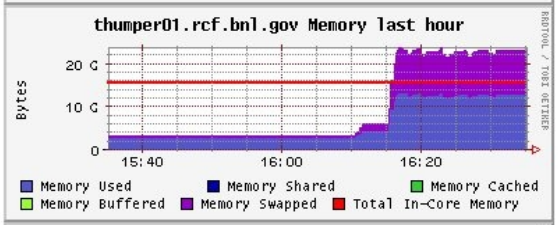
Time and String Metrics

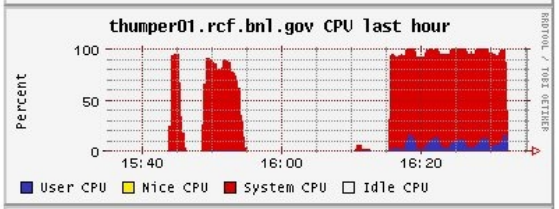
Name	Value
boottime	Fri, 8 Dec 2006 19:57:13 -0500
gexec	OFF
machine_type	i86pc
os_name	SunOS
os_release	5.10
uptime	11 days, 20:37

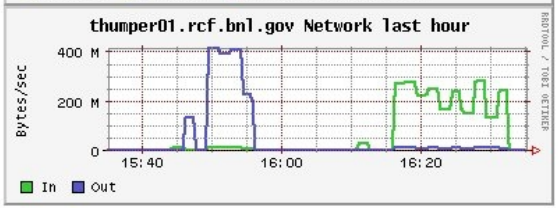
Constant Metrics

Name	Value
cpu_num	4
cpu_speed	2593 MHz
mem_total	16314992 KB
swap_total	13428168 KB









O. Rind

Test 1: 3:45PM: 30 nodes read data (same file) from dCache at 400MB/sec.
Test 2: 4:15PM: 30 nodes write data to dCache at ~250MB/sec.

Thumper dCache Tests (ZFS - dccp)

RHIC Computing Facility Grid > Phenix SUN NFS Servers > thumper01.rcf.bnl.gov

Overview

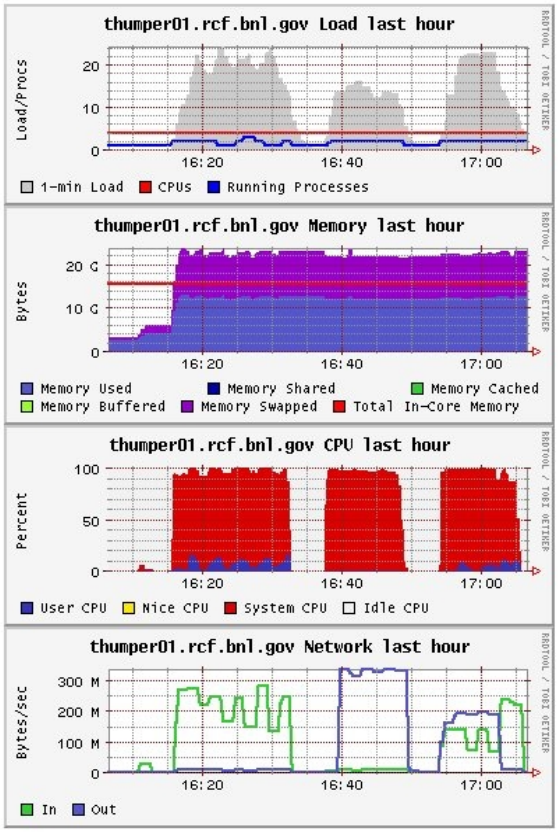
This host is up and running.

Time and String Metrics

Name	Value
boottime	Fri, 8 Dec 2006 19:57:13 -0500
gexec	OFF
machine_type	i86pc
os_name	SunOS
os_release	5.10
uptime	11 days, 21:9

Constant Metrics

Name	Value
cpu_num	4
cpu_speed	2593 MHz
mem_total	16314992 KB
swap_total	13428168 KB



O. Rind

Test 3: 4:40PM: 30 nodes read mixed data from dCache @ ~350MB/sec.
Test 4: 4:50PM: 30 nodes simultaneous read and write @ 200MB/sec. Expect that this would be average mixed performance.

Thumper dCache Tests (ZFS - dccp)

RHIC Computing Facility Grid > Phenix SUN NFS Servers > thumper01.rcf.bnl.gov

Overview

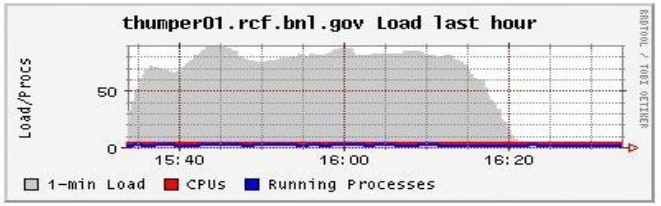
This host is up and running.

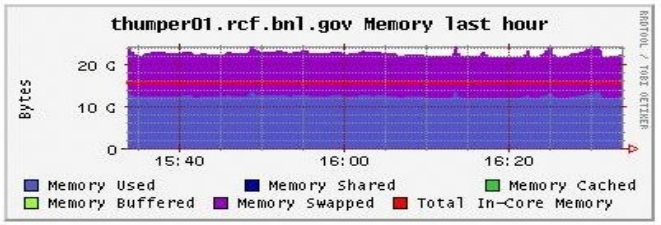
Time and String Metrics

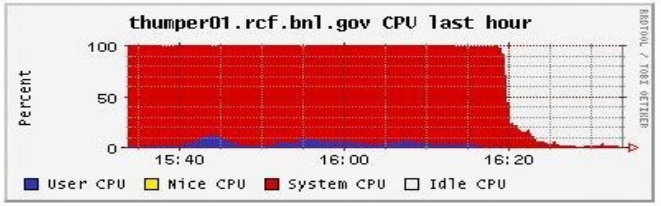
Name	Value
boottime	Fri, 8 Dec 2006 19:57:13 -0500
gexec	OFF
machine_type	i86pc
os_name	SunOS
os_release	5.10
uptime	18 days, 20:36

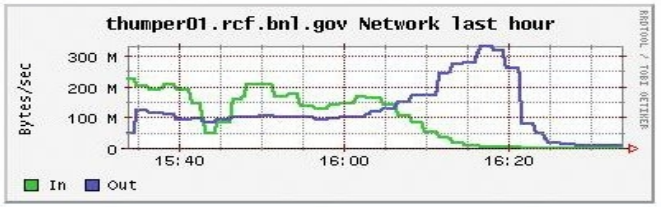
Constant Metrics

Name	Value
cpu_num	4
cpu_speed	2593 MHz
mem_total	16314992 KB
swap_total	13428168 KB









O. Rind

Test 5: 75 clients sequentially writing 3x1.5G files (green line) + 75 clients sequentially reading 4x1.5G randomly selected files (blue line) @ 250-300MB/sec.

Thumper dCache Tests (ZFS - GridFTP)

RHIC Computing Facility Grid > Phenix SUN NFS Servers > thumper01.rcf.bnl.gov

Overview

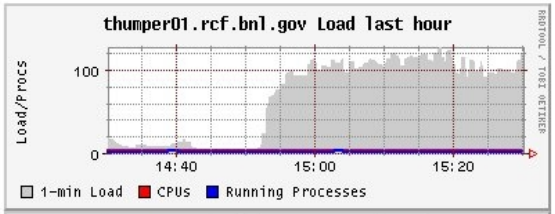
This host is up and running.

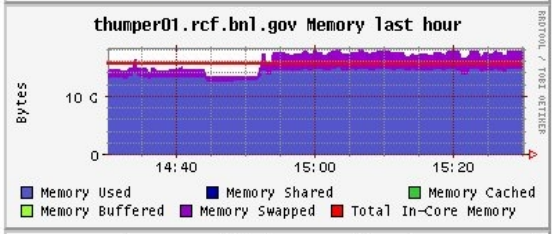
Time and String Metrics

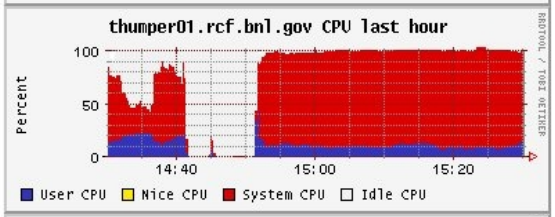
Name	Value
boottime	Fri, 8 Dec 2006 19:57:13 -0500
gexec	OFF
machine_type	i86pc
os_name	SunOS
os_release	5.10
uptime	41 days, 19:32

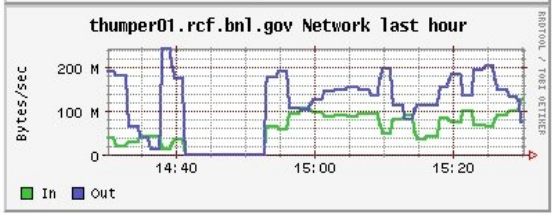
Constant Metrics

Name	Value
cpu_num	4
cpu_speed	2593 MHz
mem_total	16314992 KB
swap_total	3604148 KB









O. Rind

Test 6: 75 clients sequentially writing 3x1.5GB files (green line) + 75 clients sequentially reading 4x1.5G randomly selected files (blue line)

Thumper Test Observations

- CPU and buffer cache effects clearly seen in IOZONE graphs.
- Thumper delivers stunning I/O throughput in the 1 – 10GB file size range (dCache, Xrootd).
- Read performance parity between Linux XFS/EXT3 and Solaris10/ZFS
- SolX/ZFS dominant in sequential write performance, XFS on par for files > 4GB, abysmal EXT3 performance
  - EXT3 not 64-bit (4KB block size limited), need to retest with the 64-bit EXT3 extents patchset (EXT4)
  - EXT3 would be preferable if performance was equalized (inclusion in kernel, active development)
- ZFS performance drop in random writes for files > 4GB
  - Problem with write sequentialization for very large files?
- ZFS consumes all available physical memory which isn't released even after I/O activity has ceased
  - No convenient knob to throttle memory consumption (need to use mdb)
  - However, memory is freed when needed

Thumper Test Observations

- File operation tests (up to 50 files):
  - ZFS wins for close, access, stat, and chmod
  - XFS for read and delete
  - XFS == ZFS for writes
  - ZFS: need to test stat for thousands of files – recent dialogue on dCache list regarding slow pool start-up using ZFS – A problem with Java - ZFS?
- ZFS recommended if compatible with the software stack (yes for dCache, no for Lustre)
  - ZFS easy to set-up and administer, integrated volume management
  - Fault tolerance: end to end checksums, no RAID-5 hole
- Architectural point of failure: storage will not be externally accessible if the CPU/memory module fails. Need external hyper-transport SATA controller? Does this really matter for dCache, xrootd read nodes?

Thumper Test Observations

- Scalability concerns: will managing many racks be a hassle? At the PB scale, does a SAN backend make more sense?
- Looking forward to a quad-core Thumper and hopefully SAS Thumpers.
- Need to test:
  - 10GE – Reap the benefits of TOE and eventually RDMA/iWARP on supported cards
  - Performance cost of RAID-6?
  - XFS dCache performance
  - EXT4
  - GridFTP in – PFTP out
  - Xrootd

Other Test Systems

- We extensively tested the Scalable Informatics Jackrabbit. This is a similarly dense 5U, 48 disk (46 750 GB data disks), with (2) dual-core Opteron processors, 12GB RAM (our test unit) and (2) Areca SATA II RAID controllers, each with 1GB cache
  - Our evaluation system was not a polished product but performed well
  - There were several evaluation flaws that the vendor claims disadvantaged their system in our tests including:
    - We installed SL4.4 (2.4x) over a 2.6x kernel
    - The system was configured RAID-6 vs. RAID-5 on Thumper
- More systems in-house for testing:
  - Xstor 16 bay SAS JBOD, 4U 48 bay SAS array on the way
  - Aberdeen "Terrazilla" 6U 32 bay SATA II array with integrated server
  - Nexsan SATAbeast 4U 48 bay SATA II, dual controller
  - DDN SAF4248 4U 48 bay SATA II plus S2A controllers on the way

Questions, Comments?