An Evaluation of Panasas at BNL

HEPiX
Autumn 2004

Robert Petkus

RHIC Computing Facility
Brookhaven National Laboratory

Centralized File Service

- Single, facility-wide namespace for files.
- Uniform, facility-wide “POSIX-like” access to data.
- Require no changes to user work process to utilize resources.

Current Implementation

- Data store, home directories and scratch space accessed via NFS
- SAN based backend architecture
  - 225 TB fibre-channel disks in RAID 5 arrays
  - 13 TB IDE storage
  - 24 Brocade fibre-channel switches
- 37 Sun Solaris 9 servers (E450, V480, V240) running Veritas 4.0 (VxVM, VxFS)
- NFS transfer rate: 70MB/sec/server
- I/O throughput to disks: 70 – 90 MB/sec writes, 75 MB/sec reads

Current view of the facility

Current view of the facility

Issues

- Load balancing
- Scaling Issues
  - Horizontal (management)
  - Vertical (performance)

Issues continued

Veritas

- Pros
  - Ability to shrink file systems
  - Easy to import/deport volumes among different servers
  - Dynamic multipathing
- Cons
  - Quotas don't work on file systems > 1 TB
  - Expensive
  - Poor customer support
  - Product documentation does not match reality
  - If one element of an underlying striped volume goes offline, the filesystem continues to remain available ???

What constitutes a better system?

- Fast, scalable, reliable and fault tolerant
- Load balancing (Efficient resource utilization)
- Security and centralized management
- Incremental growth
- Global namespace
- Economic benefits of IDE disk

Spectrum of Solutions

- POSIX-like vs Non POSIX-like
- Hardware vs Software
- Existing Protocol vs New Protocol
- Ethernet vs Fibre Channel
- “Unlimited” Scalability vs “Just Fast Enough”
- Dedicated vs Non-Dedicated Resources
- Proprietary vs “Open”

Many Implementations

- Exanet: inter-nodal transfer on private network. Hardware/software solution
- Isilon - similar to Exanet. Somewhat stripped-down filesystem
- BlueArc – ASIC chips dedicated to NFS, network, and filesystem
- Ibrix - Meta servers assigned to segments in a disk pool
- Dcache – Management of heterogeneous storage repositories
- Lustre - Free, Object-based storage. Software only

Panasas Highlights

- An integrated hardware/software solution
- Single global namespace
- Direct and parallel data access
- Dynamic load balancing
- Distributed Metadata
- Seamless expansion
- POSIX compliant
- Ethernet based

Panasas Architecture

Director Blades (The brains)

- File Namespace server(s)
- Manages Metadata object map
- Coordinates between clients and Storage Blades
- Determines “RAID” characteristics of a file.
- Determines distribution of file objects over OSD's

Storage Blades (Object Storage Devices or OSD's)

- Store and retrieve data objects
- Handles I/O to client

Panasas Architecture

- ActiveScale Operating System
  - Runs on Director Blade
  - Divides files into data objects, which are arbitrary in size, and stripes them across storage blades
  - Dynamically distributes workload across storage blades
  - Each storage blades is only filled to 90% capacity. The remaining 10% is reserved for rebuilding parity.
- Direct Flow Software
  - Installed on the Linux compute node
  - Direct data path from client to storage blades
  - Optimizes data layout, caching and prefetching
  - File is reconstructed at the compute node

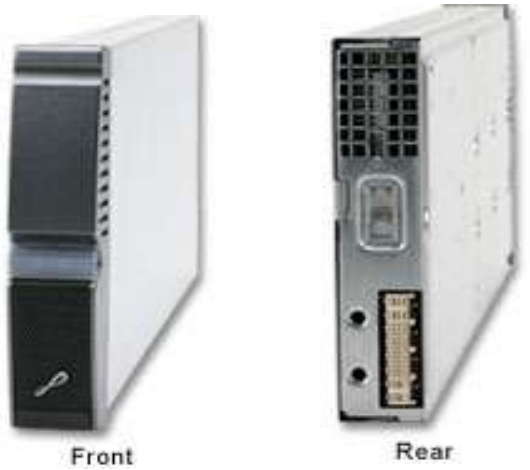
The Panasas evaluation system

- One fully configured shelf (10 Storage blades and 1 Director blade)
- Director Blade
  - 2.4 GHz Intel Xeon
  - 4 GB RAM
  - 2 100/1000BaseT
  - FreeBSD

Front

The Panasas evaluation system

- Each Storage Blade (10)
  - 500GB storage (2 IDE 250GB HDD)
  - 1.2 GHz Intel Celeron
  - 512 MB RAM
  - 100/1000BaseT
  - FreeBSD
- Switches
  - 11 Gigabit Ethernet ports for blades
  - 4 Gigabit links to network
  - Up to 4Gbps Full-Duplex – Jumbo Frames



Testing Expectations

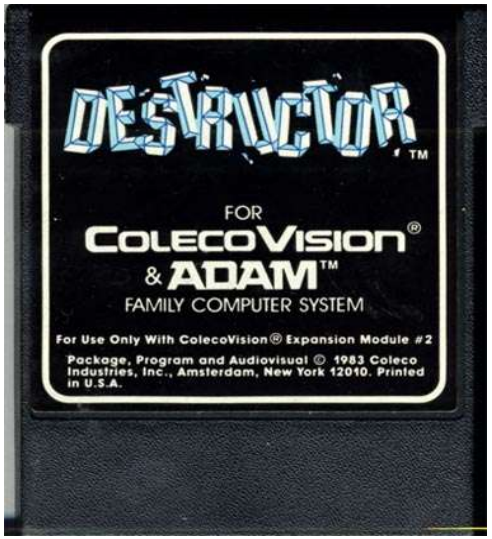
- Scaling: a linear increase in performance as more compute nodes are added
- Bandwidth: should be able to saturate the network
- Random I/O performance
- Ease of management
- NFS support

Testing methodology

- Tools: Iozone and Ioperf
- 2 trunked gigabit links
- Stage 1: write I/O testing on 10 nodes first using NFS then DirectFlow
- Stage 2: 200 nodes using DirectFlow simultaneously; randomized write I/O
- Stage 3: Read I/O on all nodes using DirectFlow. True user analysis.

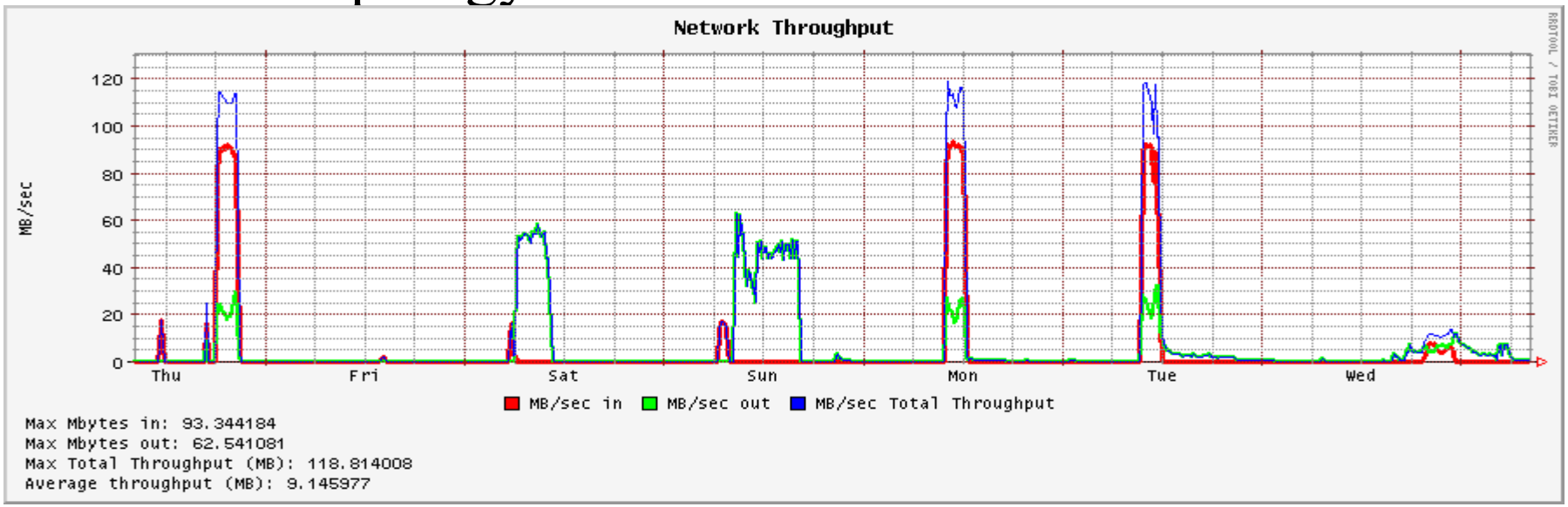
Initial Tests: Not Good

- Run Jerome Lauret's “The Destructor” -- try to find out when the system would collapse. For example:
- 512 blocks per I/O, 80000 I/O ops, 50Nseeks, 10 loops on Client A
- 8192 blocks per I/O, 5000 I/O ops, 50Nseeks, 10 loops on Client B
- Kernel panic on both nodes
- Server side data corruption



Test Chapter 2: Success

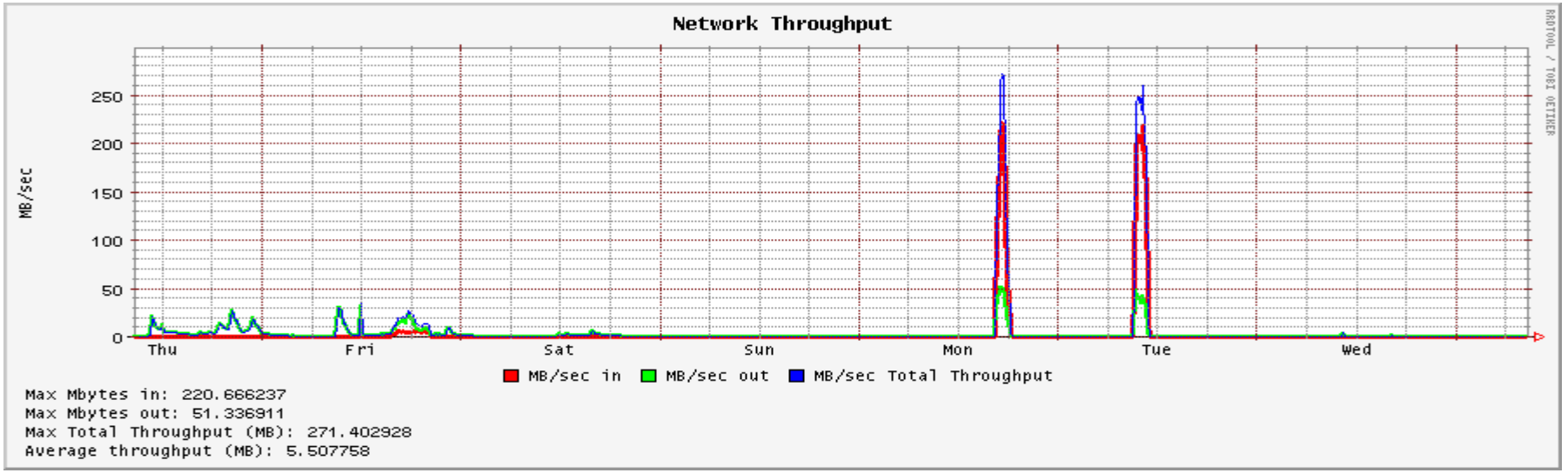
- Panasas identified the problem quickly and released a new Direct Flow client
- I/O tests resume and progress to Stage 2
- Network bottleneck at 1GB/sec – Problem with our network topology

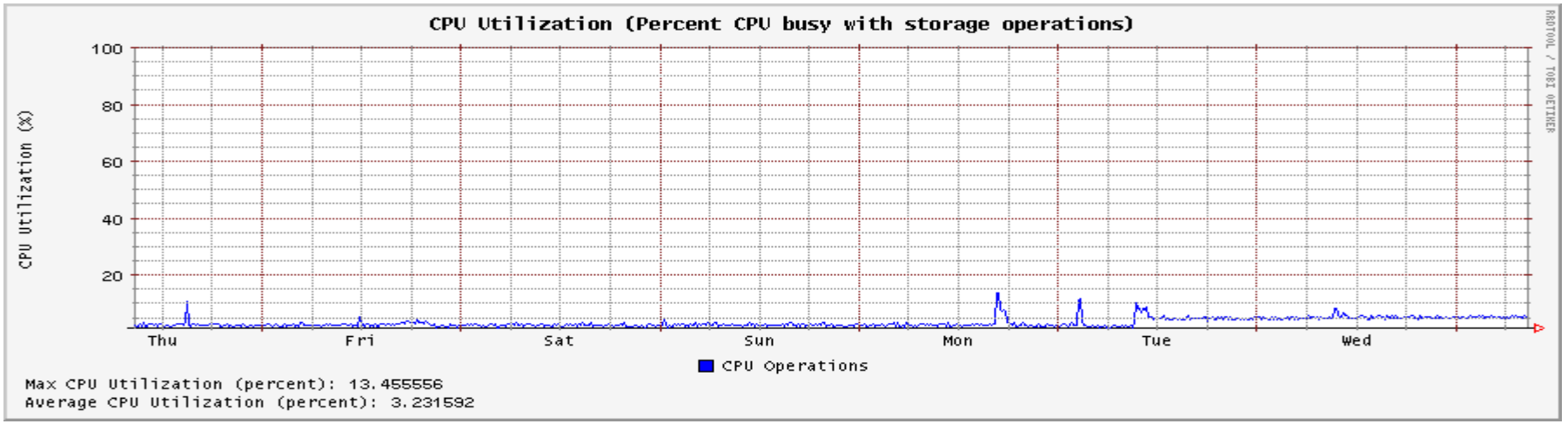


Test Chapter 2: Success

- Able to saturate the network at 2Gb/sec
- I/O limited to available network bandwidth
- Low CPU usage compared to NFS
- No crashes or corruptions

Test Chapter 2: Success





Critiques and snags along the way

- At first, no on-line reconstruction of storage blade (fixed)
- DirectFlow and ActiveScale versions were constantly changing
- No LDAP integration yet
- Issues with performance monitoring tools
- RedHat kernel patch for memory management (RedHat specific?)

Future Tests

- Move to Phase 3 – production usage by an experiment
- Fully test user / group based quotas

Conclusions

- Panasas scales well
- The latest implementation is ready for primetime
- Show stoppers?
- Why not use now?
- Is Centralized storage necessary in a grid environment ?