The InTENsity PowerWall: A SAN Performance Case Study

Presented by

Alex Elder

Tricord Systems, Inc.
elder@tricord.com

Joint NASA/IEEE Conference
On Mass Storage Systems
March 28, 2000

Talk Outline

- The LCSE
- Introduction
- InTENsity Applications
- Performance Testing
- Lessons Learned
- Future Work

Laboratory for Computational Science and Engineering (LCSE)

- Part of University of Minnesota Institute of Technology
- Funded primarily by NSF/NCSA and DoE/ASCI
- Facility offers environment in which innovative hardware and software technologies can be tested and applied
- Broad mandate to develop innovative high performance computing technologies and capabilities
- History of Collaboration with Industrial Partners (in Alphabetical Order)
  - ADIC/MountainGate, Ancor, Brocade, Ciprico, Qlogic, Seagate, Vixel
- Areas of focus include CFD, Shared File System Research, Distributed Shared Memory

The InTENsity PowerWall

- What is the InTENsity PowerWall?
- Display Component
- Computing Environments
  - Irix
  - NT/Linux Cluster
- Storage Area Network

What is the InTENsity PowerWall?

- Display system used for visualization of large volumetric data sets
- Very high resolution, for detailed display
- Very high performance–displays images at rates that allow for “movies” of data
- Driven by two computing environments with common shared storage

InTENsity Design Requirements

- Very high resolution–beyond 10 million pixels
- Physically large, semi-immersive format
- Rear-projection display technology
- Smooth frame rate (over 15 frames per second)

- Driven by SGI Onyx and PC cluster platforms
- High performance/high capacity disk system
- Significant processing capability, large memory

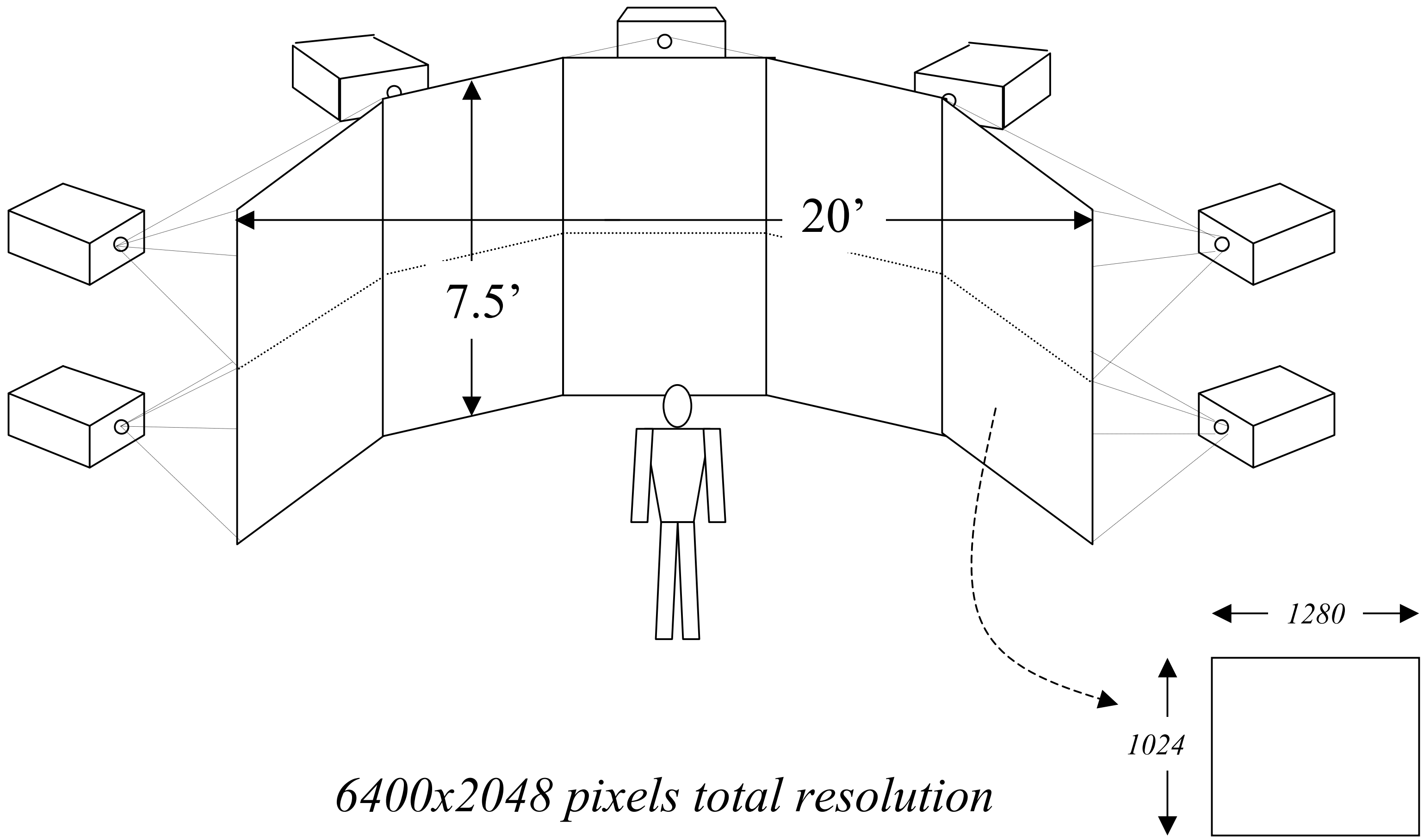
Planned Uses

- Presentation display environment
- Collaborative working environment
- Visualization post processing engine
- CFD Computational cluster
- Storage Area Network research

Display Characteristics

- Five 7’5” tall ½” thick plexiglas screens, oriented in a quarter-circle arc
- Two panels per screen
  - 1280x1024 pixel resolution each
- Each panel rear projected by Electrohome HAL Series DLV1280
- Backed by a video switching network to allow flexibility in source for display

Physical Layout



6400x2048 pixels total resolution

The Real Thing


Front View


Rear View

Computing Environments

- "Large" legacy systems
  - SGI Onyx2 and Onyx running Irix
- "Small" cluster-based systems
  - Intel based systems running Windows NT and/or Linux

Large Computer Environment

⌘Silicon Graphics Onyx

- 4 R10000 190MHz Processors
- 2 GB Main Memory
- 2 Infinite Reality graphics engines with 4 Raster Managers
- 2 Dual Channel Prisa HIO Fibre Channel (FC) Adapters
- IRIX 6.2 Operating System
- Used almost exclusively for support of older software, original PowerWall technology

Large Computer Environment, cont.

- **Silicon Graphics Onyx2**
  - 8 R10000 195MHz Processors
  - 2 GB Main Memory
  - Two Infinite Reality graphics engines with 6 Raster Managers
  - Two Dual Channel SGI Adaptec Emerald-based FC Adapters
  - Four Dual Channel Prisa PCI64 XIO FC Adapters
  - Silicon Graphics IRIX 6.5.5 Operating System
  - Used for both old and new PowerWall technology applications

Small Computer Environment

- 12 SGI Model 540 Visual PC Workstations
  - 10 display drivers
  - 1 additional designated for control
  - 1 additional designated for development
- All are connected to fabric and video network
- Can act individually or as a clustered unit

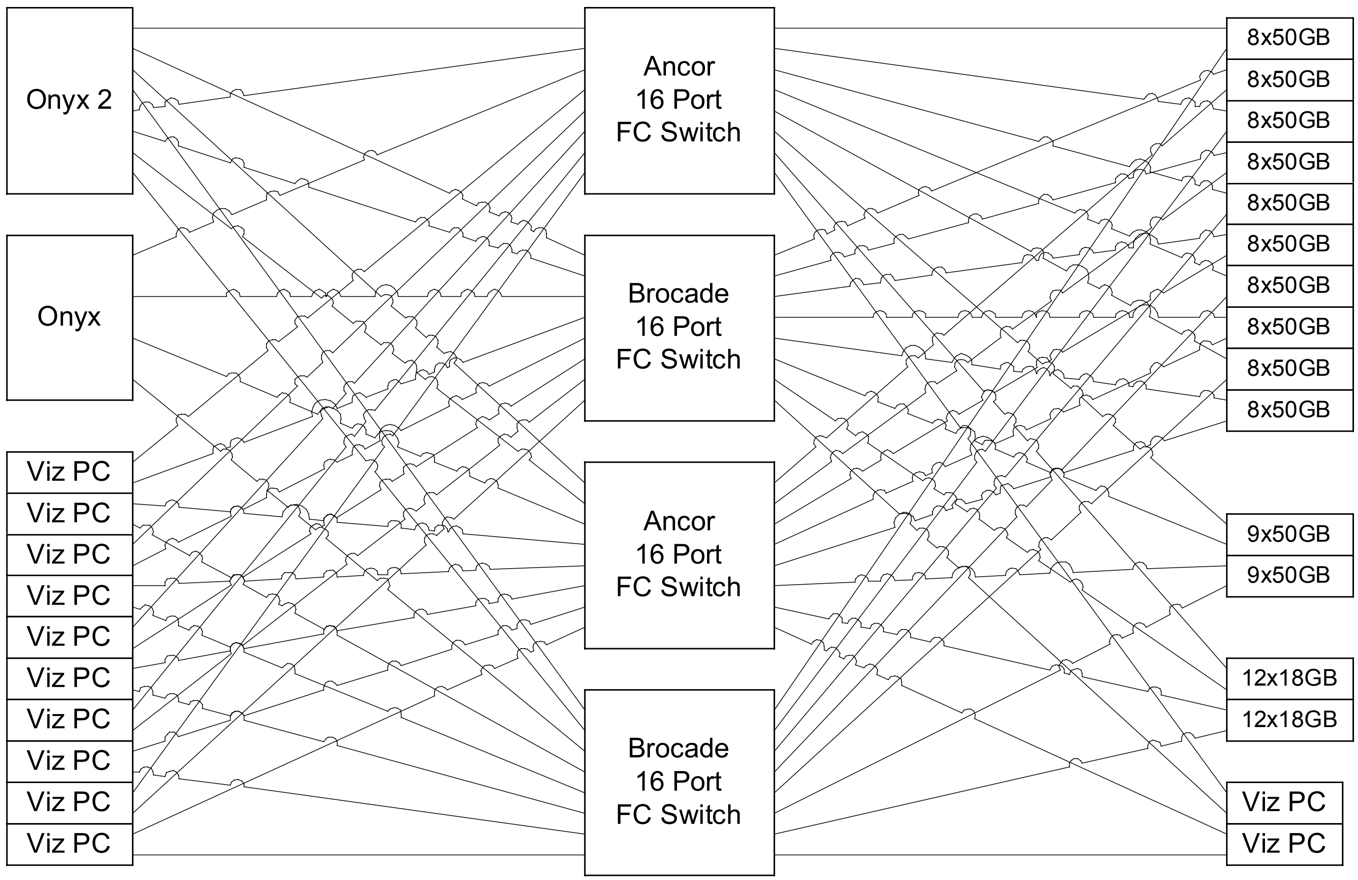
Small Computer Environment, cont.

- Visual PC Configuration (each)
  - Four 550 MHz Pentium-III Xeon Processors
  - 1 GB ECC Memory, 100 Base T Ethernet
  - A Dual-Channel Qlogic QLA2202F PCI64 FC HBA
  - Three System disks (dual boot plus scratch)
  - SGI Cobalt Graphics
  - Microsoft Windows NT 4.0 SP 4 Operating System

InTENsity Storage Area Network

- All Fibre Channel based
- Multi-vendor fabric interconnect comprised of four 16-port switches
  - Two Ancor MKII Switches
  - Two Brocade Silkworm Switches
- ~100 Seagate Barracuda 50 disk drives in twelve 8-drive JBOD enclosures

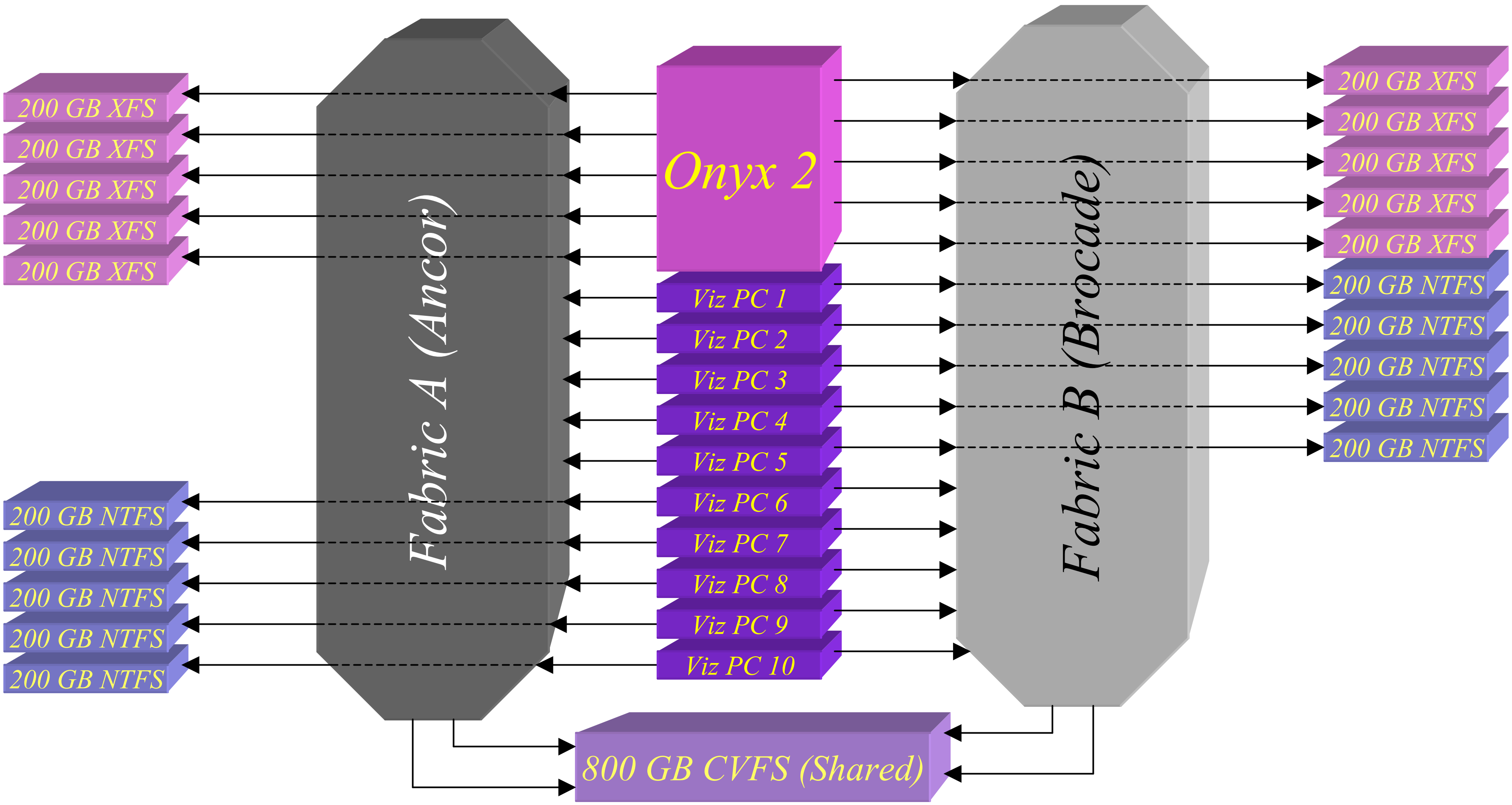
Storage Area Network Connectivity



Storage Usage

- Disks are arranged as:
  - Ten 200 GB IRIX XFS/XLV logical volumes (4 disks/volume)
  - Ten 200 GB Windows NT Logical volumes (4 disks/volume)
  - One 800 GB ADIC CentraVision File System volume
- Each XFS volume comprises a dedicated Irix file system
- Each NT volume is dedicated to one of the Viz PC's
- CentraVision (CVFS) volume is shared by all
  - Heterogeneous shared file system between NT & IRIX
  - Designed for the movement of large files (video)
- Everything is on the fabric

Logical Disk Assignments



InTENsity Applications

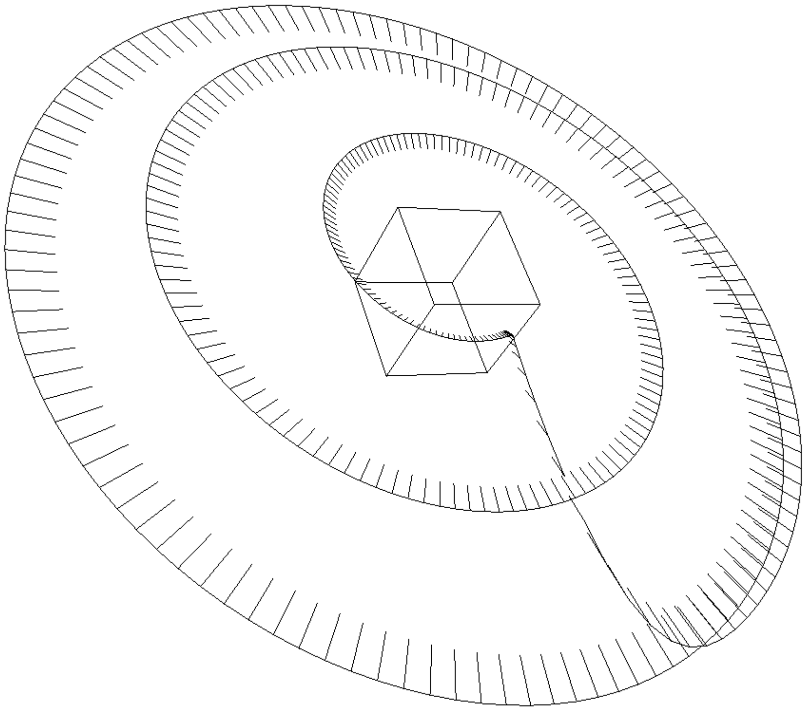
- Two Principle applications that stress the SAN
  - Movie Generation from scientific data sets
  - Movie Playback
- Other applications include use of a Distributed Shared Memory computing model that extends shared memory using shared disks

Movie Generation

- Movies are generated to visualize data representing a physical volume as it evolves over time
- View of volume is determined interactively, using a low resolution approximation of the volume
- This yields a series of key frames, which define a "flight path" around the volume

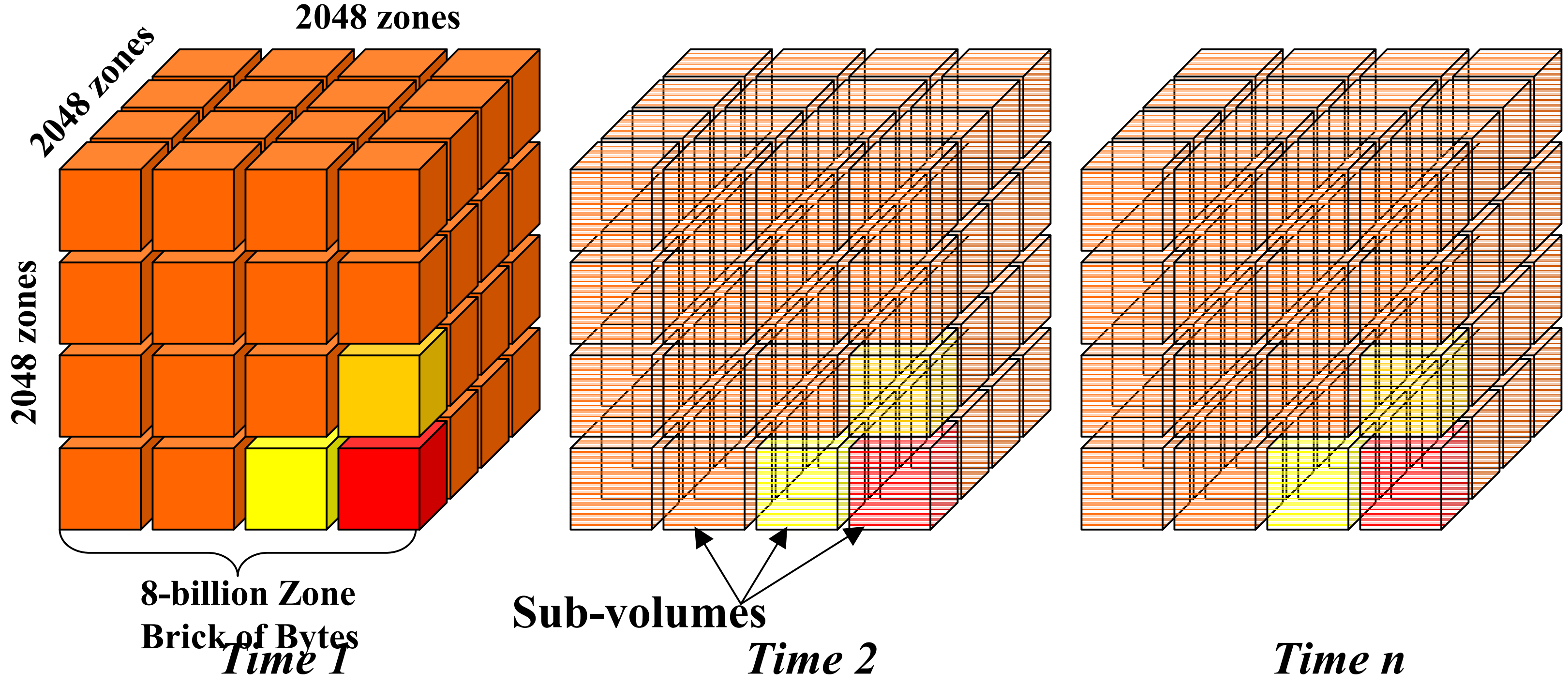
Movie Paths

- Movie frames are defined by interpolating between key frames along the flight path
- Each movie frame will require an image to be rendered for each of the ten wall panels



Decomposing the Data for Rendering

- Volume data is too large (1-10 GB/instance) to be rendered in memory all at once
- Data is broken into a hierarchy of sub-volumes



Distributed Rendering

- Shared storage makes possible distributed rendering of movie frames
  - Large data size demands high performance of direct access to I/O devices (SAN)
  - Rendering of separate movie frames is independent, so can be done in parallel
- SAN-attached systems read sub-volumes from shared storage
- 5 MB rendered movie frames written back to same shared storage

Movie Playback

- Movie playback amounts to synchronized playback 10 streams of movie frames to the display panels
- The Onyx2 is able to play all 10 streams at a rate of ~10 frames/second
- By distributing the task, 10 VizPC's are able to sustain ~20 frames/second rate

Movie Playback, continued

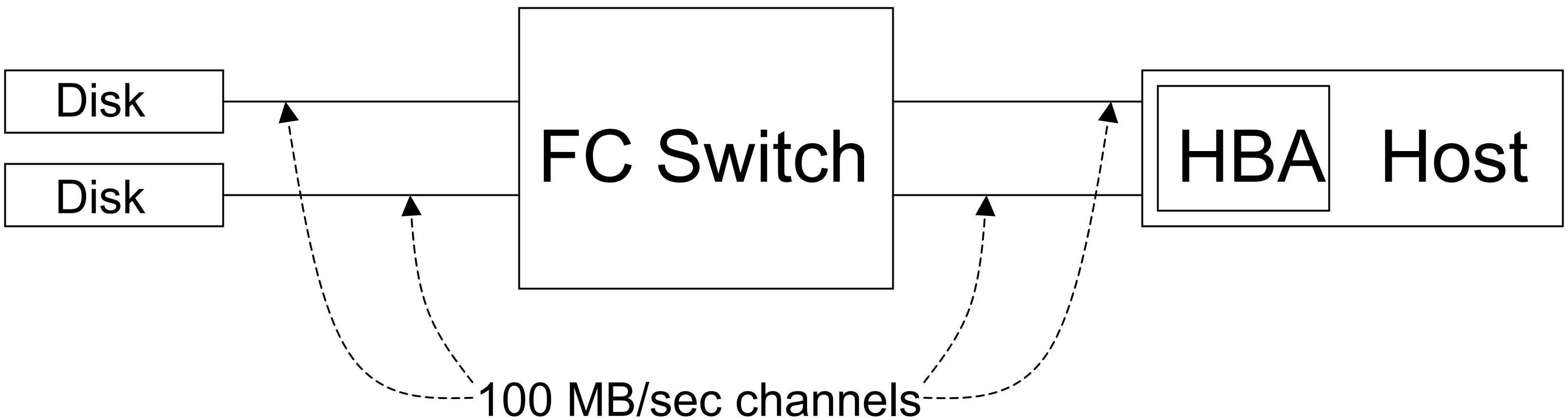
- For VizPC environment, a master Movie Player coordinates synchronization and control of separate movie streams
- Synchronization makes use of a high resolution clock and a common “clock daemon” (described in detail later)

Part II: Performance Testing

- Understanding the behavior of system components yields a better understanding of the performance of the whole system
- We approached the SAN performance testing by first evaluating individual system performance, then evaluating the performance impact of multiple-system use of the SAN

Single System Overview

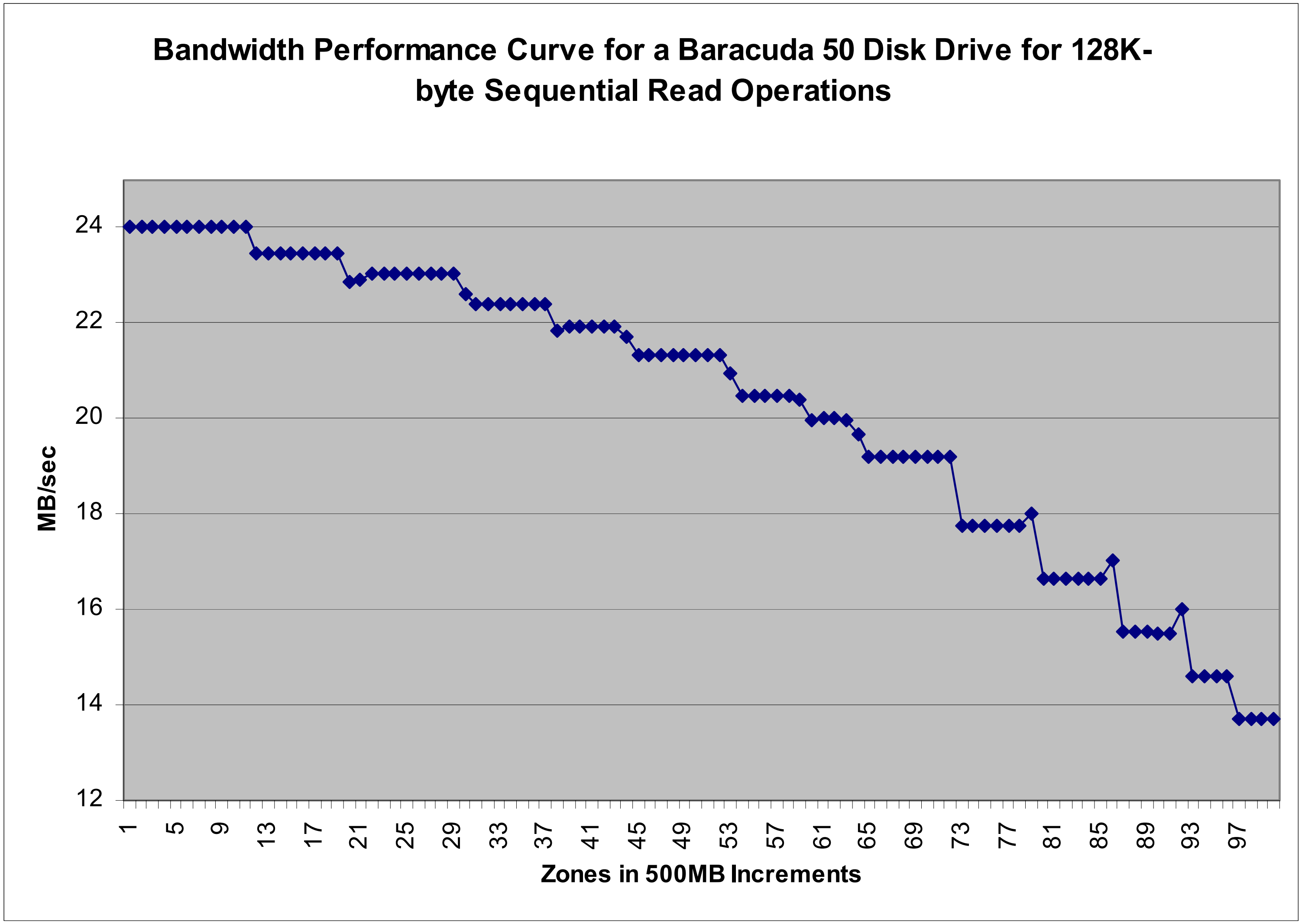
- Remainder of talk will be Viz PC oriented
- Bandwidth is primary performance criterion



Component Performance

- Individual Disk Performance
- Channel Performance
- Switch Performance
- HBA Performance
- Host System Performance

Individual Disk Performance



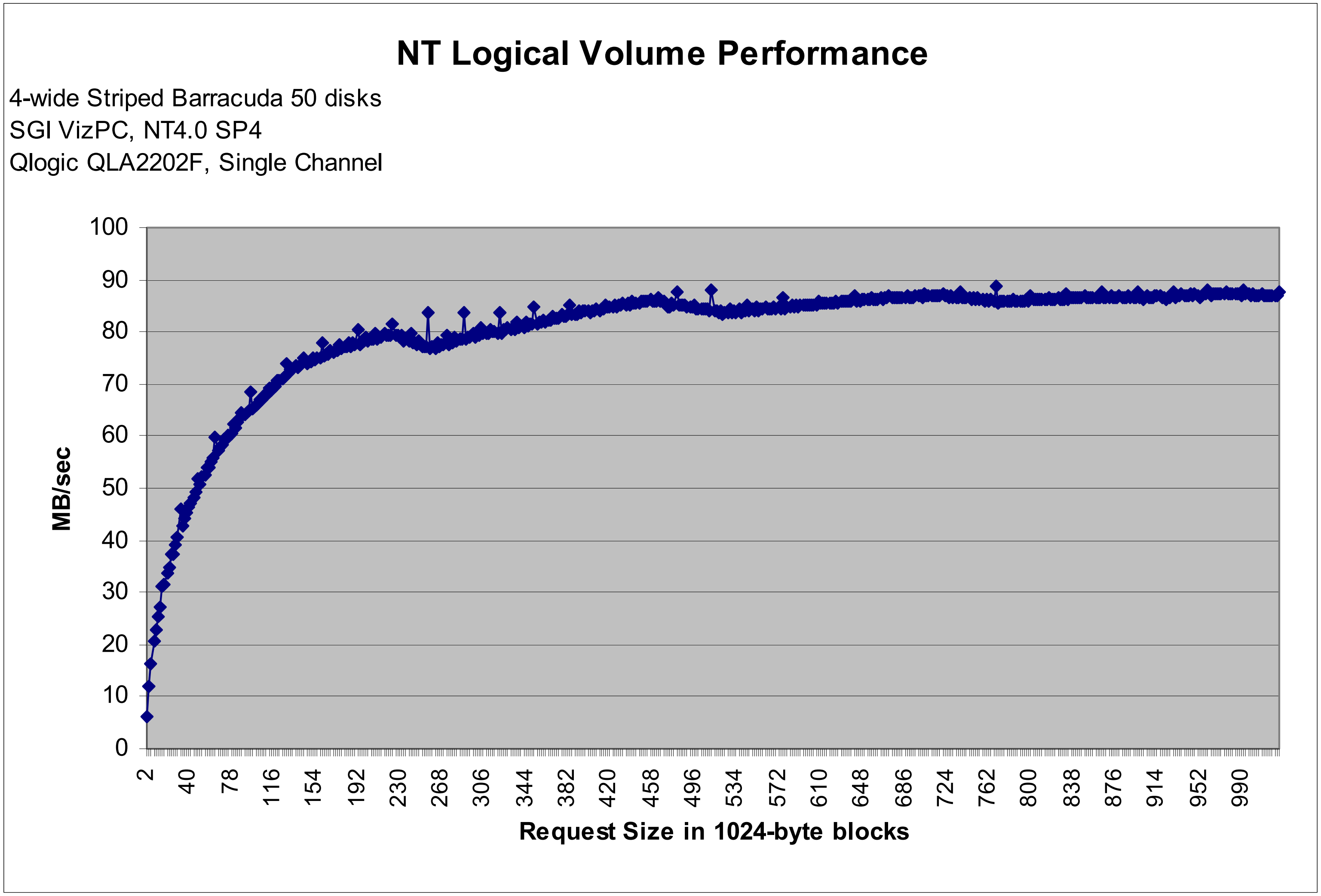
Channel and Switch Performance

- The channels (cables) are known to have a bandwidth capacity of approximately 100 MB/sec, especially for the large transactions we use
- Earlier switch testing at the LCSE showed our switches made no significant impact on the end-to-end bandwidth performance

HBA Performance

- Testing showed the Qlogic QLA2200F on the SGI 540 PC could transfer:
  - 180+ MB/sec “raw” read transfer rate from two Qlogic FC ports connected to 16 individual disks, one adapter
  - 160+ MB/sec “raw” read transfer rate through a single logical volume 14 disks wide

NT Logical Volume Performance Curve



Single System Bandwidth Performance Summary

- Seagate Barracuda 50 Disk Drive
  - 24 MB/sec transfer rates for read/writes (outer cylinders)
  - 13 MB/sec transfer rates (inner cylinders)
  - Sustained up to 88 MB/sec reading from a raw 4-wide striped logical volume using 512-Kbyte requests to a single process, non overlapped
- Could perform 880 MB/sec using 40 disks configured as 10 NT volumes
- Translates to 14 movie frames/second; better if all 80 drives and both channels were used

Multiple System Testing

- To test multiple two additional functions had to be added to the existing testing facilities
  - Accounting for existence of multiple clocks
  - Coordinating the initiation of tests to run concurrently on multiple hosts

Reference Clock

- Each host has an internal sense of time
- Each provides a high frequency clock register that can be read
- High frequency clock is used to determine time interval between “local” time and the time on a separate host whose time is taken to be the “global” time
- Time stamps all translated to global time

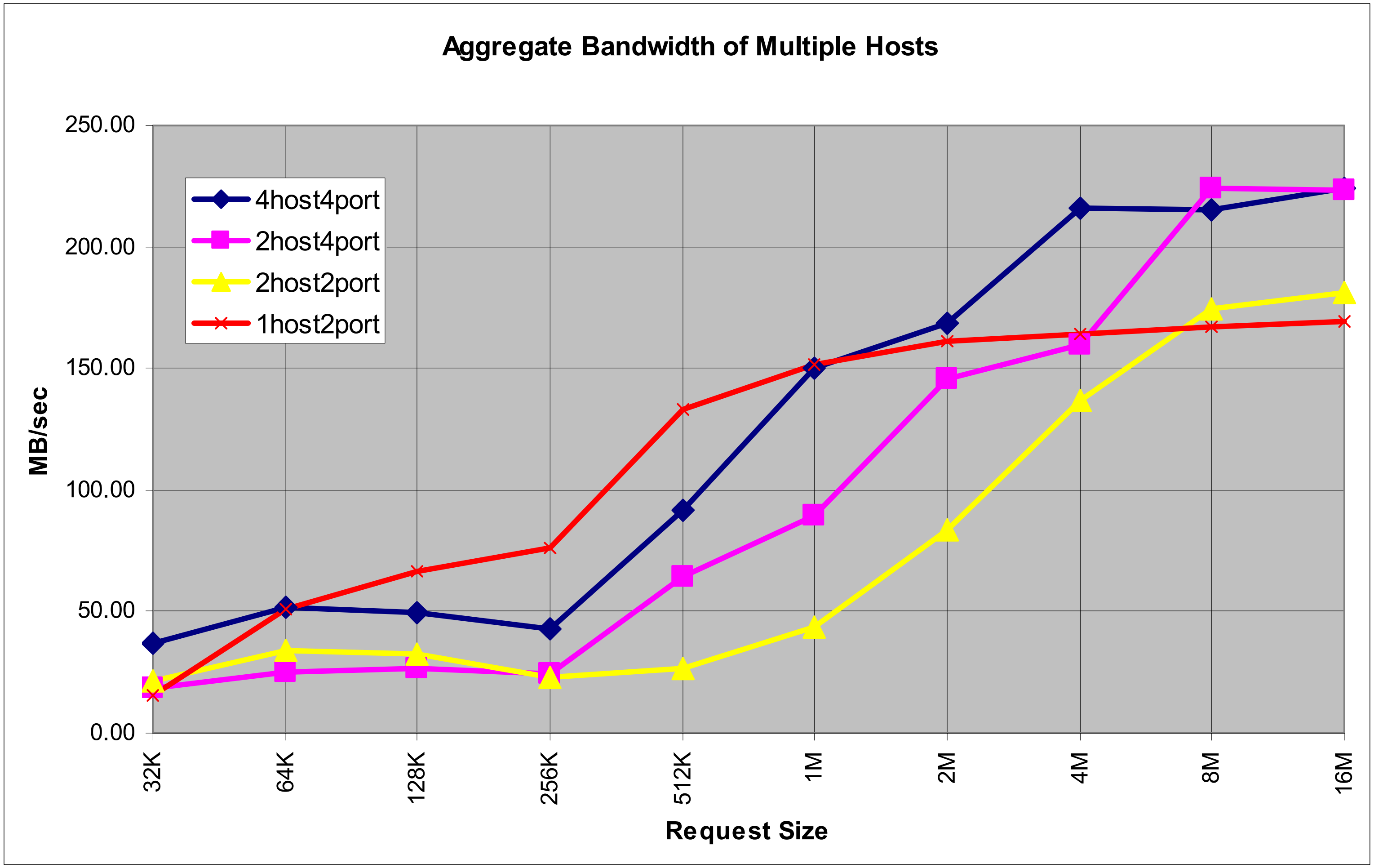
Synchronization

- Establishing a global time allows results of concurrent tests to be correlated
- Also allows for synchronization by polling the local clock until a predetermined (global) time has been reached
- This synchronization technique is used by the test framework as well as the movie player

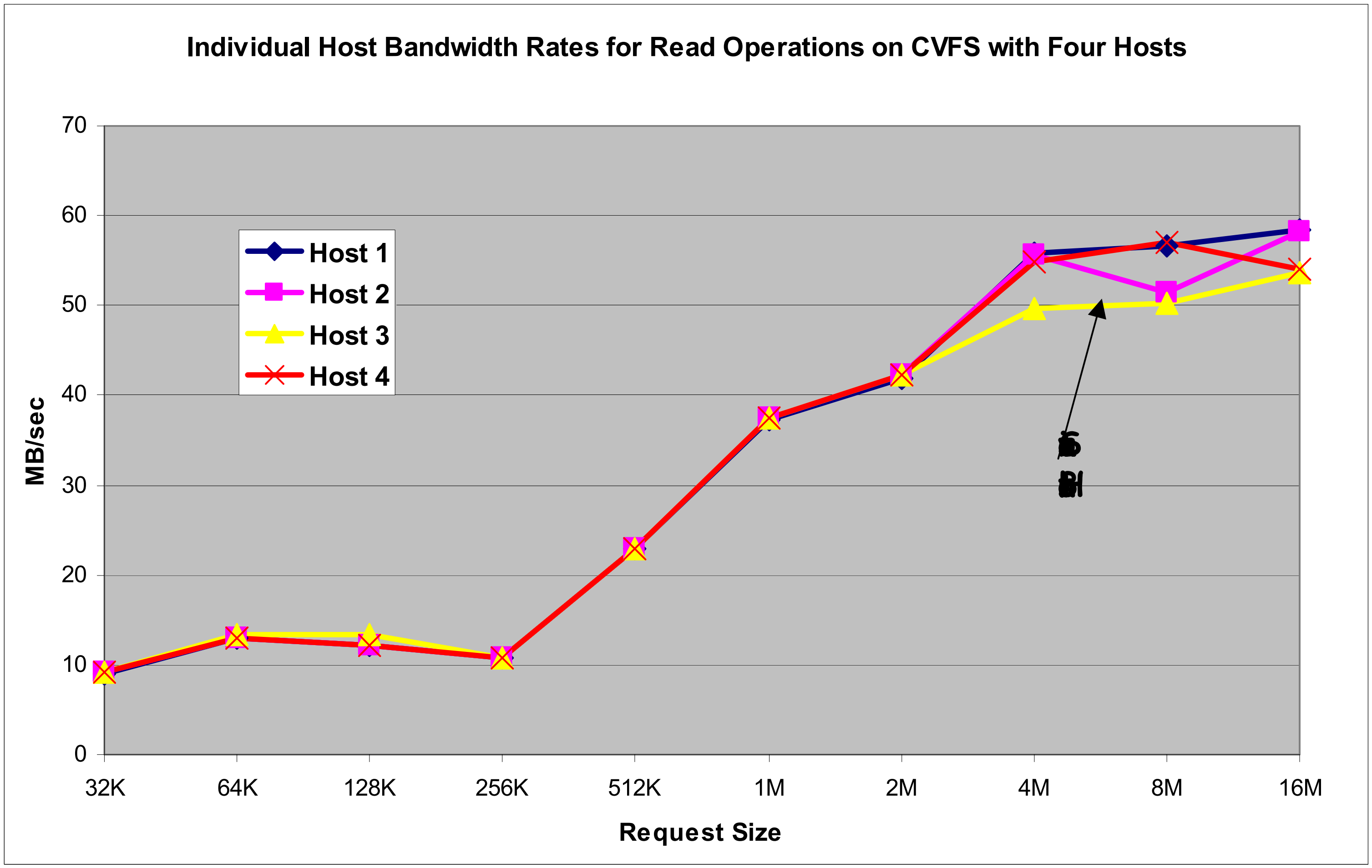
A Few Interesting Results

- CentraVision File System (CVFS) Read and Write performance
  - Single host (2 channels): up to 151 MB/sec write
  - Single host (4 channels): up to 170 MB/sec read
  - Two hosts (2 channels each): up to 222 MB/sec read
  - Four hosts (1 channel each): up to 222 MB/sec read
- We see some more interesting anomalies in the individual performance of the shared disks

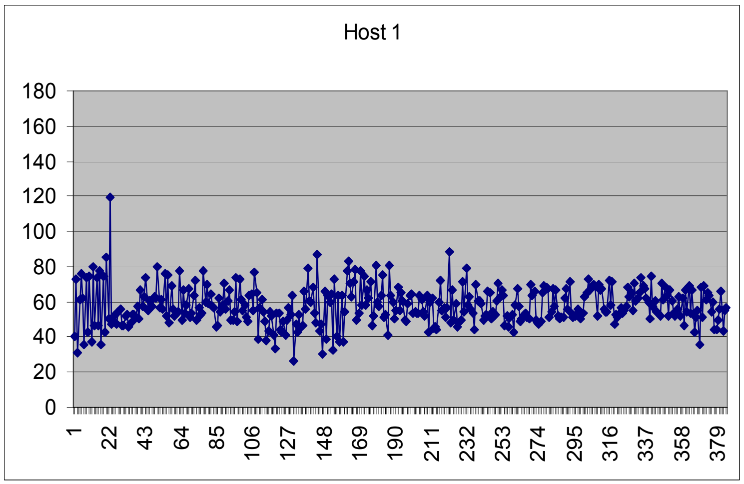
CVFS Aggregate Bandwidth

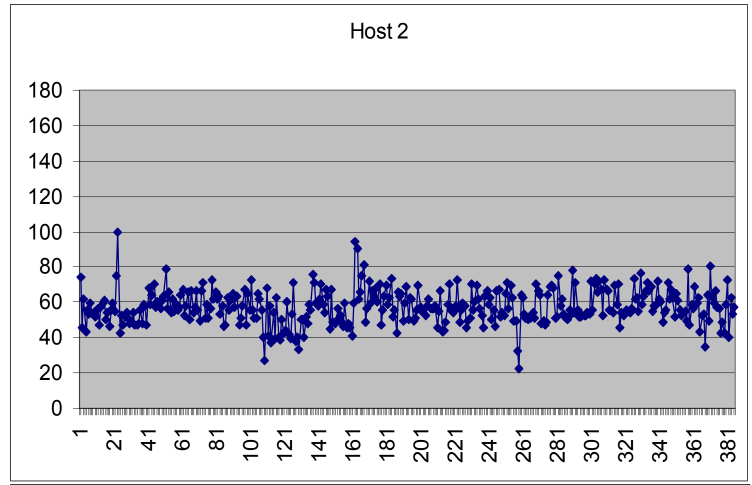


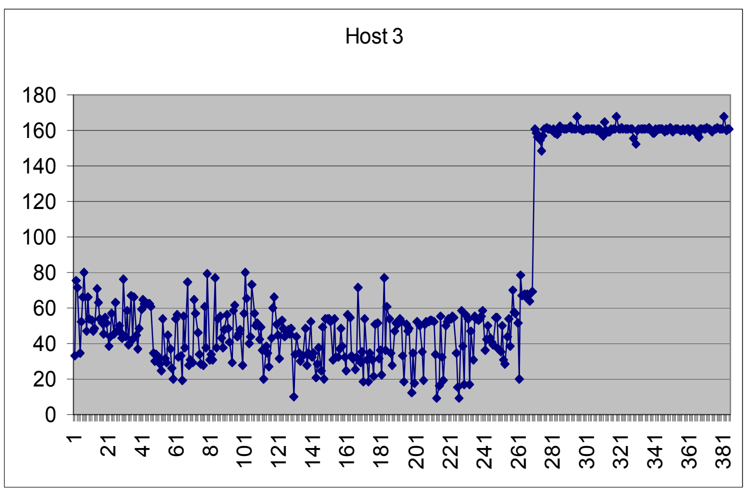
Bandwidth Distribution

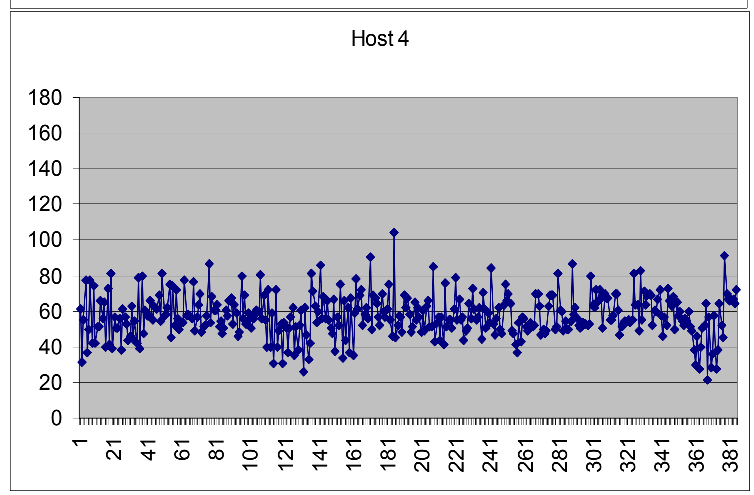


Individual Host Bandwidth (Misleading)

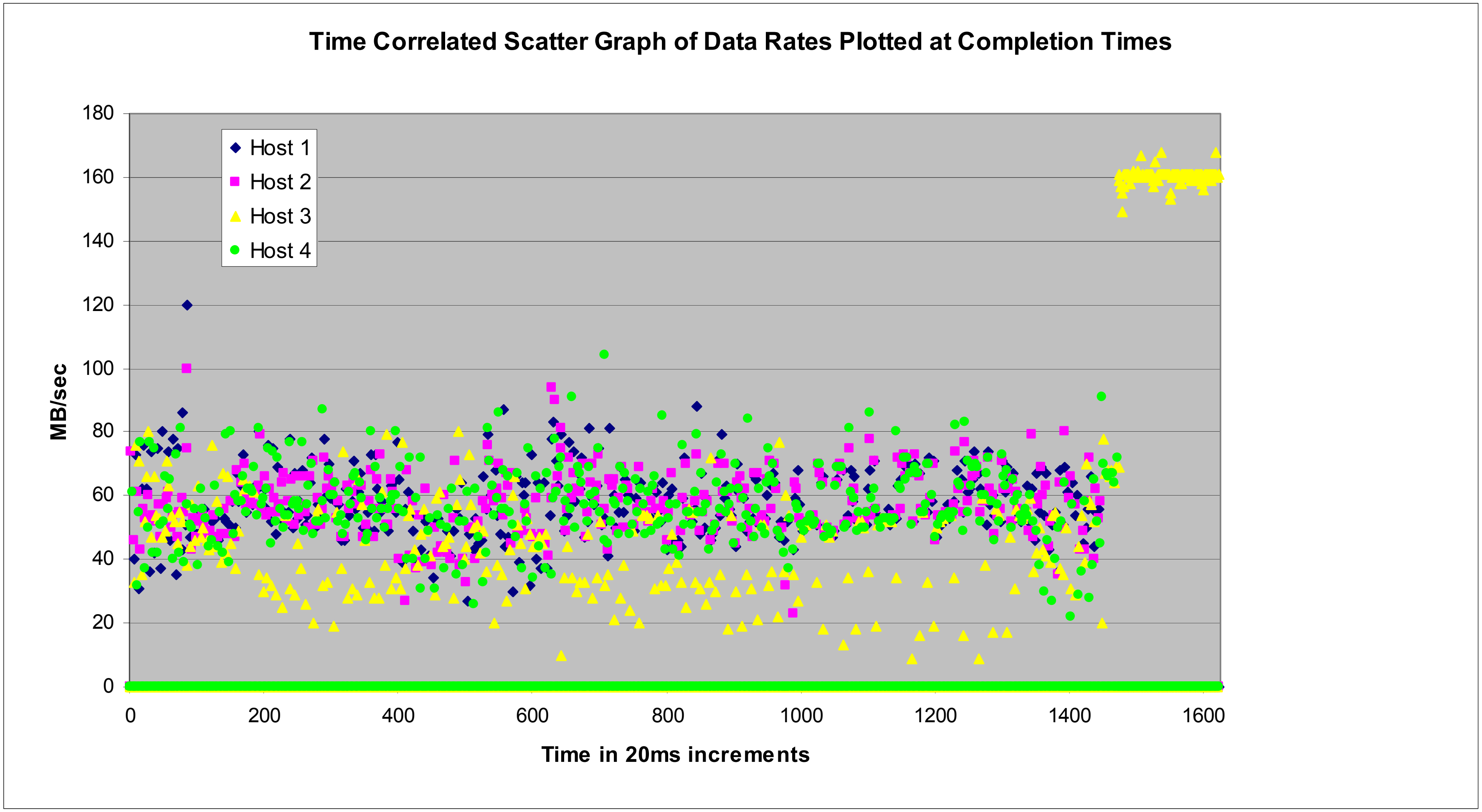








Bandwidth Performance Time Correlated View



Miscellaneous

- This display system was at SC99 in the ASCI booth
- It is available for use by DoE Researchers and Industrial Collaborators
- Incorporated into research on Storage Area Networks and “Heterogeneous” Shared File Systems
- Very flexible Presentation device because it can be configured into many different operating modes
- Useful for truly “collaborative” work: multiple people can operate multiple screens simultaneously

Future Work: InTENsity Powerwall

- Linux support – dual boot with NT
- Experiment with other Intel-based platforms
- Incorporate load-balancing Distributed Shared Memory Computing model to the PC and SGI clusters
- Seamless simulation to visualization to presentation environment
- The Digital Technology Center – 1/2001

Future Work: Performance Testing Framework

- Continued analysis of shared/distributed test results
- Applying test framework on other file systems
- Extending test framework to emphasize other aspects of performance (I/O's per second, request latency)
- Porting test framework to other platforms (OS and hardware)

Lessons Learned

- SAN Management software is sorely needed: Ability to look at a switch and see exactly what nodes are connected to which ports
- Need the ability to examine and test *components* of a SAN individually: i.e. Disks, GBICs, switch ports, cables, host adapters, …etc.
- Better fail-over capability in the upper level software layers such as the File System, logical volume device drivers, …etc.
- Logical volumes with large numbers of individual disks can have performance problems
- Need better tools to distribute and maintain firmware and driver releases on all the nodes in a SAN
- NT needs to learn more about SANs and shared disks

The InTENsity PowerWall: A SAN Performance Case Study

Presented by

Alex Elder

Tricord Systems, Inc.
elder@tricord.com

Joint NASA/IEEE Conference
On Mass Storage Systems
March 28, 2000