Disk Subsystem Performance Evaluation: From Disk Drives to Storage Area Networks

8th NASA/GSFC Conference on Mass Storage Systems and Technologies
17th IEEE Symposium on Mass Storage Systems

Thomas M. Ruwart
University of Minnesota
March 27, 2000
tmr@tc.umn.edu

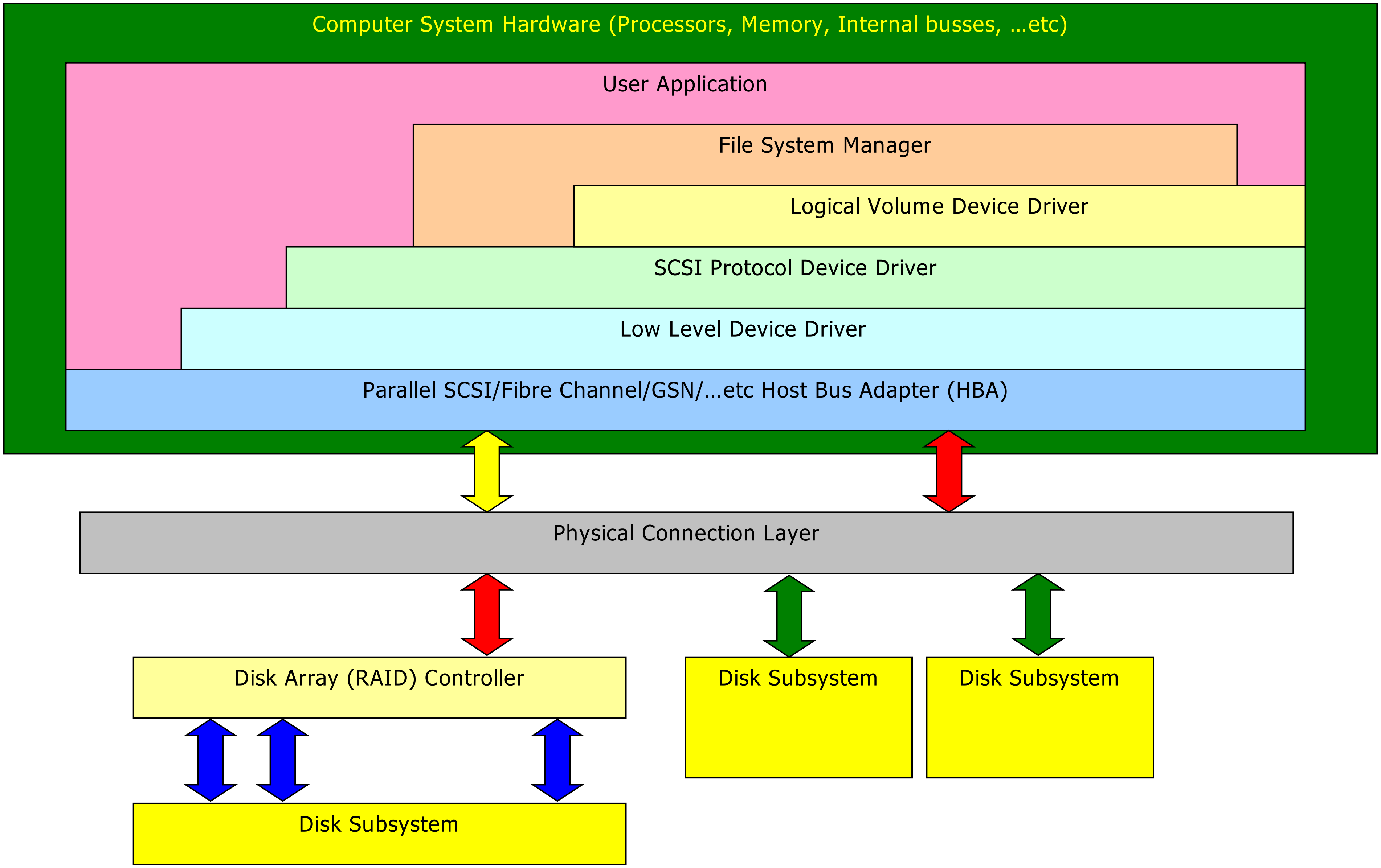
Overview

- Purpose and Motivation
- The Storage Subsystem Hierarchy
- Performance Implications: The Impedance Matching Problem
- Testing Philosophy and Methods
- Testing Framework
- Examples of Performance Results & Anomalies
- Conclusions and Future Work

Purpose and Motivation

- To develop a testing methodology and framework that can generate accurate, meaningful, and reproducible performance measurements of disk storage I/O subsystems
- Apply this testing framework to individual systems as well as clusters
- Identify and quantify "scalability" issues in disk storage subsystems

The I/O Subsystem Hierarchy



Components of the Storage Subsystem Hierarchy

- User Application – generates a request for data transfer between a buffer in the application data space and the storage media.
- File System Manager – translates the User Application request into a series of requests to transfer logical blocks of data from a logical device.
- Logical Volume Manager – used to aggregate multiple storage devices together and provides a "single device" image to the upper layers in the hierarchy. This layer issues data transfer requests to single devices.
- I/O Protocol Device Driver – translates a data transfer request into a SCSI command that is sent to a specific target device.
- Low Level Device Driver – responsible for managing the transfer of data between the host computer memory and the storage device.

Components of the Storage Subsystem Hierarchy cont'd…

- Physical Connection Layer – defines the physical data path from the storage device to the host-bus-adapter including switches, hubs, storage directors, …etc.
- Storage Device – includes disk array controllers and disk drives.
- **Each I/O request must traverse some number of layers in this hierarchy**
- **The more layers an I/O request traverses, the more susceptible it is to an:**

Impedance Mismatch

The Impedance Matching Problem

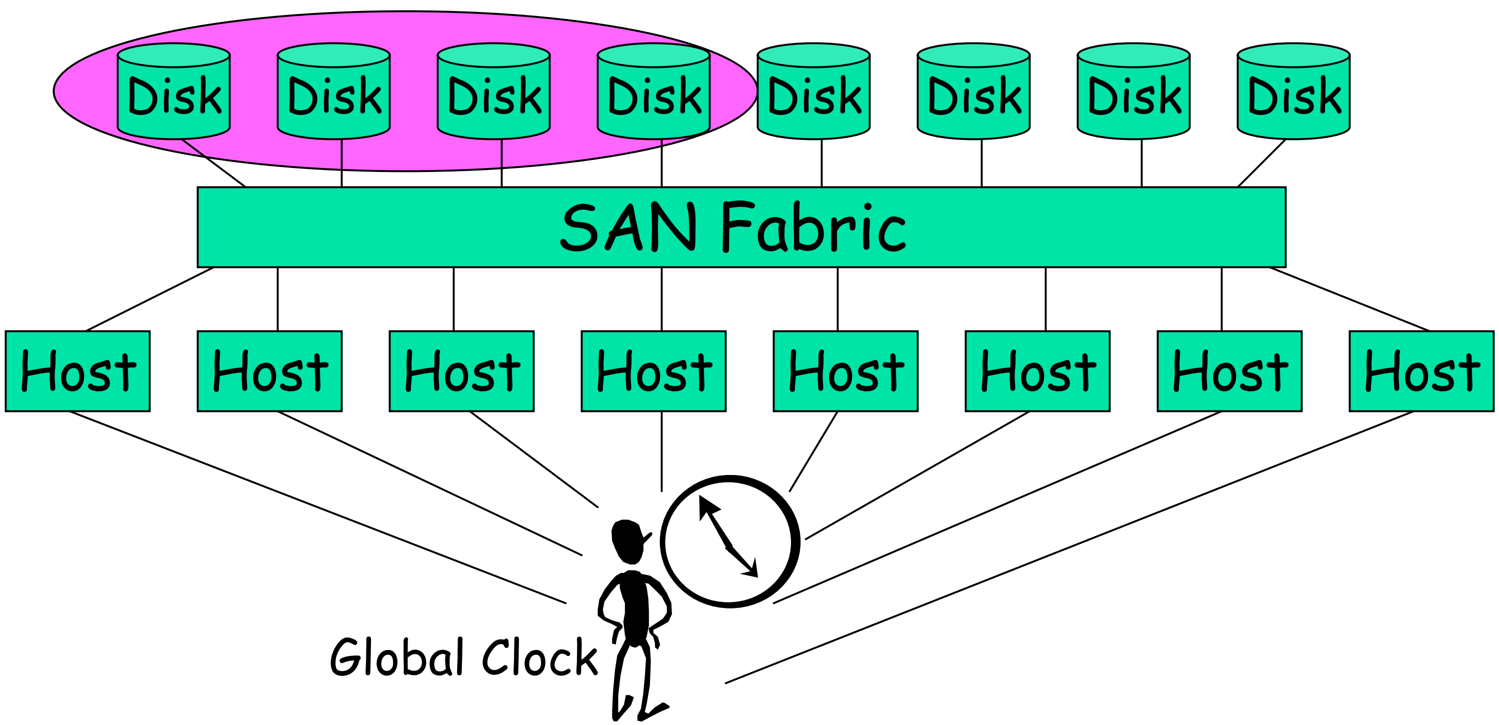
- A general term used to identify a class of problems related to the performance of the flow of data from the storage media to the application memory
- This problem persists independent of the architecture, configuration, data layout, ..etc
- Artifact resulting from the *interaction* of the components in the Storage Subsystem Hierarchy
- In this context, an “impedance mismatch” is related to things like I/O request size and alignment mismatches
- The overall effect is a
  - Decrease in observed performance delivered to the application
  - Decrease in efficiency of the storage subsystem

Testing Philosophies and Methods

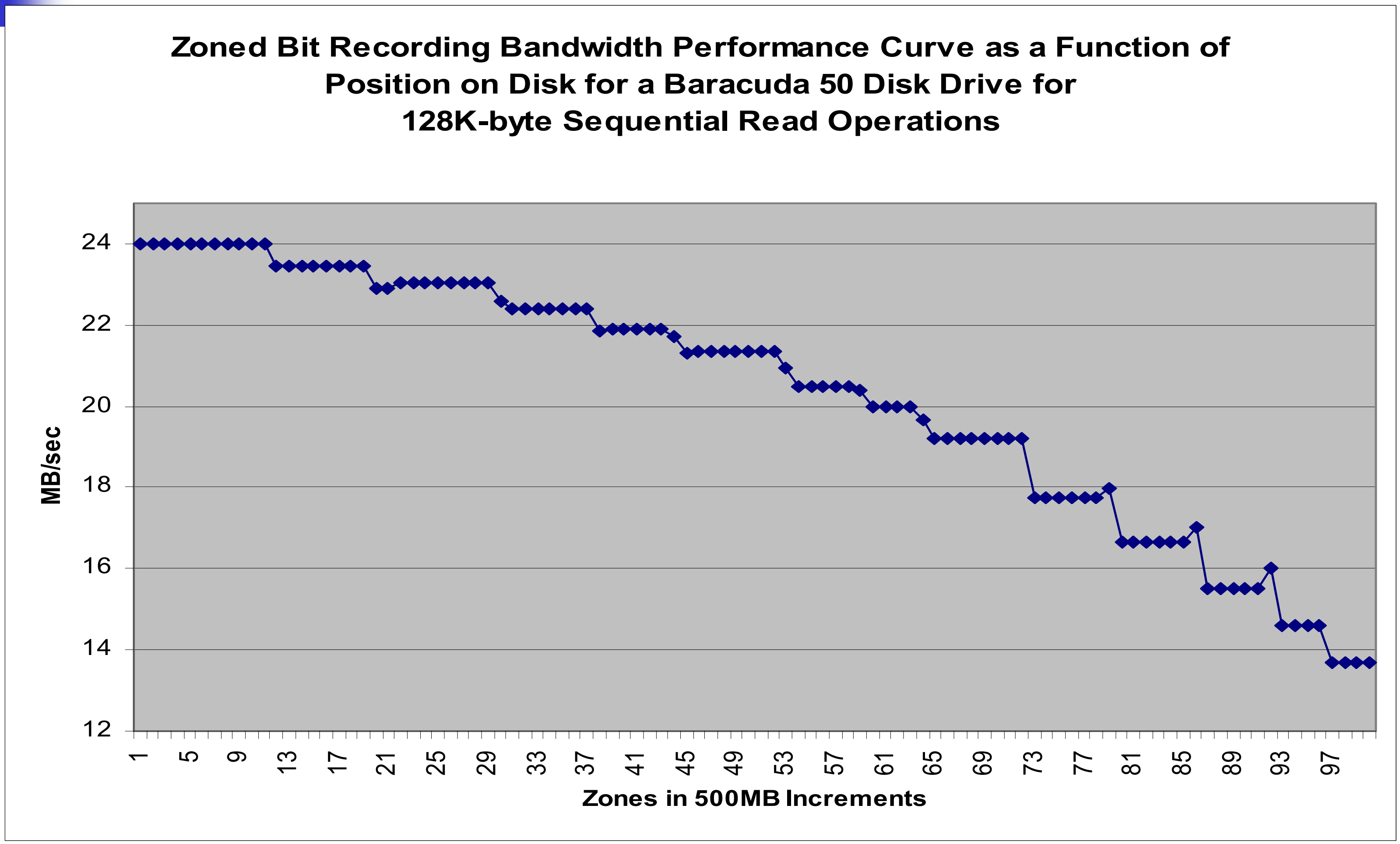
- Test individual components of the Storage Subsystem Hierarchy
- Test meaningful configurations or these components
- Tightly control all variables so as not to confound the results
- Collect as much measurement data as possible without having a significant impact on the "actual" performance (Uncertainty Principle)
- Generate reproducible results
- Generate results that can be correlated across systems and/or platforms

SAN Performance Testing Framework

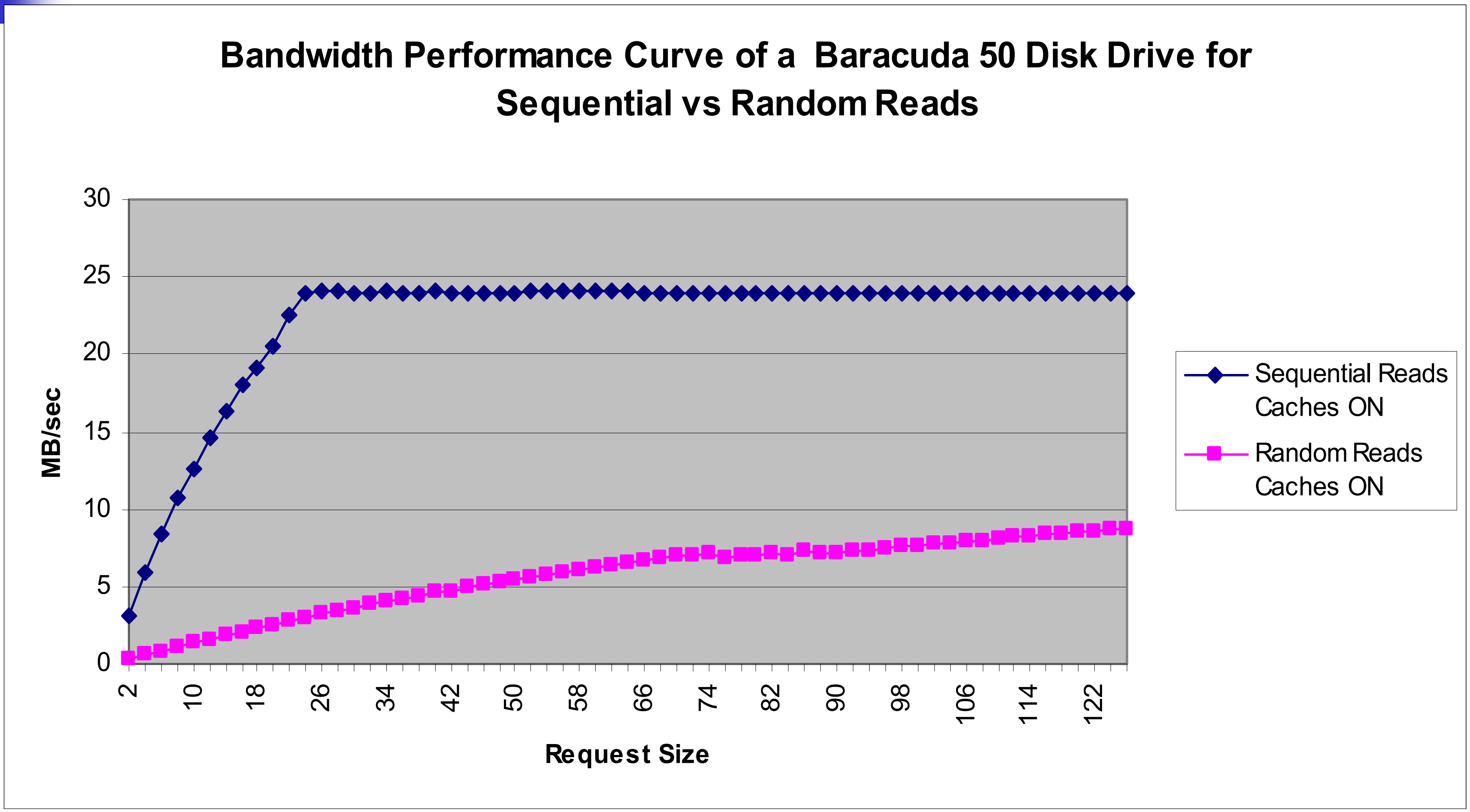
- Designed to allow for isochronous performance testing of a storage subsystem on a SAN from multiple hosts



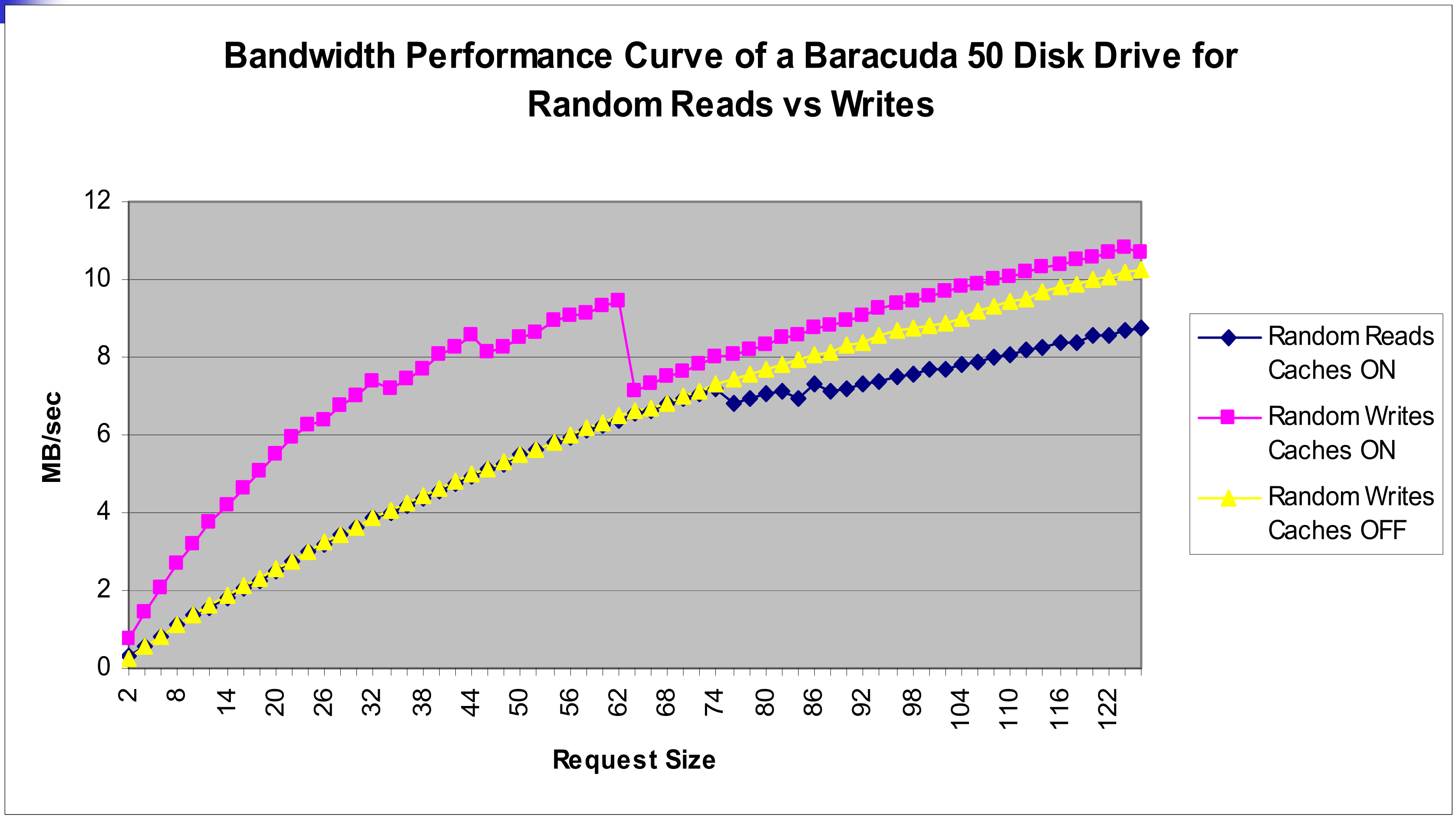
Disk Drive Performance Example – Zoned Bit Recording



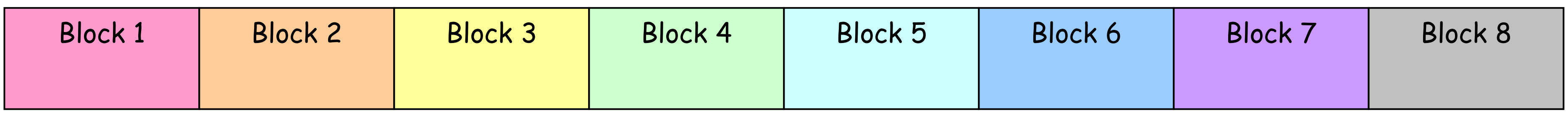
Disk Drive Performance Example – Caching Effects



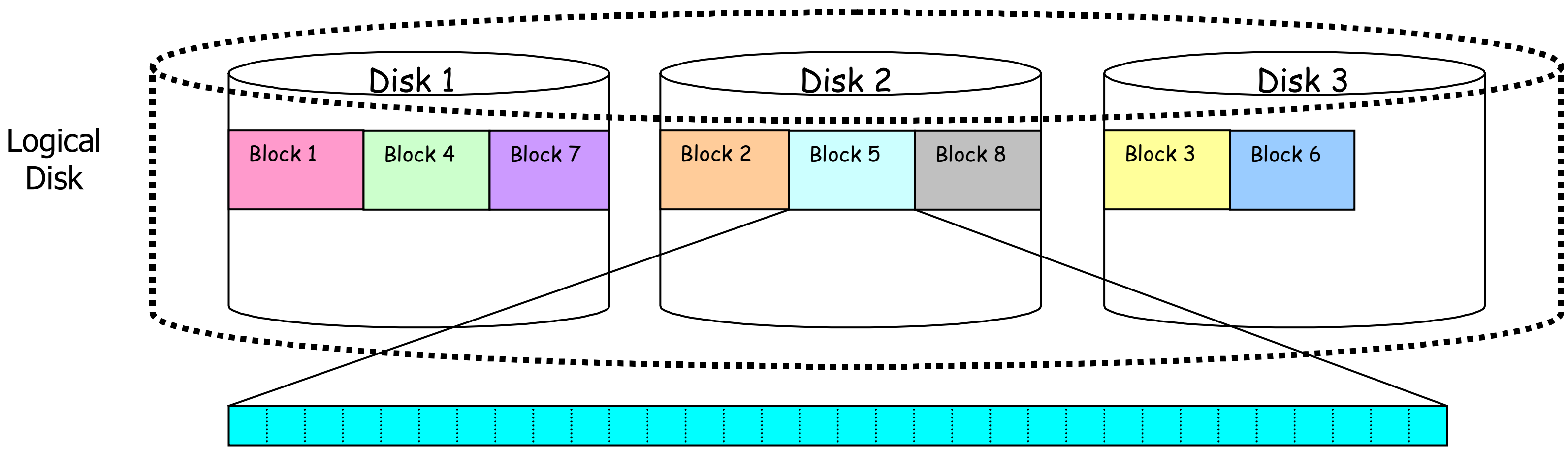
Disk Drive Performance Example – Caching Effects continued



Logical Volume Diagram

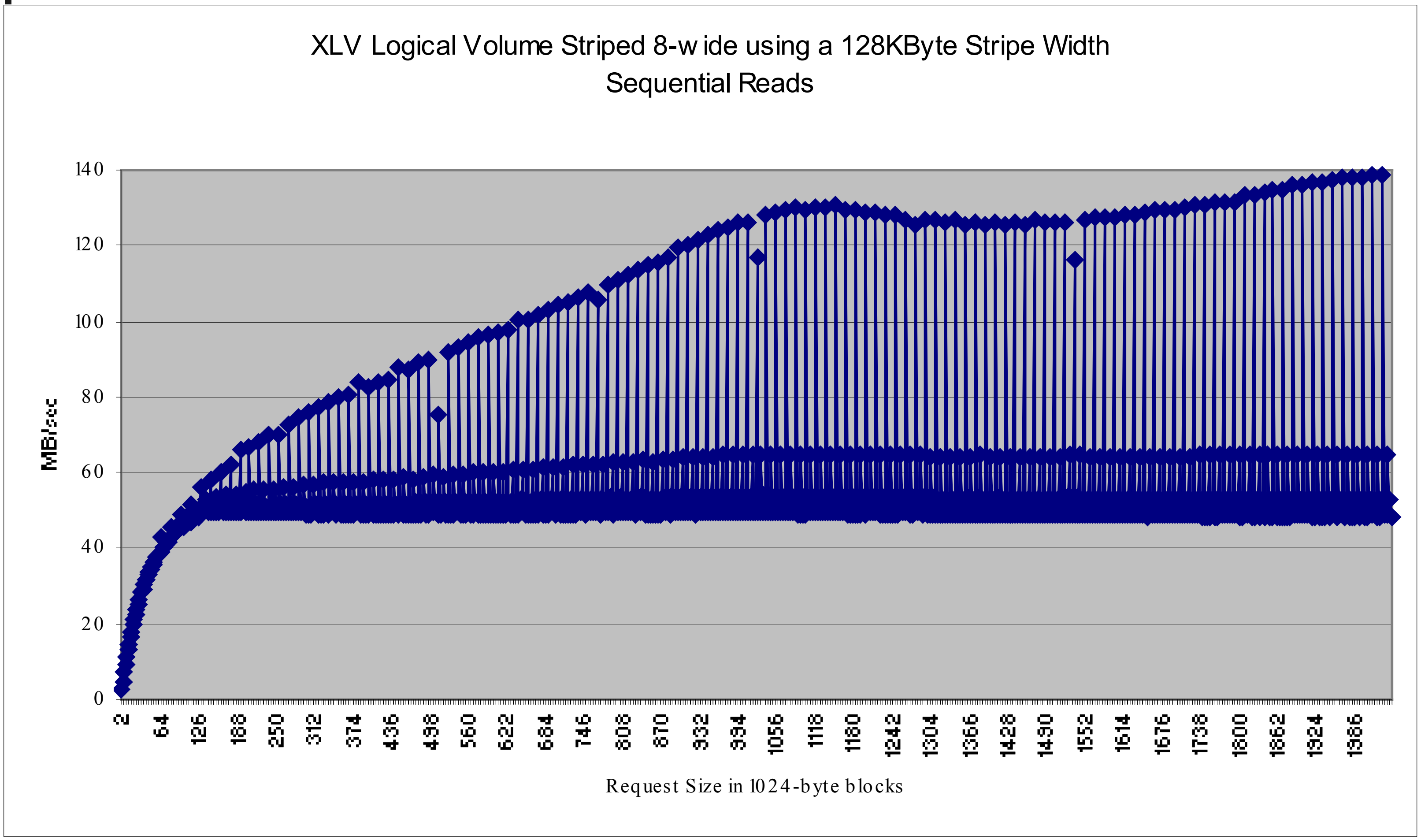


A sequence of 8 consecutive 16384-byte blocks on a “logical” disk. Blocks distributed across the physical disks as shown.

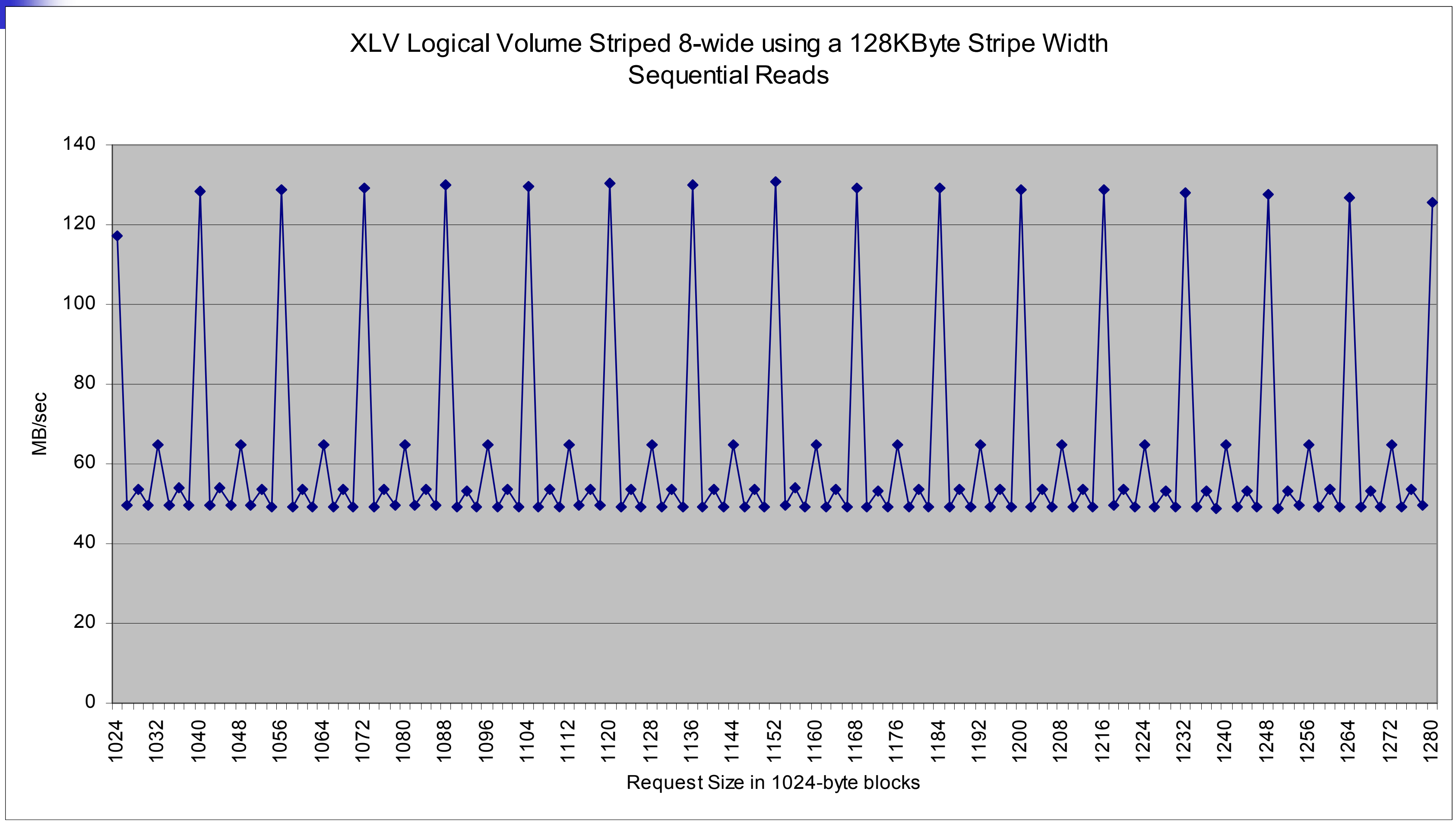


A single 16384-byte “block” consists of 32 consecutive disk sectors, 512-bytes per sector.

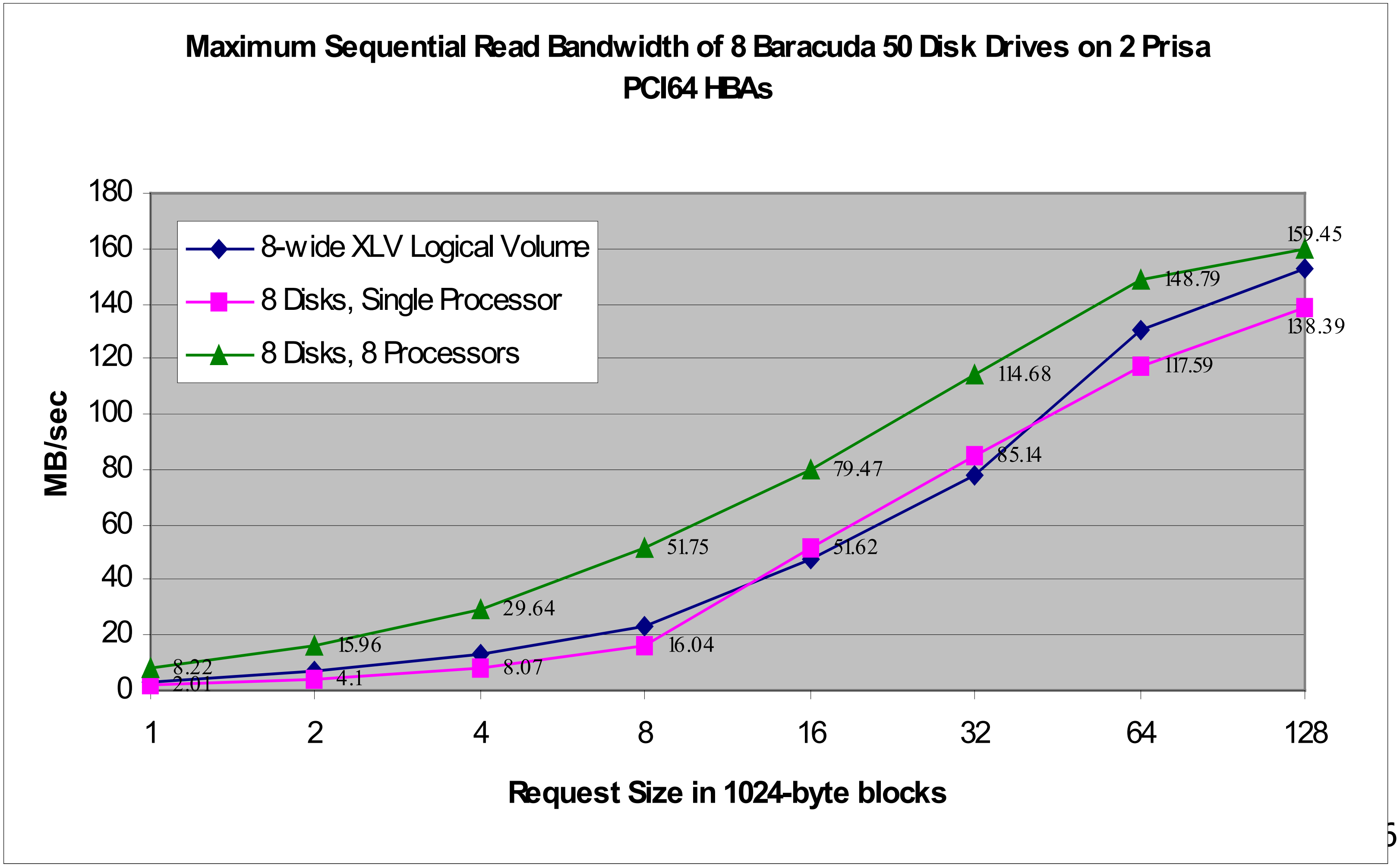
Logical Volume Impedance Matching Example



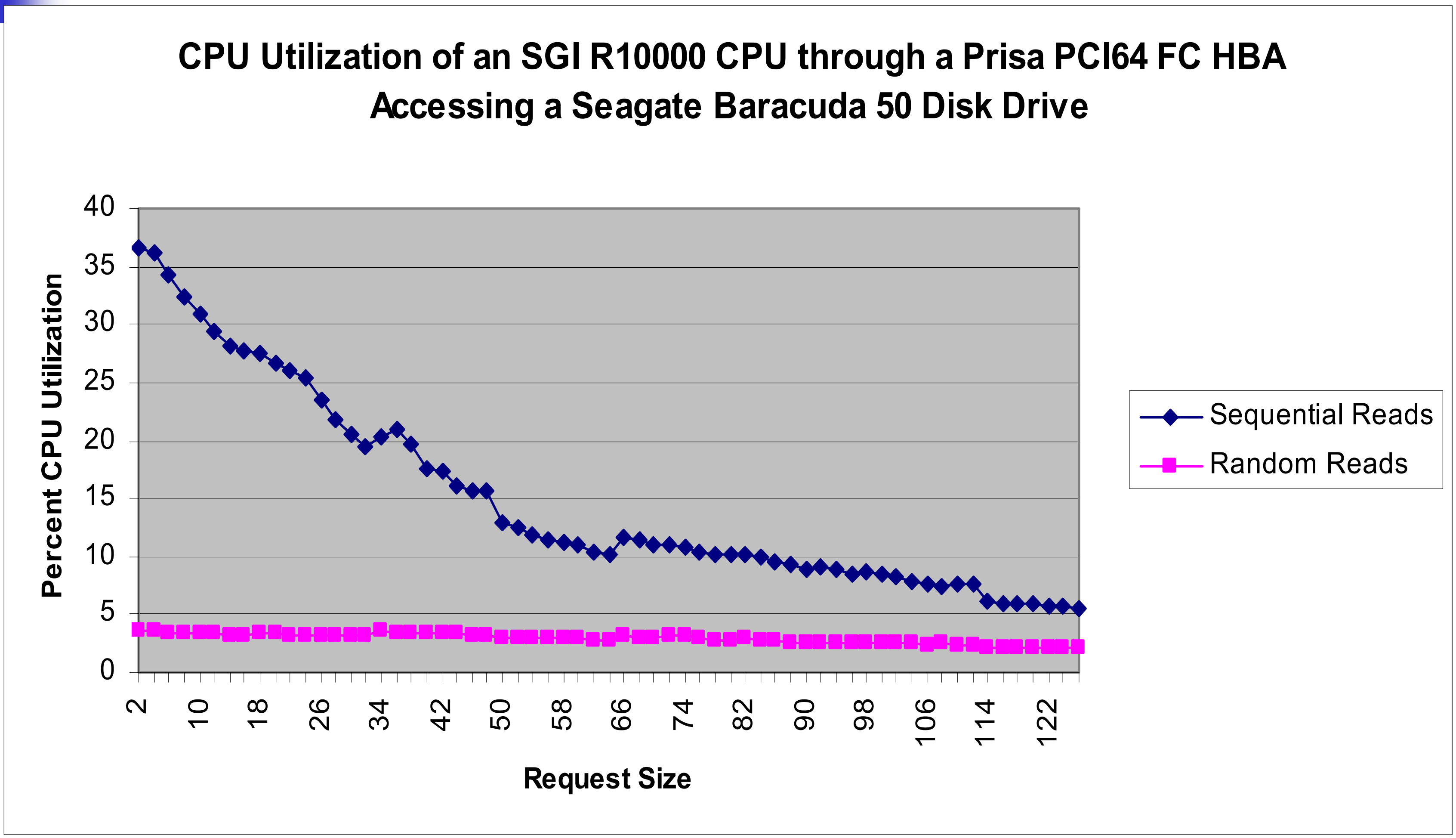
A Closer Look at the Logical Volume



Processor “Impedance” Example

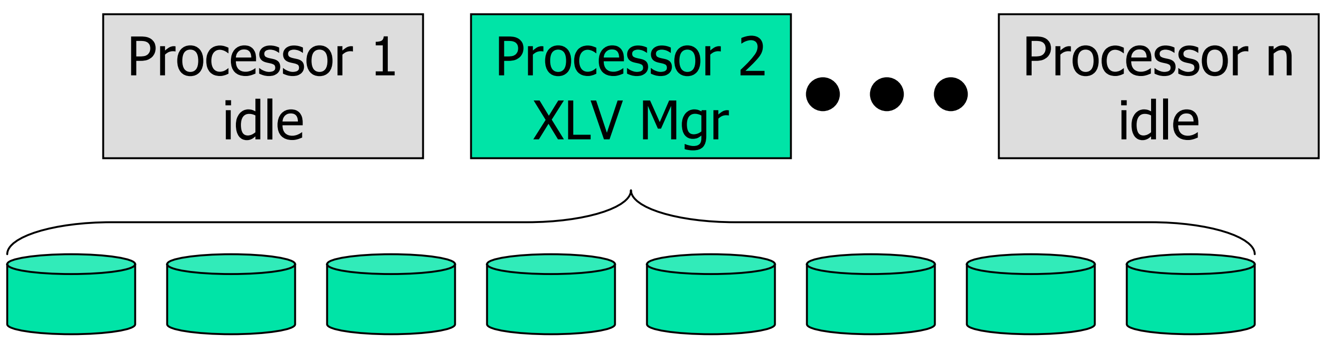


Processor Utilization Example



The Reason for the XLV Problem

Processor 2 must handle all the I/O request processing for the each of the disks in the logical volume and quickly becomes overwhelmed after about 8-10 disks while all the other processors in the system remain essentially idle.



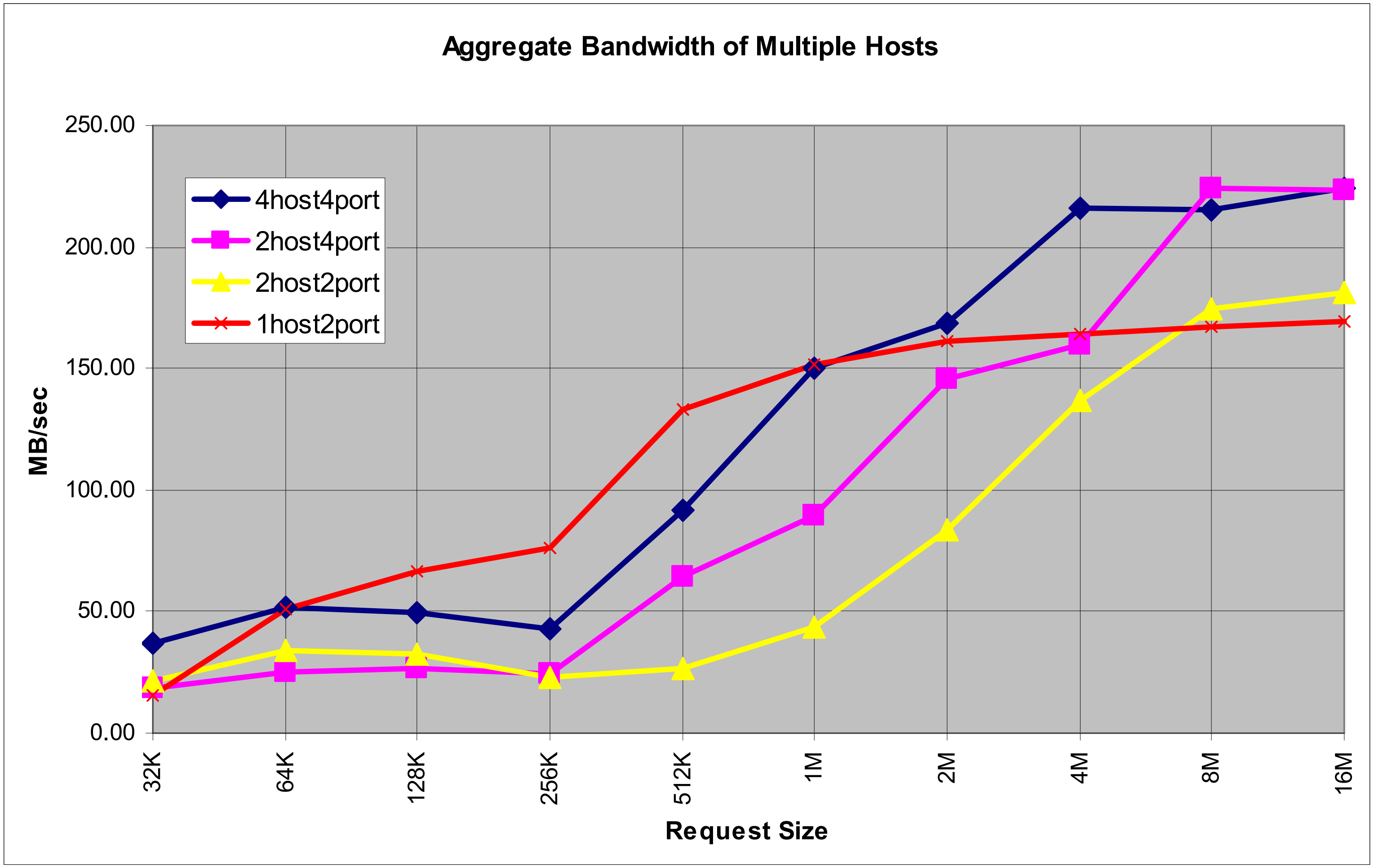
Example of a Shared File System test in a Shared Network Attached Environment

- The Shared File System was configured across
  - 16 Seagate Barracuda 50 Fibre Channel disk drives
  - Four SGI 540 Visual PC Workstations running NT 4.0 SP5
  - Through an Ancor MKII Fibre Channel Switch
  - Using 2-4 Qlogic 2200-based Fibre Channel interfaces per workstation
- The results demonstrate the testing framework as well as some interesting performance anomalies

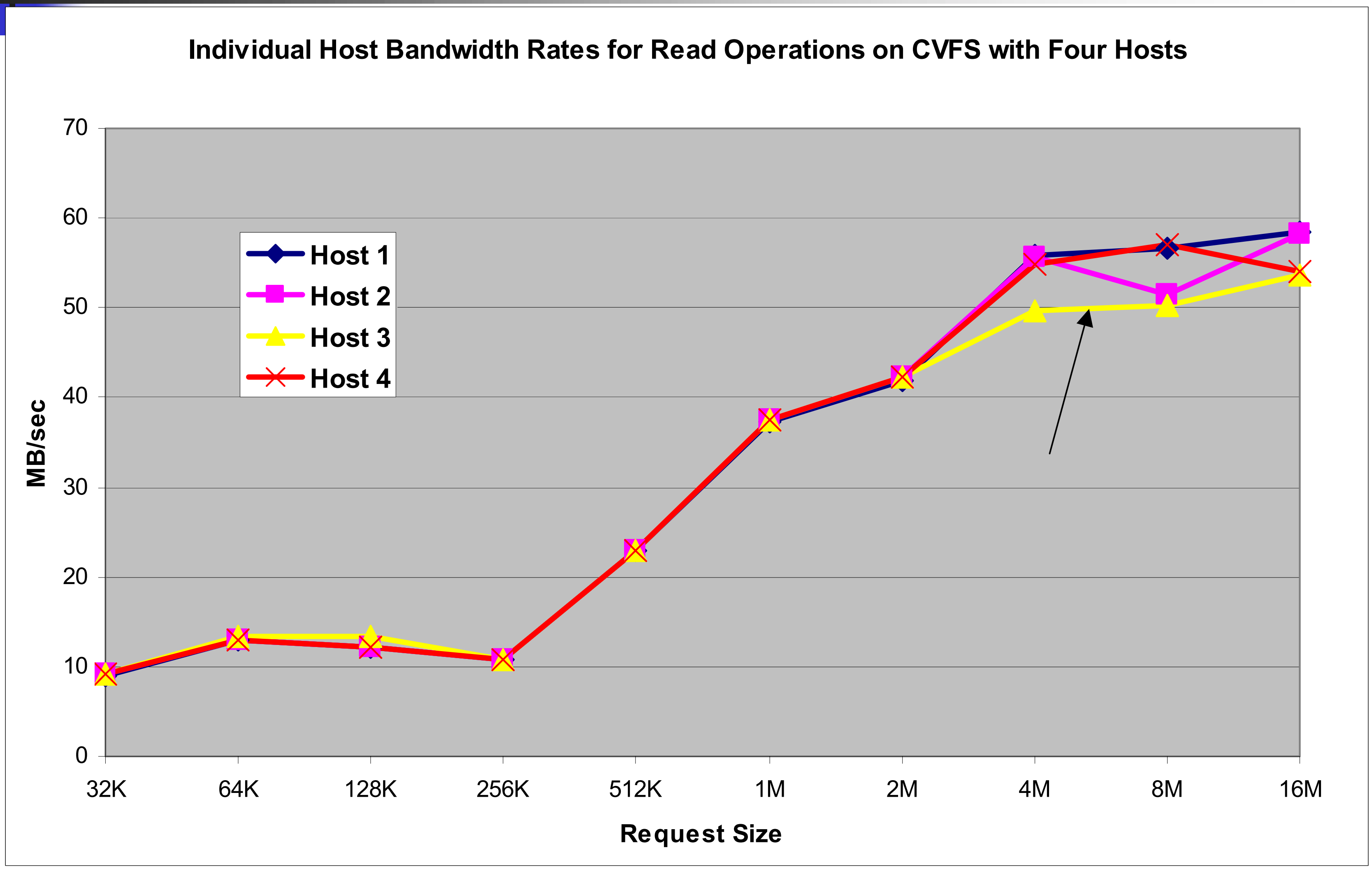
CentraVision File System (CVFS) Read and Write Bandwidth Performance

- Up to 151 MB/sec write performance from a single host
- Up to 170 MB/sec read performance to a single host
- Up to 222 MB/sec read performance across 4 Channels, 2 hosts
- Up to 222 MB/sec read performance across 4 Channels, 4 hosts

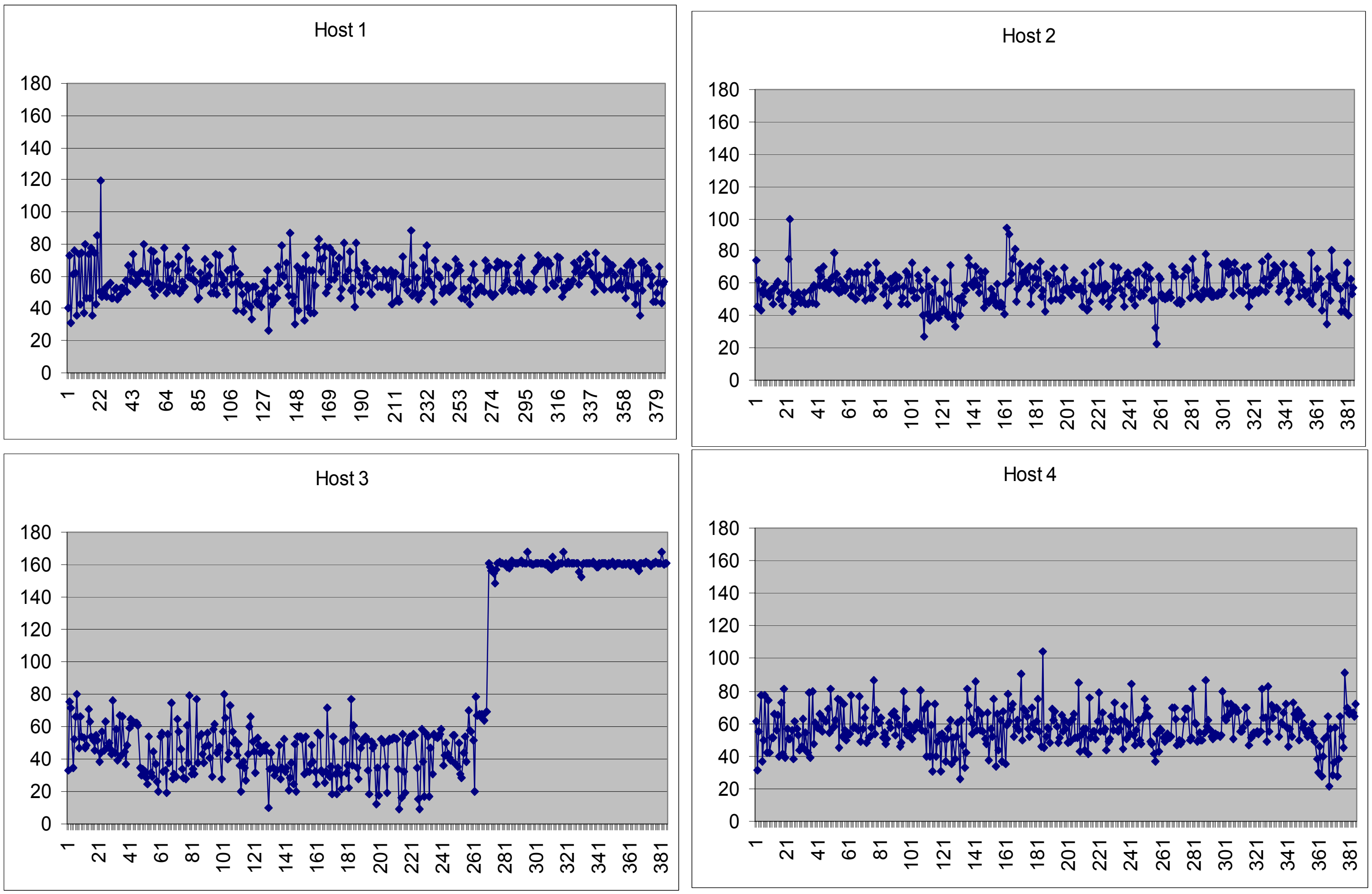
CVFS Bandwidth Distribution



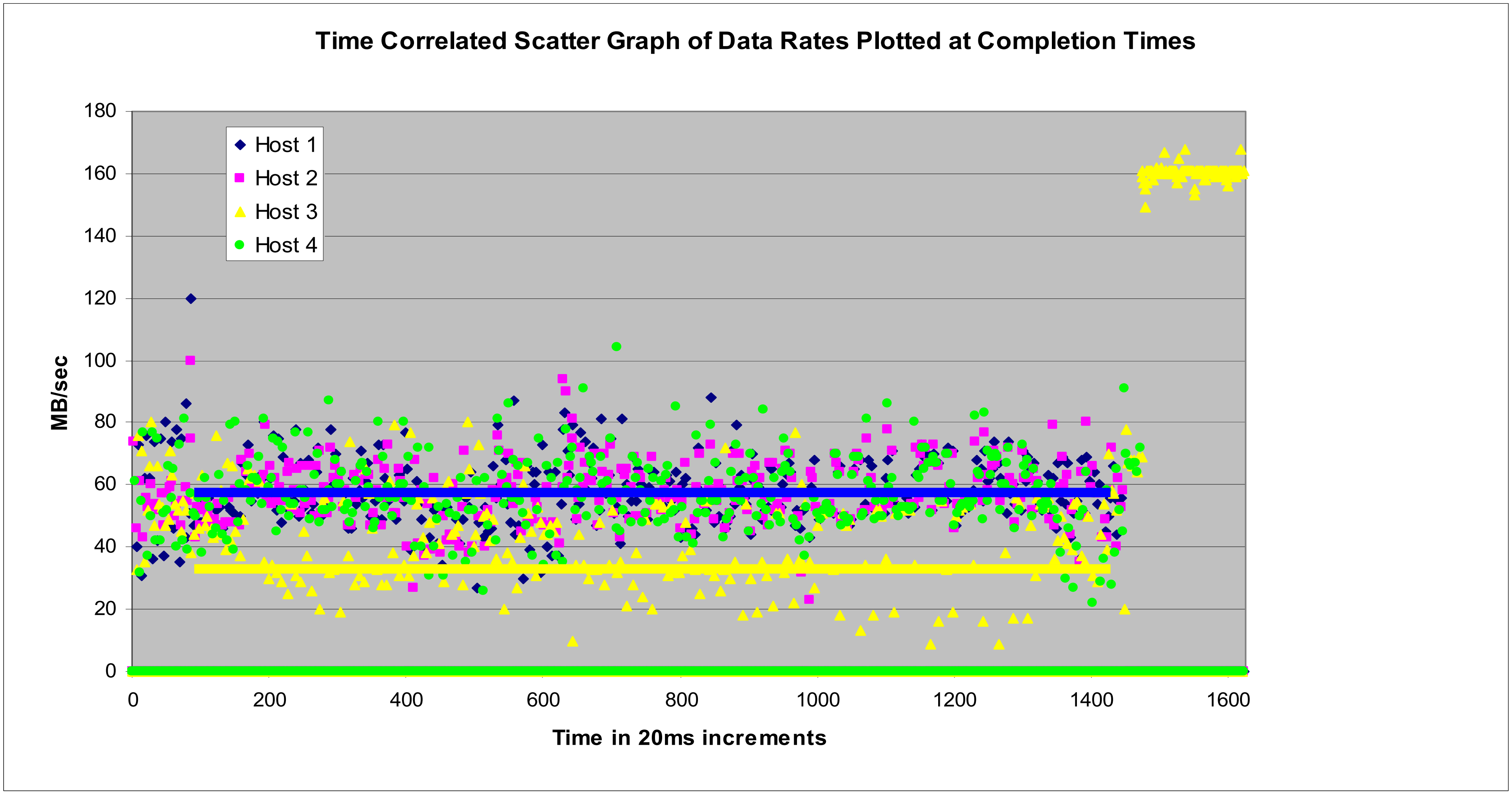
Host Bandwidth Distribution



Individual Host Bandwidth Distorted View



Bandwidth Performance Time Correlated View



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
- OS needs to learn more about SANs and shared disks