File System Benchmarks Then, Now, and Tomorrow

18th IEEE Symposium on Mass Storage Systems and 9th NASA Goddard Conference on Mass Storage Systems and Technologies

April 18, 2001

Presented by

Thomas M. Ruwart

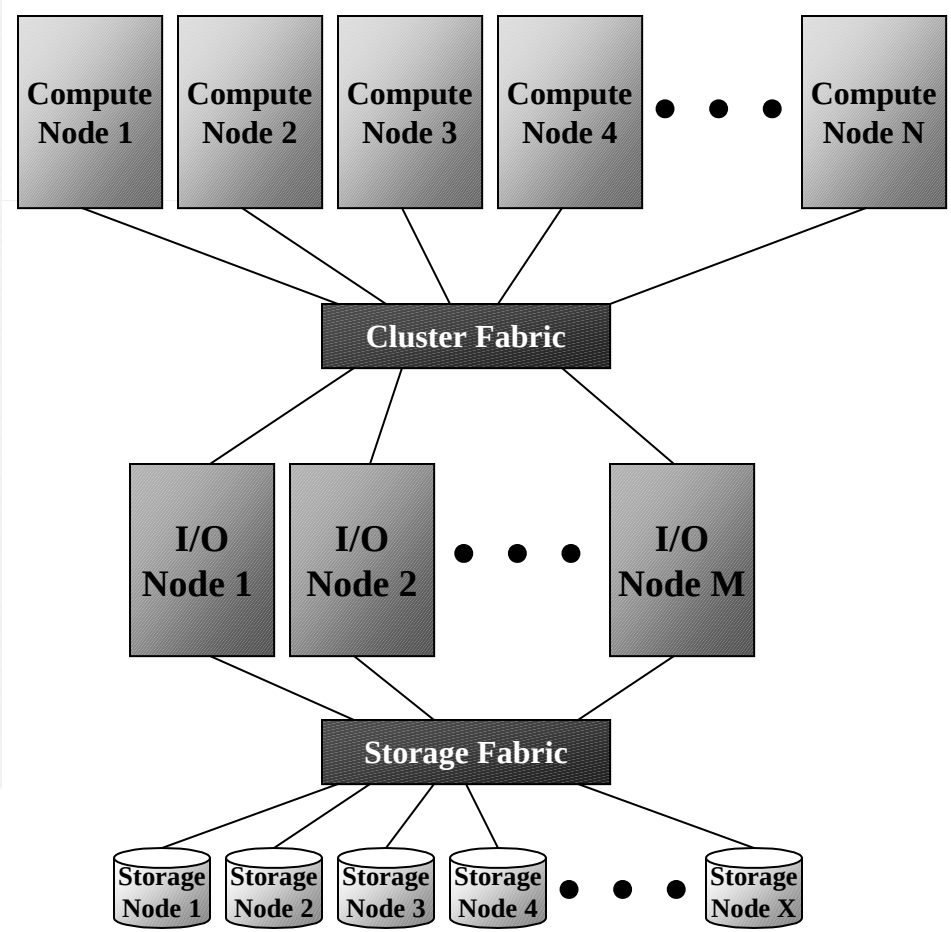
Ciprico, Inc.

Overview

- **If measuring the performance of I/O subsystems was not complicated enough, it is further complicated by SANs and Clusters**
- **SANs and emerging clustering technologies add a *distributed* aspect to the file systems themselves**
- **As the cluster/SAN grows in size, so does the task of performance measurement**
- **The objective of this study is to identify some of the more significant issues involved with file system benchmarking in a highly scalable clustered environment**
- **This research is based on work being done at Los Alamos National Labs on the ASCI 30T machine**

The ASCI 30TeraOp Machine

- **~300 Compute nodes**
- **~64 I/O Nodes**
- **32 processors per node**
- **8 Cluster Fabric connections per compute/I/O node**
- **32 FC connections per I/O node into Storage Fabric – ~2048 2.5Gb FC connections *into* SAN**
- **~700 TB disk storage**



Hardware Infrastructure Issues

- **Compute nodes – How many, how large, and how many connections into the cluster fabric**
- **System Area Network or Cluster Fabric – Latency, bandwidth, and overhead per link**
- **I/O Nodes – How many, how large, and how many connections into the Storage Area Network fabric**
- **Storage Area Network or Storage Fabric – Latency, bandwidth, and overhead per link**
- **Storage Devices – How many, latency, bandwidth, command overhead, number of connections into the SAN per "device"**

System level issues

- **The number of measurement points has increased from one computer system to many computer systems**
- **All the computer systems *share* access to the disk subsystem, or more importantly the *data***
- **Sharing occurs at many levels**
  - **File data**
  - **Metadata (I.e. directories)**
  - **Host bus adapters**
  - **Switches**
  - **Disk controllers**
  - **Disk media**
- **Important to separate the performance of the underlying hardware from the file system software**

Other effects

- **Caching Effects**
  - **Distributed File System data and metadata caching**
  - **Local file system caching**
  - **Device data caching**
  - **Caching policies**
    - **Read versus write**
    - **Temporal (LRU, …etc)**
    - **Data size (I.e. don’t cache large files)**
- **File System Aging effects**
  - **Fragmentation effects on performance**
  - **Monitoring and defragmentation impact on performance**

Benchmarking versus Characterization

- **<u>Benchmarking</u> generally yields a limited set of values that represent the performance of a file system under a specific set of operational parameters**
- **<u>Characterization</u> provides detailed graphs that describe the performance of a file system under a continuum of operational parameters**

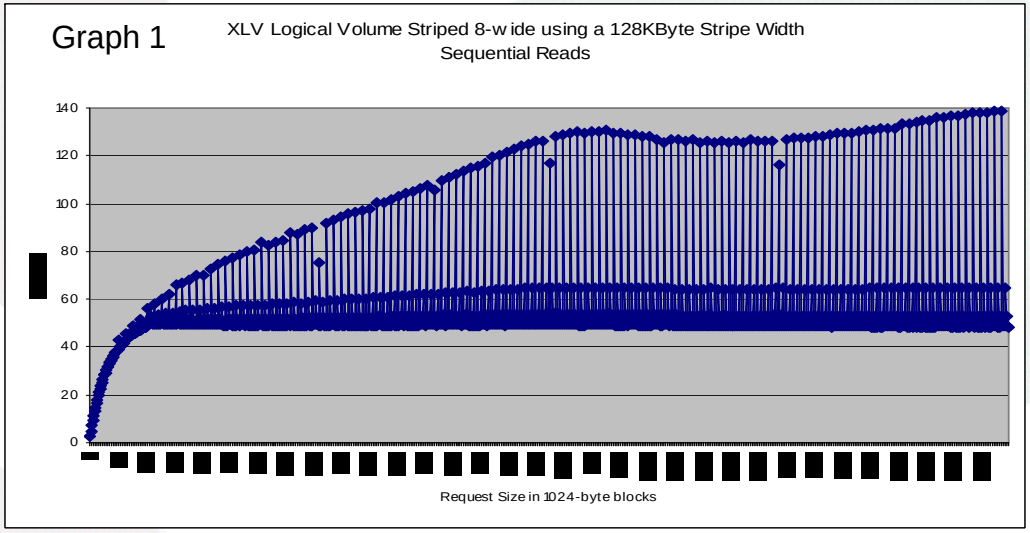
Benchmark Result

120 MB/sec

Or

400 I/O Ops per sec

Characterization Result



Benchmark I/O Permutations

- **Given N Compute Nodes and M I/O nodes there are three permutations of concurrent (parallel) access at the extremes**
  - **1 to M – a single process accessing M compute nodes**
  - **N to 1 – N process threads (across N compute nodes) accessing a single I/O node (file)**
  - **N to M – N compute nodes accessing M I/O nodes**
- **Measure the performance of each of the file system levels in isolation (if possible) and then as a composite of all file system levels**
- **Report results from different *perspectives***

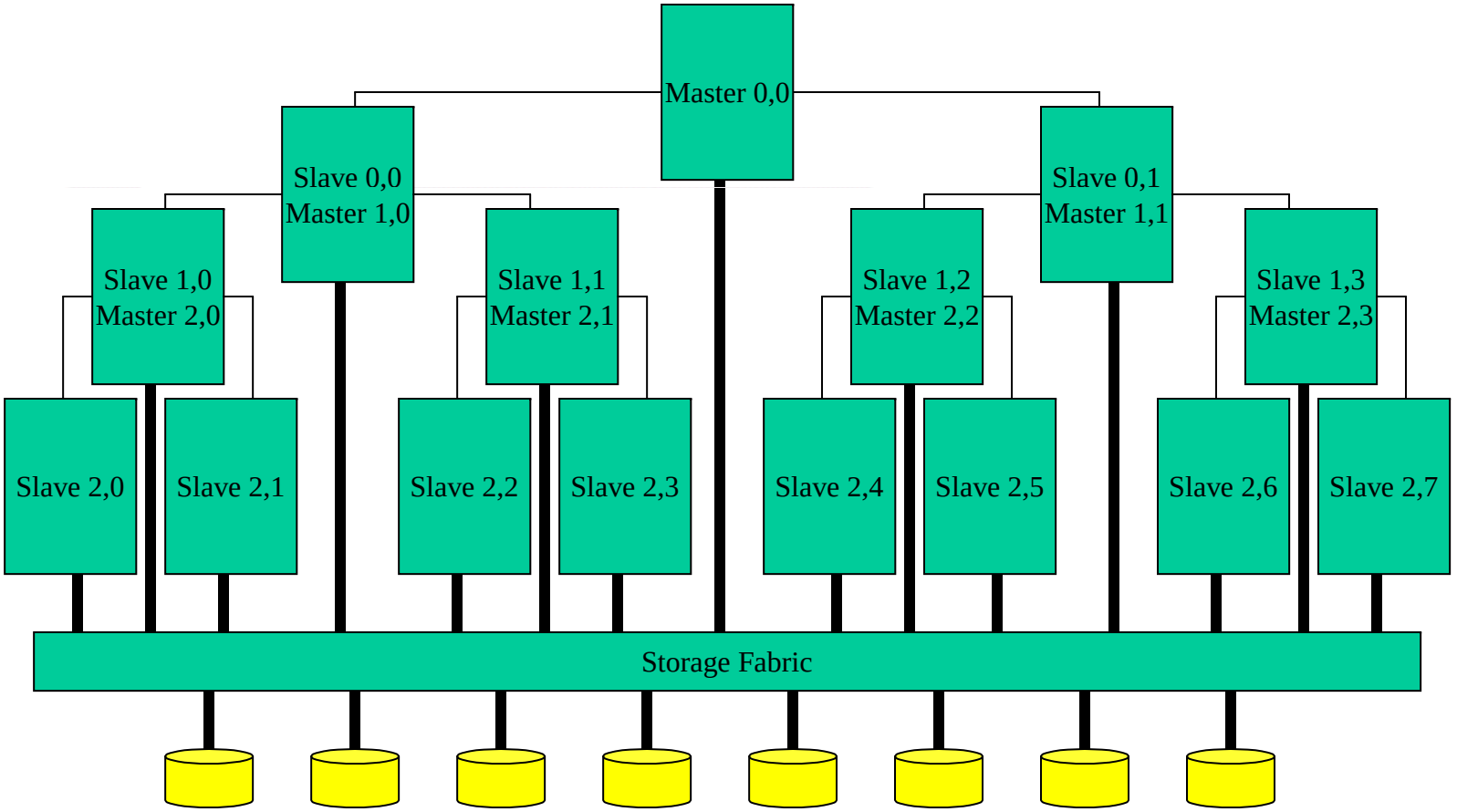
Perspectives

- **Application perspective – Meta Data and User Data**
  - **On a single node**
  - **Distribute application across all compute nodes**
- **System perspective- Composite of all applications**
  - **Single compute node**
  - **Cluster**
  - **File System on a single compute node**
  - **File system distributed across multiple nodes**
- **Device perspective – Composite off all applications and all systems**
  - **Host bus adapter**
  - **Storage Area Network**
  - **Disk Array**
  - **Disk Drive**

Workload generators

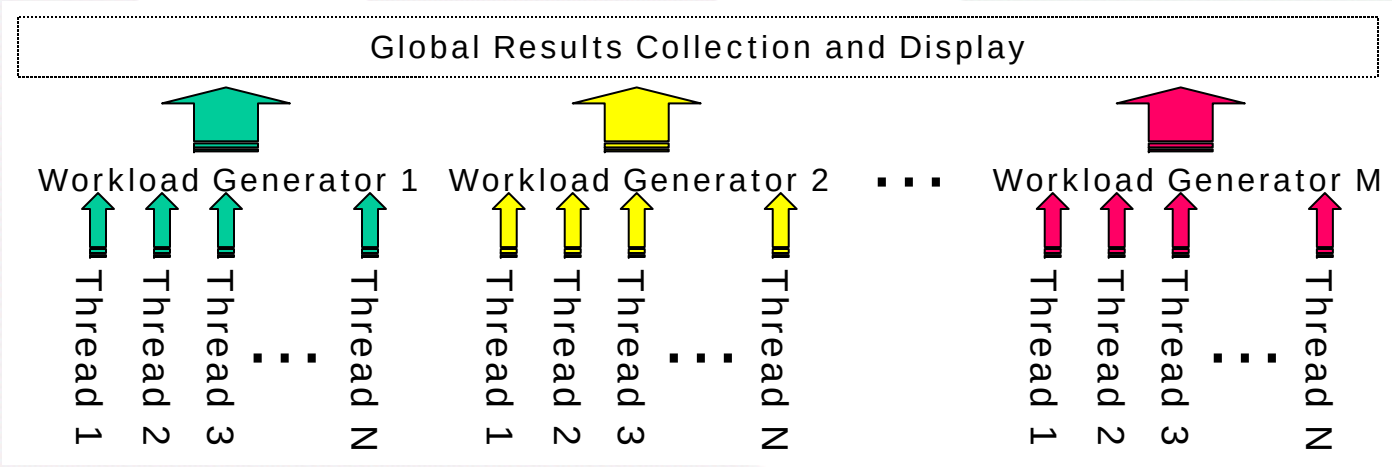
- Control mechanisms
  - Flat versus Hierarchical
- Operational Parameters
  - Request size
  - Number of transfers
  - Spatial access patterns
  - Temporal access patterns
  - Read/write ratios
- Indirect operational parameters
  - Memory allocation
  - Processor allocation
  - Process priority
  - Synchronization
- Synchronized / Unsynchronized workload generators
  - High resolution common reference clock

Benchmark Control Hierarchy



Performance Data Collection

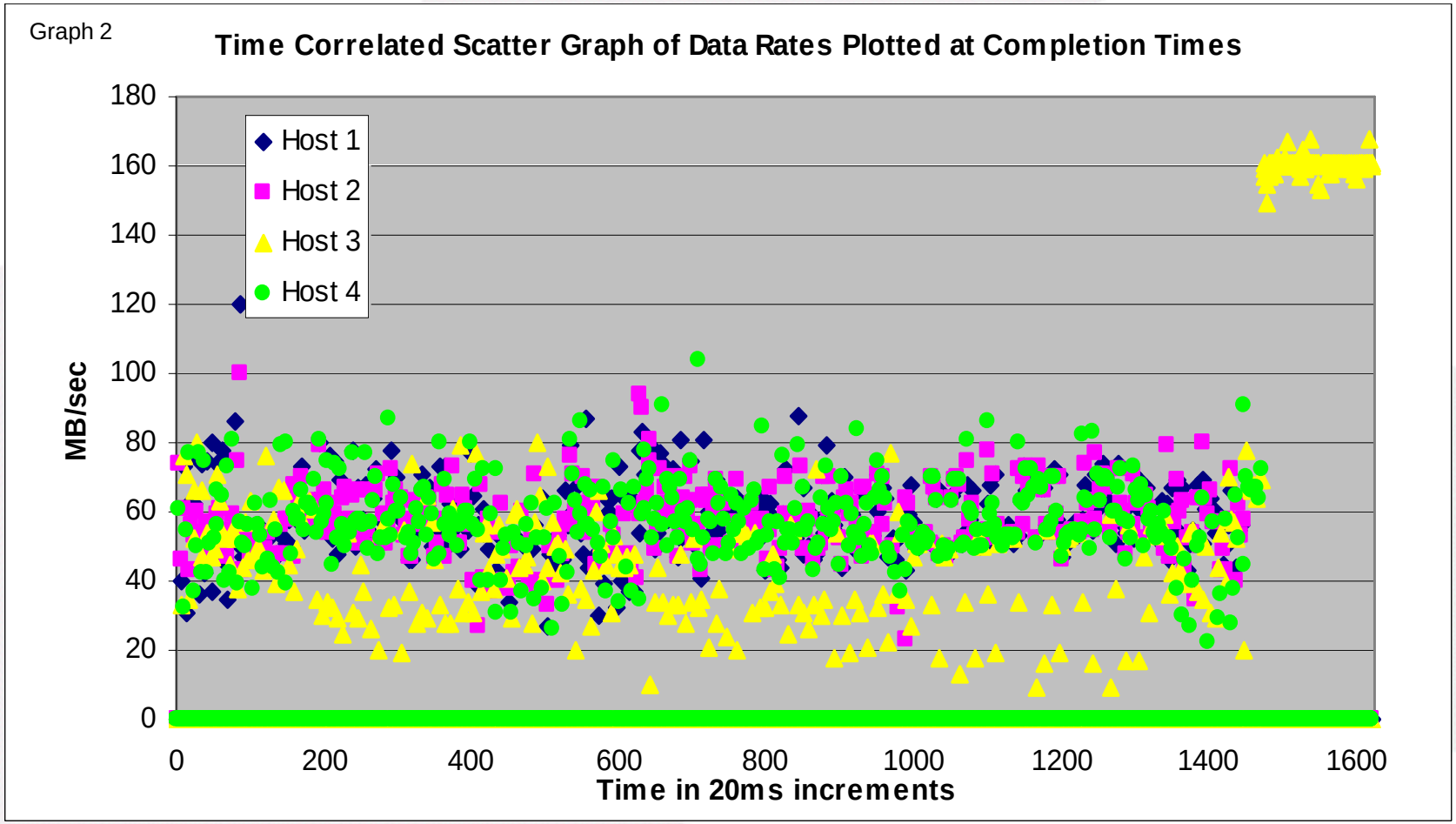
- **A fully deployed I/O benchmark would need to run nearly 10,000 I/O threads, each generating results that need to be collected, condensed, and displayed**
- **The network I/O traffic for collecting the results in real time and/or post mortem is significant**
- **The performance results data collection process cannot interfere with the data transfer for the benchmark**



Run-time Monitoring and Report Generation

- **Detailed reports from**
  - **Each thread – performance of an individual thread**
  - **Each node – aggregate performance of all threads on a node**
  - **entire system – aggregate performance of all threads on all nodes across the system**
- **Real-time (run-time) time-correlated reports – interactive displays and visualization of traffic, performance, and bottlenecks**
- **Trace data analysis tools – Post mortem analysis and visualization**
- **Bottleneck Isolation tools – real-time and post-mortem**
- **Summary reports for “benchmark” purposes**

Example of Time-Correlated Performance Data



Summary

- **The *process* of designing and running an I/O benchmark program that is attempting to**
  - **mimic the behavior of an application or a class of applications,**
  - **interpreting the results**
  - **Provide information that can be used to fine tune the I/O subsystem and/or file system(s)**
- **Provide detailed, real-time system-, application-, and node-wide perspective I/O monitoring capability for identifying performance bottlenecks of the benchmark**
- **Tight management of the variables that influence the I/O performance during a benchmark run**

CIPRICO

Protecting your image