

SEMPAR: High-Performance Remote Parallel I/O over SRB

Nawab Ali and Mario Lauria

Department of Computer Science and Engineering

The Ohio State University

Columbus, OH 43210

{alin, lauria}@cse.ohio-state.edu

- Introduction
- Remote I/O
 - Staging
 - High-Performance Parallel Remote I/O
- Storage Resource Broker
- SEMPLAR
 - Design
 - Implementation
- Experimental Setup
- Results
- Conclusion

- **Application Trends**

- Big Science projects increasingly require processing of large data sets
 - Sloan Digital Sky Survey, Large Hadron Collider, National Earthquake Engineering Simulation Grid, etc
- Large data sets stored in repositories at specialized facilities (supercomputer centers)

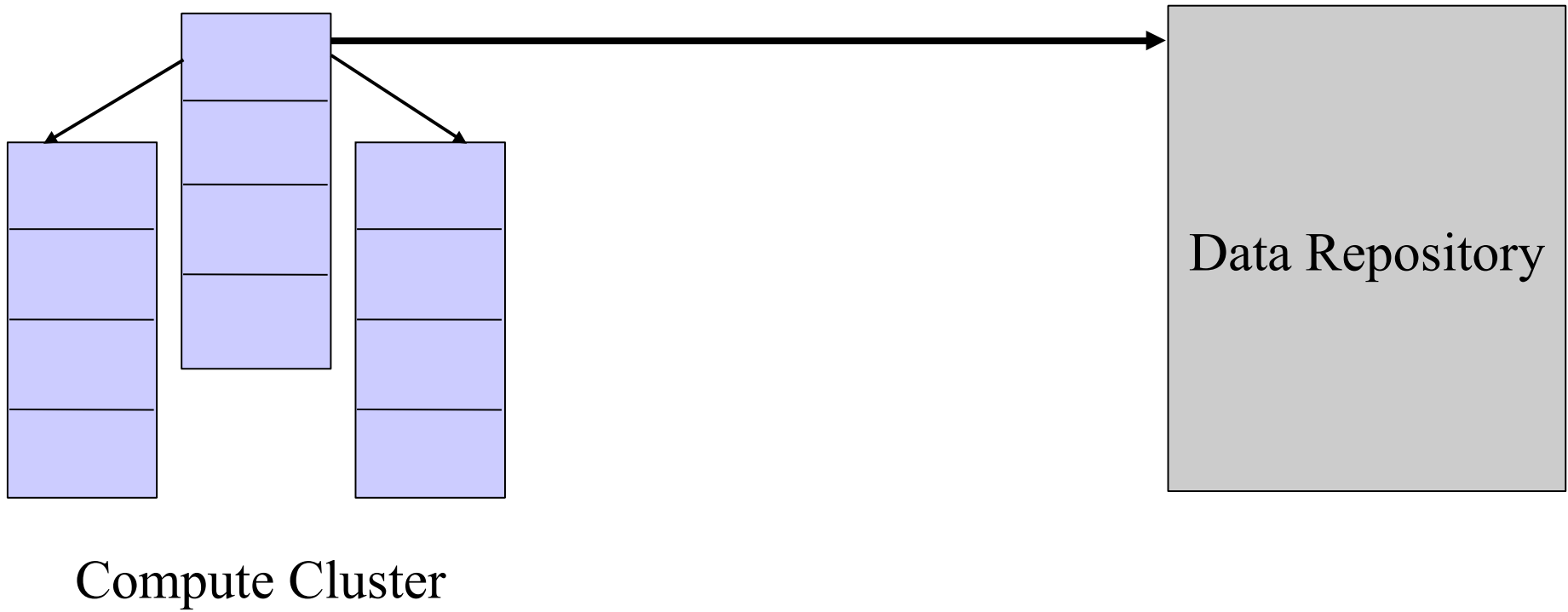
- **Technological Trends**

- Bandwidth of WAN and Internet backbones growing at a rate that makes it comparable to local interconnect speed
 - TeraGrid, LambdaRail ~40Gb/s
 - Infiniband ~ 10 Gb/s

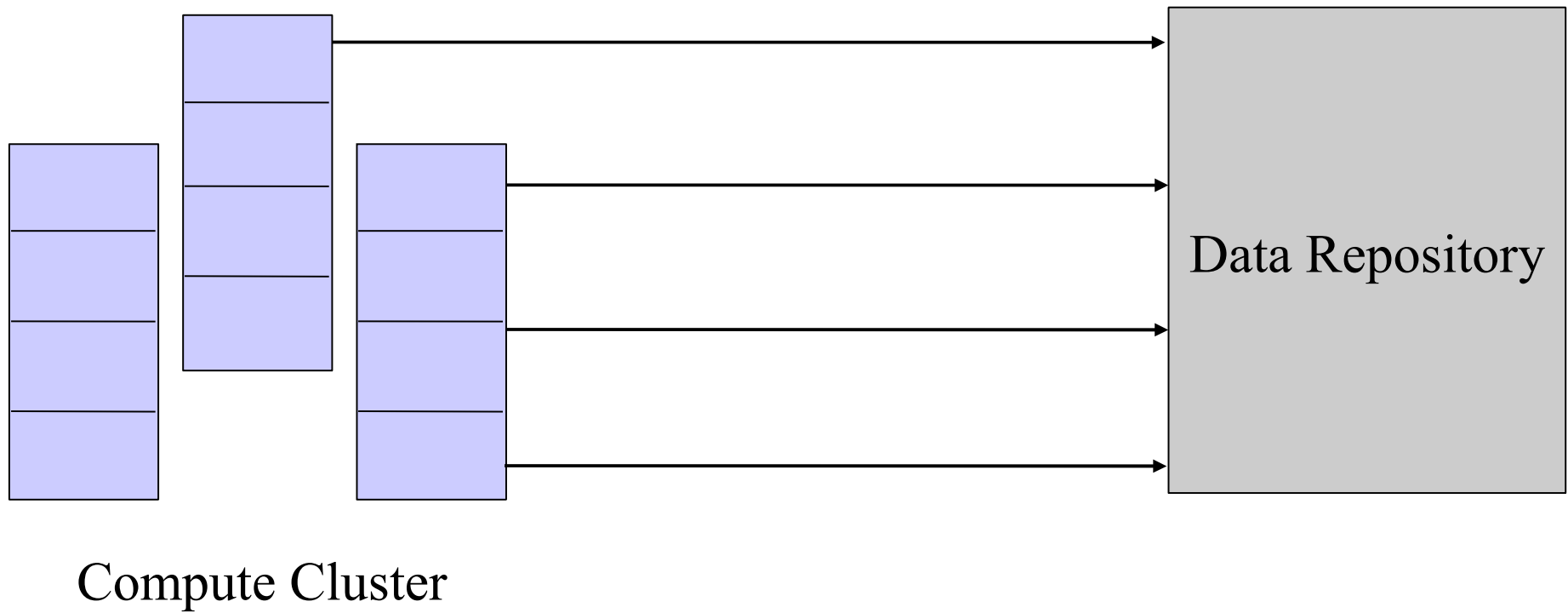
- Common approach to processing large data sets
 - Local staging vs. Direct remote access
- Problems with staging
 - Detailed knowledge about the remote filesystem
 - Overlapping of data transfer and computation not possible
 - Dynamic data sets require frequent refreshes of the local copy
- Previous attempts at building remote data access tools (RIO, GridFTP) based on incremental improvement of legacy tools haven't proven competitive with local staging

Computing on the Grid

- The notion of a well integrated compute/data Grid really hinges on the availability of high performance access to remote data sets
- Our approach is to identify current performance bottlenecks and then build a solution around them



High-Performance Parallel Remote I/O

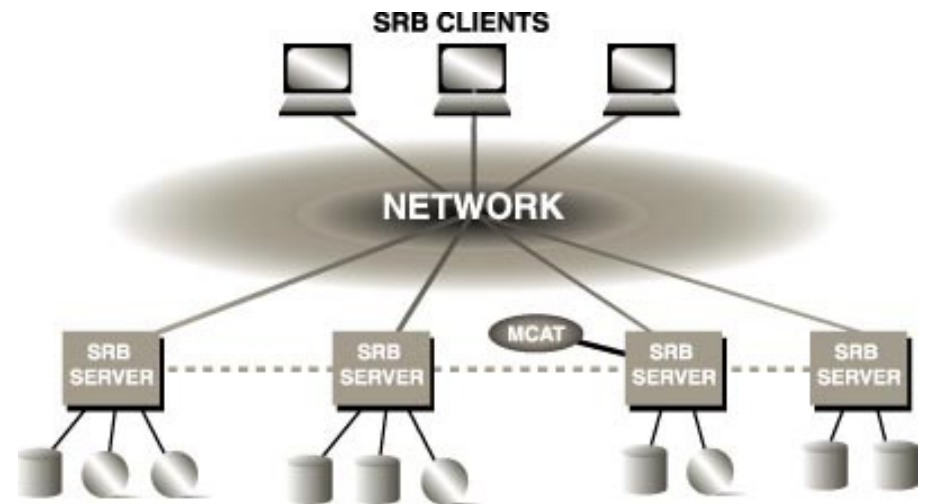


Storage Resource Broker

- SRB was developed at SDSC to provide remote I/O to massive volumes of data in a production environment
- It provides transparent access to heterogeneous storage resources
 - Filesystems
 - Database Systems
 - Archival Storage Systems
- Other services offered
 - Authentication, location transparency

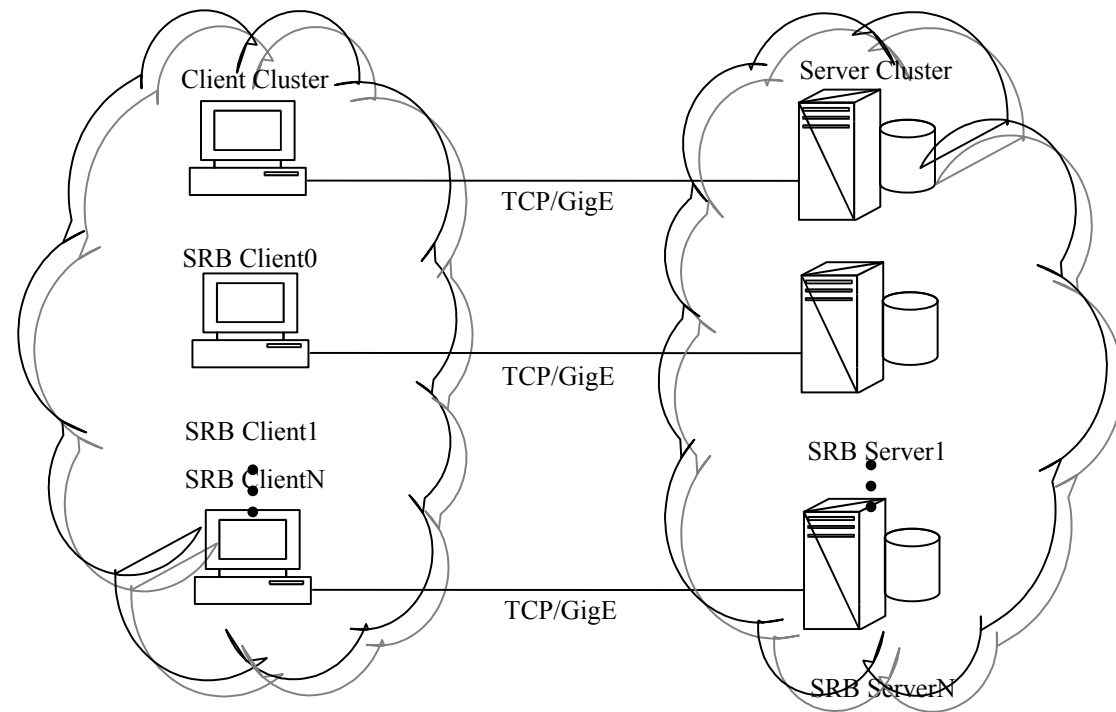
SRB Architecture

- SRB Servers
 - Control distinct set of physical resources
- Metadata Catalog Service
 - Stores file metadata
 - Access Control
 - File location
- SRB Clients
 - Connect to the SRB servers using client API
 - C high-level API
 - C low-level API



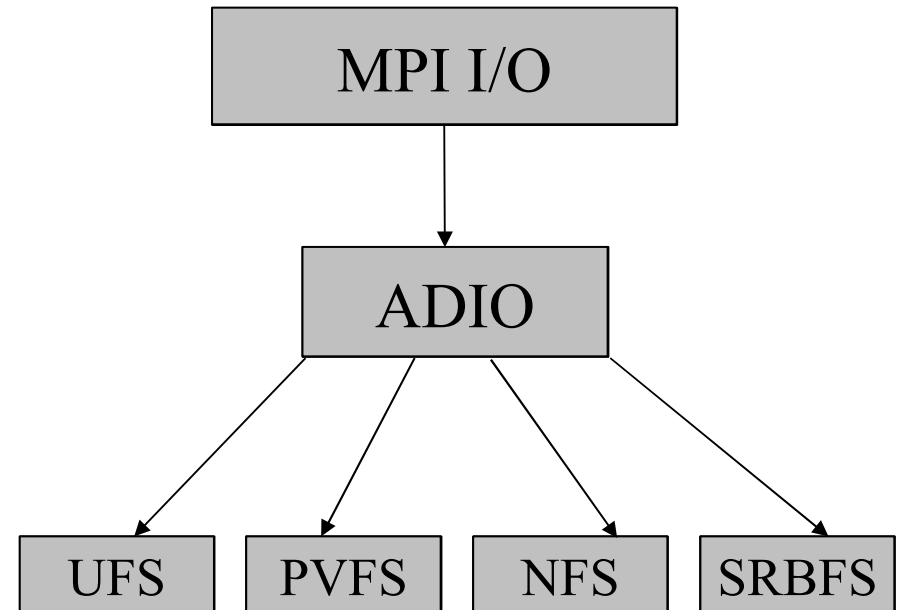
SEMPLAR: SRB Enabled MPI I/O Library for Access to Remote Storage

- I/O over the Internet
- Storage Virtualization
 - SRB
- High I/O bandwidth
 - Multiple TCP Streams
 - Multiple I/O nodes
- Standard Application Interface
 - MPI I/O



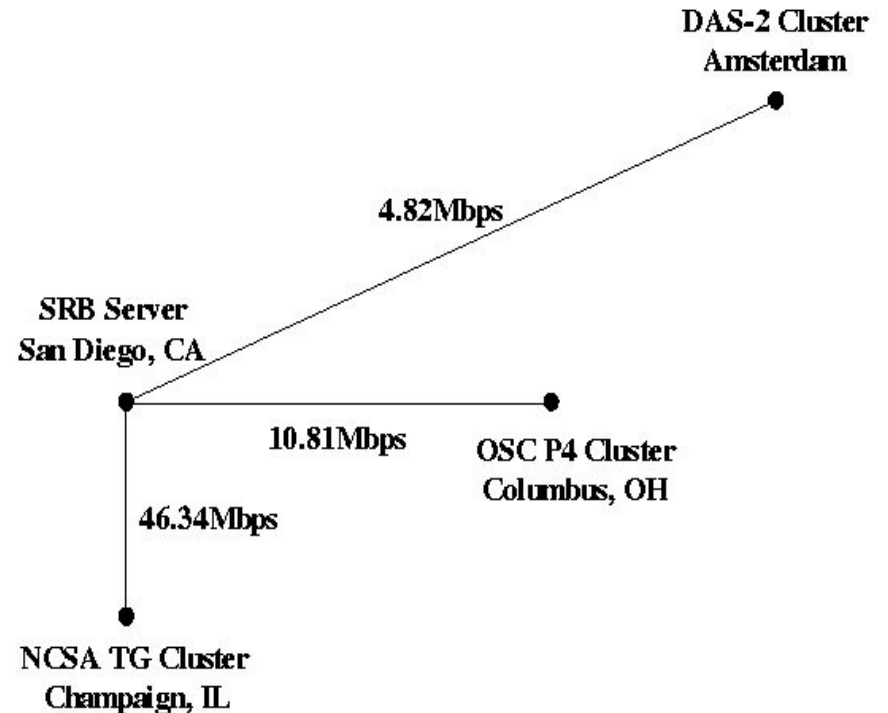
SEMPLAR Implementation

- MPI I/O implementations such as ROMIO use the portable ADIO interface
- ADIO implementations are optimized for a particular filesystem
- We have provided an ADIO implementation for the SRB filesystem



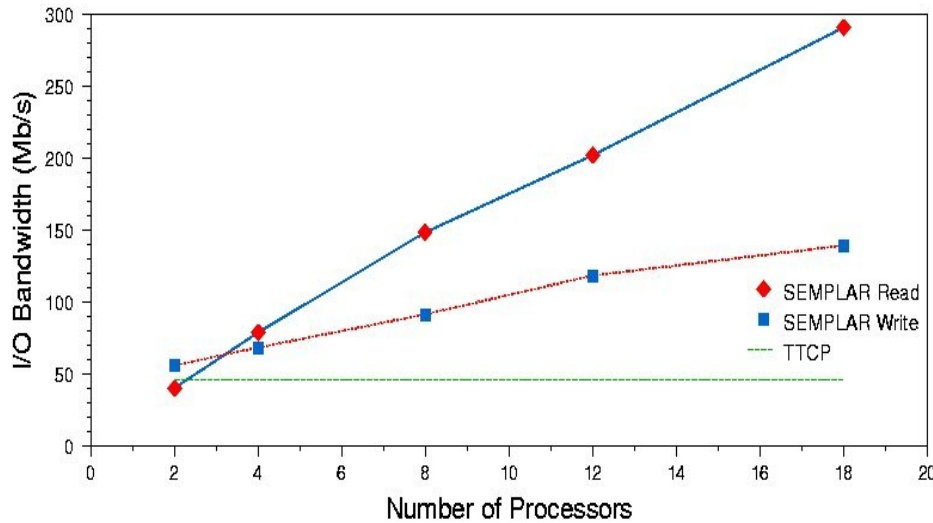
Experimental Setup

- SRB server v3.2.1 running on orion.sdsc.edu
- NCSA TeraGrid cluster
 - High bandwidth, Low latency
- DAS - 2
 - Low bandwidth, High Latency
- Intel Pentium 4 Xeon cluster at OSC
 - High bandwidth, Low latency
 - Private I/P addresses

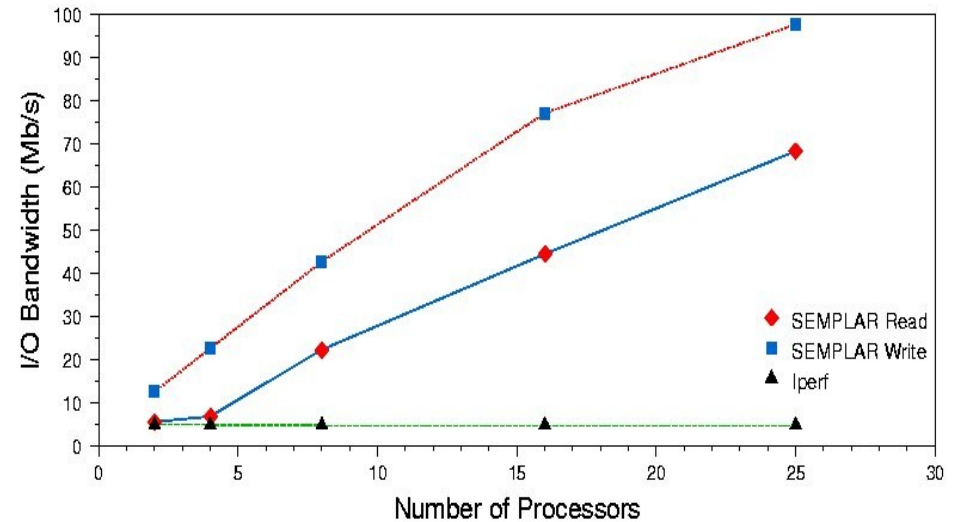


- ROMIO perf
 - Measures the read and write performance
 - Synchronous and Contiguous I/O
 - Upper-bound on the MPI I/O performance
- NAS btio
 - Non-contiguous data access pattern
 - Class C full version
 - Collective I/O
- Dynamic Fault Model
 - Earthquake simulation
 - Collective MPI I/O

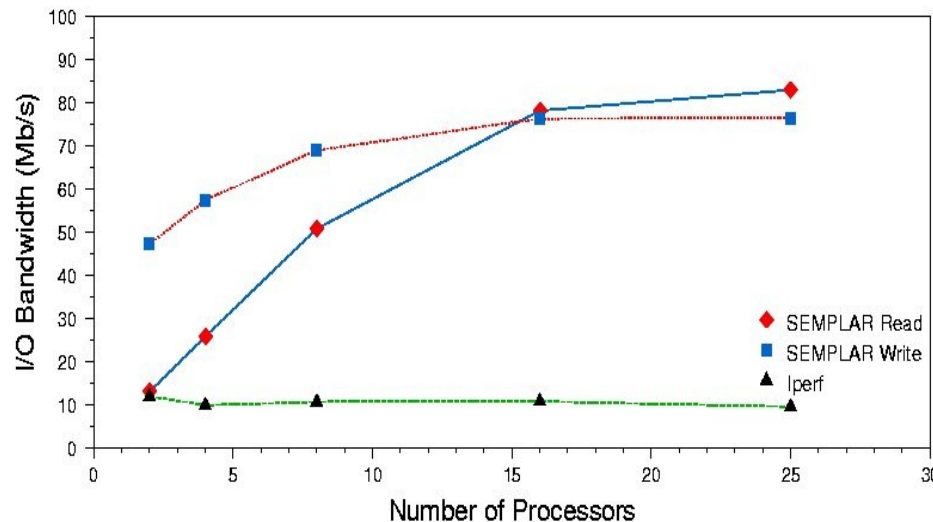
perf I/O Performance



NCSA TeraGrid Cluster



DAS-2 Cluster



OSC P4 Cluster

NCSA TeraGrid

Read: 290.88Mbps. Write: 139.44Mbps. Ttcp: 46.34Mbps

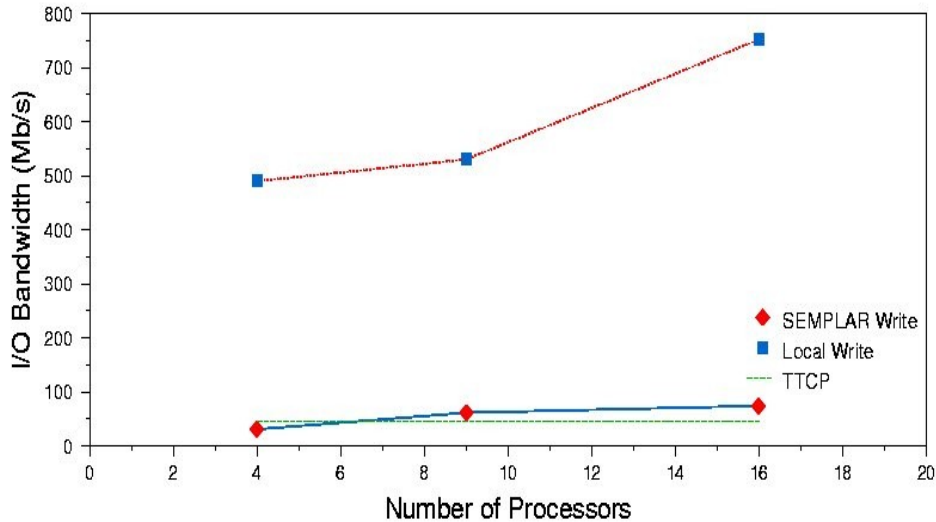
DAS - 2 Cluster

Read: 68.32Mbps. Write: 97.68Mbps. Iperf: 4.82Mbps

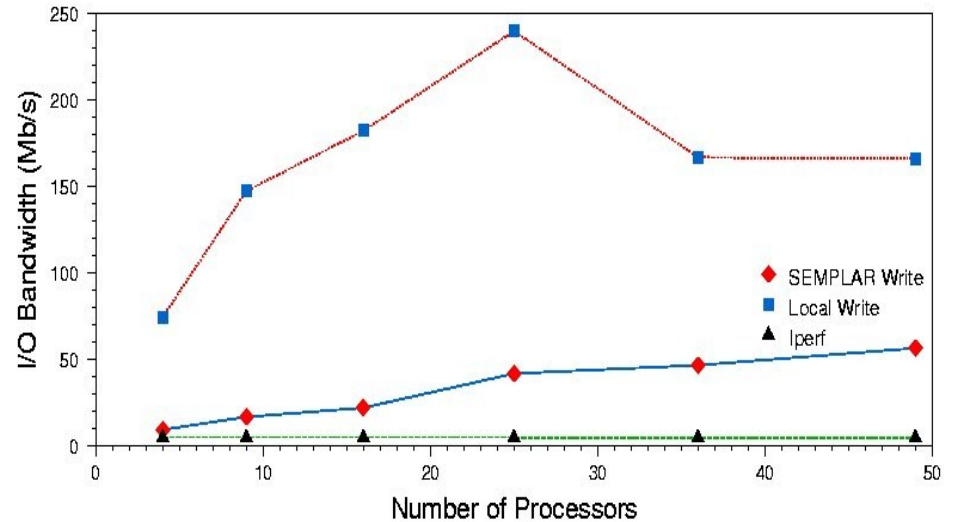
OSC Xeon Cluster

Read: 82.96Mbps. Write: 76.48Mbps. Iperf: 10.81Mbps

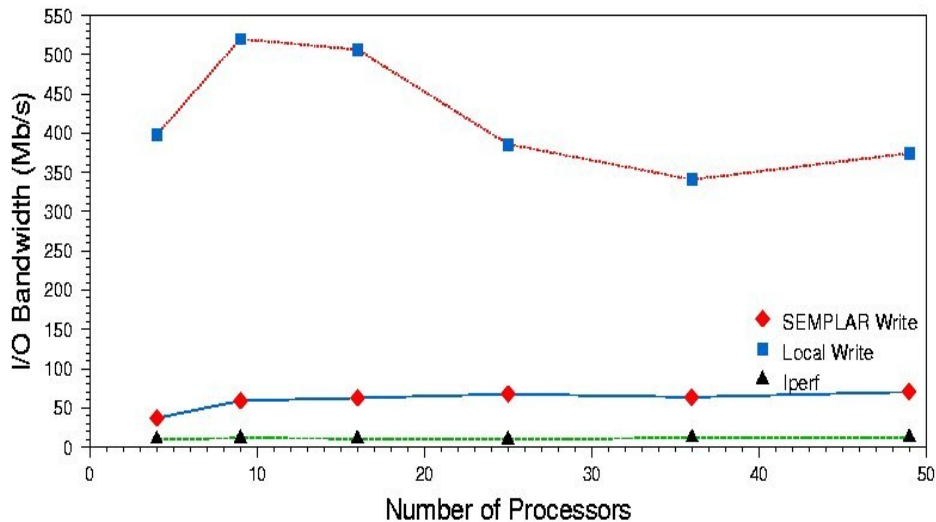
btio Class C Write Performance



NCSA TeraGrid Cluster



DAS-2 Cluster



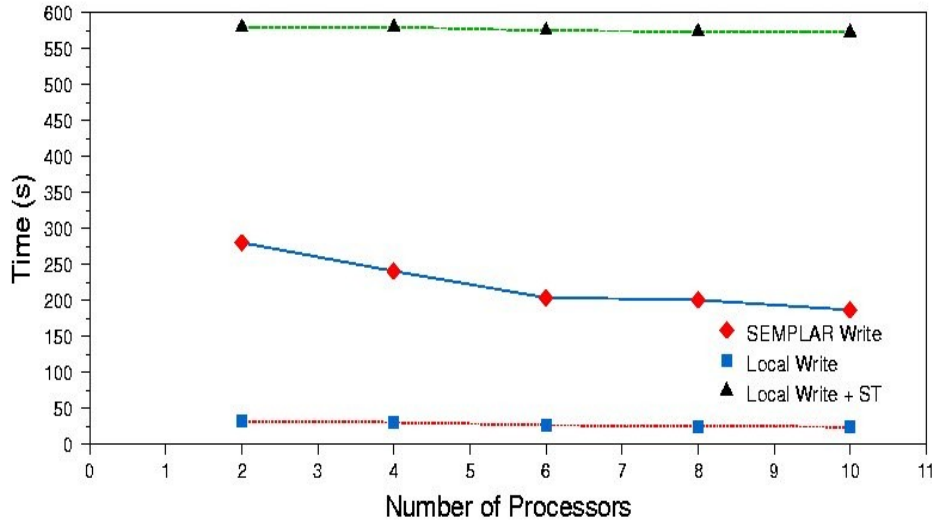
OSC P4 Cluster

NCSA TeraGrid
 btio Class C Write: 74.04Mbps. Ttcp: 46.34Mbps

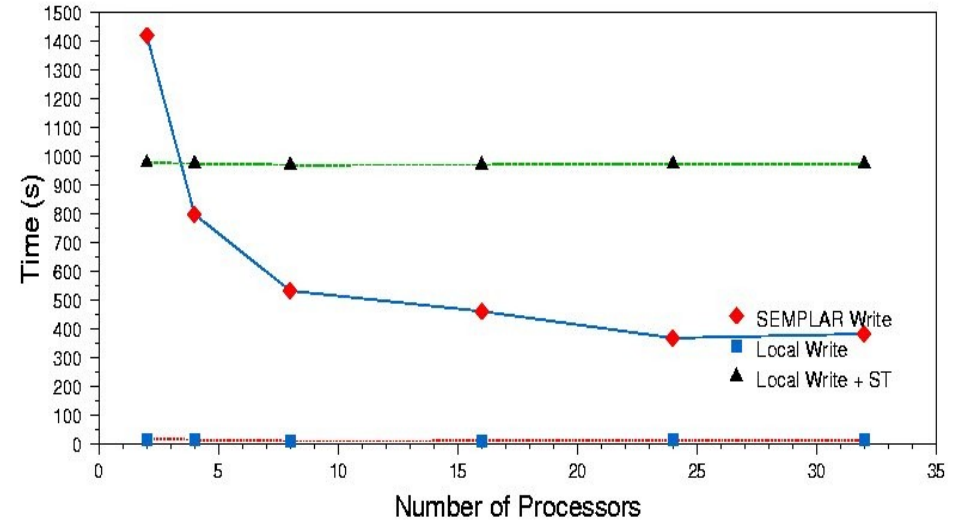
DAS – 2 Cluster
 btio Class C Write: 56.49Mbps. Iperf: 4.82Mbps

OSC Xeon Cluster
 btio Class C Write: 70.28Mbps. Iperf: 10.81Mbps

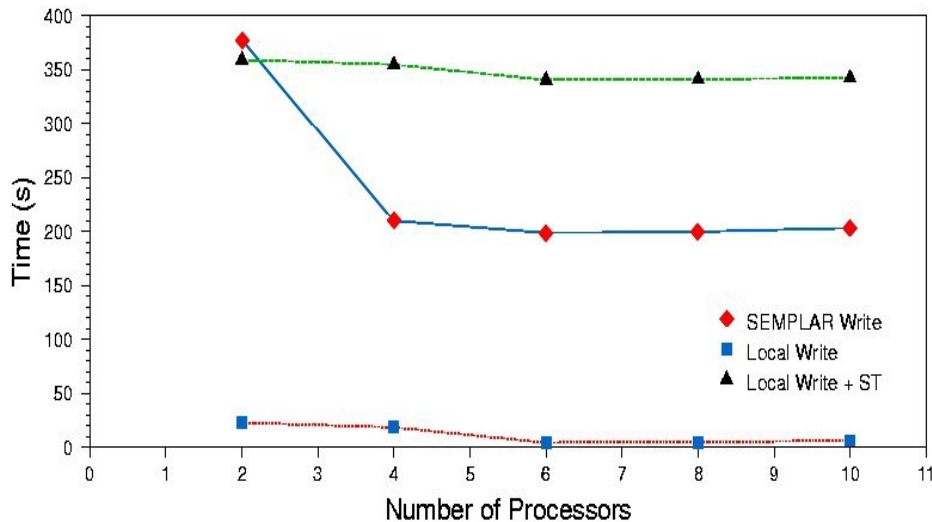
DFM I/O Performance



NCSA TeraGrid Cluster



DAS-2 Cluster



OSC P4 Cluster

<u>NCSA TeraGrid</u>	
SRB: 222.07s	Local + ST: 575.92s
<u>DAS - 2 Cluster</u>	
SRB: 660.48s	Local + ST: 973.08s
<u>OSC Xeon Cluster</u>	
SRB: 237.47s	Local + ST: 354.08s

Results Summary

- MPI-IO + remote I/O
 - RIO: single client-server connection
- Client side enhancements
 - GASS: aggressive client-side caching
 - BMT: write behind approach to remote I/O
- Multiple connections
 - DPSS – striping of data across multiple remote servers
 - GridFTP – striped connections out of a single client

- High-Performance remote I/O is an active area of research
- We have used multiple, parallel TCP streams to increase the available data bandwidth
- SRB provides a consistent interface to heterogeneous storage resources
- By integrating SRB with MPI I/O, we have developed a scalable, high-performance remote I/O library based on widely deployed tools

- Client-side Caching
 - Local Access Performance
 - Ability to access data remotely
- Asynchronous Parallel I/O
- Dynamic degree of parallelism
 - Adjust the number of connections based on the network load

Thank You