



Benefits of Quadrics Scatter/Gather to PVFS2 Noncontiguous IO





Weikuan Yu, D.K. Panda

Dept of Computer Sci. and Engg.
The Ohio State University
{yuw,panda}@cse.ohio-state.edu



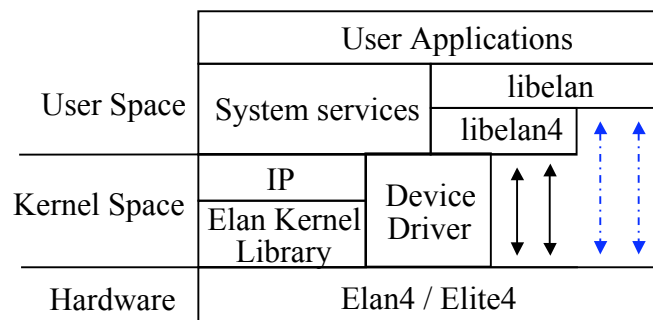
Presentation Outline

- Overview of Quadrics and PVFS2
 - Noncontiguous IO with Quadrics Scatter/Gather
 - Performance Evaluation
 - Conclusions and Future Work
- 
- 

Overview of Quadrics

- Quadrics Network: QsNet^{II}
 - High performance (10Gbps), low latency (<2us)
- Communication mechanisms
 - Queue-based DMA
 - for messages up to 2KB
 - RDMA
 - Arbitrary size messages with RDMA write/read
 - Event mechanism
 - Completion notification

Quadrics Architecture

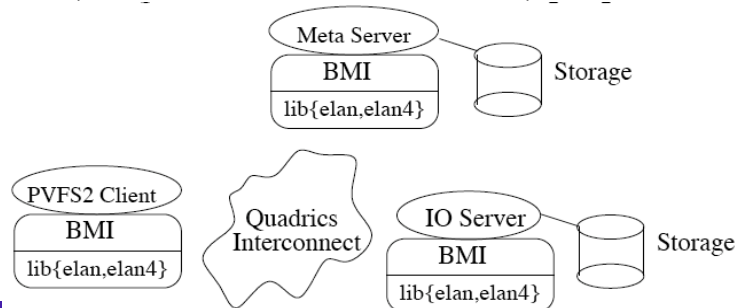


Parallel IO with PVFS2

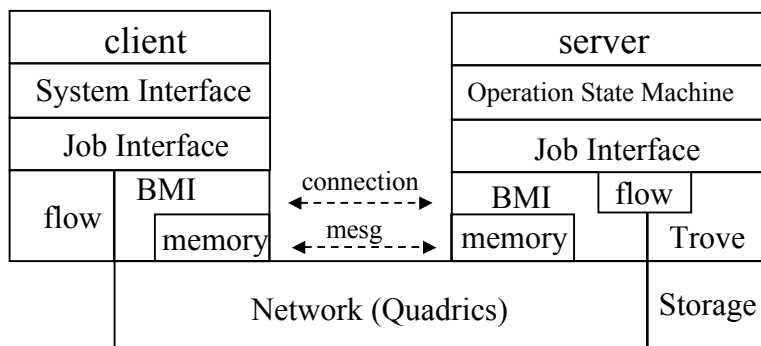
- A Parallel Virtual File system (2nd)
 - Designed for HPC, non-Posix compliant
 - List, structured IO
 - Lockless and stateless
- High speed interconnects used in PVFS2
 - Myrinet and InfiniBand
 - User-level communication and RDMA
 - Recently, we have designed a high performance implementation of PVFS2 over Quadrics
- *W. Yu, S. Liang and D.K. Panda. High Performance Support of Parallel Virtual File System (PVFS2) over Quadrics. The 19th ACM International Conference on Supercomputing (ICS '05). June, 2005*

PVFS2 over Quadrics

- Quadrics Support of PVFS2
 - Layered on top of Quadrics user-level communication, with communication being abstracted into BMI
 - Support a dynamic client/server connection model
 - Implement an efficient communication protocol



PVFS2/Quadrics Software Stack



Presentation Outline

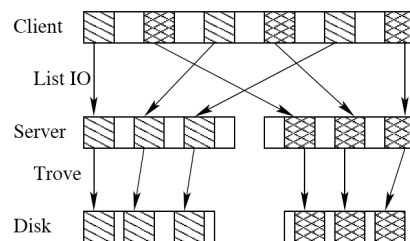
- Overview of Quadrics and PVFS2
- Noncontiguous IO with Quadrics Scatter/Gather
- Performance Evaluation
- Conclusions and Future Work

Noncontiguous IO

- A major IO pattern in scientific applications
- Used in the form of structured IO accesses: IO access to large number of noncontiguous IO regions.
- Less than 20% peak bandwidth achievable for noncontiguous IO, since File systems are typically optimized for large contiguous and sequential IO
- Solutions:
 - Support in IO libraries:
 - Two-phase IO, Data-Sieving in MPI-IO
 - Support in file systems:
 - Disk directed-IO and Server directed IO
 - PVFS2 list IO

PVFS2 List IO

- Designed for noncontiguous IO in scientific applications
- Implementation:
 - A list of IO fragments used to describe noncontiguous IO
 - Can take advantage of hardware scatter/gather support
 - If scatter/gather support is not available, fall back on basic memory packing and unpacking
 - Iterative communication for each contiguous region is also commonly used



An example of PVFS2 list IO

Quadrics RDMA and Event

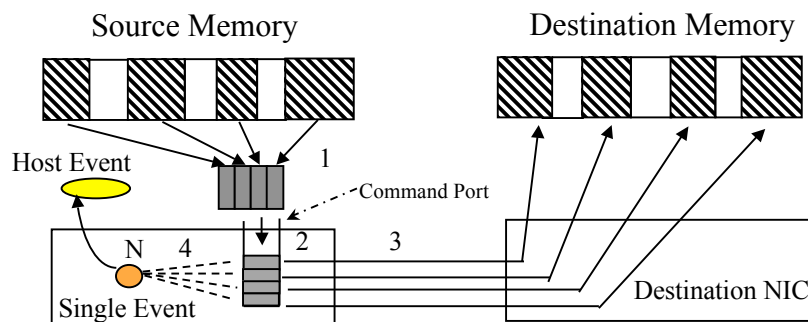
- Over Quadrics, an RDMA operation can trigger events on both the local and remote sides
- The triggered event can in turn fire another RDMA operation, when it is configured to copy a RDMA descriptor into a RDMA queue
- These steps can take place inside the network interface with little overhead

An example event to trigger another RDMA

Count	Type
	Source Address (a RDMA descriptor)
	Destination Address (a RDMA queue)

Zero-Copy Scatter/Gather with Multiple Chained RDMA

- Exchange memory address/length pairs
- Configure Multiple chained RDMA's for zero-copy PVFS2 List IO and write into the NIC



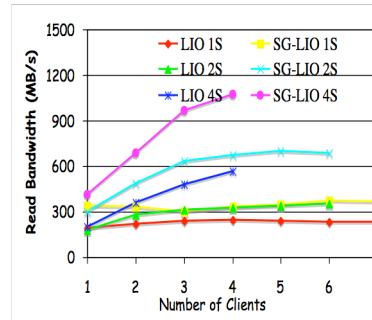
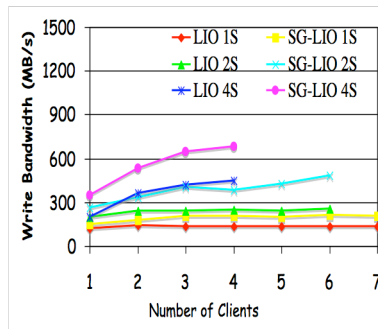
Presentation Outline

- Overview of Quadrics and PVFS2
- Noncontiguous IO with Quadrics Scatter/Gather
- Performance Evaluation
- Conclusions and Future Work

Performance Evaluation

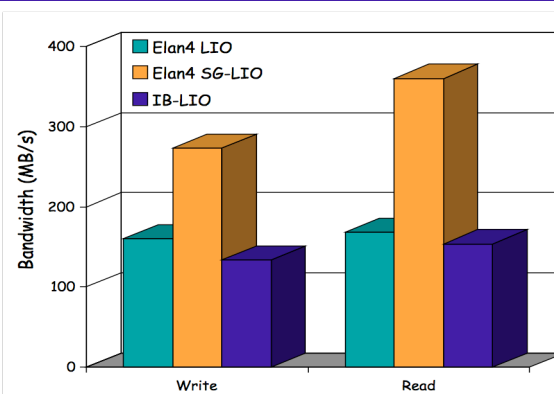
- **Experimental Testbed:**
 - A Quadrics cluster: QS-8A switch, Elan4 cards
 - Dual-SMP Intel Xeon 3.0GHz Processors
 - PCI-X 133MHz/64bit
 - 533MHz FSB
 - 1GB SDRAM memory
- **Experimental Results**
 - Concurrent Read/Write performance benefits
 - Performance of MPI-tile-IO
 - Performance of NAS BT-IO

Benefits to Concurrent Read and Write



- Improve concurrent read bandwidth by up to 89%
- Improve concurrent write bandwidth by up to 52%

Performance of MPI-tile-IO



- Improves both read and write bandwidth of MPI-Tile-IO, by 113% and 66%, respectively
- Compared to PVFS2/IBA, zero-copy list IO provides more performance benefits.

Performance of BT-IO

Table: Performance of BT-IO (seconds)

Type	Duration	IO Time
NO IO	61.71	--
BT/IO Elan4-SG-LIO	63.83	2.12
BT/IO Elan4 LIO	67.32	5.61
BT/IO IBA	69.60	7.89










- Zero-copy scatter/gather improves the IO time of BT-IO by 62% from 5.61s down to 2.12s
- PVFS2/Elan4 with zero-copy scatter/gather performs significant better than PVFS2/IBA

Presentation Outline

- Overview of Quadrics and PVFS2
- Noncontiguous IO with Quadrics Scatter/Gather
- Performance Evaluation
- Conclusions and Future Work











Conclusions

- Designed effective Quadrics scatter/gather algorithm to support PVFS2 List IO
 - Evaluated the benefits of zero-copy scatter/gather to PVFS2 list IO
 - Quadrics zero-copy scatter/gather improves PVFS2 write and read bandwidth
 - Quadrics zero-copy scatter/gather also benefits scientific applications, such as MPI-Tile-IO and BT-IO
- 
- 
- 
- 
- 
- 
- 
- 
- 



Future Work

- To study the scalability of PVFS2 over Quadrics with larger systems
 - To explore ways to take advantage of Quadrics NIC programmability and NIC memory
- 
- 
- 
- 
- 
- 
- 
- 
- 