High-Speed Block-level I/O over RDMA-capable NICs

M. Marazakis, K. Xinidis, V. Papaefstathiou, and A. Bilas

Networked Storage

Motivation:
- Primary **networked** storage subsystems
- Consolidation of storage in one subsystem
- Single interconnect for application and storage nodes

Concerns:
- Cost & Throughput between wire and host memory
- Transparent access (to storage)

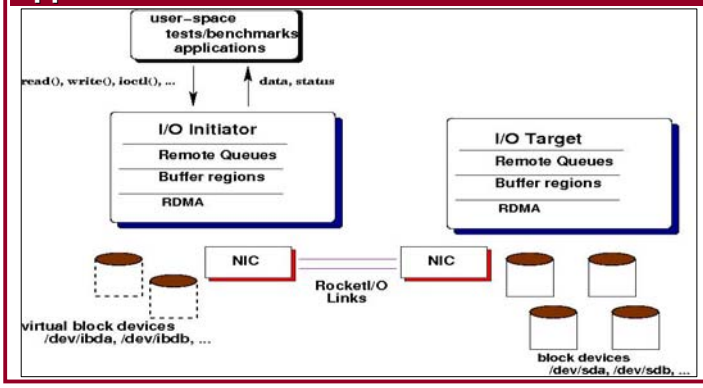
State-of-the-art:
- Specialized controllers and interconnects
- Expensive and difficult to scale
- High incremental cost for increase in performance

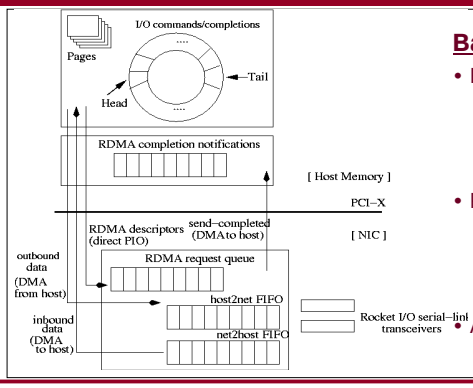
Goals:
- Efficient remote I/O
- ... using commodity components
- ... maintain transparent access
- ... identify & address overheads
- ... on a real system prototype

Approach:
- Minimal NIC architecture
  - RDMA-Write, Notification capabilities
- Design remote I/O protocol (kernel)

Communication Protocol Design:
- Reduce protocol messages
- Reduce asynchronous I/O completions
- Reduce asynchronous NIC-Host buffer mgmt
- Network-disk co-scheduling

Approach





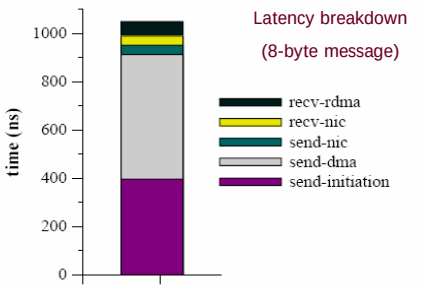
Base send-recv path
- RDMA descriptor
  - Source address
  - Destination address
  - Transfer size+notif. bits
- Packet header
  - Route (flow id)
  - Destination address
  - Transfer size+notif. bits
- Addresses are physical

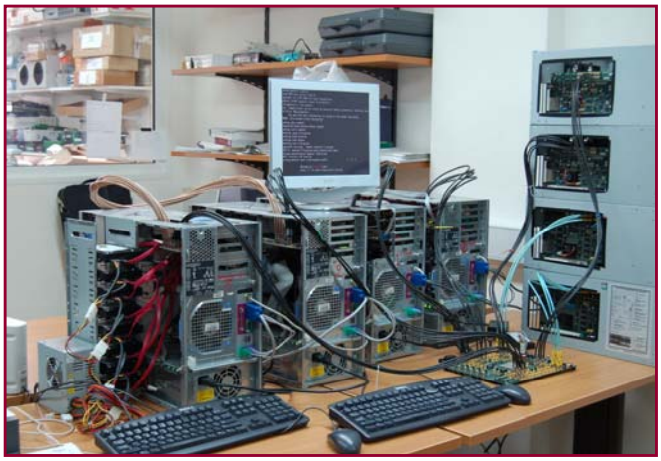
Experimental Results

Quantify overheads at high network speeds & examine various protocol optimization techniques on a real system

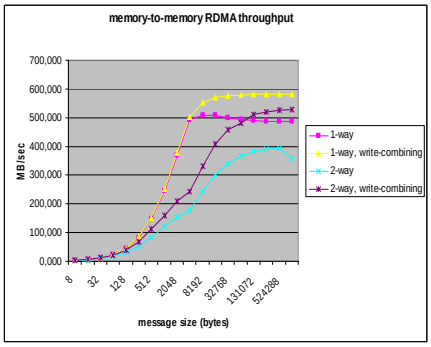
Issues:
- Number of interrupts
- CPU utilization
- RDMA initiation cost

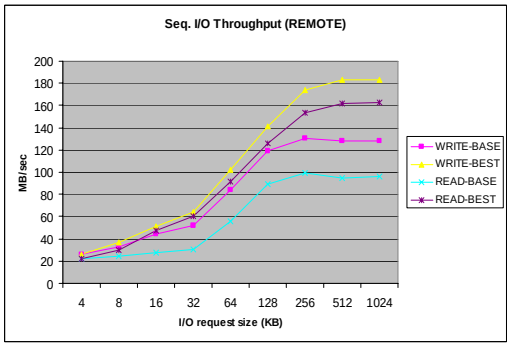
Techniques:
- I/O request batching
- Interrupt silencing
- Avoidance of small RDMAs

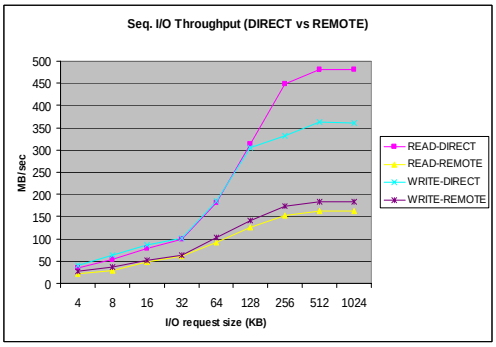




Throughput







References

- "Efficient remote block-level I/O over an RDMA-capable NIC", Proceedings of ICS'06
- "Experiences from Debugging a PCIX-based RDMA-capable NIC", Proceedings of RAIT'06 workshop (in conjunction with IEEE Cluster'06)

Foundation for Research and Technology – Hellas (FORTH)
Institute of Computer Science (ICS)

Computer Architecture & VLSI Laboratory (CARV)