Coupling advanced data movement to advanced storage

David S. Greenberg (dsgreen@cs.sandia.gov)*

Abstract

Storage architectures are poised to take a significant leap forward. The "Unix file system with processor-attached disk and tape" model will be replaced with a "data management systems and network-attached storage" view. It is important, as this critical change occurs, that advances in parallel data transfer be incorporated. In particular, the parallel computing community is moving beyond explicit OS buffering, sending, receiving, and unbuffering to protocols in which data is moved directly to or from user data structures. Future I/O systems will be able to simplify flow control and minimize buffer usage if all elements including the network-attached devices have enough power to implement these direct protocols.

The advent of desktop systems with gigabytes of RAM and of high-end systems with terabytes of RAM necessitates a rethinking of our approaches to secondary (and tertiary) storage. Fortunately, the research and vendor communities are already moving forward. The Unix, block-oriented file system is being challenged by database-like, record-oriented data management systems. The direct attachment of disks and tapes to servers is being augmented with network attached peripherals. We will be able to organize our enormous data sets in more natural ways and be able to provide the required bandwidth to storage.

Unfortunately, there is a middle area which is often ignored. In order for the high bandwidth of hundreds of NAPs to be interfaced to a distributed data management system it will be necessary to carefully coordinate the data movement. One promising solution is to leverage the lessons learned in the exploitation of massively parallel distributed-memory systems.

The MP computing community has faced a similar challenge over the last ten years. For example Sandia's Paragon has more than 1800 nodes, 16-32MByte of memory/node, interconnect link speeds of 200MB/sec, and total system bandwidth in excess of 1.3TB/sec. The Intel/Sandia Teraflop machine being built for the ASCI program will have 4500 nodes, 128Mbyte of memory/node, link speeds of 400MB/sec, and total system bandwidth in excess of 4.2TB/sec. Coordinating the transfer of data between these nodes can be difficult. However, several lessons have been learned:

1. Explicit sending and receiving of packets can help to organize and quantify necessary communication.
2. Buffer copies are very expensive. Pre-posting of receive buffers and reservation protocols can help but are often clumsy.
3. Collective operations such as broadcast and gather can be significantly optimized if library routines are used.
4. Push and pull semantics (often called put and get) are a powerful way of reducing buffering and flow control overhead.

It will be important that NAPs retain the ability to participate in state-of-the-art data movement protocols. This may require them to be able to establish pipe-lined channels directly to server or client memory and/or to be able to participate in direct put and get operations. If the capability to cooperate in data movement is not provided then storage systems are likely to experience instability, bottlenecks, and failures due to buffer overflows, handshake latencies, and other protocol overheads.

---

*Sandia National Laboratories, Mail Stop 1110, P.O. Box 5800, Albuquerque, NM 87185-5800. This work was supported by the U.S. Department of Energy and was performed at Sandia National Laboratories for the U.S. Department of Energy under contract DE-AC04-94AL85000.