SAND2006-7009P

Parallel NFS over RDMA

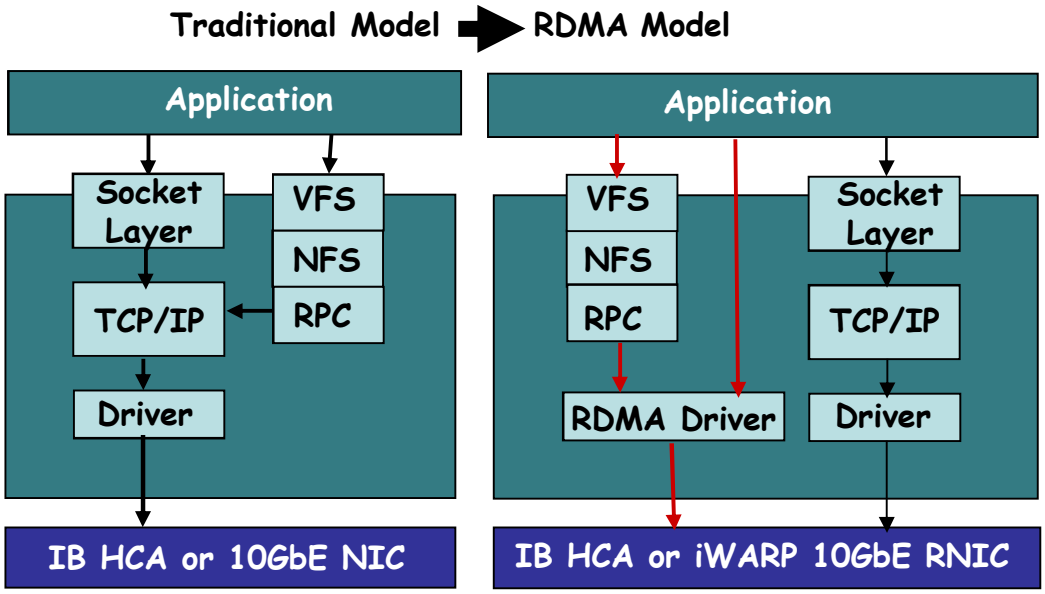
The Technology

Today's high-performance clusters use the latest 10-and 20-Gbps interconnect technologies for node-to-node communications. However, compute node to storage node performance rates have failed to keep pace with the high bandwidth made available by these new transports. This lag in advancement is of major concern as ASC applications are increasingly storage intensive. Data management challenges are becoming the dominant factor when building high performance scalable clusters.

A versatile storage architecture that is scalable and sharable is critical to meeting the performance demands for ASC and other sectors of High Performance Computing. Two emerging technologies show tremendous promise for building this next generation storage architecture: Remote Dynamic Memory Access and Parallel Storage.

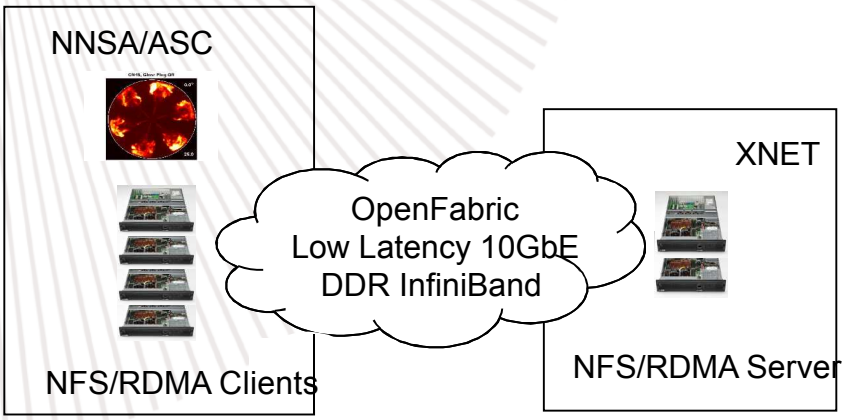
Parallel applications move large amounts of data between memory and storage. Remote Dynamic Memory Access (RDMA) can significantly reduce the CPU utilization and memory bus bandwidth required for storage to make data transfer more efficient and improve overall compute node performance. As clusters grow, the requests for storage also grows placing tremendous pressure on the storage server. Parallel Network Filesystems allow multiple storage head-ends to simultaneously participate in the data transfer of a single storage object, thereby allowing storage performance to scale with cluster size.

The ASC Scalable Computing team at Sandia National Laboratories is conducting a study to evaluate RDMA and Parallel technologies through research and industry partnerships. This effort also promotes a standardized common filesystem interface such as NFS to facilitate data sharing. Network Filesystem support for RDMA has recently become available for the Linux operating system. Its RDMA transport uses the Open Source software stack developed by the OpenFabric Alliance. Parallel NFS over RDMA is still work-in-progress. The following Figure illustrate the data path of RDMA NFS contrasting that against the traditional model.



The Demonstration

The Combustion Research Program at Sandia/CA has recently received a DOE INCITE award to execute the world's largest direct numerical simulation of a turbulent flame. This calculation will produce 5 terabytes of data vast in spatial, temporal and variable domains, creating formidable challenges for storage. We researched the effect of RDMA using a scaled down run of this simulation, writing large restart files to our RDMA enabled NFS filesystem (RNFS) [http://www.sandia.gov/D8961/Projects/IOSSN/S3D-overRNFS.pdf]. We are unable to repeat this experiment at SC06 due to the lack of compute resources, we will demonstrate RNFS's throughput and CPU advantage through benchmarking. In addition, we will perform a functional demonstration, visualizing the isosurface of a medical dataset over RNFS, a popular analysis technique used in scientific computing. Our demonstration topology is depicted below.



RDMA NFS achieved 928 MB/s
- > 3 x NFS performance
- 1/3 CPU overhead
- Much improved scalability

Better application performance!

Research and Industrial Partners:

Chelsio Communications
NetEffect
SilverStorm
Fujitsu
NetApp
Fulcrum Microsystems
OpenFabrics Alliance
System Fabric Works
Mellanox Technologies
OpenGrid Computing
University of Michigan

Team Members:

Application: Jackie Chen jhchen@sandia.gov, Craig Ulmer cdulmer@sandia.gov, and Ramanan Sankaran sankaranr@ornl.gov

Filesystem: Helen Y. Chen hycsw@sandia.gov and Dov Cohen idcohen@sandia.gob

Infrastructure: Frank Bielecki frankb@sandia.gov, Jeff Decker jcdecke@sandia.gov, Noah Fischer nfische@sandia.gov, and Mitch William mlw@sandia.gov

Sandia National Laboratories