# MPI Must Evolve or Die

Al Geist

Oak Ridge National Laboratory,
PO Box 2008,
Oak Ridge, TN 37831-6016
gst@ornl.gov
http://www.csm.ornl.gov/~geist

**Abstract.** Multicore and hybrid architecture designs dominate the landscape for systems that are 1 to 20 petaflops peak performance. As such the MPI software must evolve to effectively use these types of architectures or it will die just like the vector programming models died. While applications may continue to use MPI, it is not business as usual in how communication libraries are being changed to effectively exploit the new petascale systems. This talk presents some key research in petascale communication libraries going on in the "Harness" project, which is the follow-on to the PVM research project.

The talk will cover a number of areas being explored, including hierarchical algorithm designs, hybrid algorithm designs, dynamic algorithm selection, and fault tolerance inside next generation message passing libraries. Hierarchical algorithm designs seek to consolidate information at different levels of the architecture to reduce the number of messages and contention on the interconnect. Natural places for such consolidation include the socket level, the node level, the cabinet level, and multiple-cabinet level. Hybrid algorithm designs use different algorithms at different levels of the architecture, for example, an ALL_GATHER may use a shared memory algorithm across the node and a message passing algorithm between nodes, in order to better exploit the different data movement capabilities. An adaptive communication library may dynamically select from a set of collective communication algorithms based on the number of nodes being sent to, where they are located in the system, the size of the message being sent, and the physical topology of the computer.

This talk will also describe how ORNLs Leadership computing facility (LCF) has been proactive in getting science teams to adopt the latest communication and IO techniques. This includes assigning computational science liaisons to each science team. The liaison has knowledge of both the systems and the science, providing a bridge to improved communication patterns. The LCF also has a Cray Center of Excellence and a SUN Lustre Center of Excellence on site. These centers provide Cray and SUN engineers who work directly with the science teams to improve the MPI and MPI-IO performance of their applications.

Finally this talk will take a peek at exascale architectures and the need for new approaches to software development that integrates arcitecture design and algorithm development to facilitate the synergistic evolution of both.

A. Lastovetsky et al. (Eds.): EuroPVM/MPI 2008, LNCS 5205, p. 5, 2008.