

UD Computational Science Day 2006

February 14, 2006

# An Integrated Approach to Improving Communication Performance in Clusters

Lori L. Pollock

Department of Computer and Information Sciences  
University of Delaware  
Newark, DE 19716

## Abstract

Cluster computing has become a common, cost-effective means of parallel computing for applications programmers who require parallelism. Although adding more CPUs increases the cluster's maximum processing power, real applications can not efficiently use very large numbers of CPUs, due to lack of scalability. In regular codes, such as the case of FFT, the main impediment to achieving scalability is the communication overhead which increases as the number of CPUs increases.

Many researchers have proposed optimizations in both the compilers and the interconnecting networks. We argue that most of these optimizations target specialized hardware or programming languages, require specialized knowledge from the domain scientist, or are not enough to provide a comprehensive solution on their own. An indication of this is that most parallel applications are still written in Fortran 77 or 90, run on Linux machines connected by Fast Ethernet or Myrinet, and use MPI for the communication, as they did ten years ago. The downfall of this approach is that communication intensive parallel programs suffer in such an environment from the layers of communication software between the sender processes and the receiver processes. While network technology is currently available with the potential to achieve significantly improved performance overall for these applications, it remains largely untapped due to (1) the need for the knowledge of the context of the communication operations (i.e., the application program containing the calls to the communication library) to exploit the sophisticated network technology fully, and (2) the low level nature of programming needed within the application program context to achieve that potential. In particular, performance can often be improved through increasing the use of asynchronous communication. Unfortunately, programming with asynchronous communication is quite difficult and error prone, even for the most experienced parallel programmers.

We propose a *vertically integrated approach*, where a set of optimizations in the compiler, network and operating system, can enable legacy parallel applications to scale to a much larger number of CPUs, even if written without any knowledge of our techniques. To justify our approach, we recently built an experimental prototype and conducted some preliminary experimental studies for both a parallel application that implements a simple computation problem and FFT, examining different manually-defined communication-computation strategies. Our experimental results provide evidence that significant performance improvements are possible with a vertically integrated approach where knowledge of the context of communication operations is joined with knowledge of the network and cluster details to provide a fine-grain strategy for overlapping communication and computation. Based on our initial promising results, *the overall goal of our current research is to create a means for scalable cluster computing through enabling integrated knowledge and cooperation between the source optimizer, operating system, and network technology of the cluster, without relying on the programmer to learn about, and program in terms of, the low level details of the cluster communications system.*

This research is a collaborative effort between Martin Swany in operating systems and parallel computing and Lori Pollock in program analysis and compiler optimization for parallel architectures and combined expertise in experimental systems evaluation. The foreseen contribution of this project is a novel vertically integrated framework for gaining fine-grain overlap in communication and computation for clusters, able to be easily parameterized for various network technologies and cluster configurations.

This talk focuses on the program transformations directed toward improving communication-computation overlap in parallel programs that use MPI's collective operations. Results from a detailed study of the effect of the problem and message size, level of communication-computation overlap and amount of communication aggregation on runtime performance in a cluster environment based on an RDMA-enabled network will be presented.