Hiding Communication Overheads in High-Performance Computing

Jose Carlos Sancho, Kevin J. Barker, Darren J. Kerbyson, Kei Davis, CCS-1

Communication delays are becoming a critical factor in the performance of large-scale applications that can utilize many thousands of processors at a time. Performing communication at the same time as computation, commonly known as overlapping, can partially or even fully hide these communication delays leading to an immediate increase in achievable performance.

In recent work in the Performance and Architecture Lab (PAL), a part of the Computer Science for High-Performance Computing group at Los Alamos National Laboratory, we have developed methods to quantify the potential performance improvement from overlapping communication with computation. Insights from this work have shown that overlapping is possible and that networks in high-performance computing (HPC) could be more cost effective.

The methods are predictive and thus designed to examine the potential performance improvements prior to any application recoding. Measurements are made of the computation components that are required prior to a communication and those performed after a communication. The potential overlap is calculated by comparing the computational time available with the communication delays that will be experienced on the target network. A simple example is shown in Fig. 1 where the potential overlap is the difference in time between the point at which data is produced on the sender and the point at which it is required on the receiver.

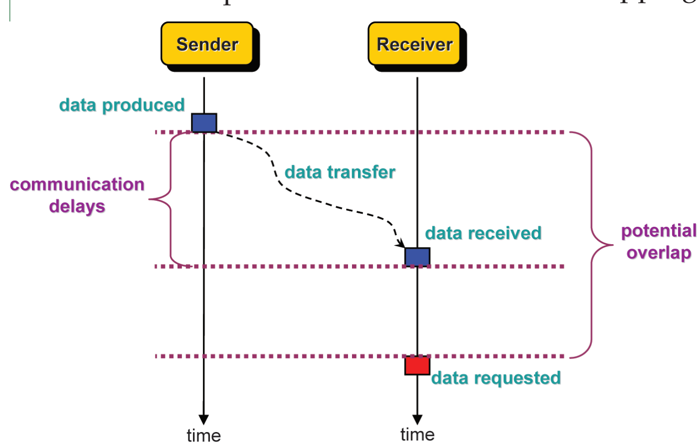

Fig. 1.
Example showing potential overlap.

Initial work considered four large-scale applications: POP (Ocean Modeling), SAGE with and without adaptive mesh refinement (shock-wave hydrodynamics), Sweep3D (deterministic transport), and HYCOM (Ocean Modeling). A medium-sized system containing 1024 processing cores with an InfiniBand network was used in the analysis.

The predictive nature of the method can be seen in Fig. 2 in which networks of varying bandwidths are considered (from 1 MB/s up to 5 GB/s). The performance is normalized to an ideal case in which communication has zero performance penalties. When overlapping is exploited the potential performance quickly attains near ideal performance even for networks with low bandwidths (in this case at ~400 MB/s). When overlapping is not exploited then the achievable performance gradually nears that of the overlapping case as the network bandwidth increases. This example is for SAGE in which message sizes can be between 10 and 100s of KB.

This result is important in that it shows that networks with lower bandwidths, and hence lower costs, could be utilized

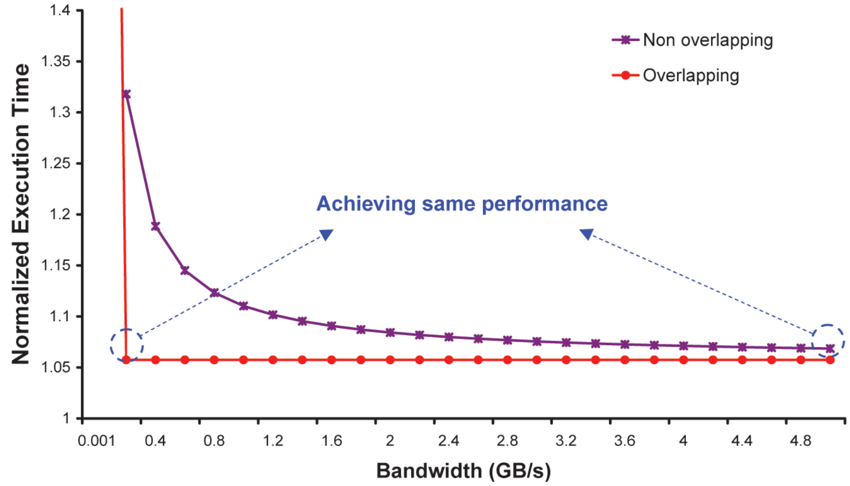

Fig. 2.
Potential of overlap in SAGE for networks of various bandwidths.

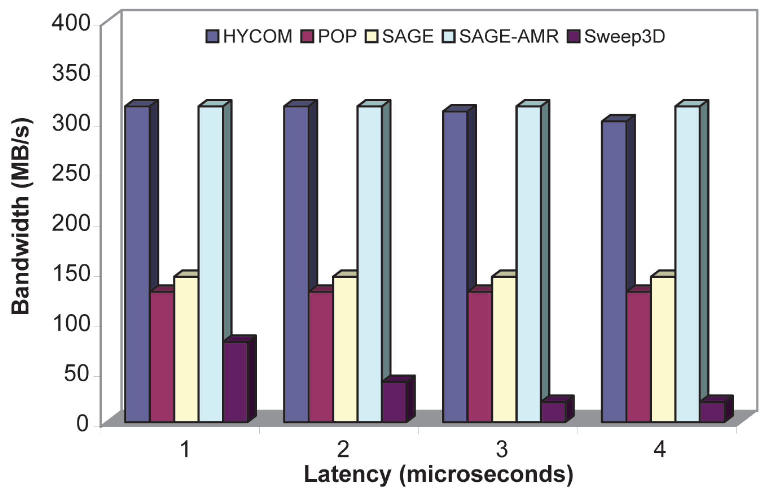

Fig. 3.
Bandwidth network characteristics when overlapping is exploited deliver the same performance as the case of nonoverlapping on a 5-GB/s network.

to achieve the same performance as more expensive networks that have higher bandwidths; that is, if overlapping can be exploited by the application. Potential overlap does exist in the four applications analyzed and shows that a reduction in network bandwidth and/or increase in latency can be tolerated. Figure 3 summarizes these results and shows the network characteristics (in terms of bandwidth and latency) that are required for each application when using overlap, compared to the case with no overlap and a network with 5 GB/s bandwidth, and 4 μs latency. This work has demonstrated that future HPC systems could benefit from the use of cost-effective networks that support overlapping of communication with computation without negatively impacting application performance.

This work was presented by PAL in the technical program at the Supercomputing Conference (SC06) in November 2006 [1].

For more information contact Jose Carlos Sancho at jcsancho@lanl.gov.

[1] J.C. Sancho, et al., "Quantifying the Potential Benefit of Overlapping Communication and Computation in Large-Scale Scientific Applications," Los Alamos National Laboratory report LA-UR-06-3109 (May 2006).

Funding Acknowledgements
NNSA's Advanced Simulation and Computing (ASC), and Defense Advanced Research Projects Agency (DARPA), High Productivity Computing Systems Program.