TN

CI

IMPROVED SCALABILITY THROUGH NODE-AWARE COMMUNICATORS

Allocation: Exploratory/50 Knh
PI: Luke Olson[1]
Co-PI: Amanda Bienz[1]
Collaborator: William Gropp[1]

[1]University of Illinois at Urbana–Champaign

EXECUTIVE SUMMARY

Sparse matrix operations abound in numerical simulations and represent significant costs at scale. As researchers anticipate future machine networks and compute units, these communication-bound operations will continue to incur significant cost. The focus of this work is on reducing communication at scale, particularly in settings where machine layout and multiple compute units can be exploited. Blue Waters is an ideal setting for developing these methods since the research team can expose node-level (as well as socket-level) parallelism. The work has identified new methods for organizing communication to reduce the overall time-to-solution.

RESEARCH CHALLENGE

Sparse matrix operations such as sparse matrix–vector multiplication and sparse matrix–matrix multiplication are key kernels in many iteration and preconditioning techniques. Yet, these operations incur a significant communication penalty in situations where the matrix patterns are highly unstructured. The goal of this work is to identify metrics for predicting communication overhead and to construct communication routines that can significantly reduce cost by utilizing certain aspects of the machine such as the node and socket layout.

METHODS & CODES

This work has led to the development of scalable solvers within the RAPtor package [1] and has helped to shape the development of a node-aware library [2] that can be used easily by application codes from a range of scientific disciplines. Both software packages rely on node-aware MPI communication, which reroutes standard internode communication to reduce the total cost of sending data through the network.

There are two variations of this work, both aggregating data at the node level to reduce the number and size of messages being sent through the interconnect. Two-step node-aware communication consists of each process gathering all data to be sent to a single node and sending these data directly to the corresponding process on the destination node. The receiving process then distributes these data to all on-node processes that need it. Alternatively, three-step communication adds another layer of aggregation, agglomerating all data to be sent between two nodes on a single process on the node of origin. These data are then sent as a single message between nodes, after which it is redistributed locally to any process on the receiving node that needs it.

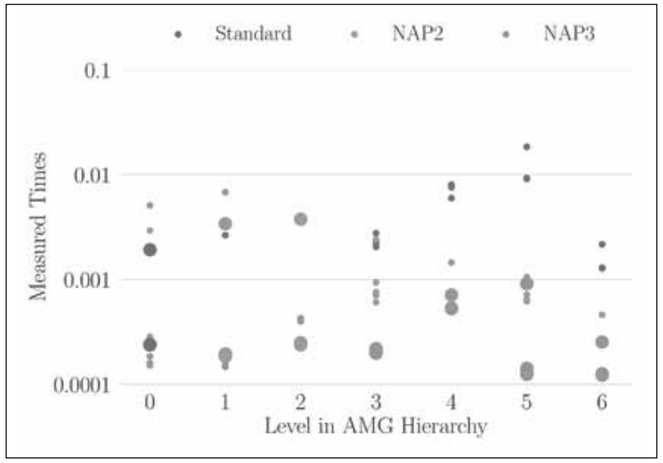

Figure 1: The matrix–matrix multiplication cost in each level of the AMG (algebraic multigrid) hierarchy using standard, two-step (NAP2), and three-step (NAP3) node-aware communication routines.

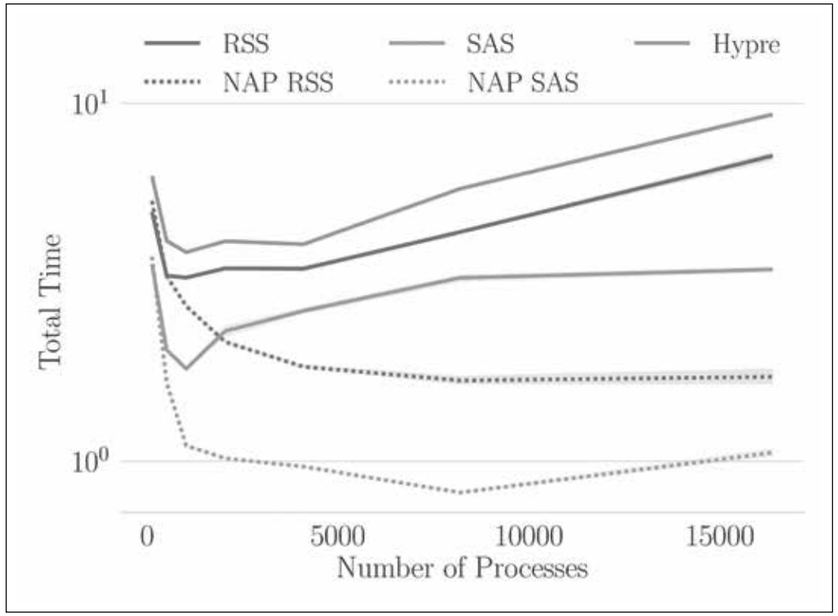

Figure 2: Total AMG times for both the Ruge–Stüben solver (RAS) and smoothed aggregation solver (SAS) for a Grad-Div problem on Blue Waters, along with the node-aware (NAP) implementation.

RESULTS & IMPACT

The optimal method of communication varies with communication pattern as well as network topology. Therefore, optimizing communication costs throughout the algebraic multigrid (AMG) requires different strategies based on the system being solved. There are several factors that impact communication costs, including the sparsity pattern on a particular level of the AMG hierarchy, the amount of data being communicated on each level, and the number of active processes. Fig. 1 displays the cost of each strategy for matrix–matrix multiplication on each level of an AMG hierarchy, a dominating kernel in the AMG setup phase. The results show that standard communication outperforms node-aware on fine levels where communication is regular and structured, while node-aware strategies are optimal on coarse levels. Similarly, while three-step communication is often optimal, two-step node-aware communication often outperforms on the first few coarse levels when messages are large.

One goal of this work on Blue Waters is to develop a method to automatically select the communication strategy. RAPtor includes a predictive performance model that identifies the optimal communication strategy for each operation. Fig. 2 shows the cost of standard AMG versus node-aware AMG where performance models select the optimal communication strategy for each operation. Communication costs can vary widely for different computational kernels and machine settings—this work automates the communication strategy leading to significant speedups and reduced time-to-solution.

WHY BLUE WATERS

Blue Waters was central to this work, providing a large scale of resources to test the models and routines developed by the research team. The node-level layout of Blue Waters provided the initial inspiration for the algorithms the team developed, and the consistency of the compute environment was a central component in reproducible testing at scale.

PUBLICATIONS & DATA SETS

A. Bienz, W. D. Gropp, L. N. Olson, "Node aware sparse matrix–vector multiplication," *J. Parallel Distrib. Comput.*, vol. 130, pp. 166–178, 2019, doi: 10.1016/j.jpdc.2019.03.016.