



Improving Performance Models for Irregular Point-to-Point Communication

Amanda Bienz
University of Illinois at
Urbana-Champaign
bienz2@illinois.edu

William D. Gropp
University of Illinois at
Urbana-Champaign
wgropp@illinois.edu

Luke N. Olson
University of Illinois at
Urbana-Champaign
lukeo@illinois.edu

ABSTRACT

Parallel applications are often unable to take full advantage of emerging parallel architectures due to scaling limitations, which arise due to inter-process communication. Performance models are used to analyze the sources of communication costs. However, traditional models for point-to-point communication fail to capture the full cost of many irregular operations, such as sparse matrix methods. In this paper, a node-aware based model is presented. Furthermore, the model is extended to include communication queue search time as well as an additional parameter estimating network contention. The resulting model is applied to a variety of irregular communication patterns throughout matrix operations, displaying improved accuracy over traditional models.

KEYWORDS

MPI, performance modeling, point-to-point communication, queue search, network contention

ACM Reference Format:

Amanda Bienz, William D. Gropp, and Luke N. Olson. 2018. Improving Performance Models for Irregular Point-to-Point Communication. In *25th European MPI Users' Group Meeting (EuroMPI '18)*, September 23–26, 2018, Barcelona, Spain. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3236367.3236368>

1 INTRODUCTION

As parallel computers advance, improvements to hardware yield potential for solving increasingly large and difficult problems. However, applications are often unable to take full advantage of state-of-the-art architectures due to scaling limitations, which result from inter-process communication costs. The cost associated with communication depends on a large number of factors, and varies across parallel systems, specific partitions, and application scale. Therefore, performance models are used to analyze the sources of communication costs among different architectures and network partitions. Accurate performance models specify whether the cost is due mainly to the number of messages communicated, number of bytes transported, distance of transported bytes, or some other factor.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

EuroMPI '18, September 23–26, 2018, Barcelona, Spain

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-6492-8/18/09...\$15.00

<https://doi.org/10.1145/3236367.3236368>

Traditional models estimate point-to-point communication as a combination message latency and the cost of transporting bytes. Irregular operations, such as sparse matrix methods, acquire costs that are not captured by traditional models. Figure 1 displays the measured and modeled communication costs acquired when performing a sparse matrix-matrix (SpGEMM) multiply on the levels of an algebraic multigrid (AMG) hierarchy for an unstructured linear elasticity matrix. These timings, as well as the associated model

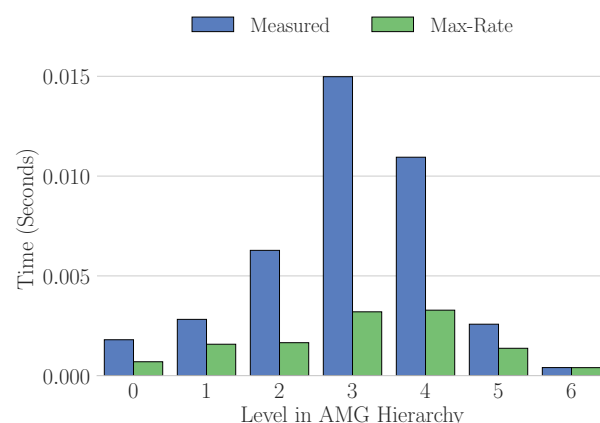


Figure 1: Measured and modeled communication costs associated with a SpGEMM on each level of a linear elasticity AMG hierarchy, on 8 192 processes of Blue Waters supercomputer.

parameters, are for 8 192 processes of the Blue Waters supercomputer¹ [4]. The traditional postal model results in nearly identical timings to more robust models, such as the max-rate model [11], which takes into account the limitations of multiple communicating processes per node. For this problem both models capture only a fraction of the measured time.

This paper extends traditional performance models to accurately model the irregular point-to-point communication that occurs in commonly used operations. This paper presents three novel contributions, including an improvement to the max-rate model [11] measurements:

- (1) node-aware model parameters;
- (2) an extension to include quadratic queue search times in communication; and
- (3) an additional parameter estimating network contention.

¹<https://bluwaters.ncsa.illinois.edu/>

The remainder of this paper is outlined as follows. Section 2 describes parallel point-to-point communication as well as corresponding traditional performance models. Node-awareness is added to traditional models in Section 3. Section 4 describes a high-volume ping-pong algorithm along with additional acquired penalties, with a queue search parameter described in Section 4.1 and a network contention penalty in Section 4.2. The improved model parameters are applied to commonly used operations and compared against measured timings in Section 5. Finally, conclusions, limitations, and future directions are described in Section 6.

2 BACKGROUND

Many common parallel operations, such as those involving sparse matrices, require MPI point-to-point communication. This category of communication consists of sending a single message between a set of processes. In a typical implementation, the pairs of communicating processes, along with the size of associated messages, vary. The point-to-point communication procedure varies with MPI implementation. Typically, each message consists of both an envelope and data, where the envelope contains a message description including the tag, MPI communicator, message length, and process of origin. There are a variety of methods for sending data, such as sending the data immediately or waiting for the receive process to allocate buffer space. In the implementations investigated in this paper, a message is communicated via a specific protocol of short, eager, or rendezvous, based on message size. The short protocol consists of sending very small messages as part of the envelope directly between processes. Messages that are too large to fit in the envelope, but remain relatively small, are communicated with eager protocol. This protocol assumes buffer space is available, and immediately communicates the data to the receiving process. Lastly, sufficiently large messages are communicated with rendezvous protocol, during which the envelope is communicated first, and the remainder of the data is only communicated after the receiving process has allocated buffer space.

The cost associated with each message is dependent on the time required to initialize communication as well as the per-byte transport cost. Therefore, the short protocol is significantly less costly than the others as only a single envelope is communicated, yielding minimal costs associated with both latency and bandwidth. Messages communicated with eager protocol have low latency costs as the messages are sent directly between processes. However, the associated per-byte transport cost increases, as these messages can require significant amounts of buffering. Finally, rendezvous messages yield low per-byte transport costs but increased latency requirements associated with initial envelope communication and synchronization.

The traditional postal model estimates the cost of communicating a message as the sum of the message startup cost and the per-byte transport, with separate parameters for each message protocol. This can be defined as

$$T = \alpha + \beta \cdot s, \quad (1)$$

where α is the latency, β is the cost to transport a byte of data, and s is the number of bytes to be transported. As the associated costs vary with message protocol, separate values for α and β are used when communicating with short, eager, and rendezvous protocols.

This model accurately analyzes the cost of a standard ping-pong test, in which two processes are sending messages to one another. However, it fails to account for a variety of penalties that occur during communication in typical operations on state-of-the-art supercomputers.

There are many alternatives to the postal model that account for many penalties that arise in standard supercomputers. The LogP model splits the α into latency, the cost required by the hardware, and overhead, the cost associated with the software [6, 10]. This addition of overhead allows this model to capture the cost of overlapping communication and computation. The LogGP model extends the LogP model to analyze the cost of long messages [2]. Network contention parameters are investigated with the LoPC and LoGPC models [9, 15]. Accurate models exist for network contention in collective communication, such as the MPI_Alltoall operation [18]. Furthermore, learning algorithms yield accurate contention prediction [13]. Topology-awareness can improve models, as hop count, or the number of links traversed by a message, affects the cost of communication [1]. The cost of various implementations of the receive queue have been investigated, and improvements to queue search cost is an active area of research [14, 19–21]. Computer simulations can accurately estimate the cost of communication, but at a significantly increased cost [12, 16, 17]. Network contention has been previously modeled for collective communication, specifically the MPI_Alltoall operation.

The max-rate model [11] improves upon the postal model by defining the cost of communication as dependent on not only the latency and inter-process bandwidth costs, but also on the maximum bandwidth by which a node can inject data into the network. Therefore, this model accounts for the fact that injection bandwidth becomes a bottleneck when communicating from four or more processes per node, as is typical with state-of-the-art parallel computers [11].

Throughout the remainder of this paper, the max-rate model is used as a baseline. All ping pong timings are collected through multiple runs of Baseenv², a topology-aware library useful for benchmarking performance. Each ping-pong test consists of four duplicate timings, with exception to the original max-rate tests, which test the various numbers of actively communicating processes-per-node one time each. Each Baseenv program is tested three different times.

The models throughout this paper are tested on Blue Waters, a Cray XE/XK machine at the National Center for Supercomputing Applications (NCSA) at University of Illinois. Blue Waters contains a 3D torus Gemini interconnect, in which each Gemini consists of two nodes. The system contains 22 636 XE compute nodes, each comprised of two AMD 6276 Interlagos processors, as well as 4 228 XK compute nodes containing a single AMG processor along with an NVIDIA GK110 Kepler GPU. All tests in this paper are performed on partitions of XE system nodes. The tests use a CrayMPI implementation that is similar to MPICH.

²<http://wgropp.cs.illinois.edu/projects/software/index.html>

3 NODE-AWARE MODELING

The max-rate model is defined as

$$T = \alpha + \frac{\text{ppn} \cdot s}{\min(R_N, \text{ppn} \cdot R_b)}, \quad (2)$$

where ppn is the number of actively communicating processes per node, R_b is the rate at which data can be sent between two processes, or the inverse of β , and R_N is the maximum rate at which a node can inject data into the network. Therefore, when the value of $\text{ppn} \cdot R_b$ is less than injection bandwidth, this model reduces to the postal model. However, with a sufficiently large number of active processes per node, the per-byte transport rate is measured as injection bandwidth. Figure 2 displays the max-rate model versus measured times when communicating a single message of various size between pairs of processes. These associated models

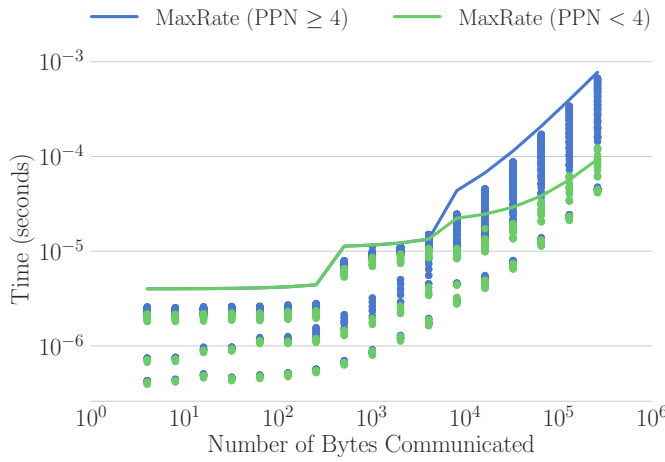


Figure 2: Ping-pong measured times (dots) versus max-rate model (lines) on Blue Waters using parameters from [?].

are computed with published Blue Waters parameters [11], and the measured times are acquired from sending messages between processes that lie on the same socket, different sockets of the same node, or on neighboring nodes of Blue Waters. While the addition of network injection limits yields large improvement over the standard postal model when communicating rendezvous messages from a large number of processes per node, the model overestimates for a large portion of these timings.

There is a large difference between intra-socket messages, intra-node messages that traverse across sockets, and communication between two nodes. Therefore, different parameters should be used for each of these cases. Furthermore, intra-node messages are not injected into the network, and therefore the simple postal model is sufficient. Figure 3 displays the measured versus modeled times after splitting the model into on-socket, on-node but off-socket, and off-node messages for both Blue Waters. The parameters corresponding to this node-aware model are listed in Table 1.

4 ADDITIONAL PENALTIES

Realistic applications that involve point-to-point communication typically require more than one message to be communicated from

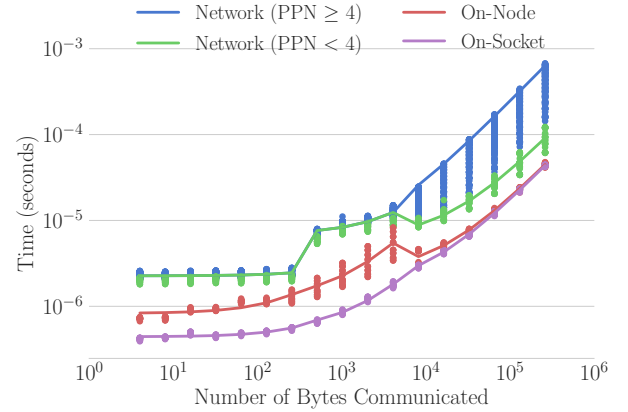


Figure 3: The max-rate model versus measured times when split into on-socket, on-node but off-socket, and off-node communication.

any process. However, standard communication models were created to analyze the cost of sending a single message, and do not extend accurately to large message counts. A ping-pong test with large message counts, described in Algorithm 1, acquires additional costs not captured by the postal or max-rate models.

Algorithm 1: HighVolumePingPong

Input:	rank:	MPI rank of current process
	p :	process with which to communicate
	n :	number of messages to communicate
	s :	size of each message (in bytes)
	send_tags:	list of n MPI send tags
	recv_tags:	list of n MPI receive tags

```

if rank <  $p$ 
  for  $i < n$  do
    MPI_Isend(...,  $s$ , ...,  $p$ , send_tags $_i$ , ...)
  MPI_Waitall( $n$ , ...)

  for  $i < n$  do
    MPI_Irecv(...,  $s$ , ...,  $p$ , recv_tags $_i$ , ...)
  MPI_Waitall( $n$ , ...)
else
  for  $i < n$  do
    MPI_Irecv(...,  $s$ , ...,  $p$ , recv_tags $_i$ , ...)
  MPI_Waitall( $n$ , ...)

  for  $i < n$  do
    MPI_Isend(...,  $s$ , ...,  $p$ , send_tags $_i$ , ...)
  MPI_Waitall( $n$ , ...)

```

Table 1: Parameters for node-aware max-rate model on Blue Waters.

	intra-socket		intra-node		inter-node		
	α	R_b	α	R_b	α	R_b	R_N
short	4.4e-07	2.2e09	8.3e-07	4.8e08	2.3e-06	1.3e09	∞
eager	5.3e-07	3.2e09	1.2e-06	9.6e08	7.0e-06	7.5e08	∞
rend	1.7e-06	6.2e09	2.5e-06	6.2e09	3.0e-06	2.9e09	6.6e09

4.1 Queue Search

Point-to-point communication conceptually requires multiple queues to be formed and traversed, including a send queue, comprised of sends that have been posted; a receive queue of similarly posted receives; and an unexpected message queue, containing messages that have been communicated for which no matching receive has been posted [5]. The function and availability of these queues, which are dependent on MPI implementation, greatly affects the performance of communicating a large number of messages.

The standard implementation of MPICH creates two separate receive queues, one for the posted messages and the other for unexpected messages [7]. When an envelope is received, the queue of posted messages is searched for a message in which all variables such as tag, datatype, communicator, and sending process match the envelope. If no associated message has been posted, the envelope and any corresponding data is added to the unexpected message queue. Similarly, when a message is posted by the application, the unexpected message queue is traversed for any corresponding envelope. If no match is found, the message is added to the posted message queue.

The CrayMPI implementation requires a receive queue to be searched linearly, yielding an additional cost when communicating multiple messages. In the worst case, the messages are received in the order opposite of which they are posted, requiring a traversal of an entire queue for each receive. Therefore, the queue search is an $O(n^2)$ operation. Methods have been created to reduce this queue search cost, such as the use of multiple queues in combination with hash maps [8]. However, as a standard queue search is currently implemented in the version of MPI on Blue Waters, the large queue search cost is investigated.

Figure 4 displays both the measured and modeled costs for performing the HighVolumePingPong described in Algorithm 1 among all 16 processes local to a single node on Blue Waters. The number of messages communicated ranges from 1 to 10 000 with the total number of bytes injected into the network remaining constant. In the ideal scenario, the variables send_tags_i is equal to recv_tags_i for all $i < n$, resulting in messages being received in the same order as they are posted. Therefore, the first message in the searched queue yields a match, resulting in an $O(n)$ queue search cost. As a result, the max-rate model accurately analyzes these measured times. In the worst-case scenario, the variables send_tags_i is equal to recv_tags_{n-i-1} for all $i < n$, posting receives in the opposite order from which messages are received. Therefore, the entire queue is traversed for each receive, resulting in an $O(n^2)$ queue search cost, yielding measured times that vary greatly from the model.

The large inaccuracies of traditional models for large message counts motivates adding an additional parameter for the time required to search the receive queues. This addition to the models is defined as

$$T_q = \gamma \cdot n^2, \quad (3)$$

where γ is the cost of stepping through either the posted or unexpected message queue. This cost is independent of message sending protocol as well as relative locations of the send and receive processes. Therefore, there is a single parameter for all combinations of on-socket, on-node, off-node, and short, eager, and rendezvous. The upper bound queue search cost is described as

$$\gamma = 8.4e - 09. \quad (4)$$

Figure 5 shows the measured versus modeled times for the HighVolumePingPong test in which the messages are received in the opposite order from which they are posted. This figure adds the queue search parameter to the original max-rate model, yielding a more accurate analysis of the cost of large message counts.

4.2 Network Contention

When a large amount of data is communicated throughout the network, multiple messages are often required to traverse the same link, yielding contention within the network. This network contention can occur on as few as eight nodes of Blue Waters when a one-dimensional partition of the network is attained. Figure 6 shows a line of four Geminis, each containing two nodes, with a one-dimension network partition. Communicating messages from all 32 processes on Gemini 0 to corresponding processes on Gemini 2, and similarly sending from Geminis 1 to 3, requires all data to traverse the network link between Geminis 1 and 2. Therefore, contention of this link occurs.

Figure 7 shows the measured and modeled costs of sending messages of various counts and sizes among all processes on the row of Geminis, where the model contains both the max-rate and queue search measures. The model underestimates the cost of communicating a large amount of data at smaller message counts, before queue search time dominates. The additional measured cost can be modeled through an extra network contention parameter. This addition measure is defined as

$$T_c = \delta \cdot \ell, \quad (5)$$

where δ is the per-byte penalty acquired waiting for a network link and ℓ is the number of bytes to traverse each link. Network contention only occurs during inter-node communication. However, the cost of contention is constant regardless of the message sending protocol. The measure for all inter-node messages on Blue Waters

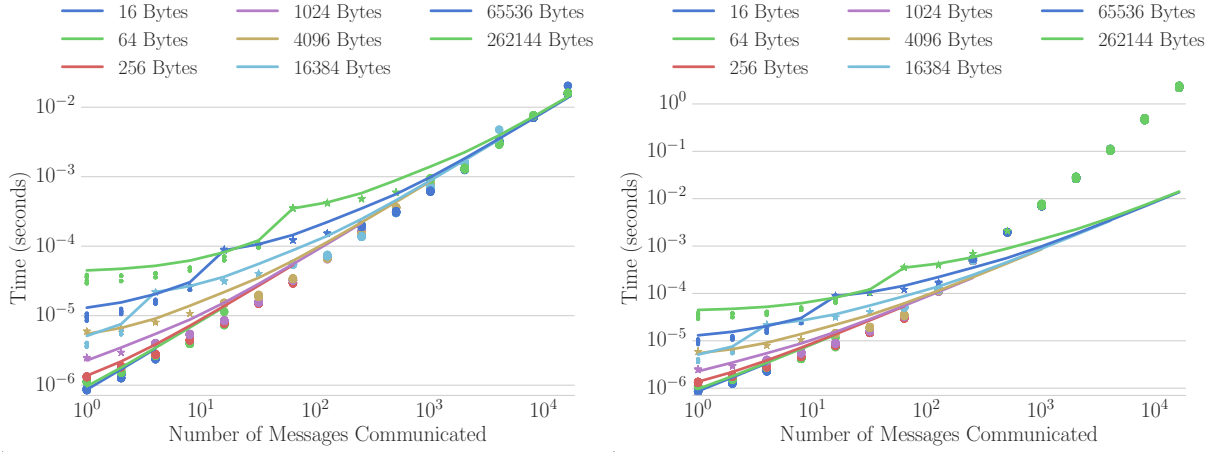


Figure 4: The measured versus modeled (max-rate) costs of sending a number of messages between two processes that lie on the same node of Blue Waters. On the left, all receives are posted in the same order that messages are received, resulting in no queue search cost. The right plot displays the cost when receives are posted in the opposite order from which they are received, resulted in a quadratic queue search cost that is not captured in the max-rate model.

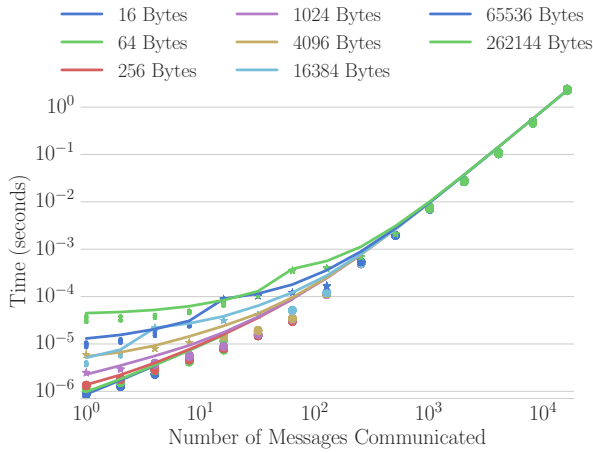


Figure 5: The measured versus modeled times for Blue Waters, where the model is a combination of the max-rate model and the contention, for HighVolumePingPong tests with a variety of message counts and sizes. The receives are posted in the opposite order of which messages are received.

is

$$\delta = 1.0e - 10. \quad (6)$$

The number of bytes to traverse any link, or ℓ , is dependent on the number of links each message traverses. Therefore, knowledge of the specific partition of the network is required to model the associated cost. This requirement is removed by assuming the nodes are connected through a perfect three-dimensional cube portion of Blue Waters' three dimensional torus, as displayed in Figure 8. Therefore, ℓ is defined as the following

$$\ell = 2h^3 \cdot b \cdot \text{ppn}, \quad (7)$$

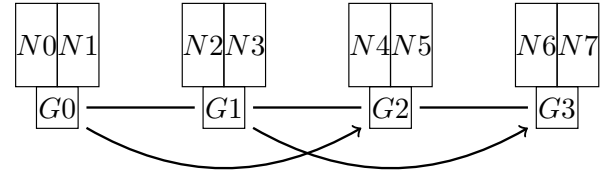


Figure 6: Four Gemini spanning a one-dimensional partition of the Blue Waters network. A HighVolumePingPong test between all processes on Geminis G0 and G3, and equivalent messages between G1 and G3 will result in a large amount of contention for the middle link in the partition.

where h is the average number of hops, or network links, traversed by each byte of data, and b is the average number of bytes to be sent from any process. Therefore, as h^3 yields the number of Geminis within h hops of a given link, this measure estimates network contention assuming all bytes that can traverse one single link do. Furthermore, $2b \cdot \text{ppn}$ calculates the average number of bytes communicated from each Gemini.

Figure 9 displays the measured and modeled costs of communicating a variety of message counts and sizes among four Geminis of Blue Waters, with a one-dimensional network contention. This figure includes the max-rate, queue search, and network contention models, yielding improved accuracy in the model.

5 APPLICATIONS

Sparse matrix-vector (SpMV) and sparse matrix-matrix (SpGEMM) multiplication are commonly used in a variety of applications such as numerical methods and graph algorithms. When matrices are sufficiently sparse, point-to-point communication is used to send only necessary values to the processes that need them.

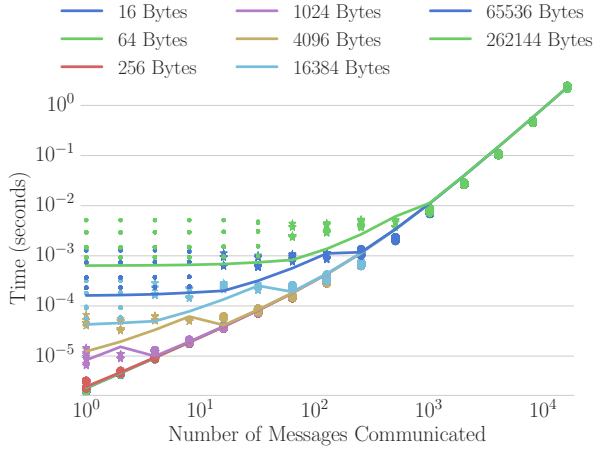


Figure 7: Measured versus modeled times for HighVolumePingPong communication among the sets of the Blue Waters Geminis described in Figure 6. The modeled times, which are a combination of max-rate models and the queue search parameters, do not capture the additional costs associated with contention.

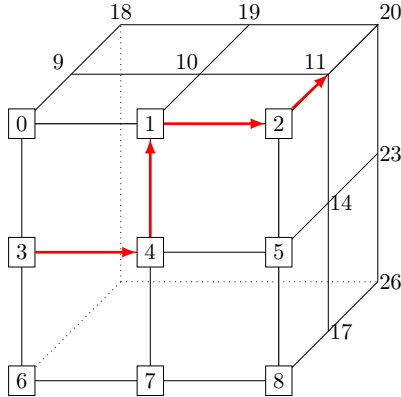


Figure 8: A perfect cube partition of Blue Waters Geminis is used to calculate the average number of hops traversed by each byte of data. In this example, a message from a process on Gemini 3 to Gemini 11 traverses 4 network links.

Algebraic multigrid (AMG) is a sparse linear solver comprised of matrix operations. An AMG hierarchy consists of successively coarser, but denser, matrices. Therefore, each level in the hierarchy decreases in dimension, but often increases in the number of non-zeros per row. The various levels require a variety of communication patterns, as the finer levels require communication of few large messages while coarse levels require communicating a larger number of small messages.

This section focuses on modeling the cost of matrix operations throughout an AMG hierarchy as communication costs vary drastically among the levels. The hierarchy is formed with classical AMG

to solve a three-dimensional unstructured linear elasticity problems formed with MFEM³. All tests are performed with RAPtor [3] on 512 nodes of Blue Waters. The original linear elasticity system consists of 840 000 unknowns and 65 million non-zeros.

Figure 10 displays the measured and modeled costs for performing a SpMV on each level of the AMG hierarchy. The modeled costs are partitioned into the max-rate model, queue search costs, and network contention penalties. The cost of each SpMV is accurately captured when all model parameters are included, with a large improvement over modeling with only the max-rate model. Furthermore, the model indicates that the majority of communication costs on coarse levels near the middle of the hierarchy are due mainly to queue search costs, motivating efforts for minimizing the number of messages received or posted at any time.

The measured and modeled costs for performing an SpGEMM on each level of the AMG hierarchy are shown in Figure 11. The models are again partitioned into max-rate, queue search, and network contention costs, displaying a large improvement in the model accuracy from the combination of queue search and contention parameters. These models show that more significant costs of the SpGEMMs are from network contention, and while queue search times could be reduced, larger savings would be acquired by reducing the number of bytes to traverse any link.

Combining the max-rate model with queue search and network contention parameters improves the accuracy of the model, but also over-predicts the timings. This over-prediction is a result of using an upper bound on the queue search parameter, corresponding to the cost of posting receives in the opposite order from which messages arrive. The upper bound assumes the $\frac{n \cdot (n+1)}{2}$ elements are searched, while only n elements are searched if the receives are posted in the correct order. The queue search cost was measured for each of these applications by probing for the first available message and computing its position in the queue. The maximum cost on any process was consistently around $\frac{n \cdot n}{3}$, which is approximated much more accurately by the upper bound than the lower. Furthermore, reversing the ordering of the receives results in a different process incurring the maximum queue search penalty, but this cost stays constant.

6 CONCLUSION

Common applications of point-to-point communication typically require a large number of messages to be communicated, with large variability in corresponding messages sizes. Furthermore, there are often combinations of intra-socket, intra-node, and inter-node messages, with the latter commonly traversing multiple links of the network. Therefore, the traditional postal and max-rate models can be improved by splitting standard parameters into on-socket, on-node, and off-node. Additional parameters for bounding the queue search cost and estimating network contention further improve the accuracy of these models. When all parameters are used, the models accurately capture the cost of irregular sparse matrix operations on Blue Waters.

The model parameters are all computed with ping-pong and HighVolumePingPong tests on few nodes, with the majority of tests requiring only a single node while network contention parameters

³<http://mfem.org>

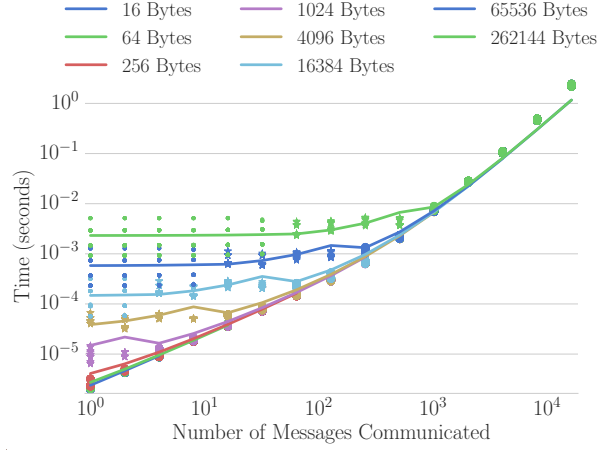


Figure 9: Modeled versus measured times for HighVolumePingPong communication about four Blue Waters Geminis spanning a one-dimensional partition of the networks. The processes on the Geminis and nodes communicate as described in Figure 6. The modeled times are a combination of max-rate models, queue search costs, and the contention parameter.

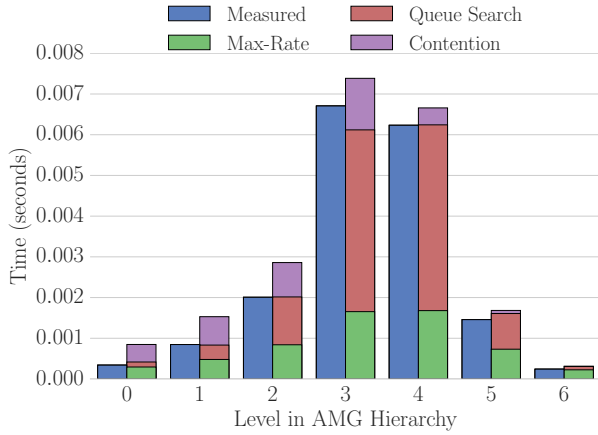


Figure 10: Measured versus modeled times for performing a SpMV on each level of linear elasticity AMG hierarchy.

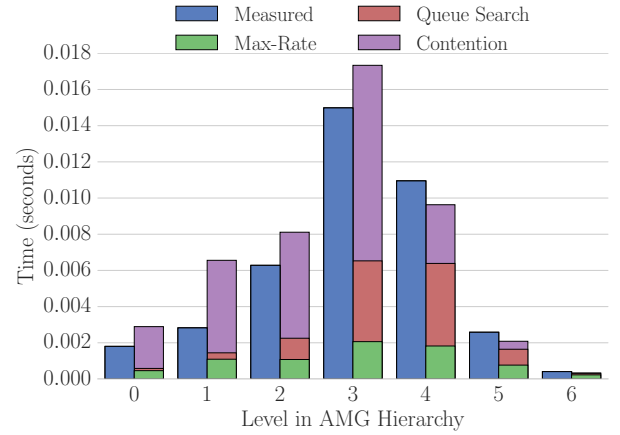


Figure 11: Measured versus modeled times for performing a SpGEMM on each level of linear elasticity AMG hierarchy.

are calculated on up to eight nodes. However, these parameters remain accurate when modeling sparse matrix operations on 512 nodes, indicating this model can be extended to large core counts with no additional work.

This model may need alterations to accurately capture the costs on different systems. For example, architectures with only a single socket per node, such as Blue Gene/Q machines, will only need to partition the max-rate model into on-node and off-node messages. Furthermore, MPI implementations with an optimized queue search will require alternative queue search penalties, while implementations with dynamic message routing will require a block of communicating nodes to capture network contention in the test in Figure 6.

Limitations for this model include using the upper bound for queue search time and assuming the process domain is mapped to

a cube for network contention. The queue search time will overestimate the actual cost, while the accuracy of the network contention penalty can vary with actual partition acquired.

These models can be further extended to include topology-aware parameters, such as additional latency required for messages traversing a large number of links. Furthermore, the models motivate future directions for tested applications, such as methods for reducing queue search time in SpMVs and network contention in SpGEMMs.

ACKNOWLEDGMENTS

This research is part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing

Applications. This material is based in part upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant Number DGE-1144245. This material is based in part upon work supported by the Department of Energy, National Nuclear, under Award Number DE-NA0002374.

REFERENCES

- [1] Tarun Agarwal, Amit Sharma, and Laxmikant V. Kalé. 2006. Topology-aware Task Mapping for Reducing Communication Contention on Large Parallel Machines. In *Proceedings of the 20th International Conference on Parallel and Distributed Processing (IPDPS'06)*. IEEE Computer Society, Washington, DC, USA, 145–145. <http://dl.acm.org/citation.cfm?id=1898953.1899075>
- [2] Albert Alexandrov, Mihai F. Ionescu, Klaus E. Schauser, and Chris Scheiman. 1995. LogGP: Incorporating Long Messages into the LogP Model – One Step Closer Towards a Realistic Model for Parallel Computation. In *Proceedings of the Seventh Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA '95)*. ACM, New York, NY, USA, 95–105. <https://doi.org/10.1145/215399.215427>
- [3] Amanda Bienz and Luke N. Olson. 2017. RAPtor: parallel algebraic multigrid v0.1. <https://github.com/lukeolson/raptor> Release 0.1.
- [4] Brett Bode, Michelle Butler, Thom Dunning, Torsten Hoefler, William Kramer, William Gropp, and Wen-mei Hwu. 2013. The Blue Waters Super-System for Super-Science. In *Contemporary High Performance Computing*. Chapman and Hall/CRC, 339–366. <https://doi.org/10.1201/b14677-16>
- [5] James Cownie and William Gropp. 1999. A Standard Interface for Debugger Access to Message Queue Information in MPI. In *Recent Advances in Parallel Virtual Machine and Message Passing Interface*, Jack Dongarra, Emilio Luque, and Tomás Margalef (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 51–58.
- [6] David E. Culler, Richard M. Karp, David Patterson, Abhijit Sahay, Eunice E. Santos, Klaus Erik Schauser, Ramesh Subramonian, and Thorsten von Eicken. 1996. LogP: A Practical Model of Parallel Computation. *Commun. ACM* 39, 11 (Nov. 1996), 78–85. <https://doi.org/10.1145/240455.240477>
- [7] Gábor Dóza, Sameer Kumar, Pavan Balaji, Darius Buntinas, David Goodell, William Gropp, Joe Ratterman, and Rajeev Thakur. 2010. Enabling Concurrent Multithreaded MPI Communication on Multicore Petascale Systems. In *Proceedings of the 17th European MPI Users' Group Meeting Conference on Recent Advances in the Message Passing Interface (EuroMPI'10)*. Springer-Verlag, Berlin, Heidelberg, 11–20. <http://dl.acm.org/citation.cfm?id=1894122.1894125>
- [8] Mario Flajslik, James Dinan, and Keith D. Underwood. 2016. Mitigating MPI Message Matching Misery. In *High Performance Computing*, Julian M. Kunkel, Pavan Balaji, and Jack Dongarra (Eds.). Springer International Publishing, Cham, 281–299.
- [9] Matthew I. Frank, Anant Agarwal, and Mary K. Vernon. 1997. LoPC: Modeling Contention in Parallel Algorithms. In *Proceedings of the Sixth ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPOPP '97)*. ACM, New York, NY, USA, 276–287. <https://doi.org/10.1145/263764.263803>
- [10] P. B. Gibbons. 1989. A More Practical PRAM Model. In *Proceedings of the First Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA '89)*. ACM, New York, NY, USA, 158–168. <https://doi.org/10.1145/72935.72953>
- [11] William Gropp, Luke N. Olson, and Philipp Samfass. 2016. Modeling MPI Communication Performance on SMP Nodes: Is It Time to Retire the Ping Pong Test. In *Proceedings of the 23rd European MPI Users' Group Meeting (EuroMPI 2016)*. ACM, New York, NY, USA, 41–50. <https://doi.org/10.1145/2966884.2966919>
- [12] T. Hoefler, T. Schneider, and A. Lumsdaine. 2010. LogGOPSIm - Simulating Large-Scale Applications in the LogGOPS Model. In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*. ACM, Chicago, Illinois, 597–604.
- [13] Nikhil Jain, Abhinav Bhatele, Michael P. Robson, Todd Gamblin, and Laxmikant V. Kale. 2013. Predicting Application Performance Using Supervised Learning on Communication Features. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC '13)*. ACM, New York, NY, USA, Article 95, 12 pages. <https://doi.org/10.1145/2503210.2503263>
- [14] Rainer Keller and Richard L. Graham. 2010. Characteristics of the Unexpected Message Queue of MPI Applications. In *Proceedings of the 17th European MPI Users' Group Meeting Conference on Recent Advances in the Message Passing Interface (EuroMPI'10)*. Springer-Verlag, Berlin, Heidelberg, 179–188. <http://dl.acm.org/citation.cfm?id=1894122.1894146>
- [15] Csaba Andras Moritz and Matthew I. Frank. 1998. LoGPC: Modeling Network Contention in Message-passing Programs. In *Proceedings of the 1998 ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '98/PERFORMANCE '98)*. ACM, New York, NY, USA, 254–263. <https://doi.org/10.1145/277851.277933>
- [16] N. Saboo, A. K. Singla, J. M. Unger, and L. V. Kale. 2001. Emulating petaflops machines and blue gene. In *Proceedings 15th International Parallel and Distributed Processing Symposium. IPDPS 2001*. IEEE, San Francisco, CA, USA, 2084–2091. <https://doi.org/10.1109/IPDPS.2001.925206>
- [17] T. Schneider, T. Hoefler, and A. Lumsdaine. 2009. *ORCS: An Oblivious Routing Congestion Simulator*. Technical Report 675. Indiana University.
- [18] L. A. Steffanel. 2006. Modeling Network Contention Effects on All-to-All Operations. In *2006 IEEE International Conference on Cluster Computing*. IEEE, Barcelona, Spain, 1–10. <https://doi.org/10.1109/CLUSTER.2006.311889>
- [19] K. D. Underwood and R. Brightwell. 2004. The impact of MPI queue usage on message latency. In *International Conference on Parallel Processing, 2004. ICPP 2004*. 152–160 vol.1.
- [20] Keith D. Underwood, K. Scott Hemmert, Arun Rodrigues, Richard Murphy, and Ron Brightwell. 2005. A Hardware Acceleration Unit for MPI Queue Processing. In *Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05) - Papers - Volume 01 (IPDPS '05)*. IEEE Computer Society, Washington, DC, USA, 96.2–. <https://doi.org/10.1109/IPDPS.2005.30>
- [21] J. A. Zounmevo and A. Afsahi. 2012. An Efficient MPI Message Queue Mechanism for Large-scale Jobs. In *2012 IEEE 18th International Conference on Parallel and Distributed Systems*. 464–471.