# Evaluating Communication Performance of BlueGene/Q and Cray XE6 Supercomputers

Huy Bui[†] (abui4@uic.edu), Venkatram Vishwanath[°] (venkat@anl.gov),
Jason Leigh[†] (spiff@uic.edu), Michael E. Papka[°*] (papka@anl.gov)

[†] Electronic Visualization Laboratory, University of Illinois at Chicago, [°] Argonne National Laboratory,
[*] Northern Illinois University

## Abstract

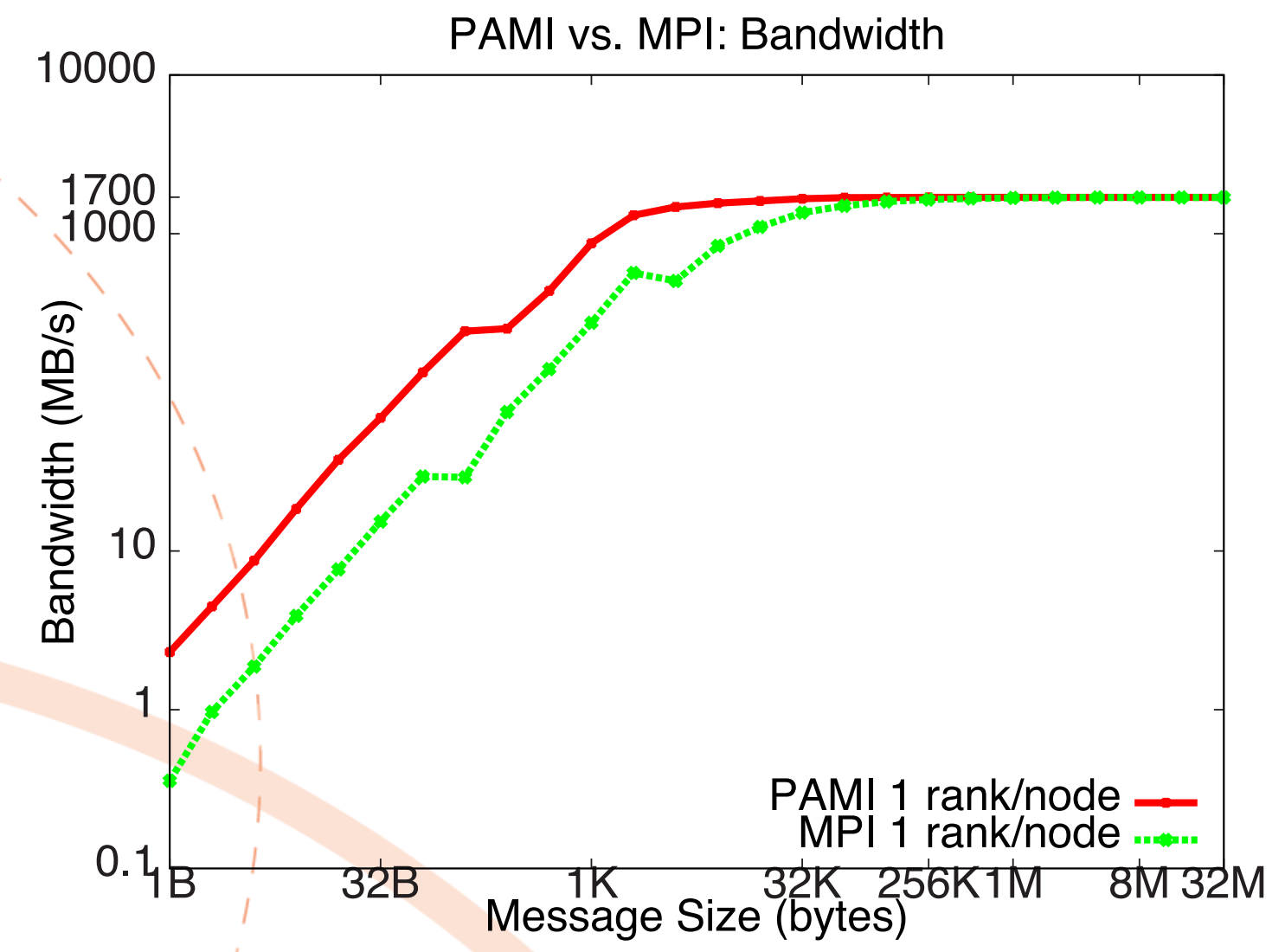
Communication performance is of paramount importance to high performance computing (HPC) applications. MPI is widely used in HPC due to its portability across platforms. However, the performance of MPI implementations on large-scale supercomputers is significantly impacted by factors including its inherent buffering, type checking, and other control overheads.

At the same time, we are witnessing the advent of intelligent network interfaces and system interconnects together with support for paradigms such as active messages. The Parallel Active Message Interface (PAMI) library for the IBM Blue Gene/Q system and the User Generic Network Interface (uGNI ) for the Cray XE6 system are communication libraries enabling one to fully exploit the features of the underlying network interconnects. In this poster, we evaluate the performance of MPI with PAMI and uGNI and demonstrate that these libraries achieve significant improvements in communication performance over MPI.

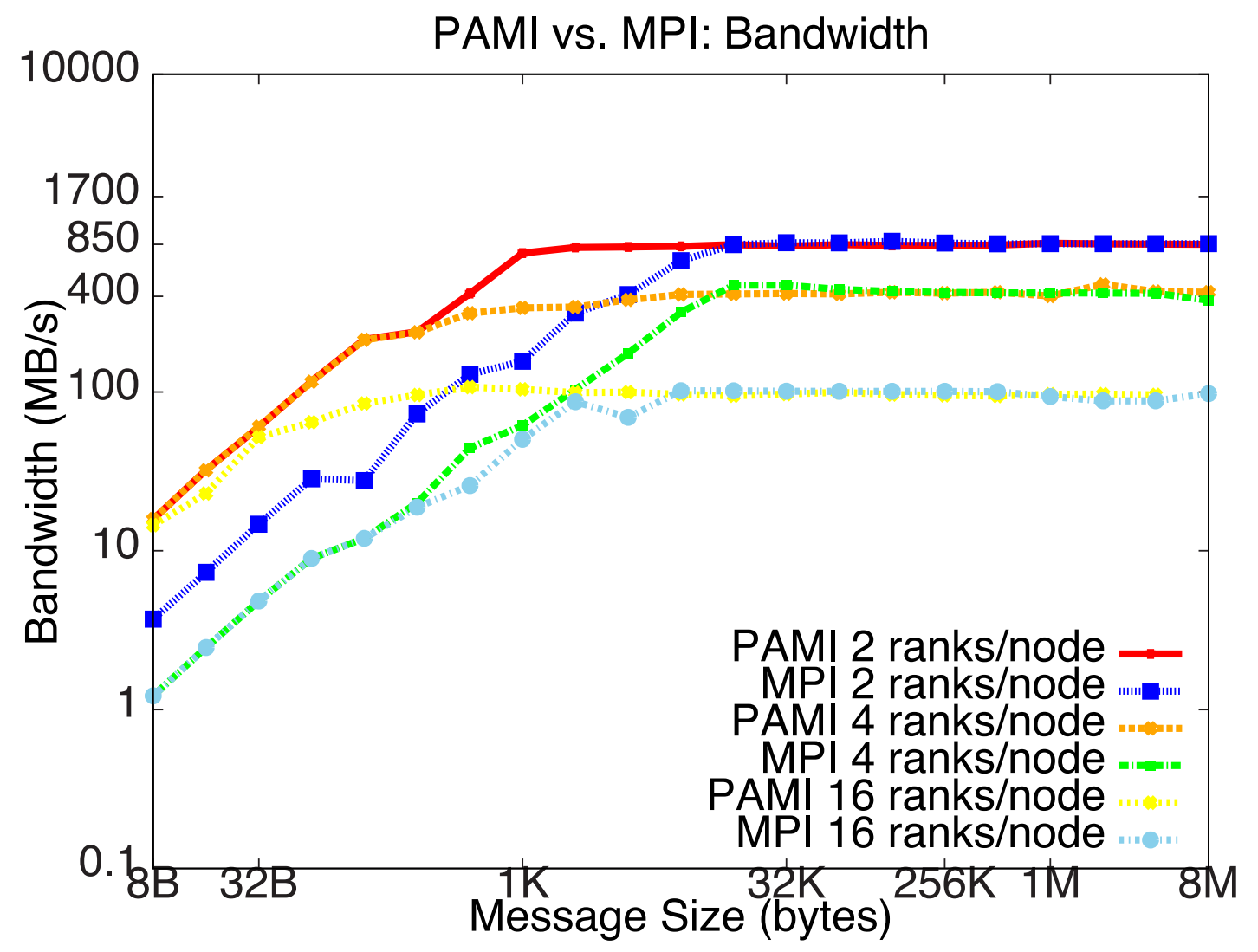
We are working towards creating abstractions that hide the complexities found in current system interconnects, and expose a simple API to enable applications and middleware developers fully exploit the underlying features.

## PAMI vs. MPI on BG/Q: Two-sided Communication

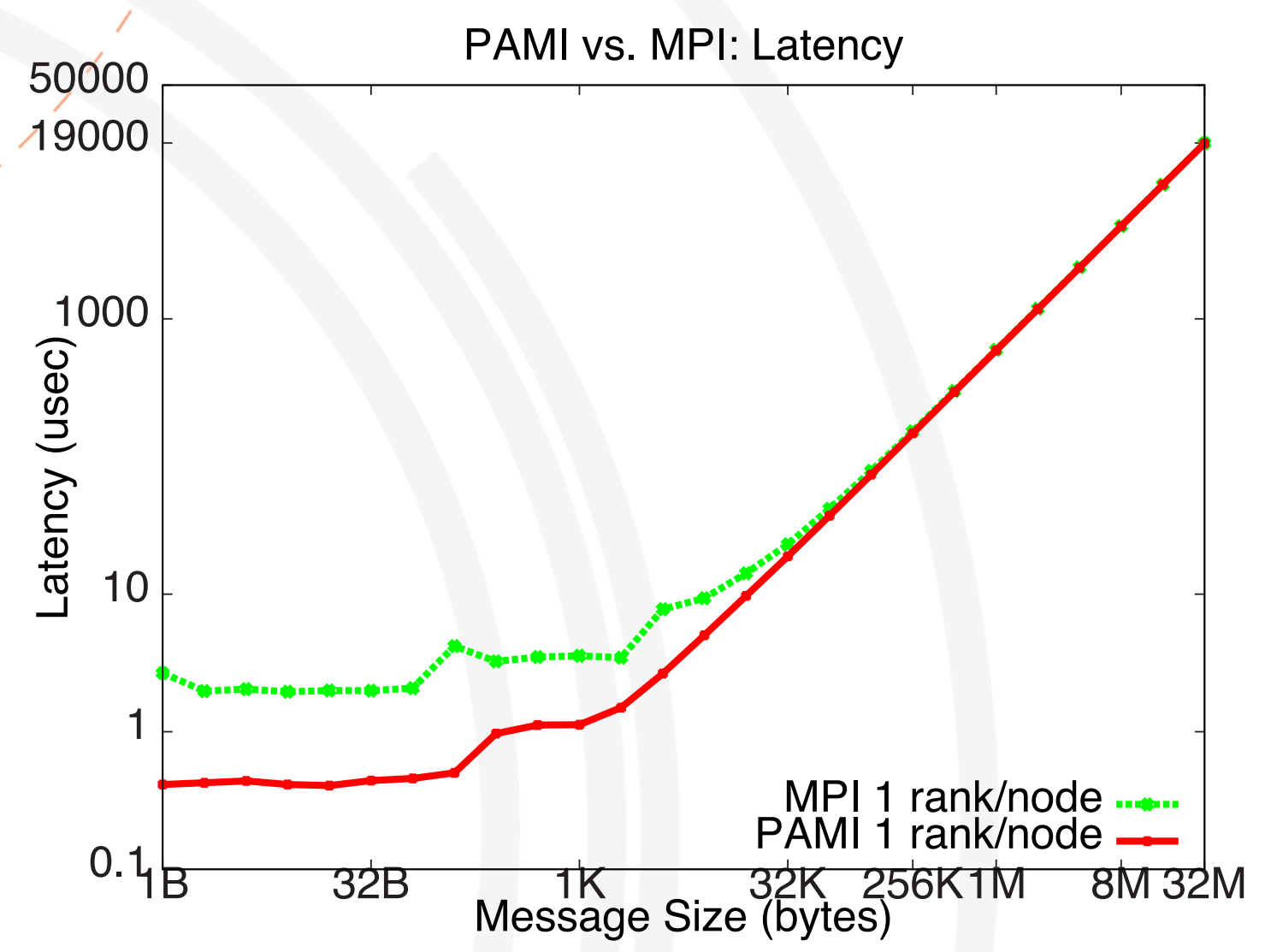
To transfer contiguous data asynchronously, PAMI uses PAMI Send for large messages and PAMI Send intermediate for messages less than 128 bytes. In case of MPI, we use ISend/Recv.



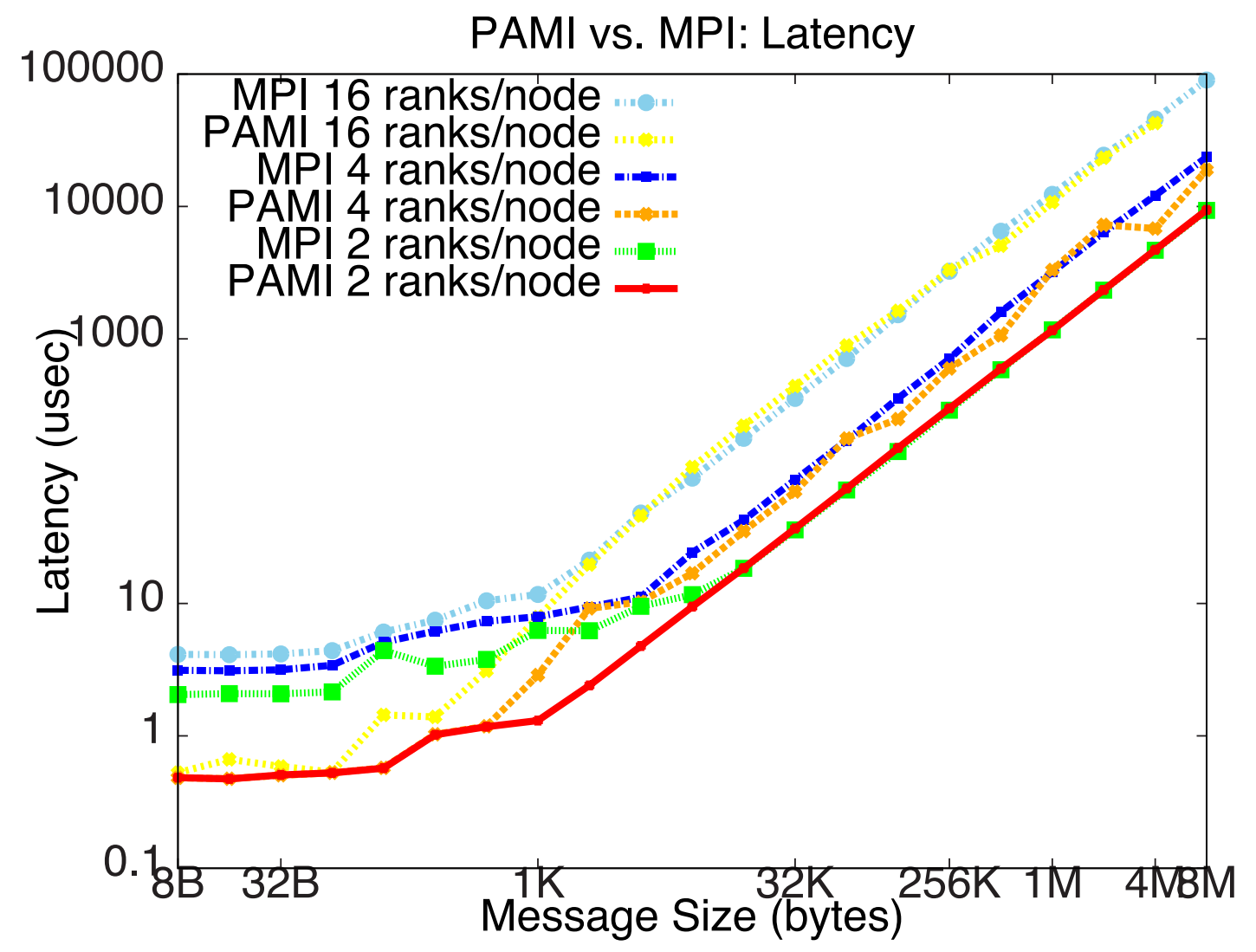
Achievable bandwidth for non-blocking point-to-point communication as we vary the message size.



Achievable bandwidth for non-blocking point-to-point communication as we vary the number of ranks per node and the message size.



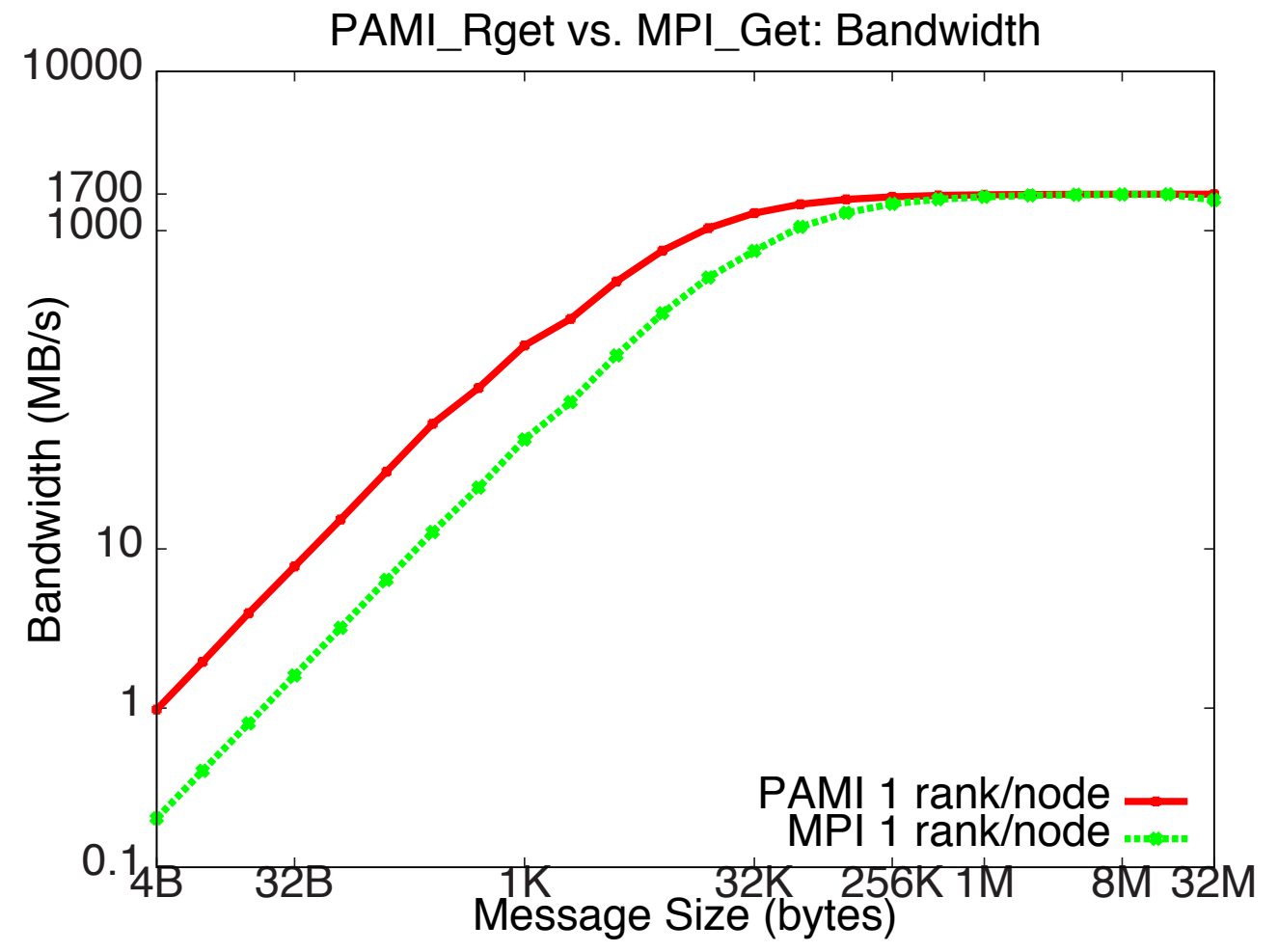
Latency of non-blocking point-to-point communication as we vary the message size.



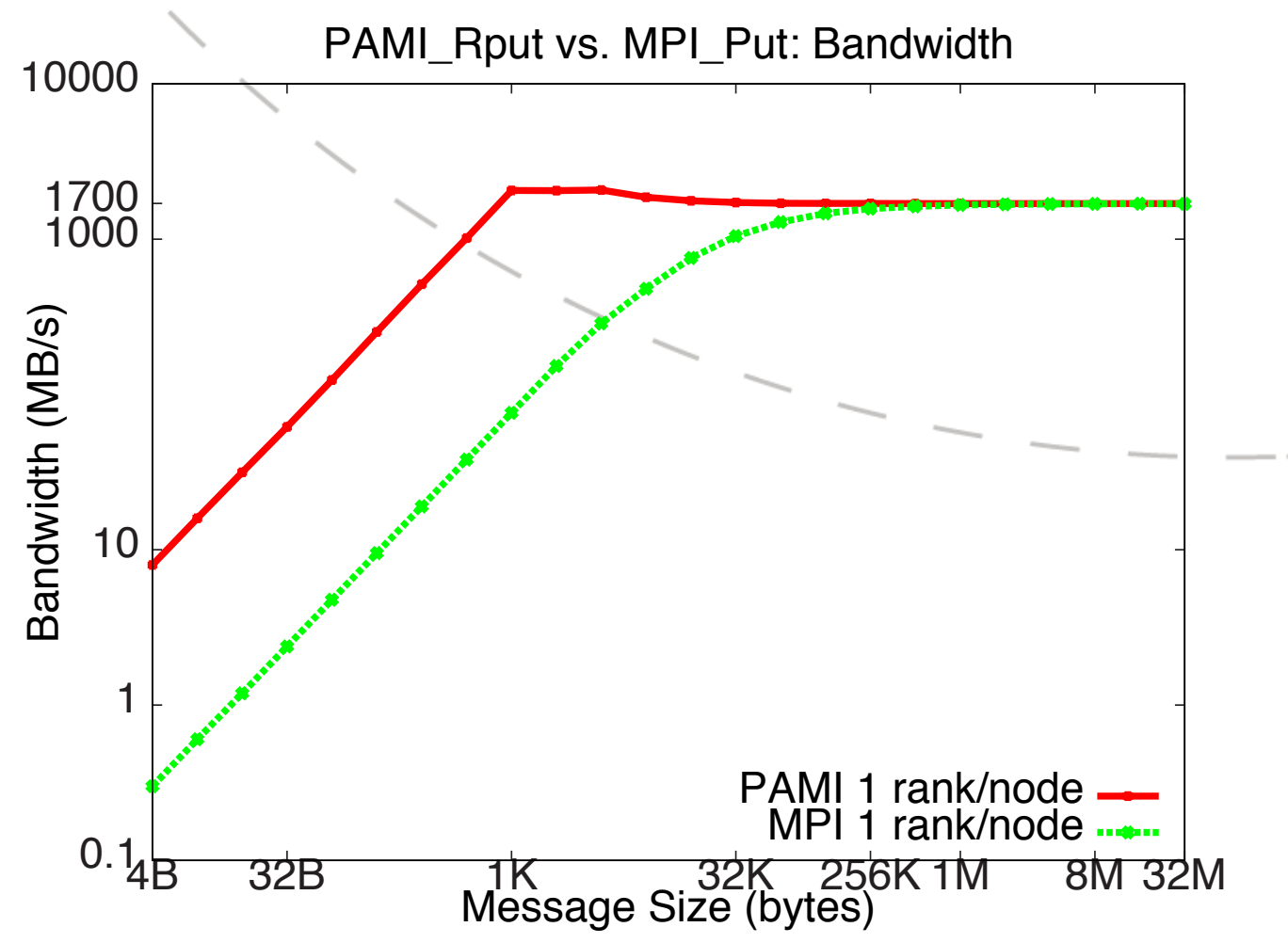
Latency of non-blocking point-to-point communication as we vary the number of ranks per node and the message size.

## PAMI vs. MPI on BG/Q: One-sided Communication

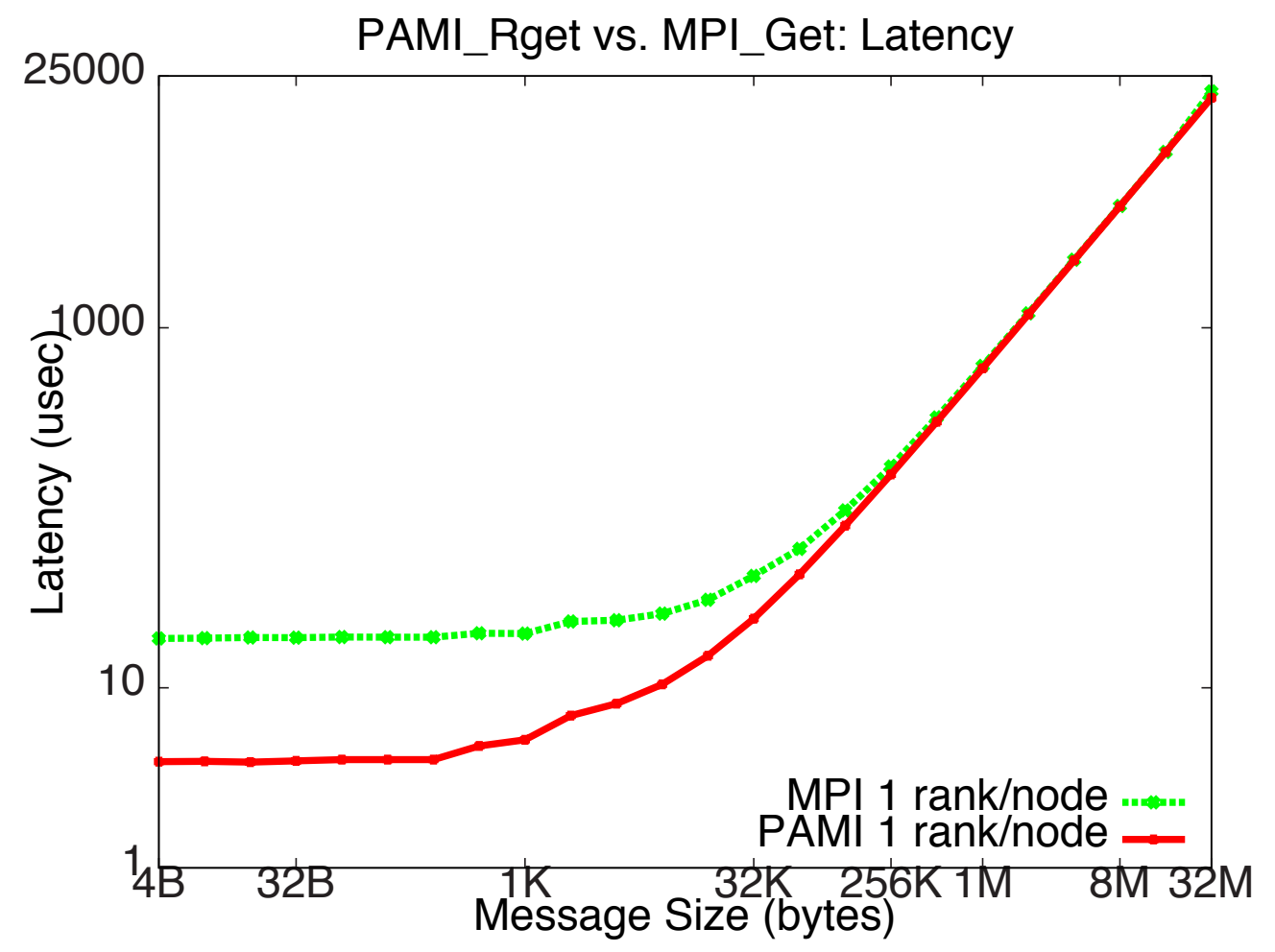
One-sided communication is becoming more ubiquitous in HPC. We compare the performance of PAMI and MPI on BG/Q for one-sided communication. We compare PAMI Rget against MPI Get, and PAMI RPut against MPI Put.



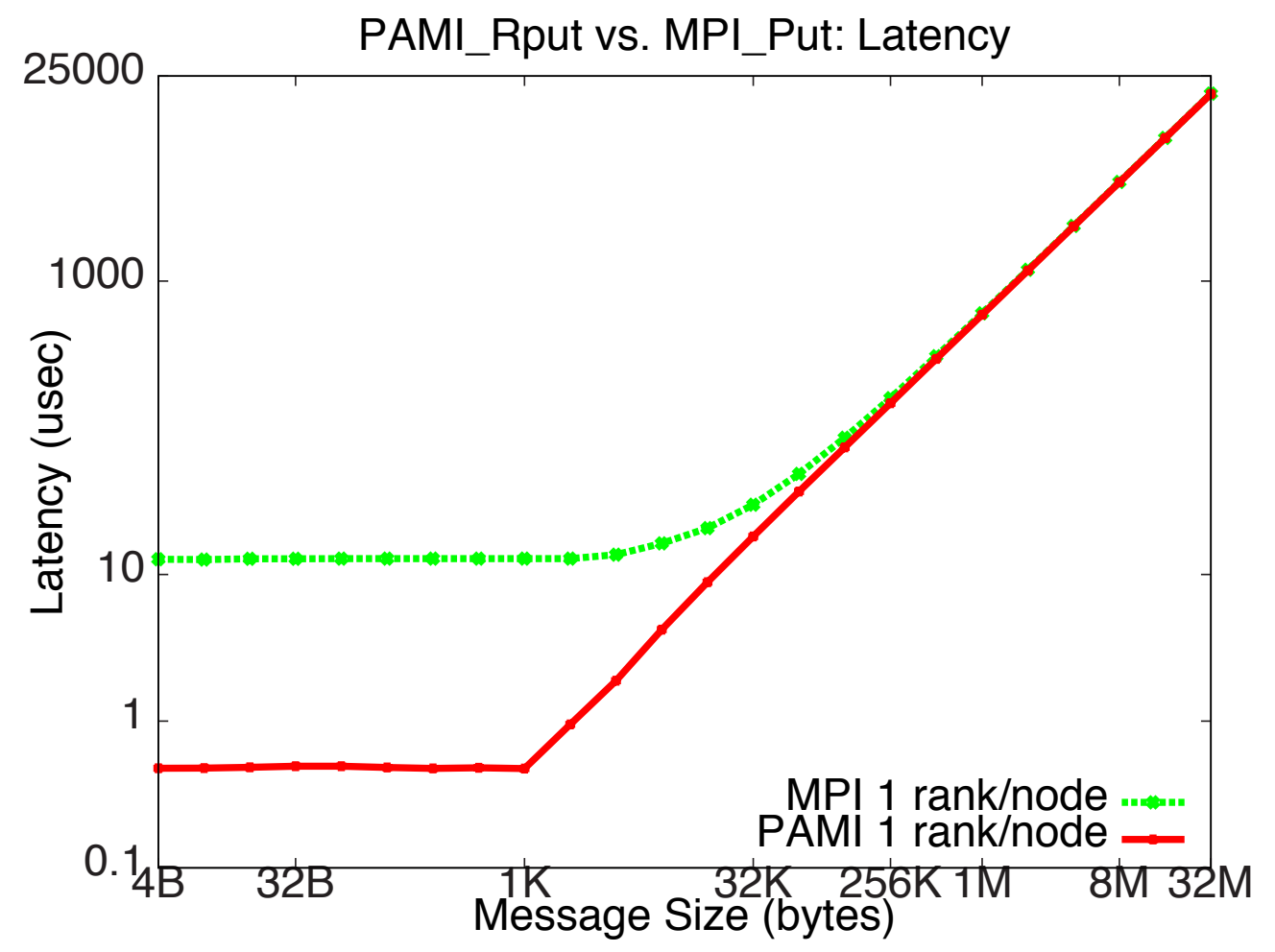
Achievable bandwidth for Get operations as we vary the message size.



Achievable bandwidth for Put operations as we vary the message size.



Latency of Get operations as we vary the message size.

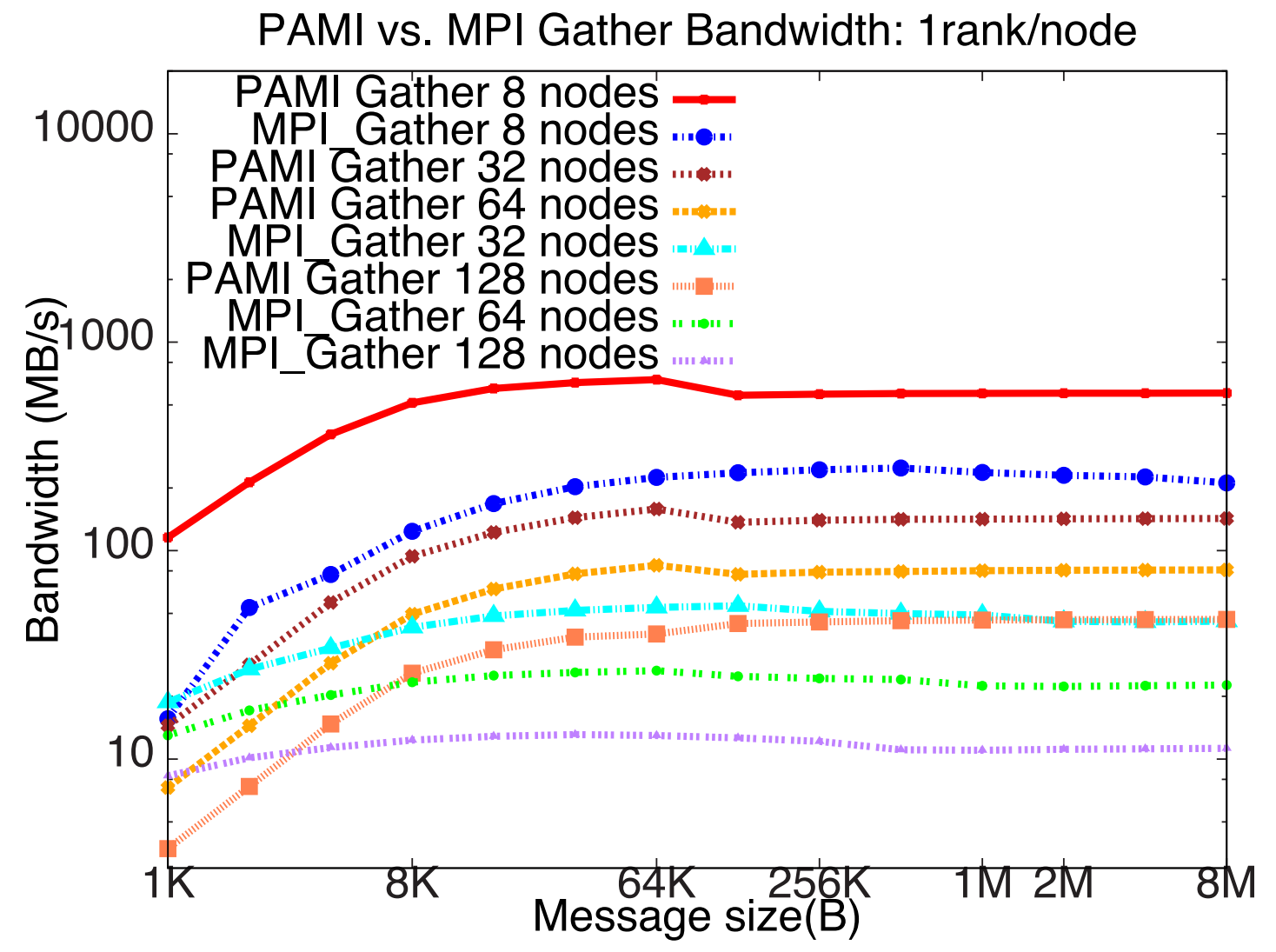


Latency of Put operations as we vary the message size.

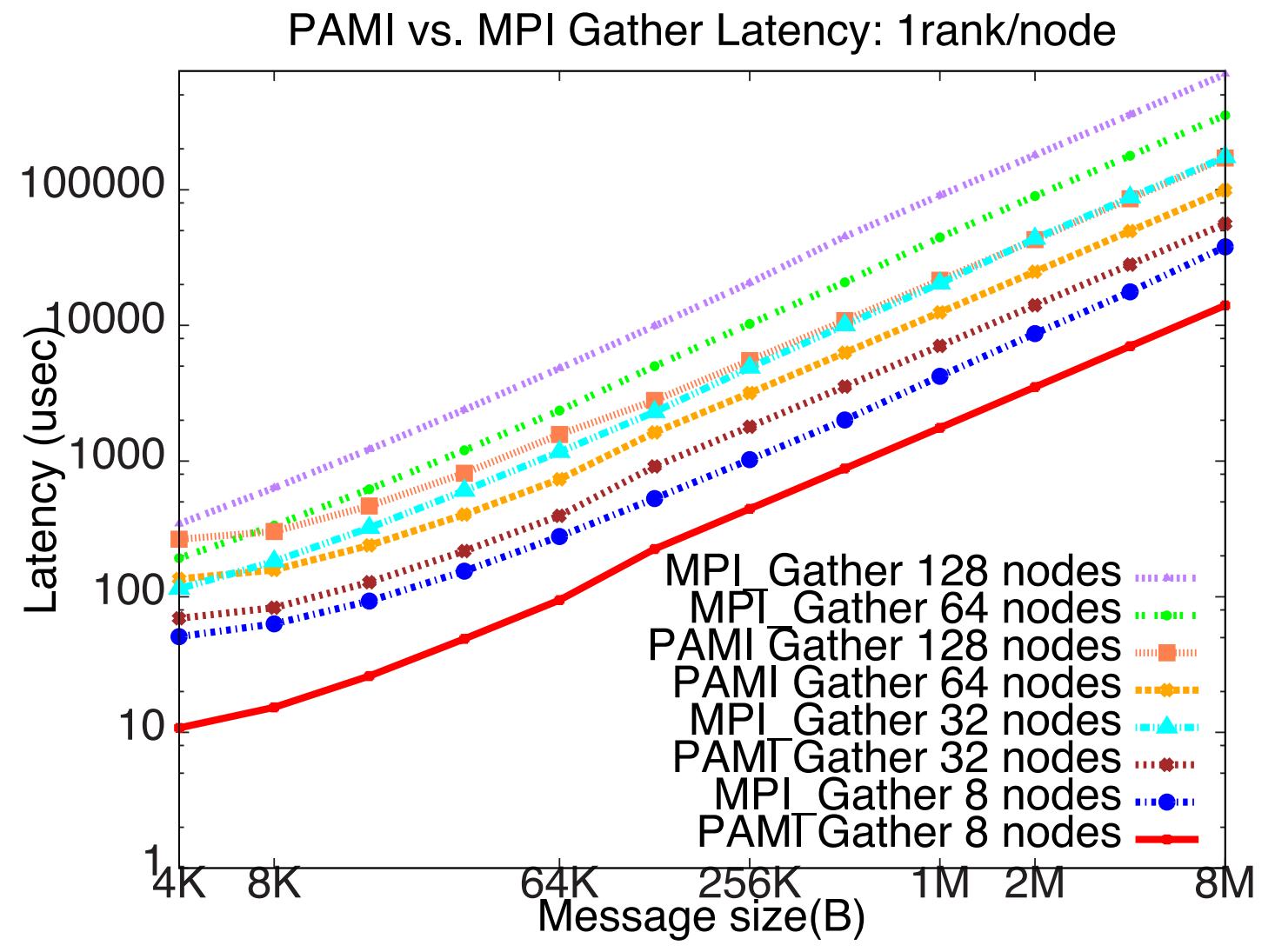
In both cases, we achieve better performance with PAMI in comparison to MPI for message sizes less than 512KB. Put operations with PAMI have lower latency and higher bandwidth.

## PAMI vs. MPI on BG/Q: Collective Communication

We compare the performance of PAMI and MPI for collective communication (Gather) on BG/Q.



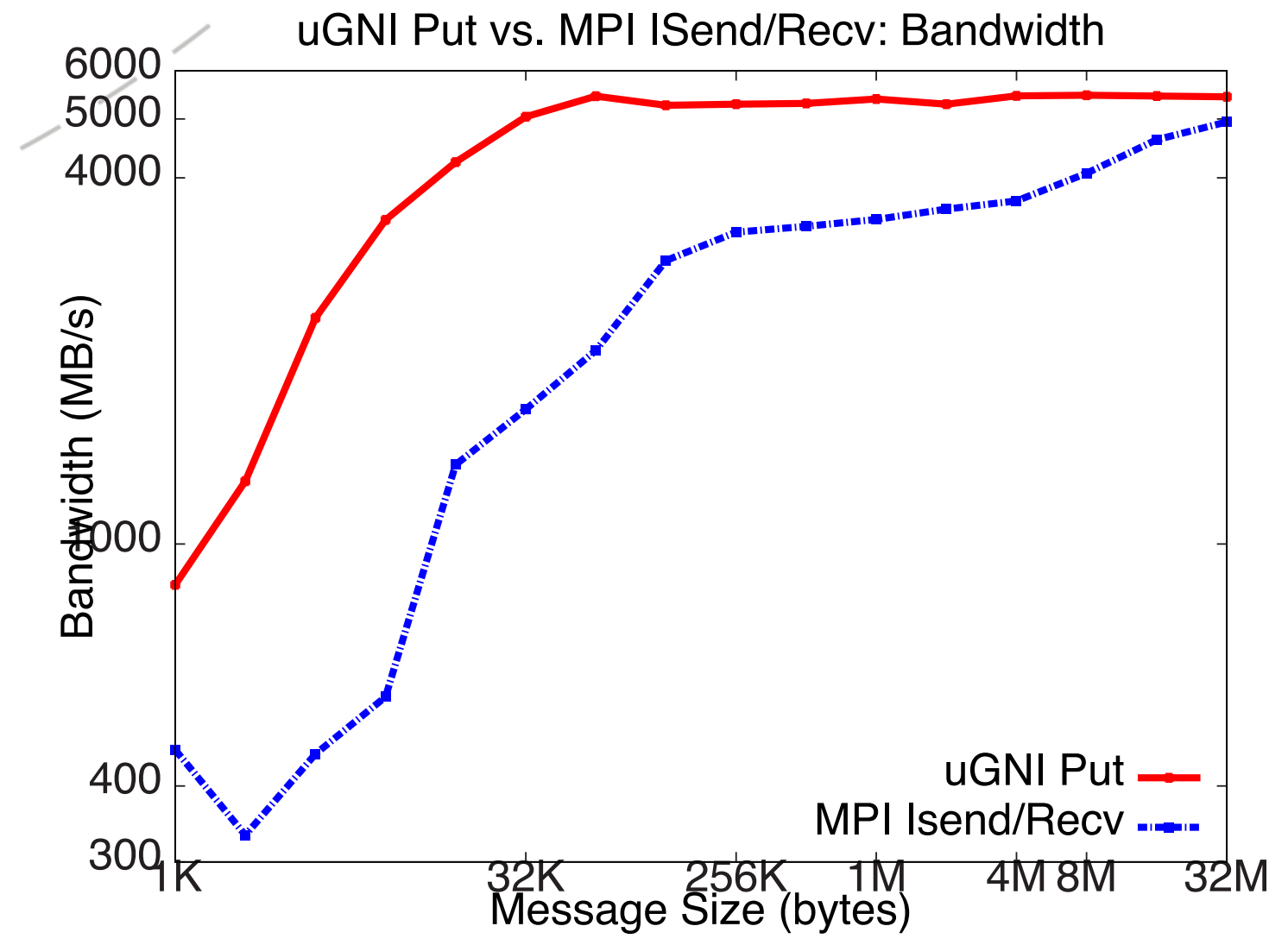
Achievable bandwidth for Gather using PAMI and MPI as we vary the number of nodes and message size.



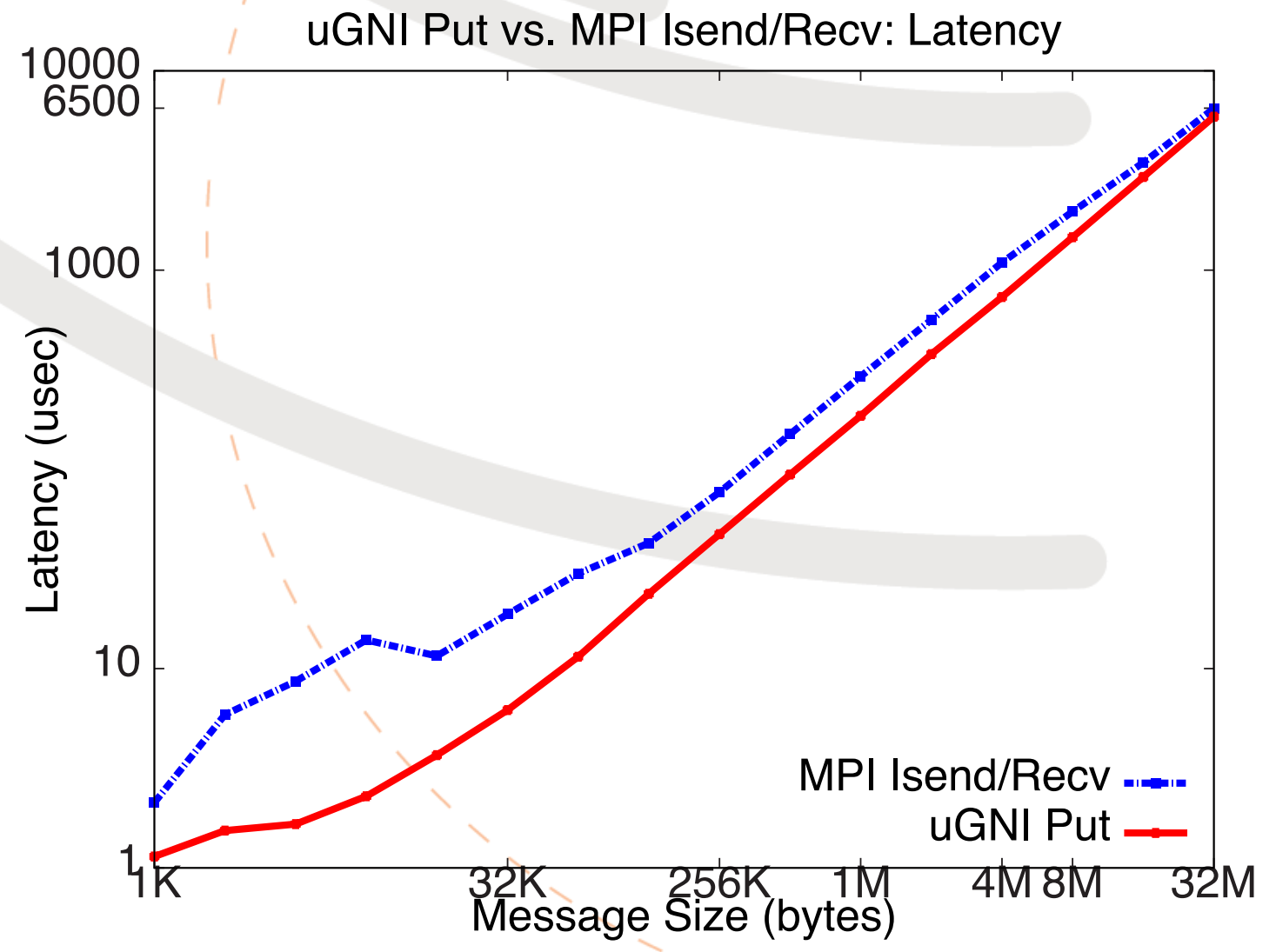
Latency of Gather using PAMI and MPI as we vary the number of nodes and message size.

## uGNI vs. MPI on Cray XE6

Cray XE6 machine uses the Gemini interconnect. On the Cray XE6 machine, MPI uses a lower-library named uGNI. uGNI does not provide two-sided communication. uGNI supports only one-sided distributed remote memory access (RMA) operations for both long and short messages.



Achievable bandwidth of MPI Isend/Recv and uGNI Put as we vary the message size.



Latency of MPI Isend/Recv and uGNI Put as we vary the message size.

uGNI achieve up to 4X improvement in performance over MPI.

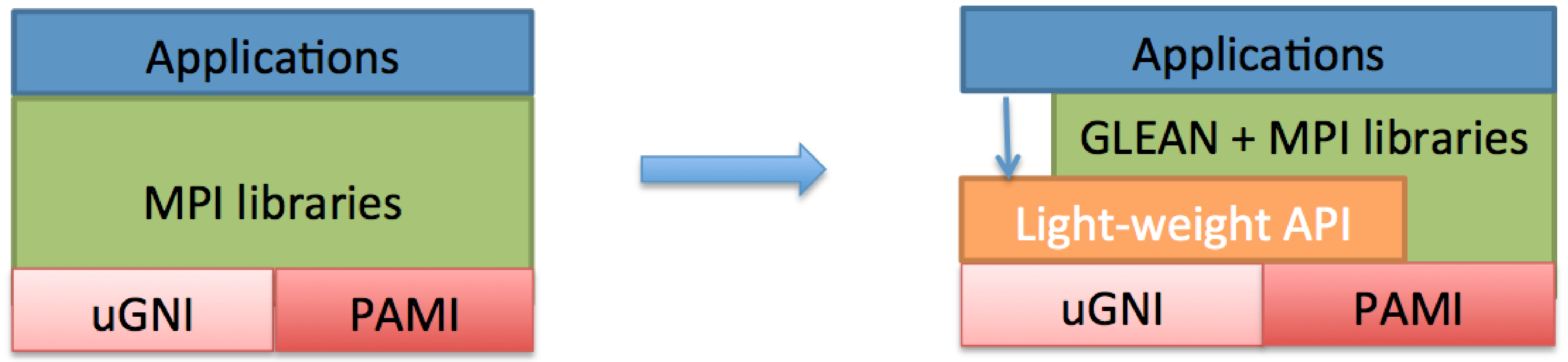
## Conclusions and Future Work

### Conclusion

Achieving low latency and higher bandwidth is of critical importance to HPC applications. Lower-level communication libraries enable one to fully exploit the novel features in the network interfaces and system interconnects, and are key for improved performance. PAMI on the IBM Blue Gene/Q and uGNI on the Cray XE/6 system demonstrate up to 3-4 times improvement for latency and 3-4 times for bandwidth over MPI. Also, these mechanisms provide for an improved collective communication performance.

### Future work

We plan to develop a light-weight API incorporating common features in current and upcoming system interconnects. A library designed with this API can be leveraged directly by applications, MPI implementations, and runtime libraries for improved performance.



MPI to light-weight API and GLEAN.

We plan to incorporate this API in GLEAN[1], a simulation-time analysis and I/O acceleration framework, to accelerate time-to-insight on on supercomputing systems.

[1] V. Vishwanath, M. Hereld, V. Morozov, and M. E. Papka, "Topology-aware data movement and staging for I/O acceleration on Blue Gene/P supercomputing systems", In Proceedings of the IEEE/ACM International Conference for High Performance Computing, Networking, Storage and Analysis (SC 2011), Seattle, USA, November 2011.