Enhancing NIC Performance for MPI using Processing-in-Memory

Arun Rodrigues, Richard Murphy, Ron Brightwell, and Keith D. Underwood
Sandia National Laboratories†
P.O. Box 5800
MS-1110
Albuquerque, NM 87185-1110
Email: {afrodri, rcmurph, rbbrigh, kdunder}@sandia.gov

Abstract— **Processing-in-Memory (PIM) technology encompasses a broad range of research leveraging a tight coupling of memory and processing. The most unique features of the technology are extremely wide paths to memory, extremely low latency to memory, and wide functional units. Many PIM researchers are also exploring extremely fine-grained multi-threading capabilities. This paper explores a mechanism for leveraging these features of PIM technology to enhance commodity architectures in the most mundane seeming of ways: accelerating MPI. Modern network interfaces leverage simple processors to offload portions of the MPI semantics, particularly the management of posted receive and unexpected message queues. Without adding cost, using PIMs in the network interface can enhance performance without having to increase clock frequency. The results are a significant decrease in latency and increase in small message bandwidth, particularly when long queues are present.**

I. Introduction

A novel technology that is currently receiving widespread attention in high-performance computing is known as Processing-in-Memory (PIM) [1], [2], [3], [4] (also called Intelligent RAM [5]). The objective of PIM technology is to merge a processor with memory to avoid the impending memory wall [6]. Researchers have begun to envision a future where there is a sea of memory with processors scattered throughout [7], [8], [2]. Those processors would be simple with wide paths to memory and wide functional units [9]. They would be inherently multithreaded [10] with extremely lightweight thread context and thread invocation mechanisms. This paper, however, is not about the way PIM technology is going to revolutionize the primary processor in supercomputers. It is not even going to discuss how PIMs may completely change the way memory systems are built. Instead, this paper will discuss how PIMs could make a difference in supercomputers in the least expected of places: the network interface.

Modern network interfaces offload much of the MPI processing [11], [12], [13]. This is a natural migration in light of research indicating that network overhead (the time the application processor spends handling messages traffic [14], [15]) has the single largest impact on application performance [16]. Unfortunately, the work that is offloaded includes traversing the posted receive queue. This queue can grow quite long [17], [18]. As the queue grows, the time to traverse it also grows due to the inherently low clock speed and simplicity of the processor on the NIC [19]. Thus, another important measure of network performance, the gap (time between when new messages can be initiated [14]) begins to grow.

One solution is to build a better NIC processor. Such a processor needs to decrease the time to search a linked list to reduce the latency of messages. It needs to overlap multiple searches of the list to increase throughput. It needs to... look a lot like a PIM. A PIM has low latency access to memory, so it traverses linked lists well. Memory access is wide and can be operated on with a wide ALU, so multiple list entries can be searched rapidly with a handful of instructions. The PIM is multithreaded and has extremely fine-grained locking

† Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

support, so multiple simultaneous queue traversals can be used to leverage the increase in bandwidth and hide what little latency exists. Best of all, PIMs are simple, so the area and the NRE devoted to them will be minimal. In short, PIMs are a natural fit to the needs of the embedded processor on the NIC.

This paper explores the details of how a PIM could be used to accelerate NIC-based MPI processing. A model of a modern NIC was created as a baseline for comparison. In the standard NIC, a PowerPC 440 embedded processor was used. A PIM based on similar technology was also used. An MPI implementation was modified to leverage the wide memory accesses, the wide ALU operations, and the multithreading capabilities of the PIM. The result was a dramatic decrease in message latency in the presence of long queues and an increase in the achievable throughput.

In Section II, related work is described. Following that the hardware (Section III) and MPI implementation (Section IV) are described. The evaluation methodology is described in Section V and the results are presented in Section VI. The paper wraps up with conclusions and future work in Section VII.

II. Related Work

This paper explores PIM-based acceleration of match processing for the MPI posted receive queue. This is a natural progression of previous work on utilizing PIM hardware to support MPI and exploring those areas of MPI that may benefit from hardware assisted processing. We have previously addressed issues in supporting MPI functionality for a variety of PIM-based system architectures [20]. We have also proposed augmenting a traditional network interface with custom hardware for accelerating MPI queue processing [18]. To our knowledge, there is no previous work on strategies for optimizing MPI queue processing in published research.

There is, however, a significant amount of previous work aimed at utilizing programmable network interface hardware, such as Quadrics [12], Myrinet [21], and InfiniBand [22], for optimizing MPI performance. These optimizations include those aimed at offloading MPI protocol processing [23], [24], [13], [25], using hardware capabilities for efficient data movement (especially collective operations [26], [27], [28], [29], [30]), and using scatter/gather functionality for handling non-contiguous data transfers [31], [32].

As a processing technology, PIM has been extensively examined for use in supercomputing for almost a decade[7], [1], [8], [4], [3], [33], [34], [2]. It has also been examined in detail in the context of embedded systems[35], [36]. However, to date, we know of no analysis of PIMs used to enhance NIC performance. Given the high bandwidth access to memory made possible by a PIM, combined with the throughput-oriented (latency tolerant) architecture proposed in the context of supercomputing, accelerating the performance of MPI processing seems like a natural match.

III. Hardware Overview

The baseline NIC (shown in Figure 1) is representative of numerous modern network interfaces, including the Quadrics Elan3 [12] and Elan4 networks, the Myrinet network [13], [21], and, most recently, the Red Storm system developed by Cray and Sandia [11], [37]. These NICS interface to the network fabric on one side and a high-performance interface to the microprocessor (such as HyperTransport or PCI-Express) on the other. What delivers the high-performance is the logic between the two interfaces: two DMA engines (one in each direction) driven by local processing on the network interface. The local processing on the network interface is a microprocessor (sometimes two) with either internal RAM (as shown in Figure 1(a) and found in Red Storm [11], [37]) or external RAM (as typically found in commodity networks like Myrinet [13], [21] and Quadrics [12]).

A. PIM-Based NIC Enhancement

Commodity processors are limited by their architecture. They can only process one input at a time (they are single threaded), and they have narrow functional units requiring many instructions to process a single list entry. In contrast, PIMs can use fine-grain locking and multi-threading support to concurrently perform list walks for multiple incoming messages simultaneously. They also have wide functional units that allow them to compare multiple list entries simultaneously. PIMs use reduced core complexity (at the expense of slightly

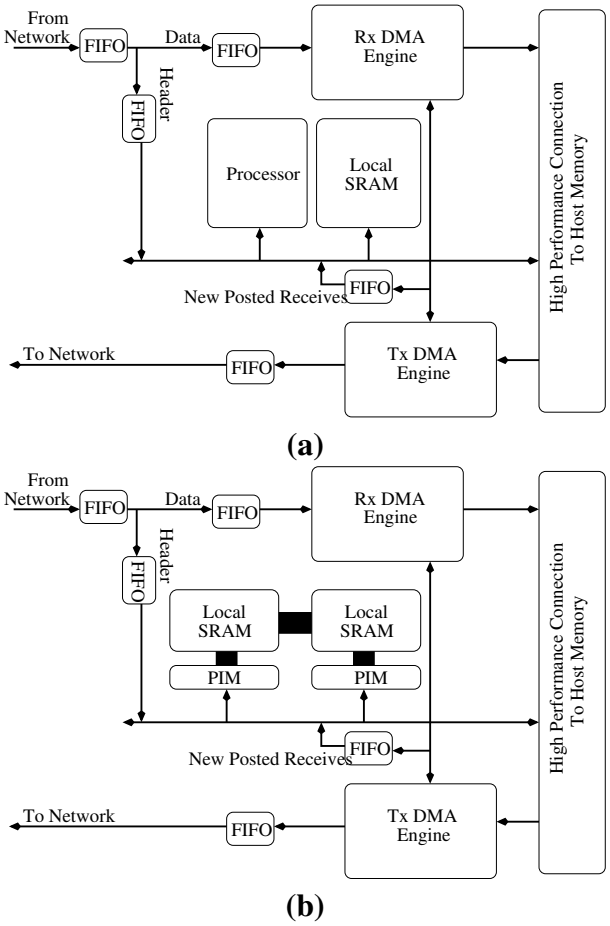

Fig. 1. A high-performance network interface enhanced with **(a)** a processor, and **(b)** multiple PIMs

lower instruction per cycle (IPC) rate for some workloads) to dramatically reduce core size. Thus, PIMs have the potential to bring significant gains in message latency and message throughput when the posted receive queue in MPI is long.

The proposed modification is shown in Figure 1(b). The embedded SRAM becomes part of the PIMs and is accessed in extremely wide words. The PIMs are interfaced to the rest of the NIC in exactly the same way as the conventional microprocessor would be with the exact same amount of memory backing it. A single PIM, however, is much smaller. Thus, it is possible to include two PIMs. Each of these PIMs has equal wide access to both memories (contention is now possible) as indicated by the wide connection between the SRAMs. Virtually all PIM architectures consider multiple cooperating PIMs in a single memory, so two PIMs is a reasonable design point to consider.

B. Relative Area Requirements

Silicon is still a precious resource and is the standard for comparing two NIC designs. It is only fair to compare two similar designs if they are of comparable area. Thus, it is critical to understand the area trade-off when comparing a PIM-based NIC to a conventional NIC. As the first significant PIM system to incorporate the essential features examined, PIM Lite [9] provides the hard data needed for comparison. Brockman [38] compares the relative area of conventional processor cores, PIM Lite, and conventional memories in a process-independent fashion. The 128-bit PIM Lite implementation has a process-independent equivalent area to 10.3 KB of SRAM. Since this paper uses a 256-bit data path, the area is doubled to 20.6 KB of SRAM. Assuming that additional features might be desired in the PIM core, we double this estimate and add a 4 KB instruction cache to yield a 45.2 KB SRAM equivalence. The design evaluated here uses two PIM cores (90.4 KB of SRAM equivalence) and 384K of SRAM.

By comparison, the PowerPC 440 integer core is equivalent to 33.3 KB of SRAM [38]. It includes 32 KB of instruction cache and 32 KB of data cache for a total of 97.3 KB of SRAM equivalence. Thus, the PIM solution is only 93% of the area that is required by the PowerPC solution.

IV. MPI Implementation

To explore MPI acceleration, a prototype MPI implementation that was split between the host processor and the NIC processor was created. A single, split implementation of MPI rather than an implementation of MPI over a more general purpose API was created to maximize performance. The prototype MPI implements a subset of MPI-1.2 [39]. With the exception of `MPI_Barrier()`, only basic point-to-point communication and basic support functions were implemented(Figure 2). Only support for basic MPI Datatypes is included and `MPI_COMM_WORLD` is the only group. The MPI was implemented in roughly 1600 lines of C++ and compiled with GNU g++ 3.3[1].

[1]gcc version 3.3 20030304 (Apple Computer, Inc. build 1495)

MPI_Comm_rank()	MPI_Isend()
MPI_Comm_size()	MPI_Recv() †
MPI_Finalize()	MPI_Send() †
MPI_Init() †	MPI_Wait()
MPI_Irecv()	MPI_WaitAll() †
MPI_Barrier() †	

Fig. 2. Subset of MPI implemented. †indicates functions that are built from other MPI functions.

A. Host Software

The host portion of the MPI implementation is minimal. The three basic communications functions (`MPI_Irecv()`, `MPI_Isend()`, `MPI_Wait()`) offload virtually all work onto the network interface. The functions built on top of those as well as the non-communication functions are implemented on the host. To perform an `MPI_Irecv()`, the host checks for available space and posts an item to a queue in the network interface. The network interface manages the posted receive queue and unexpected message queue. The implementation of `MPI_Isend()` is similar. It only needs to place information about the message to be sent in a queue on the network interface. To implement an `MPI_Wait`, the host can spin on a memory location in host memory.

B. NIC-Based MPI Implementation

Most of the work is done by the processor on the NIC. It manages all NIC resources, drives the DMA engines, and handles most of the MPI semantics (including progress and matching). To accomplish this, the NIC maintains five linked lists of MPI state. The following lists contain requests and the state required to advance them.

- `postedRecvQ`: Posted receive buffers for incoming messages to match against
- `activeRecvQ`: Active receives requiring processing (i.e. rendezvous requests which need a reply, requests waiting for a DMA engine, etc.)
- `unexpectedQ`: List of unexpected messages. Compared to new posted receives.
- `unexpectedActiveQ`: Active unexpected messages which must be advanced (i.e. unexpected messages requiring DMA transfer).
- `sendQ`: Queue send requests for processing.

The core of the software on the network interface processor is a loop that continually checks for work. If a new message is waiting to be processed, the header is read by the NIC processor. The posted receive queue is traversed and the header is compared to each posted receive. Upon finding a match, the processor moves the receive request to the active list and either a DMA is setup (a short message) or a rendezvous reply is sent (a long message). If no match is found, the message is placed on the `unexpectedQ`, to be compared to future receives as they are posted.

New send requests cause either a DMA to start (for short messages) or a rendezvous request to be sent (for long messages). The send request is then added to the list of active requests. New receive requests are first compared to the existing unexpected message queue (`unexpectedQ`) for a match. A match causes the message to be copied appropriately or, for long messages, for the data to be requested from the remote processor. Failure to match the receive with the `unexpectedQ` will cause the receive to be placed on the `postedRecvQ`.

The processor also checks the active queues (`activeRecvQ` and `unexpectedActiveQ`) for messages that need to be progressed. Progressing messages can include such things as providing additional information to the DMA engine and responding to rendezvous replies. This enables the MPI implementation to meet the strictest interpretation of the independent progress rule while using a rendezvous protocol and without involving a host processor thread. As items on the active queues complete, the host processor is notified by writing a location in host memory.

C. PIM Feature Exploitation

PIMs have several features that offer keys to a higher performing NIC. The most prominent of these is the extremely wide path between the memory and the processor core (256 bits rather than 32 or 64 bits). This path enables the processor to load large pieces of data to be matched with a single instruction. Moreover, PIMs do not have traditional caches, which typically force several sequential transfers from memory to cache (regardless of access size) and cause unneeded data to be loaded along with the requested data.

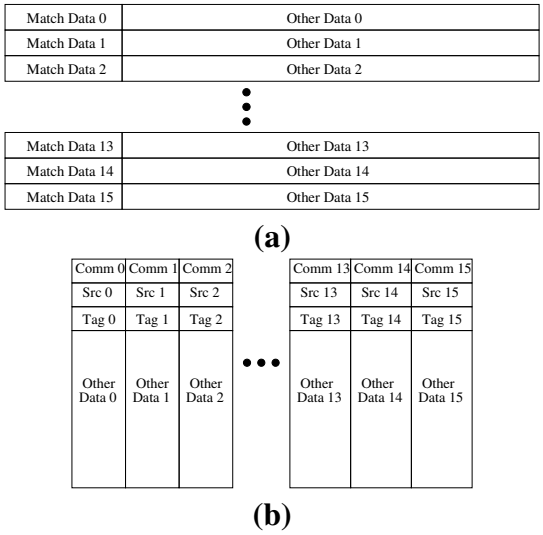

Fig. 3. **(a)** A conventional data layout; **(b)** a PIM data layout

The second useful feature of PIMs is the wide ALU that is typically provided. Since PIMs have wide paths between processor and memory, they typically provide an ALU that matches that width. Thus, the entire match (often multiple bytes) can be performed in a smaller number of instructions. Since MPI matches typically use relatively little data, this is a seemingly small advantage; however, a change in the organization of the list data structures enables a much more powerful usage of the wide ALU. Figure 3 illustrates this concept.

In Figure 3(a), the typical data layout of an MPI posted receive queue is shown. Memory addresses increase from right to left and from top to bottom. Although the queue entries are often sequential (due to the way list items are allocated), that has no impact on this example. If the MPI match data is 48 bits (16 bits for each of the three fields — tag, source, and communicator — is sufficient to meet the specification), then six bytes of each cache line are useful. Since cache lines are 32 (or more) bytes and processors fetch a full cache line for each cache "miss", much of the data retrieved from memory is wasted. Sixteen memory accesses would be required to examine sixteen queues. With the data organization shown in Figure 3(a), the same would occur for the wide words accessed by the PIM.

The best data layout for a PIM, however, is shown in Figure 3(b), where the data structures are interleaved. A single list entry has sixteen list items. All 256 bits of the PIM access to memory is used and only three memory accesses are needed to examine sixteen list items. This utilizes the wide ALU more effectively because all of the data bits are needed for matching. The obvious question is: could a processor leverage the same data layout? The realistic answer is no. In real usage, the list becomes fragmented and any given list entry (with sixteen list items) would contain only a few valid items. The PIM does not suffer from this limitation because of the combination of multithreading and extremely fine-grained synchronization. These two features enable a small fraction of the processing power to be continuously devoted to "list clean-up". A "list clean-up" thread traverses the list using locking at the level of the list entry and compacts out the holes that have accumulated. Thus, at any given time, only a small amount of fragmentation exists anywhere in the list.

The ability to have multiple threads traversing the list using locking at the level of the list entry yields advantages in message throughput as well. Each time a message arrives, a commodity processor must access the header over a bus with a 20 ns latency. Then, it must traverse a linked list. The processor is only free to move onto the next header when the list traversal is done. In a multithreaded environment, however, one thread can be responsible for reading headers and spawning additional threads to traverse the lists. This effectively hides the latency of accessing the queue to retrieve a header.

V. Methodology

This research is focused on the solution to an interesting problem: how can the latency and throughput impacts of long posted receive queues be reduced? Exploring that question required that a benchmark be developed to quantify the problem. After developing a quantitative understanding of the issues, it was necessary to create an environment where solutions to the issue could be explored. This environment included a model of the baseline NIC as well as a model of the enhanced NIC.

A. Benchmarks

The primary motivation for this design was to reduce the latency of and increase the throughput for messages when long posted receive queues were

present. The magnitude of the problem was revealed in an earlier study [19] using a newly designed benchmark. This benchmark was extended to study the impacts of the using a PIM in the NIC.

The benchmark designed to measure the impact of changes in the pre-posted receive queue length provides three degrees of freedom to enable the user to measure the impacts of both the receive queue length (affects caching) and of actual queue traversal (affects processing time). This benchmark was also extended to measure the impact of long posted receive queues on message throughput. Using a traditional processor on a NIC, only one incoming message can be handled at a time. Thus, as the length of a queue grows, the number of messages that can be handled per second decreases. This shifts the standard bandwidth curve to the right — at a given message size, the bandwidth is decreased.

The message throughput can actually be decreased independently of decreasing latency. A $1\mu s$ latency that is concentrated in one atomic block implies a limit on message throughput of one million messages per second. If that $1\mu s$ latency can be broken into five independent pipeline stages, then five million messages per second can be achieved. Alternatively, if five parallel processing units can be provided, the same effect can be achieved. To measure this effect, it is necessary to have more than one message in flight. Thus, the benchmark from [19] was modified to have multiple messages (25) in flight. The modifications to the benchmark are shown in Figure 4. This scenario is reasonable for applications that have long posted receive queues, especially if they would normally communicate with multiple neighbors.

```
prepost_traversed_receives();
post_25_latency_receives();
prepost_untraversed_receives();
barrier();
begin_timer();
send_25_messages();
wait_for_25_responses();
end_timer();
clear_receives();
```
(a)

```
prepost_traversed_receives();
post_25_latency_receives();
prepost_untraversed_receives();
barrier();
wait_for_25_messages();
send_25_responses();
clear_receives();
```
(b)

Fig. 4. Pseudo-code for pre-posted queue impact benchmark: (a) node 0, and (b) node 1

B. Simulation Environment

System-level simulation used a simulator based on Enkidu [40], a component-based discrete event simulation framework. This simulator was originally designed to model a homogeneous sea of PIMs as a supercomputer. To simulate a more traditional system with commodity processors for both the host and the NIC, `sim-outorder` from the SimpleScalar [41] tool suite was integrated into this framework. In addition, a network model was needed. For a simple two node benchmark, the network was modeled as a point-to-point connection with a fixed latency. For the NIC, components representing DMA engines were added. Properly modeling the interaction with the host processor also required a memory controller, DRAM chips, and a simple model of the interface between the host processor and NIC. To maximize fidelity, the memory hierarchy was modeled to include contention for open rows on the DRAM chips.

Both the NIC and the host processor used the PowerPC ISA, augmented with a basic subset of the Altivec [42] vector instruction set. The semantics of this Altivec subset were changed to allow 256-bit vectors. Only six instructions were used (load vector, store vector, copy vector, compare equal-to, and vector bitwise AND) and only 8 vector registers were used.

C. Processor Models

The main processor was parameterized to be similar to a modern high-performance processor, such as an AMD Opteron. Although the Opteron is only six way issue, the SimpleScalar tool suite prefers powers of two and so an aggressive 8 way issue processor was modeled. The NIC processor was parameterized to be similar to a processor in higher-end network cards, such as the PowerPC 440 (see Table I) that is used in Red Storm. A simple

TABLE I

PROCESSOR SIMULATION PARAMETERS

Parameter	CPU	Conv. NIC	PIM
Fetch Q	4	2	1
Issue Width	8	4	1
Commit Width	4	4	1
RUU Size	64	16	NA
Integer Units	4	2	1
Memory Ports	3	1	1
L1 (Size/Assoc.)	64K/2	32K/64	4K/8 (I)
L2	512K	none	none
Clock Speed	2GHz	500MHz	
Main Memory Lat.	140-160 cyc.	30-32 cycles	
ISA	PowerPC		PPC/Altivec
Network Wire Lat.	200 ns		
Network Wire BW	3 GB/s		

bus on the NIC connected the main processor with the DMA engine, SRAM, and matching structure. This bus was simulated with a 20 ns delay to access any components connected to it. Overall, this model attempts to provide as reasonable of a match to a real network as possible.

VI. RESULTS

Figure 5 compares the performance of a NIC using a conventional processor with a NIC using two PIMs for the message processing. The left side of the graph compares bandwidth (a product of message throughput) while the right side compares zero byte message latency. The PIM-based solution shows dramatic improvements in bandwidth when multiple messages are in flight. Even with a short posted receive queue, the overlap achieved by having two simple processors and multiple threads with fine-grained locking can be 10-20% on short messages. Moving to longer queues rapidly exposes the advantage of a wide SIMD unit and wide memory accesses and offers an order of magnitude better performance. As the length of the message grows, the time to DMA the message begins to hide the message processing time; thus, both approaches have the same asymptotic bandwidth. In general, the PIM-based approach ramps much more quickly.

The pure short message latency performance (the right side of Figure 5) is a much more mixed result. At short posted receive queues, the PIM loses dramatically. With short queues, the processing of a single message cannot be multithreaded effectively and cannot leverage wide memory accesses or wide ALU capabilities. Furthermore, there is no significant parallelism (between multiple arriving messages) to exploit the capabilities of PIMS. As the queue length grows, however, PIMs are able to better utilize their unique architectural capabilities.

The relatively poor latency performance of the PIM is explained by the simplicity of the PIM processor. With only a single message, the PIM is unable to spawn multiple threads to mask latency. Profiling indicates that during the latency tests the PIM spends 75% of its cycles with only the main thread active. The operations this thread performs cannot be parallelized due to the ordering semantics of MPI and to avoid deadlock. With only a single thread, the PIM is essentially a single pipeline without branch prediction or the ability to hide memory latency, competing against a dual-issue pipeline with a single-cycle cache. This disadvantage can negate much of the benefit of wide word comparison.

VII. CONCLUSIONS AND FUTURE WORK

Novel features that are core to many modern PIM architectures offer a dramatic advantage for the types of processing found in an MPI library. Exploiting these processors in the embedded environment found on high-performance network interfaces is a natural progression from current NIC architectures that use a traditional microprocessor. The benefits are numerous. The latency for a single message with long queues drops dramatically. More importantly, small message bandwidth grows by 10% with extremely short queues, $2\times$ with moderate queue lengths, and $20\times$ with extremely long queues. These advantages are conferred by the inherent advantages in PIM architectures.

Going forward, we plan to further explore how to leverage the capabilities of PIMs and how to enhance PIMs in the network interface. One avenue of exploration is more PIMs on the NIC. A second potential avenue of exploration is enhancement to the PIM core to make single threaded performance more competitive with PowerPC performance.

REFERENCES

[1] P. M. Kogge, J. B. Brockman, and V. Freeh, "Processing-In-Memory Based Systems: Performance Evaluation Considerations," in *Workshop on Performance Analysis and its Impact*

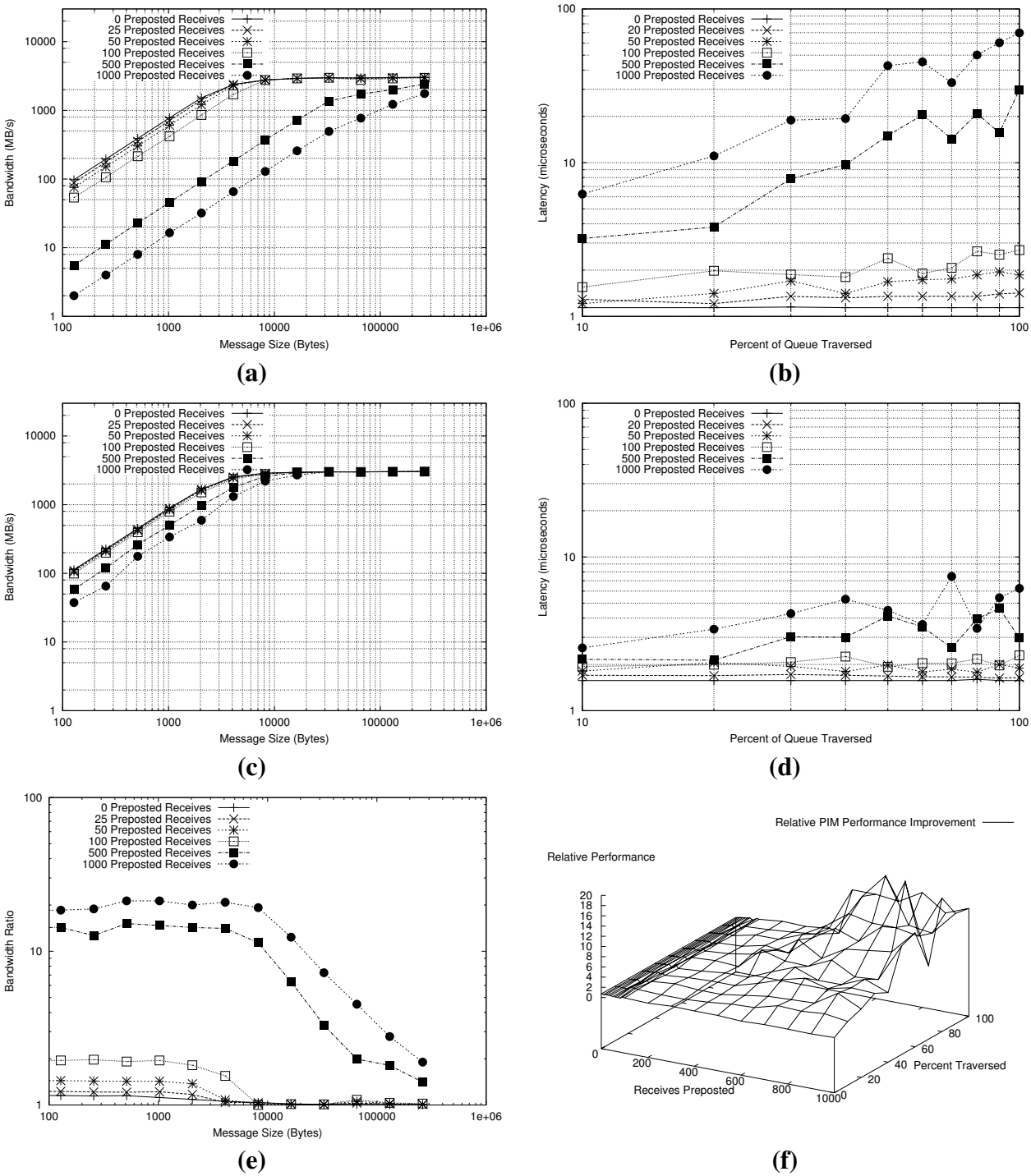

Fig. 5. Conventional processor-based bandwidth and latency curves ((**a**) and (**b**)); PIM based-bandwidth and latency curves ((**c**) and (**d**)); Ratio between the two ((**e**) and (**f**))

on Design held in conjunction with ISCA, Barcelona, Spain, June 27-28, 1998.

[2] J. B. Brockman, P. M. Kogge, V. Freeh, S. K. Kuntz, and T. Sterling, "Microservers: A new memory semantics for massively parallel computing," in *ICS*, 1999.

[3] R. C. Murphy, P. M. Kogge, and A. A. Rodrigues, "The characterization of data intensive memory workloads on distributed PIM systems," in *Proceedings of the Second Workshop on Intelligent Memory Systems, held in conjunction with ASPLOS-IX, Cambridge, MA*. ACM Press, November 12-15, 2000.

[4] M. Hall, P. Kogge, J. Koller, P. Diniz, J. Chame, J. Draper, J. LaCoss, J. Granacki, A. Srivastava, W. Athas, J. Brockman, V. Freeh, J. Park, and J. Shin, "Mapping irregular applications to DIVA, a PIM-based data-intensive architecture," in *Supercomputing, Portland, OR*, November 1999.

[5] D. Patterson, T. Anderson, N. Cardwell, R. Fromm, K. Keeton, C. Kozyrakis, R. Thomas, and K. Yelick, "A case for intelligent DRAM: IRAM," *IEEE Micro*, April, 1997.

[6] W. A. Wulf and S. A. McKee, "Hitting the memory wall: Implications of the obvious," *Computer Architecture News*, vol. 23, no. 1, pp. 20–24, March 1995.

[7] P. M. Kogge, S. C. Bass, J. B. Brockman, D. Z. Chen, and E. H. Sha, "Pursuing a petaflop: Point designs for 100TF computers using PIM technologies," in *6th Symosium on Frontiers of Massively Parallel Computation*, Annapolis, MD, Oct. 1996.

[8] P. M. Kogge, J. B. Brockman, and V. W. Freeh, "PIM architectures to support petaflops level computation in the HTMT machine," in *3rd International Workshop on Innovative Architectures, Maui High Performance Computer Center, Maui, HI*, November 1-3, 1999.

[9] J. Brockman, P. Kogge, S. Thoziyoor, and E. Kang, "PIM lite: On the road towards relentless multi-threading in massively parallel systems," Computer Science and Engineering Department, University of Notre Dame, 384 Fitzpatrick Hall, Notre Dame IN 46545, Tech. Rep. TR-03-01, February 2003.

[10] T. Sterling and H. Zima, "Gilgamesh: A multithreaded processor-in-memory architecture for petaflops computing," in *In Proceedings of the SC 2002 Conference on High Performance Networking and Computing*, Baltimore, MD, November 2002.

[11] R. Alverson, "Red Storm," in *Invited Talk, Hot Interconnects 10*, August 2003.

[12] F. Petrini, W. chun Feng, A. Hoisie, S. Coll, and E. Frachtenberg, "The Quadrics network: High-performance clustering technology," *IEEE Micro*, vol. 22, no. 1, pp. 46–57, January/February 2002.

[13] Myricom, Inc., "Myrinet Express (MX): A high performance, low-level, message-passing interface for Myrinet," July 2003. [Online]. Available: http://www.myri.com/scs/MX/doc/mx.pdf

[14] D. E. Culler, R. M. Karp, D. A. Patterson, A. Sahay, K. E. Schauser, E. Santos, R. Subramonian, and T. von Eicken, "LogP: Towards a realistic model of parallel computation," in *Proceedings 4th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, 1993, pp. 1–12.

[15] A. Alexandrov, M. F. Ionescu, K. E. Schauser, and C. Sheiman, "LogGP: Incorporating long messages into the LogP model," *Journal of Parallel and Distributed Computing*, vol. 44, no. 1, pp. 71–79, 1997.

[16] R. P. Martin, A. M. Vahdat, D. E. Culler, and T. E. Anderson, "Effects of communication latency, overhead, and bandwidth in a cluster architecture," in *Proceedings of the 24th Annual International Symposium on Computer Architecture*, June 1997.

[17] R. Brightwell and K. D. Underwood, "An analysis of NIC resource usage for offloading MPI," in *Proceedings of the 2004 Workshop on Communication Architecture for Clusters*, Santa Fe, NM, April 2004.

[18] R. Brightwell, S. Goudy, and K. D. Underwood, "A preliminary analysis of the MPI queue characteritics of several applications," *submitted*, May 2004. [Online]. Available: ftp://ftp.cs.sandia.gov/pub/papers/bright/mpi-queue-apps.pdf

[19] K. D. Underwood and R. Brightwell, "The impact of MPI queue usage on message latency," in *Proceedings of the International Conference on Parallel Processing (ICPP)*, Montreal, Canada, August 2004.

[20] A. Rodrigues, R. Murphy, P. Kogge, J. Brockman, R. Brightwell, and K. Underwood, "Implications of a PIM architectural model for MPI," in *Proceedings the 2003 IEEE Conference on Cluster Computing*, December 2003.

[21] N. J. Boden, D. Cohen, R. E. F. A. E. Kulawik, C. L. Seitz, J. N. Seizovic, and W.-K. Su, "Myrinet: A gigabit-per-second local area network," *IEEE Micro*, vol. 15, no. 1, pp. 29–36, Feb. 1995.

[22] *http://www.infinibandta.org*, Infiniband Trade Association, 1999.

[23] B. Tourancheau and R. Westrelin, "Support for MPI at the network interface level," in *Recent Advances in Parallel Virtual Machine and Message Passing Interface: 8th European PVM/MPI Users' Group Meeting*. Santorini (Thera) Island, Greece: Springer - Verlag, Septermber 2001.

[24] A. B. Maccabe, W. Zhu, J. Otto, and R. Riesen, "Experience in offloading protocol processing to a programmable NIC," in *IEEE International Conference on Cluster Computing*, September 2002.

[25] J. Liu and D. K. Panda, "Implementing efficient and scalable flow control schemes in MPI over InfiniBand," in *Proceedings of the 2004 Workshop on Communication Architecture for Clusters*, April 2004.

[26] D. Buntinas, D. K. Panda, and P. Sadayappan, "Fast NIC-based barrier over Myrinet/GM," in *Proceedings of the International Parallel and Distributed Processing Symposium*, April 2001.

[27] D. Buntinas and D. K. Panda, "NIC-based reduction in Myrinet clusters: Is it beneficial?" in *Proceedings of the SAN-02 Workshop (in conjunction with HPCA)*, February 2002.

[28] A. Moody, J. Fernandez, F. Petrini, and D. K. Panda, "Scalable NIC-based reduction on large-scale clusters," in *Proceedings of the ACM/IEEE SC2003 Conference*, November 2003.

[29] W. Yu, D. Buntinas, and D. Panda, "Scalable and high performance NIC-based allgather over Myrinet/GM," in *Proceedings of the 2004 IEEE International Conference on Cluster Computing*, September 2004.

[30] A. Mamidala, J. Liu, and D. K. Panda, "Efficient barrier and allreduce on IBA clusters using hardware multicast and adaptive algorithms," in *Proceedings of the 2004 IEEE International Conference on Cluster Computing*, September 2004.

[31] J. Wu, P. Wyckoff, and D. K. Panda, "High performance implementation of MPI datatype communication over InfiniBand," in *Proceedings of the 18th International Parallel and Distributed Processing Symposium*, April 2004.

[32] G. Santhanaraman, J. Wu, and D. K. Panda, "Zero-copy MPI derived datatype communication over InfiniBand," in *Proceedings of the 11th European PVM/MPI Users' Group Meeting*, ser. Lecture Notes in Computer Science, D. Kranzlmueller, P. Kacsuk, and J. Dongarra, Eds., no. 3241. Springer Verlag, September 2004, pp. 47–56.

[33] R. C. Murphy and P. M. Kogge, "Trading bandwidth for latency: Managing continuations through a carpet bag cache," in *Proceedings of the International Workshop on Innovative Architecture 2002 (IWIA02)*. IEEE Computer Society, January 10-11, 2002.

[34] S. K. Kuntz, R. C. Murphy, M. T. Niemier, J. Izaguirre, and P. M. Kogge, "Petaflop computing for protein folding," in *Proceedings of the Tenth SIAM Conference on Parallel Processing for Scientific Computing, Portsmouth, VA*, March 12-14, 2001.

[35] D. Patterson, T. Anderson, N. Cardwell, R. Fromm, K. Keeton, C. Kozyrakis, R. Thomas, and K. Yelick, "A case for intelligent RAM," *IEEE Micro*, vol. 17, no. 2, pp. 34–44, March/April 1997.

[36] Y. Kang, W. Huang, S.-M. Yoo, D. Keen, Z. Ge, V. Lam, P. Pattnaik, and J. Torrellas, "FlexRAM: Toward an advanced intelligent memory system," in *Proceedings of 1999 IEEE International Conference on Computer Design*, Austin, Texas, USA, Oct. 1999.

[37] W. J. Camp and J. L. Tomkins, "Thor's hammer: The first version of the Red Storm MPP architecture," in *In Proceedings of the SC 2002 Conference on High Performance Networking and Computing*, Baltimore, MD, November 2002.

[38] J. B. Brockman, S. Thoziyoor, S. K. Kuntz, and P. M. Kogge, "A low cost multithreaded processing-in-memory system," in *3rd Workshop on Memory Performance Issues (WMPI)*, 2004.

[39] Message Passing Interface Forum, "MPI: A message-passing interface standard," *The International Journal of Supercomputer Applications and High Performance Computing*, vol. 8, 1994.

[40] A. Rodrigues, "Enkidu discrete event simulation framework," University of Notre Dame, Tech. Rep. TR04-14, 2004.

[41] D. Burger and T. Austin, *The SimpleScalar Tool Set, Version 2.0*, SimpleScalar LLC.

[42] IBM Microelectronics Division, *PowerPC Microprocessor Family: Altivec Technology Programming Enviornments Manual*, 2nd ed., IBM, July 2003.