MPI Collective Communications on The Blue Gene/P Supercomputer -Algorithms and Optimizations

Ahmad Faraj, Sameer Kumar, Brian Smith,
Amith Mamidala, John Gunnels, Philip Heidelberger
{faraja, sameerk, smithbr, amithr, gunnels, philiph}@us.ibm.com

Presented By: Amith Mamidala
IBM T. J. Watson Research Center

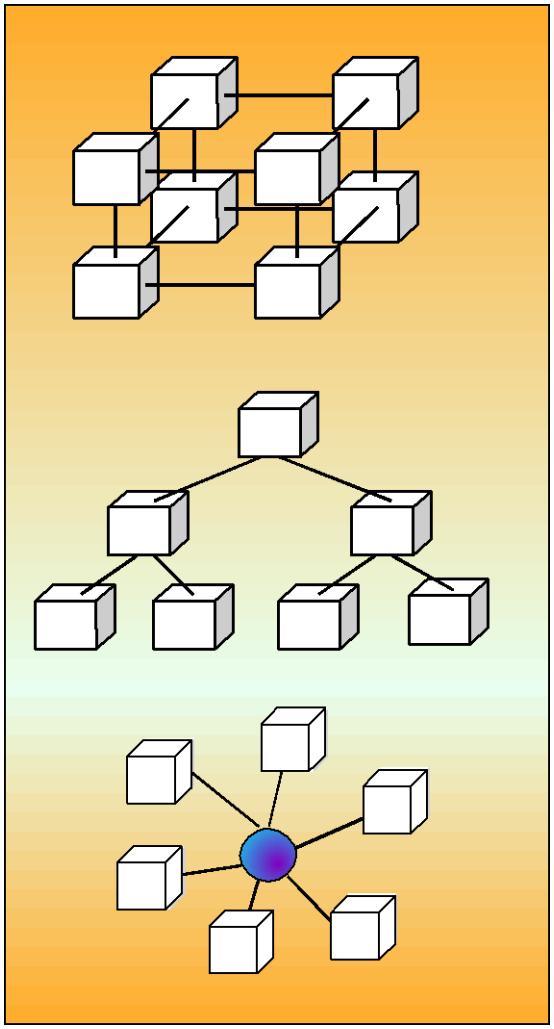
Presentation Outline

- Introduction
- MPI Collectives on BGP
  - Challenges
  - Algorithms & Optimizations
- Performance Evaluation
- Conclusion

Introduction

- MPI
  - Well established standard for coding distributed parallel applications
  - Collective operations
    - Used Extensively in applications
    - Support rich set of semantics
      - Barrier
        - > Process synchronization
      - Broadcast
        - > Data moves from a root process to all other processes
      - Allreduce
        - > Operation across all processes (e.g. Global Sum)
      - Alltoall
        - > Personalized data movement between each process pair
      - Allgather
        - > All to all broadcast of data
- BlueGene/P
  - Significantly large system size and efficient power consumption
  - Scalability very important

BlueGene/P Overview



3 Dimensional Torus – DMA responsible for handing packets

- Interconnects all compute nodes (73,728)
- Deposit bit feature for broadcast along a line
- DMA can saturate all the 12 node links (425 MB/s per link, 5.1 GB/s per node)

Collective Network – core responsible for handling packets

- Broadcast and Reductions
- 6.8 Gb/s (850 MB/s) of bandwidth per link
- Latency of one way network traversal 1.3 µs

Low Latency Global Barrier and Interrupt

Four Cache Coherent SMP cores/node

Challenges designing Efficient Collectives

- Leverage BGP Hardware/System features
- Algorithms to suit the Network topology (torus, mesh)
- Semantics of the Collective Interface
- Algorithm Decision
- Process skew

Leveraging BG/P Features

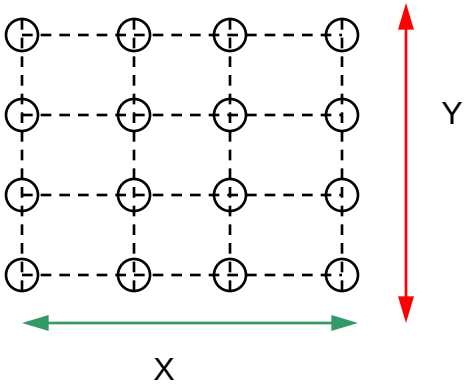
- Torus and Collective Networks
  - Utilize all the six incoming and outgoing links together with the deposit bit feature
  - Collective networks for barrier, broadcast, allreduce
- DMA Features
  - Send/Recv: packets are copied to buffer in main memory
  - DMA Write/Read: packet payload is directly moved into user buffer
  - Frees processing cores from packet management resulting in better computation/communication overlap

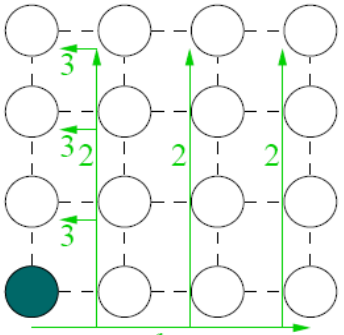
Leveraging BG/P Features

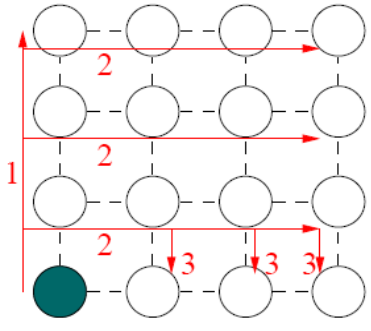
- Intra-node communication
  - Using DMA
  - Using Shared Memory
    - Dedicated segment
    - Direct access to process address space
- Atomic Operations
  - Fetch and Increment
    - LL/SC instructions on memory locations
    - Hardware Lockbox

Algorithms for BGP Torus topology

- Short-rectangle for barrier and allreduce
  - Nodes perform line broadcasts along X torus dimension and process all incoming X packets. Repeated for Y and Z dimensions
  - Nodes send three messages, optimized for short messages
- Multi-color rectangle for broadcast and allreduce
  - Uses multiple edge-disjoint routes from root to all nodes to perform collectives simultaneously, optimized for large messages







Semantics of the Interface

- Non-blocking algorithms based on data movement (DMA, shared memory, collective tree)
- Synchronous algorithms
  - Built-in barrier ensuring all participants' arrival at collective
  - Participants preallocate network resource before messages arrive
- Asynchronous algorithms
  - Data moves in the background before all participants arrive
  - Multiple instances of a collective can be active simultaneously

Choosing the correct algorithms

- Self Tuned Adaptive Routines (STAR)
  - Contains set of communication algorithms
  - Runtime empirical selection mechanism
  - Searches through the algorithm space until it finds the best algorithm for a given call site

Classifying Collective Algorithms

- Three classes of algorithms
- Global algorithms (MPI_COMM_WORLD)
  - Exploit global interrupt and collective tree network to optimize barrier, broadcast, reduction
- Rectangular algorithms
  - Target torus network with efficient line broadcast
  - Short-rectangle, multicolor-spanning tree
- Binomial algorithms (Irregular communication)
  - Uses torus point-to-point links
  - $Log_2(N)$ complexity

Barrier

TABLE I: Barrier Time (micro-sec) in SMP mode

	Communicator	Cores	Optimized	MPICH
BG/P	MPI_COMM_WORLD (GI barrier)	512	1.16	38.0
		2048	1.36	50.0
		16384	1.54	67.2
	Rectangular (short-rect. barrier)	512	10.1	38.0
		2048	14.9	50.0
		16384	22.6	67.2
	Irregular (binomial barrier)	511	20.0	37.4
		2047	27.2	50.0
		16383	36.1	67.2
BG/L	MPI_COMM_WORLD	512	0.95	30.0

TABLE II: BG/P Barrier Time (micro-sec) in DUAL and QUAD modes over MPI_COMM_WORLD

Cores	Algorithm	Optimized
1024 DUAL	GI + Lockbox	2.83
	GI + LL/SC	2.51
	GI + Memory	1.98
2048 QUAD	GI + Lockbox	3.01
	GI + LL/SC	3.16
	GI + Memory	2.09

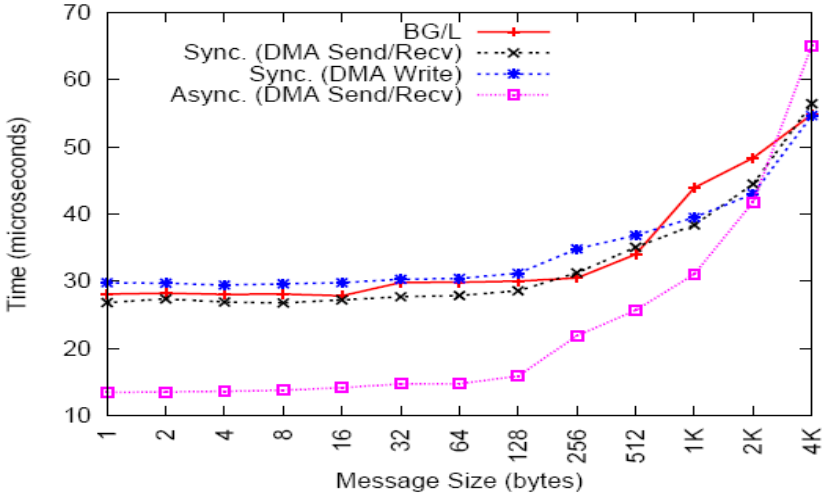
Memory barrier is explored for intra-node

- Cores store 1 byte key in shared memory
- master checks for cumulative 4 byte region and sets completion flag

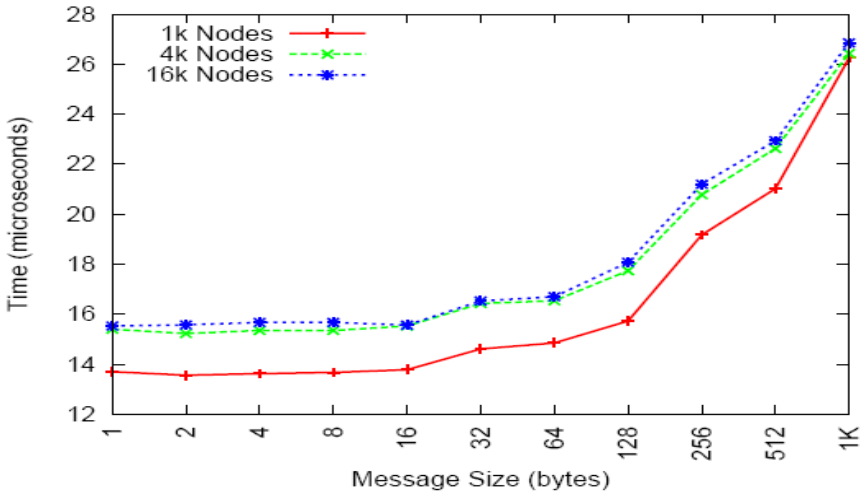
Broadcast

- Global broadcast
  - Collective tree allreduce with the “OR” operator
- Rectangular broadcast
  - Use either DMA Send/Recv or DMA Write with line broadcast
  - Six-color algorithm: six routes from root on the 3D torus are simultaneously used
- Shared address space scheme for intra-node broadcasts
  - Slave cores map master’s broadcast buffer into own address space. A *memcpy* is used to directly copy data into core’s space

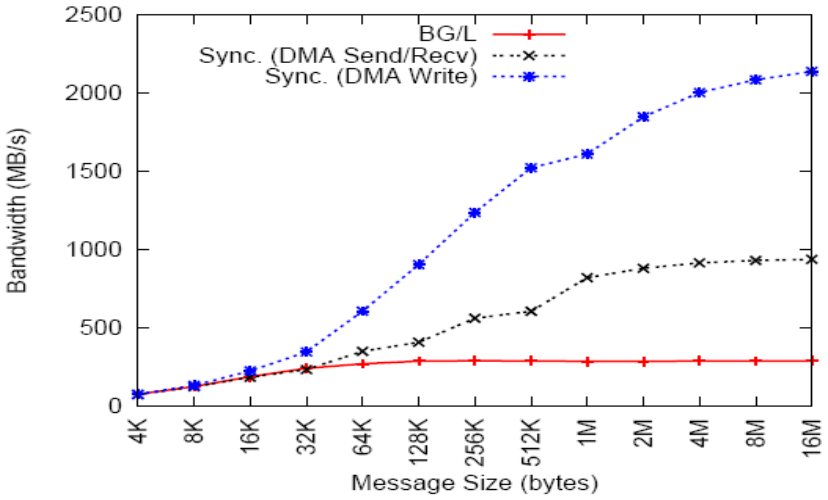
Performance of Broadcast



(a) Latency

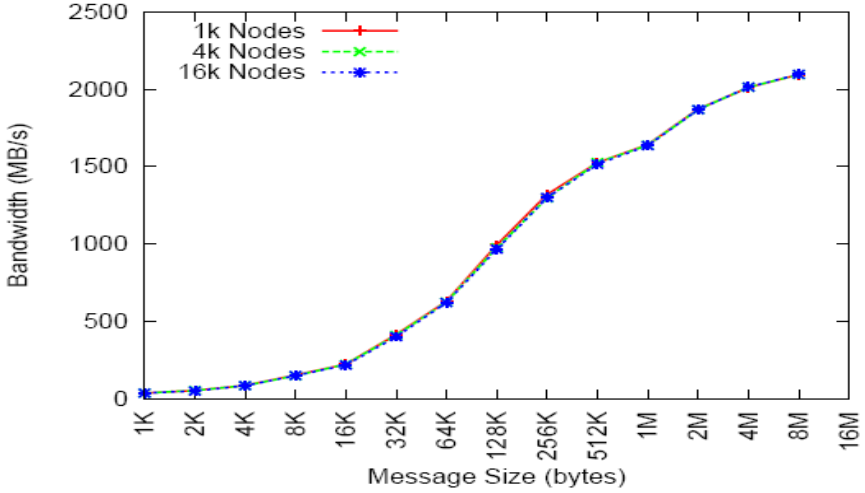


(a) Async. DMA Send/Recv



(b) Bandwidth

Rectangle algorithms over 16K nodes



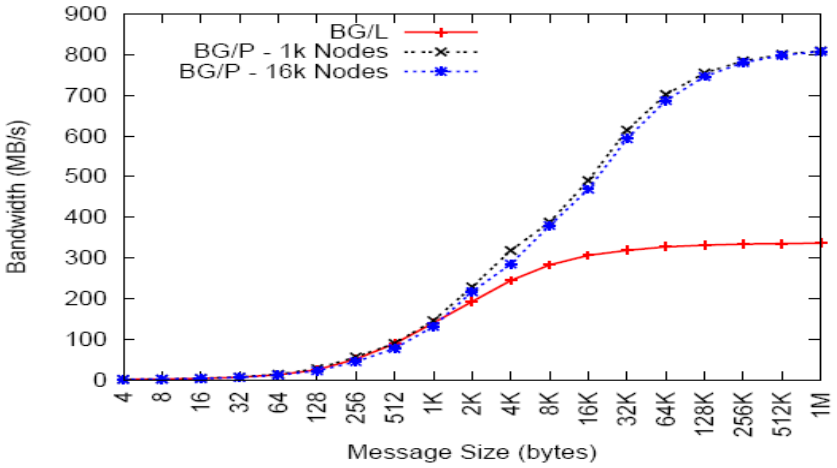
(b) Sync. DMA Write

Scalability of Rectangle algorithms

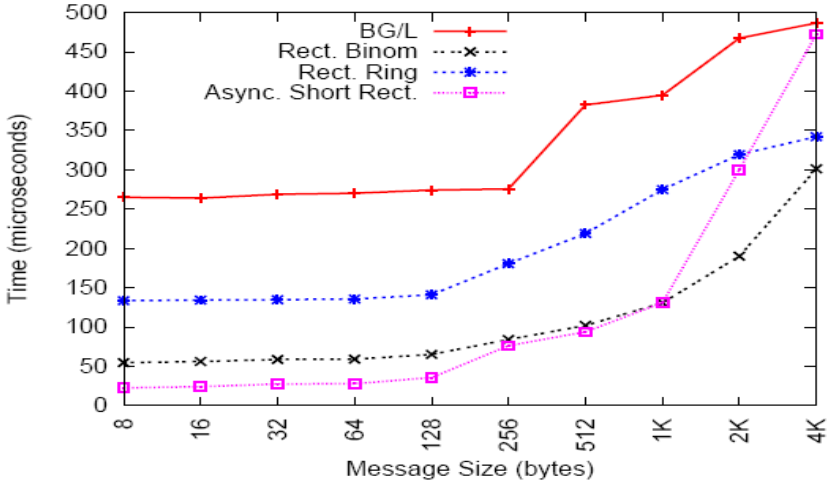
Allreduce

- Global allreduce
  - Collective network uses fast math units to perform reduction efficiently
- Short rectangle (short messages)
  - Using deposit bit feature on each X, Y, Z dimension
- Rectangle-binomial (medium messages)
  - Binomial reduce to the root
  - Root does a Rectangle broadcast
- Multi-color rectangle-ring (large messages)
  - Line reductions using ring algorithm along different torus dims, followed by a broadcast from the root

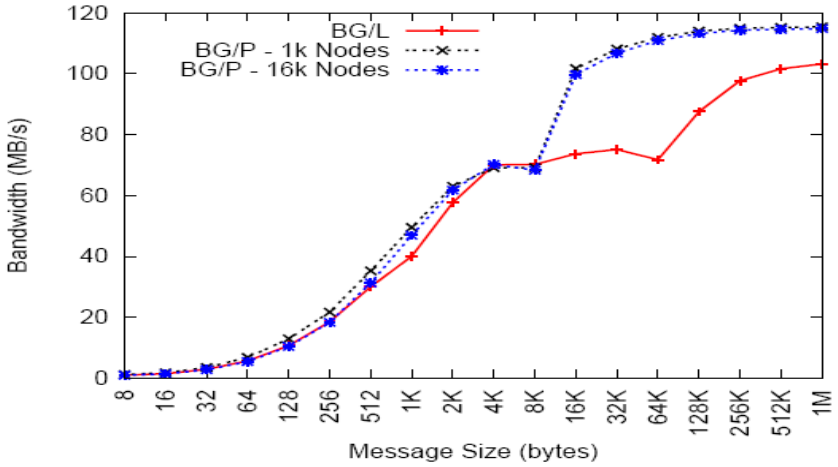
Performance of Allreduce



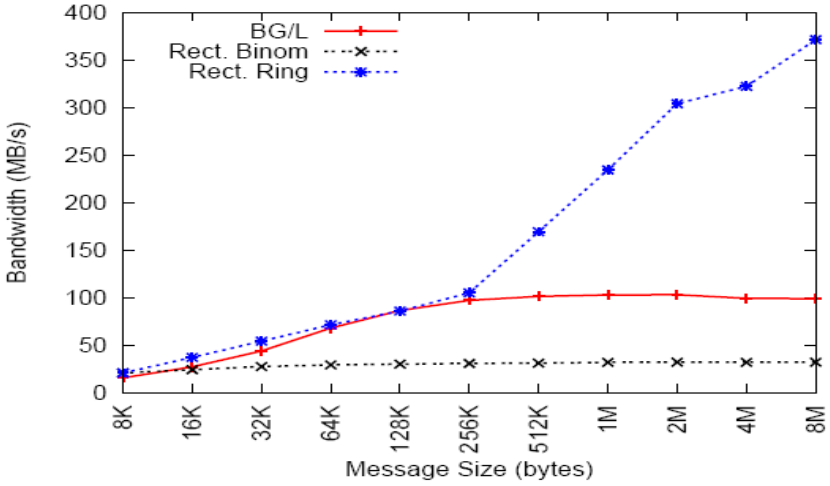
(a) Integers



(a) Latency



(b) Doubles



(b) Bandwidth

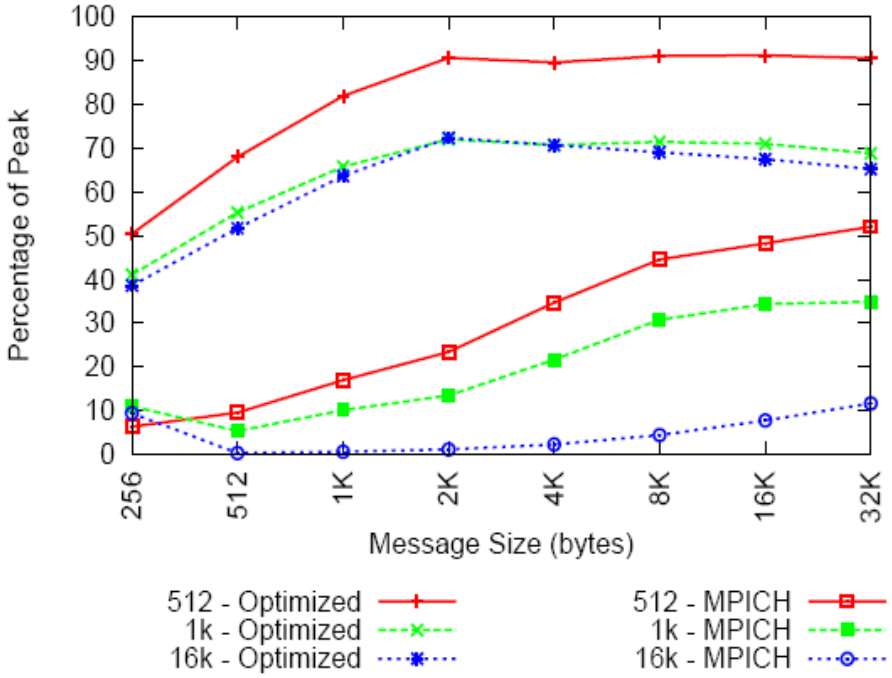
Collective tree allreduce on 16K node COMM_WORLD

Rectangle allreudce over 16K node rectangle comm.

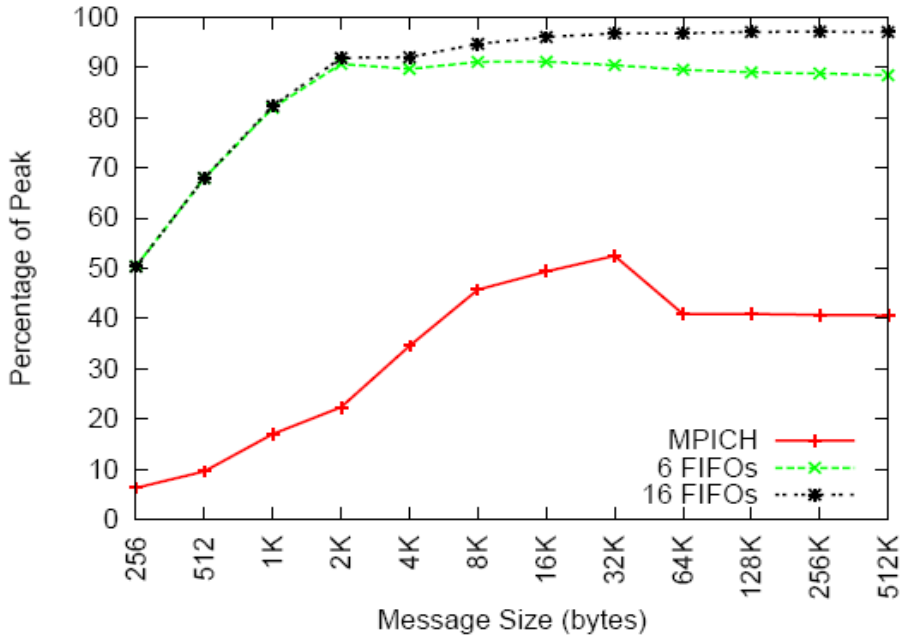
AlltoAll

- A randomized algorithm is used to send packets over each node's six links
- Default Algorithm uses 6 DMA injection FIFOs for different destinations
- Optimized mode uses 16 injection FIFOs for higher throughput

Performance of Alltoall



(a) Scalability of alltoall torus algorithm

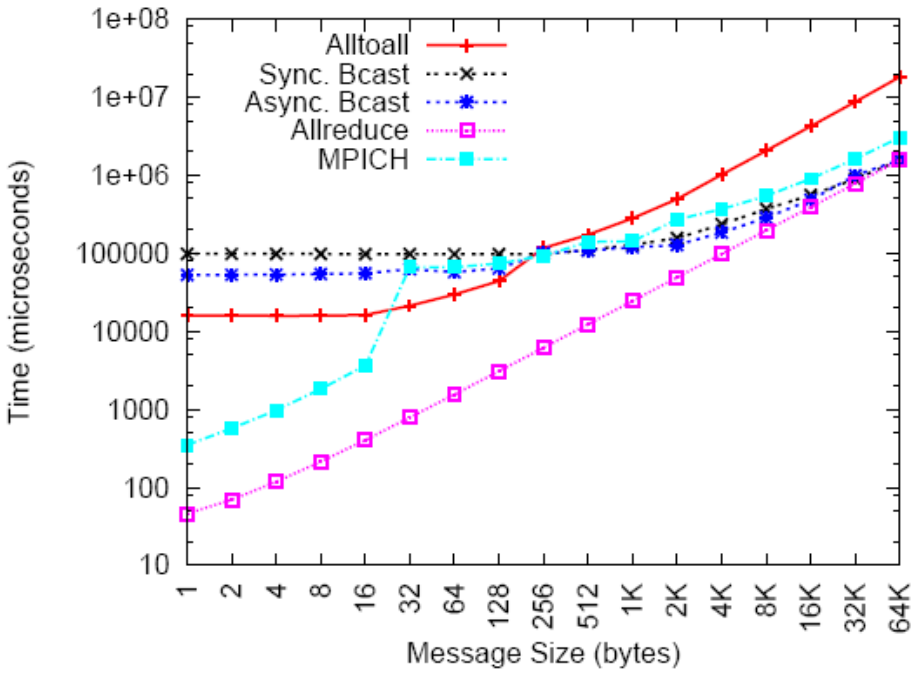


(b) Performance over 512 nodes with different FIFOs

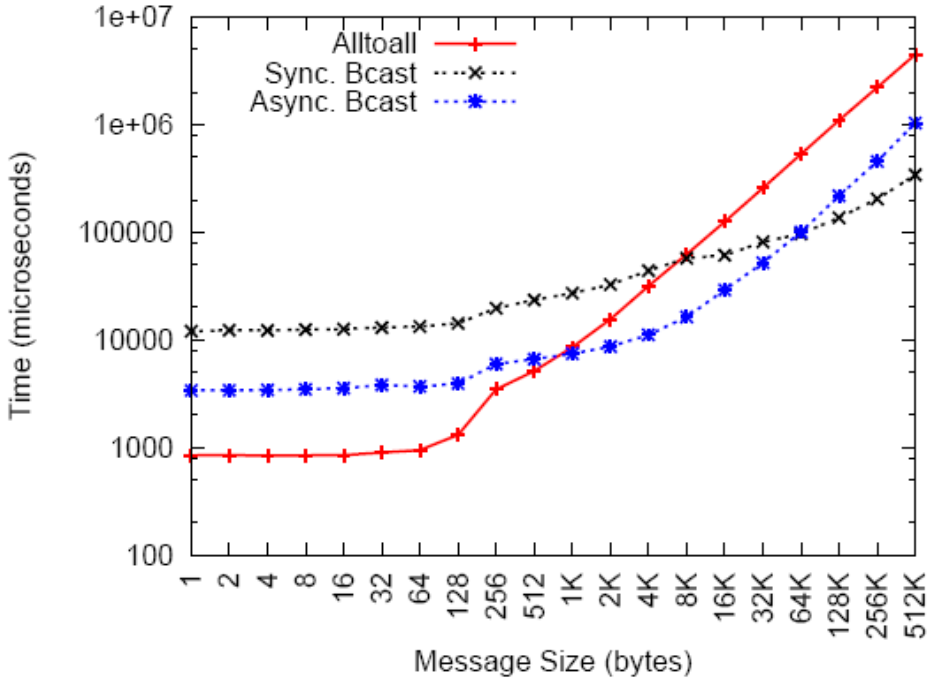
Allgather

- Allgather via Alltoall (short messages)
  - data sent from one node to others is the same
- Allgather via broadcast (medium-large messages)
  - Series of Broadcasts from different roots (nodes take turns)
  - Asynchronous broadcasts exploits torus bandwidth as multiple broadcasts from different roots can be active
- Allgather via allreduce (optimized for MPI_COMM_WORLD)
  - Global OR operation over collective network on the buffers of participants

Allgather



(a) MPI COMM WORLD



(b) Rectangular communicator

Process Skew

```
(1) r = rand() % MAX_IMB_FACTOR;
(2) for (i=0; i<ITER; i++) {
(3)    MPI_Barrier (...);
(4)    for (j=0; j<r; j++) {
(5)       ... /* computation time equal to one msg time */
(6)    }
(7)    t0 = MPI_Wtime();
(8)    MPI_Bcast(...);
(9)    elapsed += MPI_Wtime() - t0;
(10)}
(11) elapsed /= ITER;
```

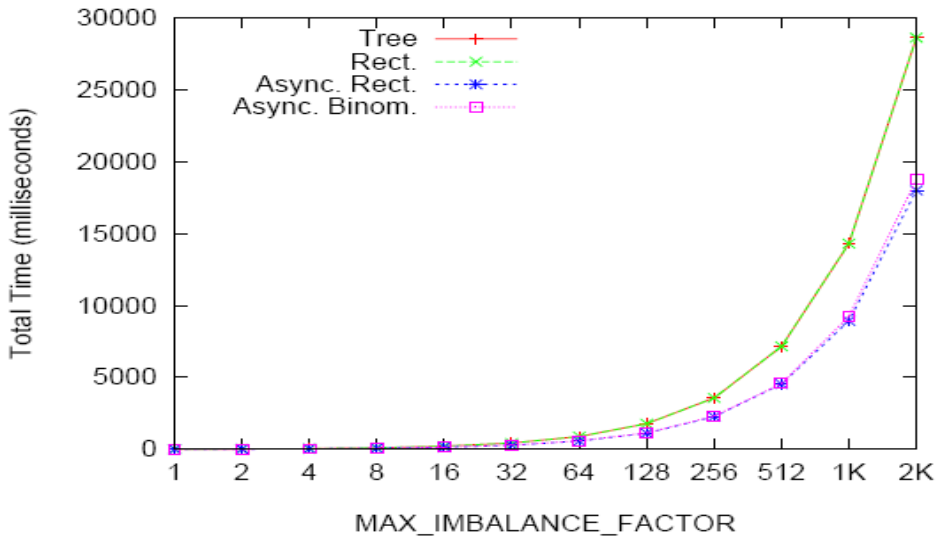


Fig. 13: Impact of arrival patterns on broadcast algorithms on 64KB message (2K nodes)

Application Evaluation

TABLE IV: Summarized description of applications and their MPI collective usage

Bench.	Description	MPI Collectives	Freq.	Msg. size	num. cores
Moldy	a general-purpose molecular dynamics simulation	Allreduce	10000	98400B	8192
Paratec	performs quantum-mechanical total energy calculations	Allreduce	9124	997447B	1024
L5VFO	simulates large networks of detailed model neurons	Allgather	120000	80B	1500
HPL	solves a dense linear system in double precision arithmetic	Bcast	4864	5640992B	4096
		Scatterv	5288	38682B	
		Allgatherv	5288	38682B	

TABLE V: Performance of applications (time in sec)

Bench.	Algorithm	comm.	total
Moldy	collective-allreduce	13.10	317.0
	Async. Binomial	50.00	351.0
	Async. Rect-Binomial	49.70	355.0
	Sync. Rect-Ring (DMA Send/Recv)	49.60	356.0
	Sync. Rect-Ring (DMA Put)	11.40	317.0
	MPICH	49.60	355.0
Paratec	collective-allreduce	142.0	3752
	Async. Binomial	606.0	4278
	Async. Rect-Binomial	351.0	3939
	Sync. Rect-Ring (DMA Send/Recv)	150.0	3750
	Sync. Rect-Ring (DMA Put)	75.00	3658
	MPICH	151.0	3758
L5VFO	Alltoall	15368	20826
	Allreduce	148.7	3012
	Async. Bcast	959.0	3847
	Sync. Bcast	1863	4953
	MPICH	959.0	3847

Summary

- BG/P collective algorithms extensively exploit hardware features to achieve near-peak performance

- Different data movements allow us develop algorithms of different objectives and performance characteristics
  - DMA allow near-peak broadcast and allreduce performance
  - Cache coherent SMP and collective network enable near-optimal integer allreduce performance