An Analysis of the Impact of MPI Overlap and Independent Progress

Ron Brightwell and Keith Underwood

Scalable Computing Systems

Sandia National Laboratories

Albuquerque, New Mexico, USA

18th Annual ACM International Conference on Supercomputing

June 28, 2004

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.

Introduction

- Analyze the impact of MPI overlap, independent progress, and offload from an application perspective
- Network perspective analyzes the ability of the network to provide/offer these features
- Need to understand the ability of applications to take advantage of these features

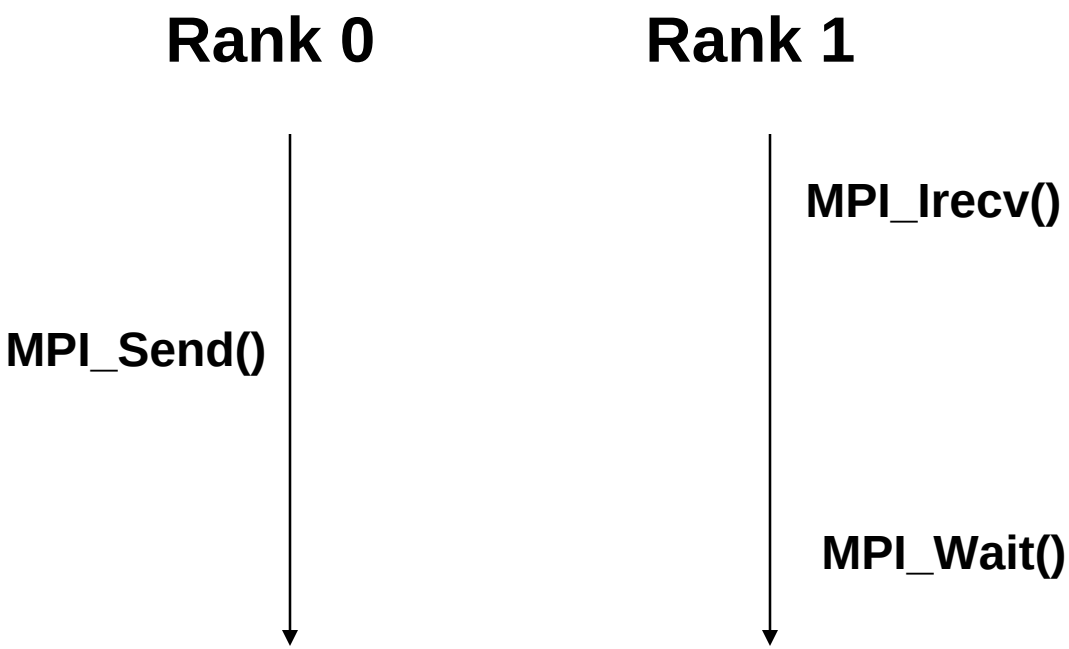
Overlap of Computation and Communication

- Most micro-benchmarks measure the overlap potential of a network and/or MPI implementation
- How much overlap potential is available from the application?
- For applications that cannot benefit, providing overlap may actually decrease performance

MPI Progress Rule

- Determines how a point-to-point communication operation completes once it has been “enabled”
- Room for interpretation
  - Strict: once a communication has been enabled, no subsequent MPI calls are needed to complete it
  - Weak interpretation: application must make library calls in order for an operation to make progress

Example



Impact of Progress

- Important for large messages
- Rendezvous protocol makes performance dependent on how often the application makes MPI library calls
  - Depends on the structure of the code
  - Depends on the OS
- Independent progress
  - Progress is independent of library calls
  - Application is not the MPI progress engine
- Understand the potential of an application to benefit from independent progress

Overlap and Progress Can Be Separate

- OS bypass can overlap data transfer without progress
  - Although potential for application overlap may be much less
- Interrupt-driven networks can provide progress without overlap
  - Although potential for network overlap may be much less

Offload

- MPI matching operations and/or protocol processing are handled by a separate processor

Platforms

- Accelerando cluster at LANL
  - 32 nodes
  - Dual 1 GHz Intel Itanium-2 processors
  - 2 GB main memory
  - 2 Quadrics Elan-3 NICs
  - Linux 2.4.21
- ASCI Red at SNL
  - 4000+ nodes
  - Dual 333 MHz Intel Pentium-2 processors
  - 256 MB main memory
  - CNIC
  - Cougar lightweight kernel

MPI Implementations

- Elan-3 cluster
  - MPI/Tports
    - Vendor-supplied
  - MPI/SHMEM
    - "A New MPI Implementation for Cray SHMEM", EuroPVM/MPI 2004
- ASCI Red
  - Eager / Rendezvous
  - Heater (P0) / Co-processor (P1) modes

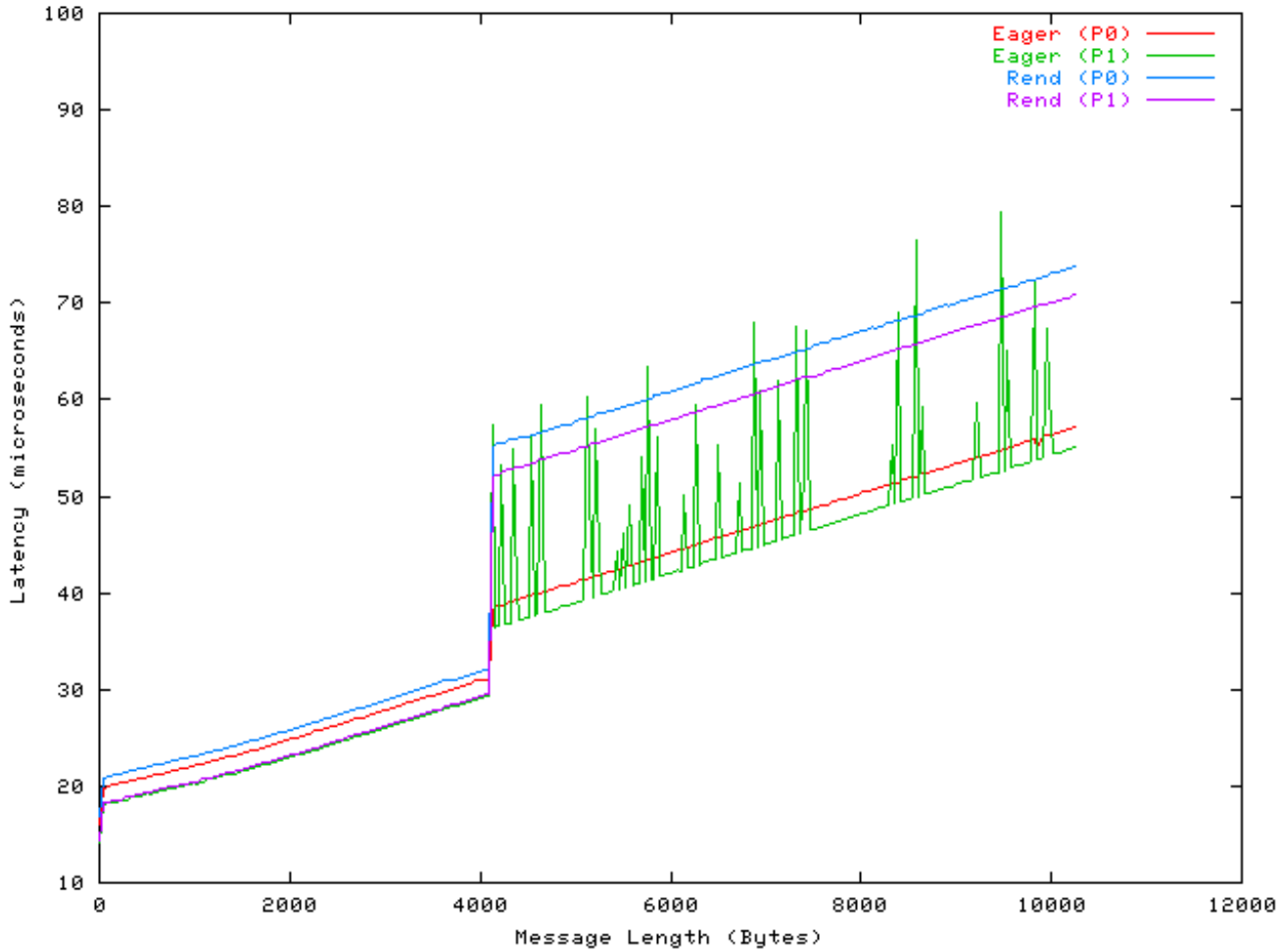
MPI Implementation Characteristics

	ASCI Red				Quadrics	
	Eager		Rendezvous		Tports	SHMEM
	P0	P1	P0	P1		
Progress	✓	✓			✓	✓
Overlap		✓		✓	✓	
Offload		✓		✓	✓	

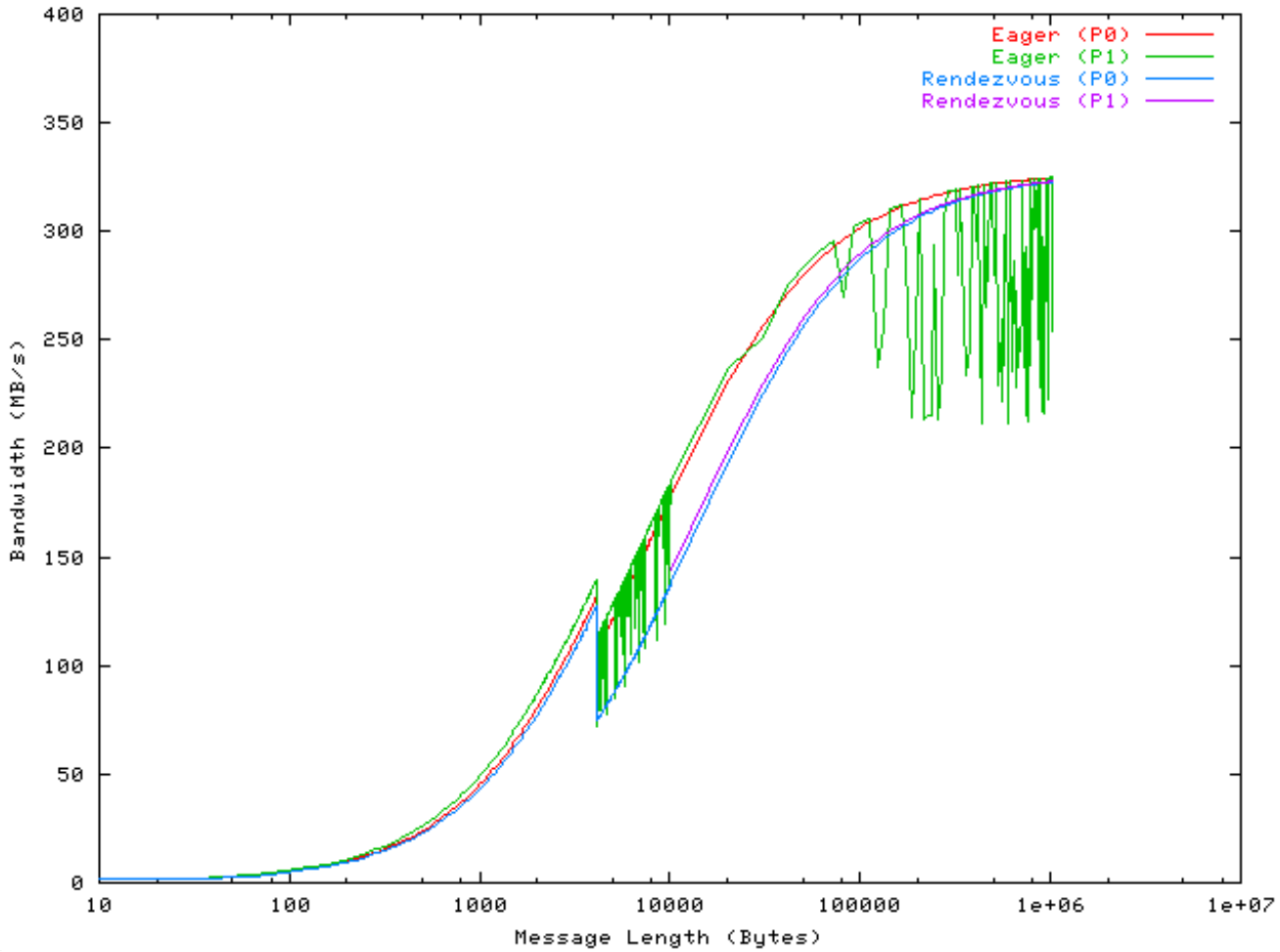
Communication Micro-Benchmarks

- Ping-pong
  - Latency and bandwidth
- Implementations show nearly identical performance
  - Except for Eager/P1 on Red

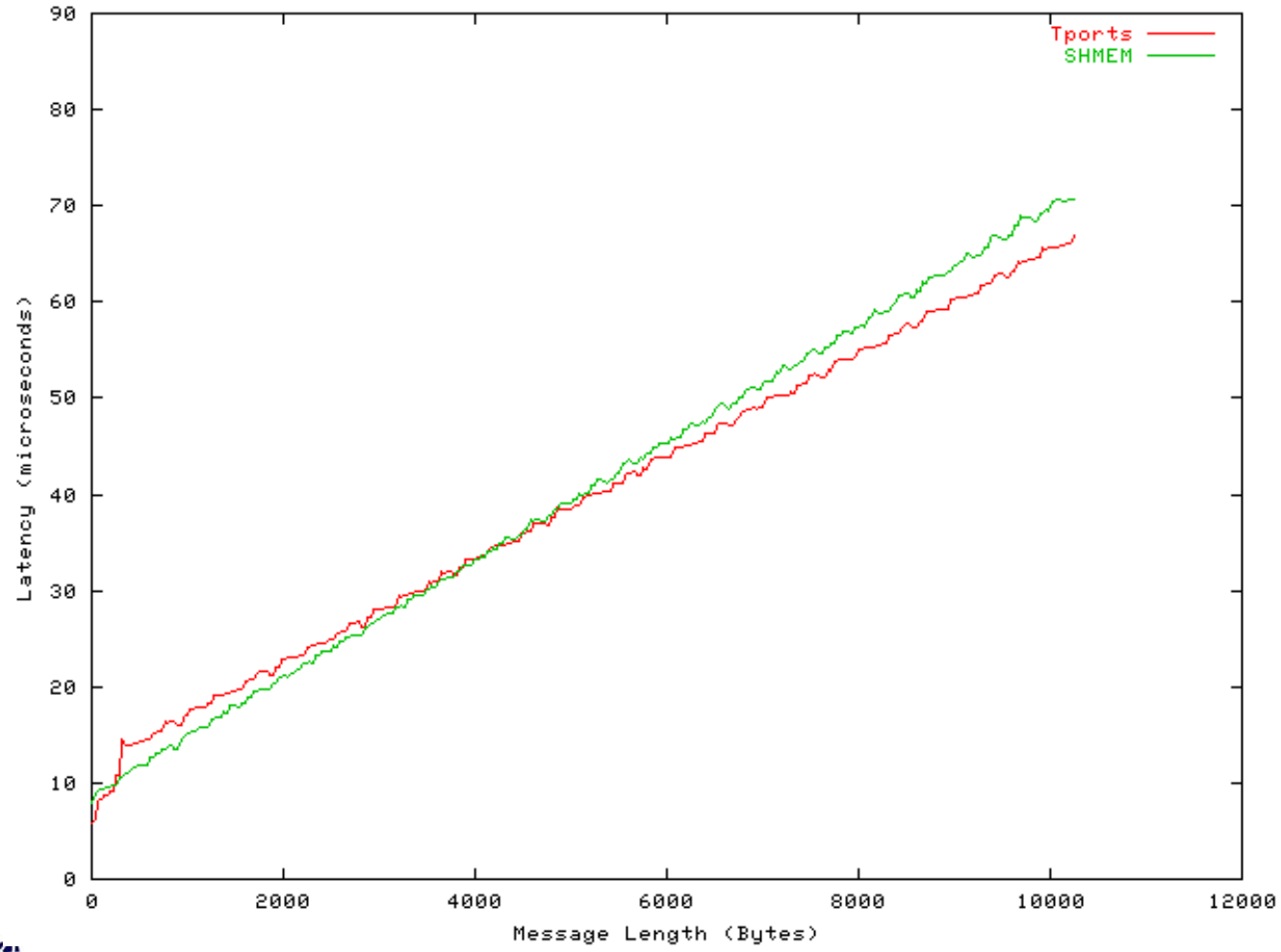
Latency – ASCI Red



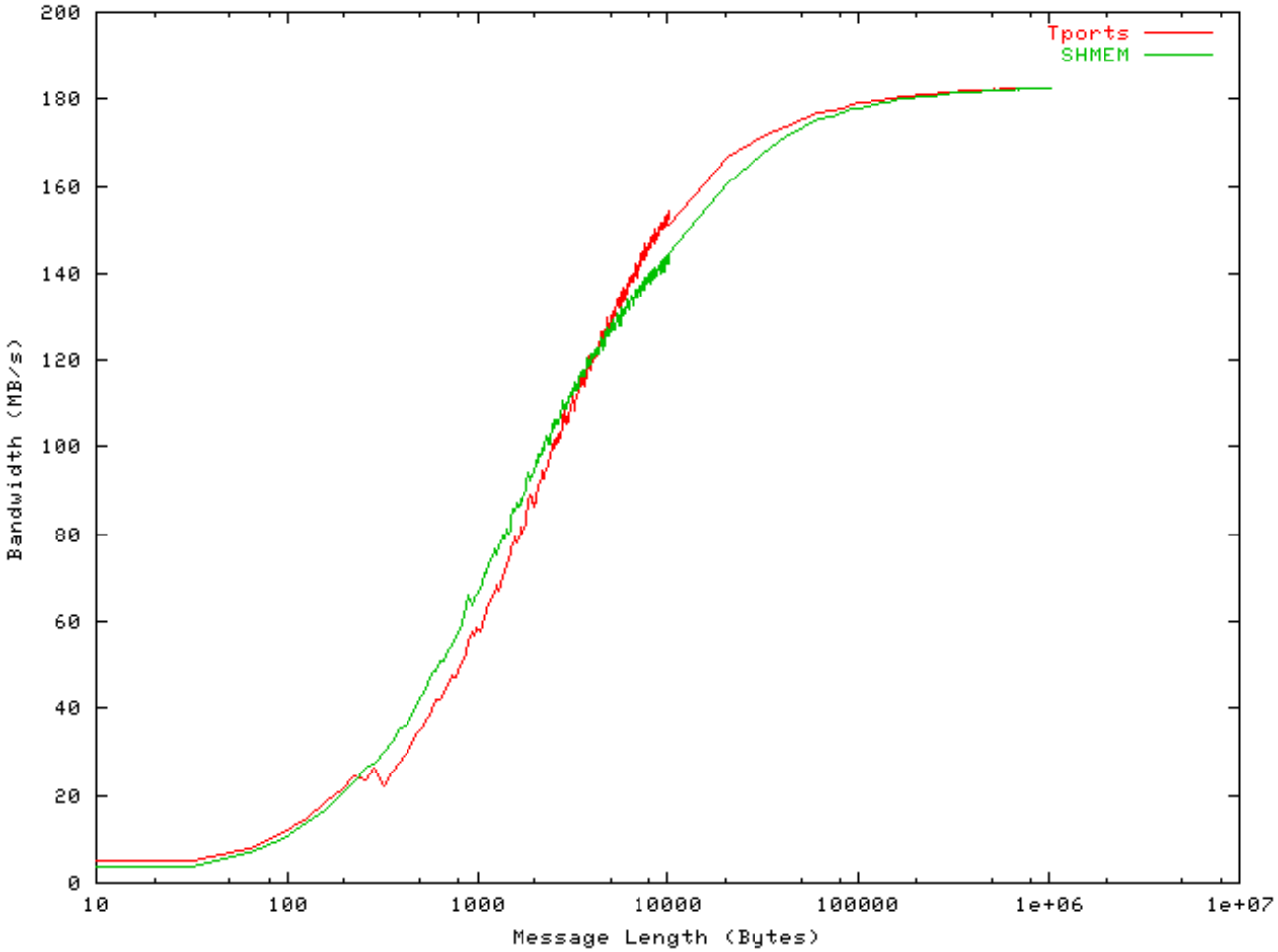
ASCI Red Bandwidth



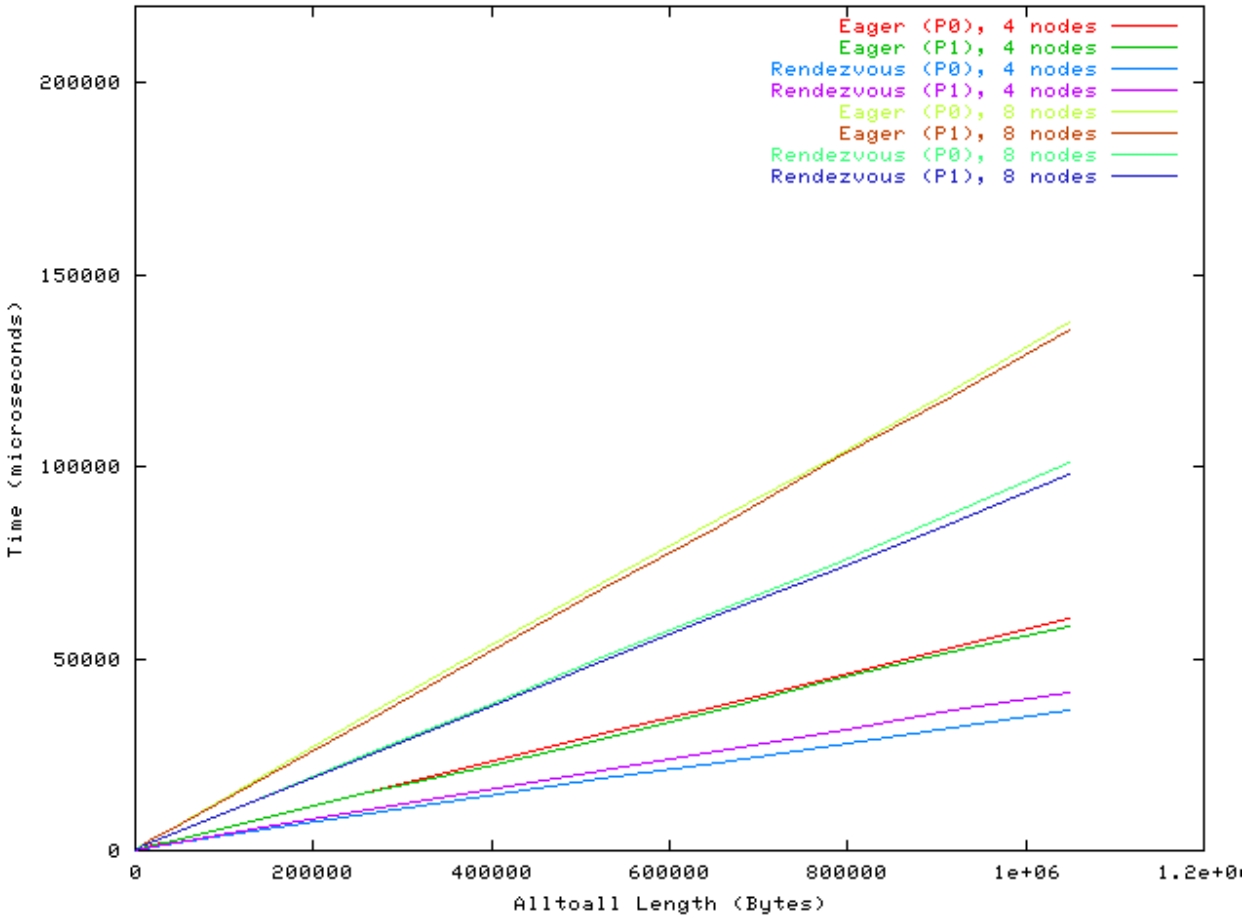
Elan-3 Latency



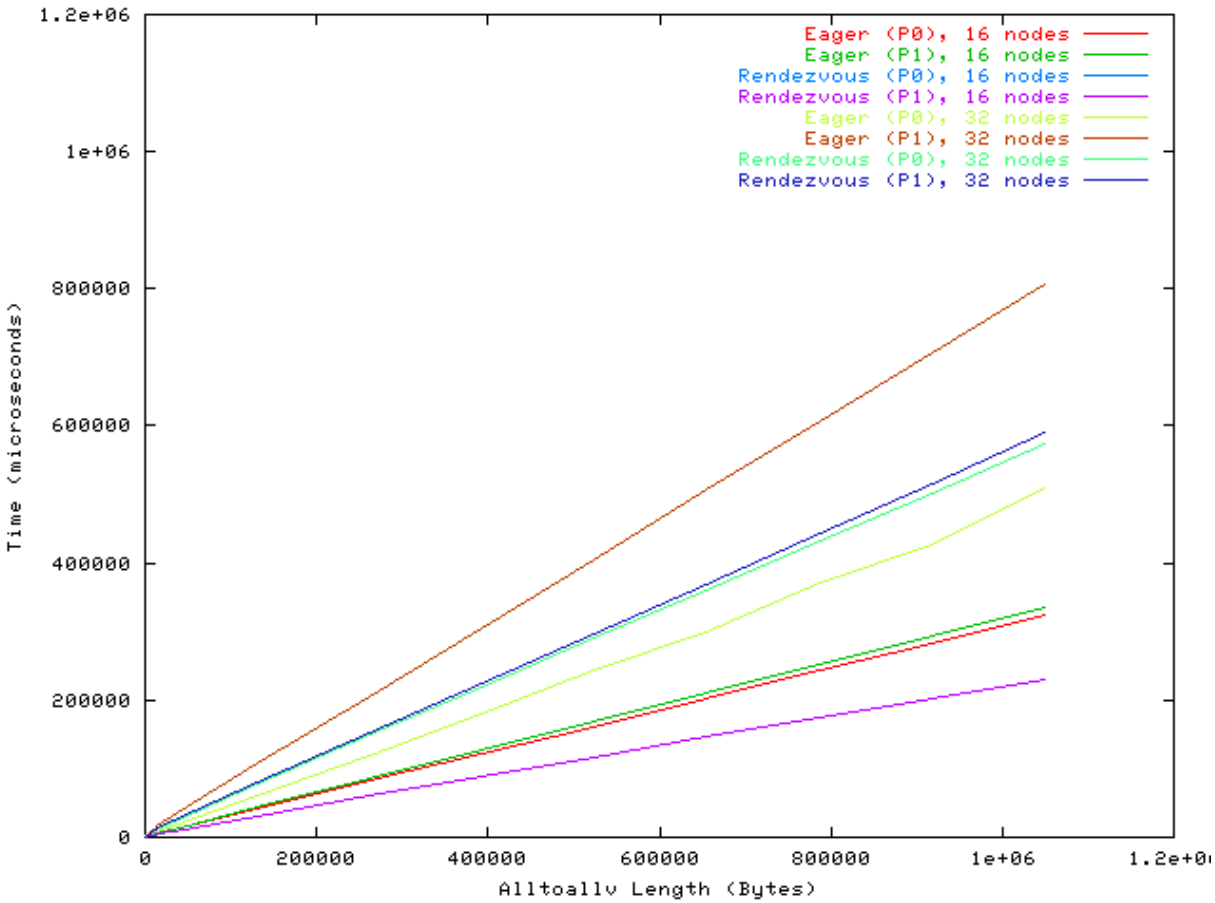
Elan-3 Bandwidth



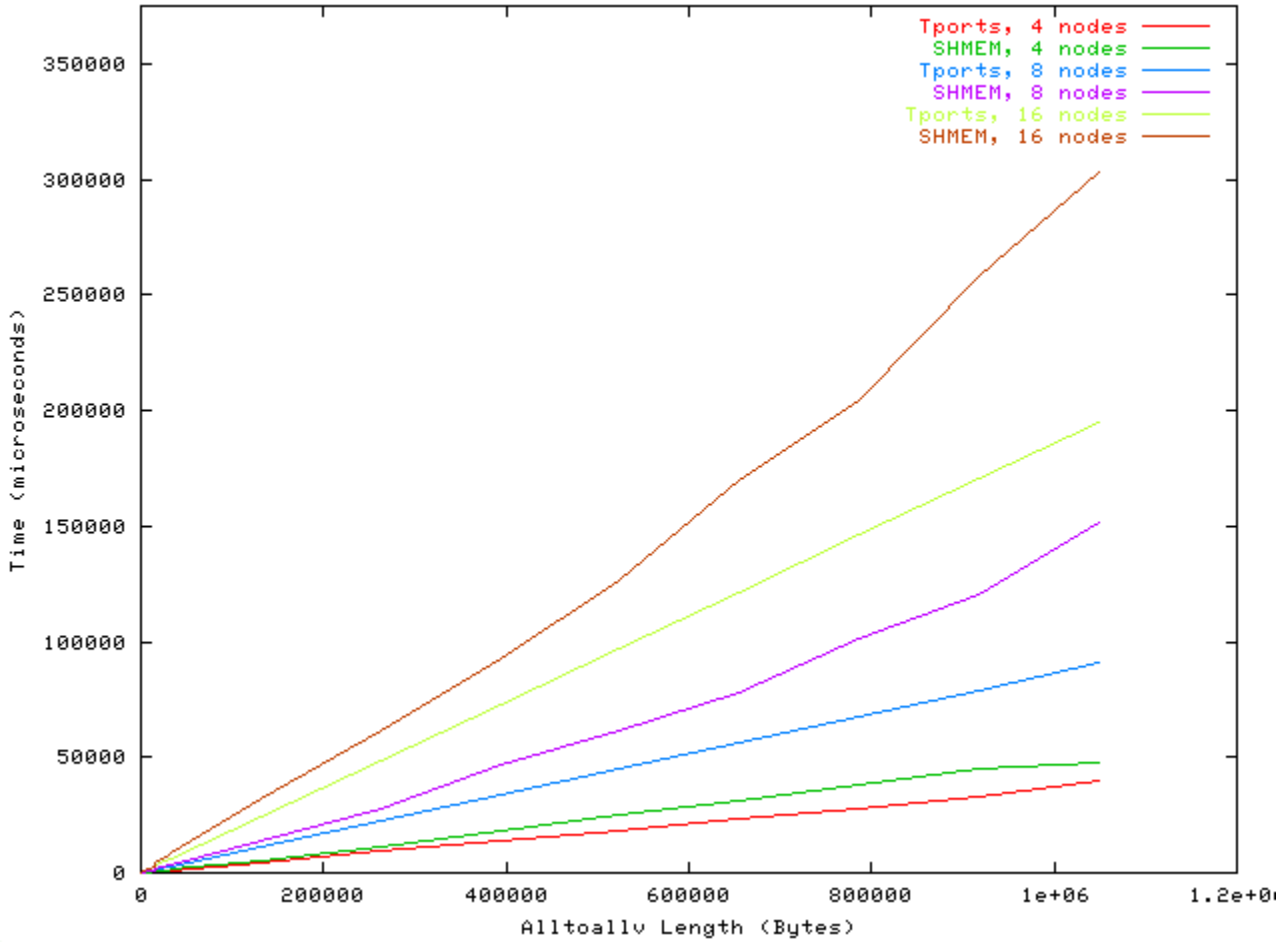
ASCI Red - MPI_Alltoallv() (1)



ASCI Red - MPI_Alltoallv() (2)



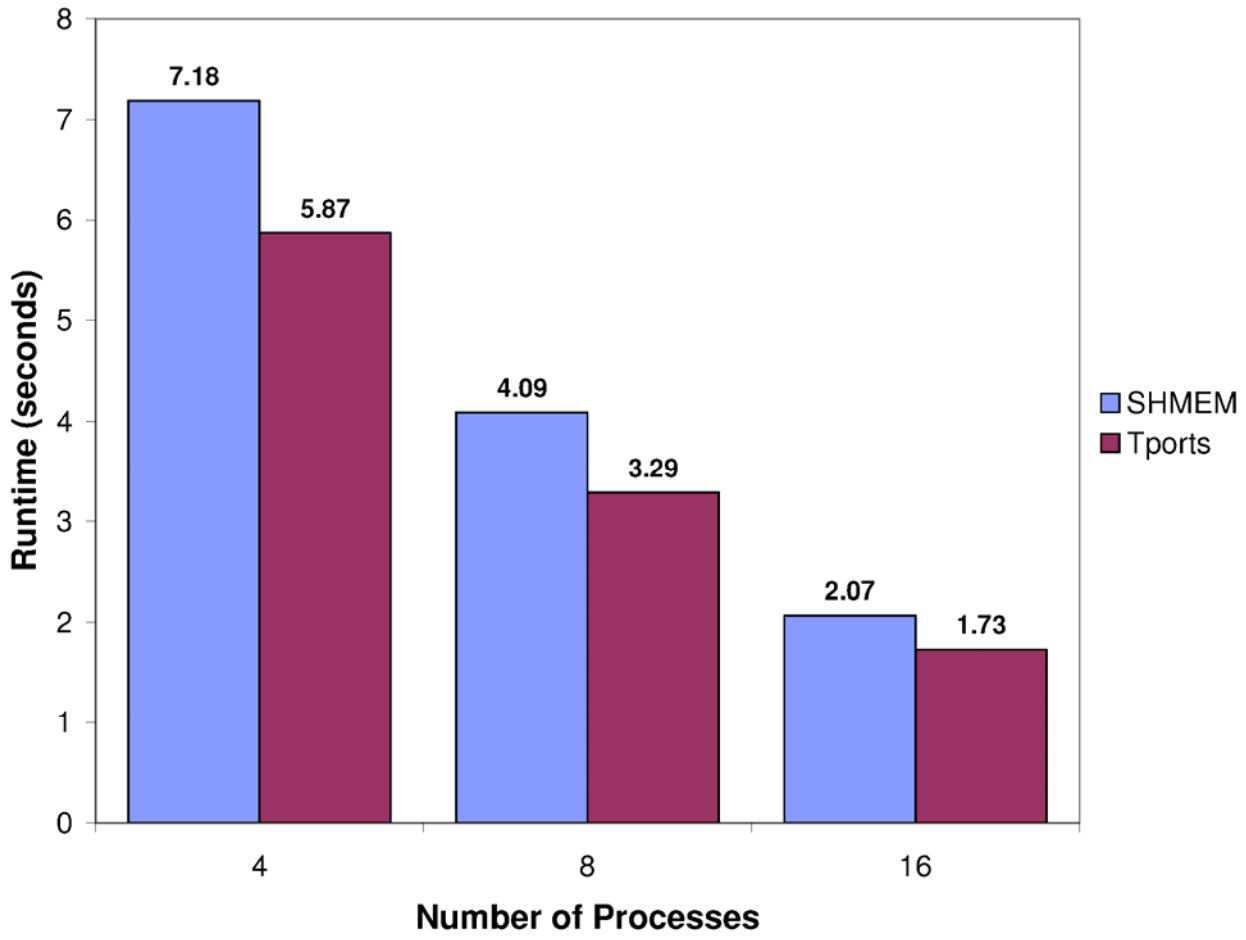
Elan-3 - MPI_Alltoallv()



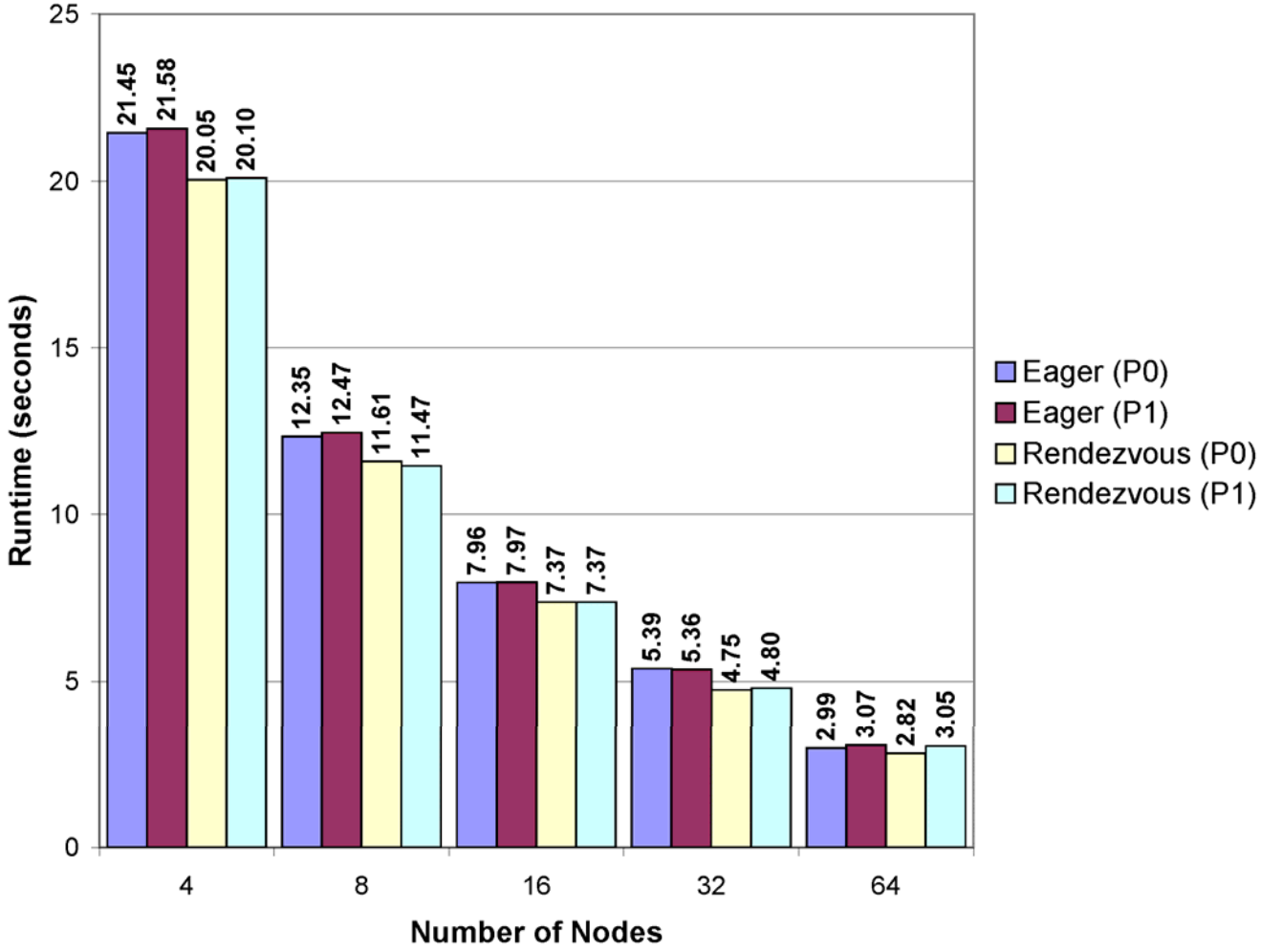
NAS Benchmarks 2.3 Class B

- IS, BT, and SP are the only ones that showed a significant difference in performance
- Times are an average of four runs
- Times from ASCI Red varied by less than 0.5%
- Elan-3 numbers do not use optimized collective operations

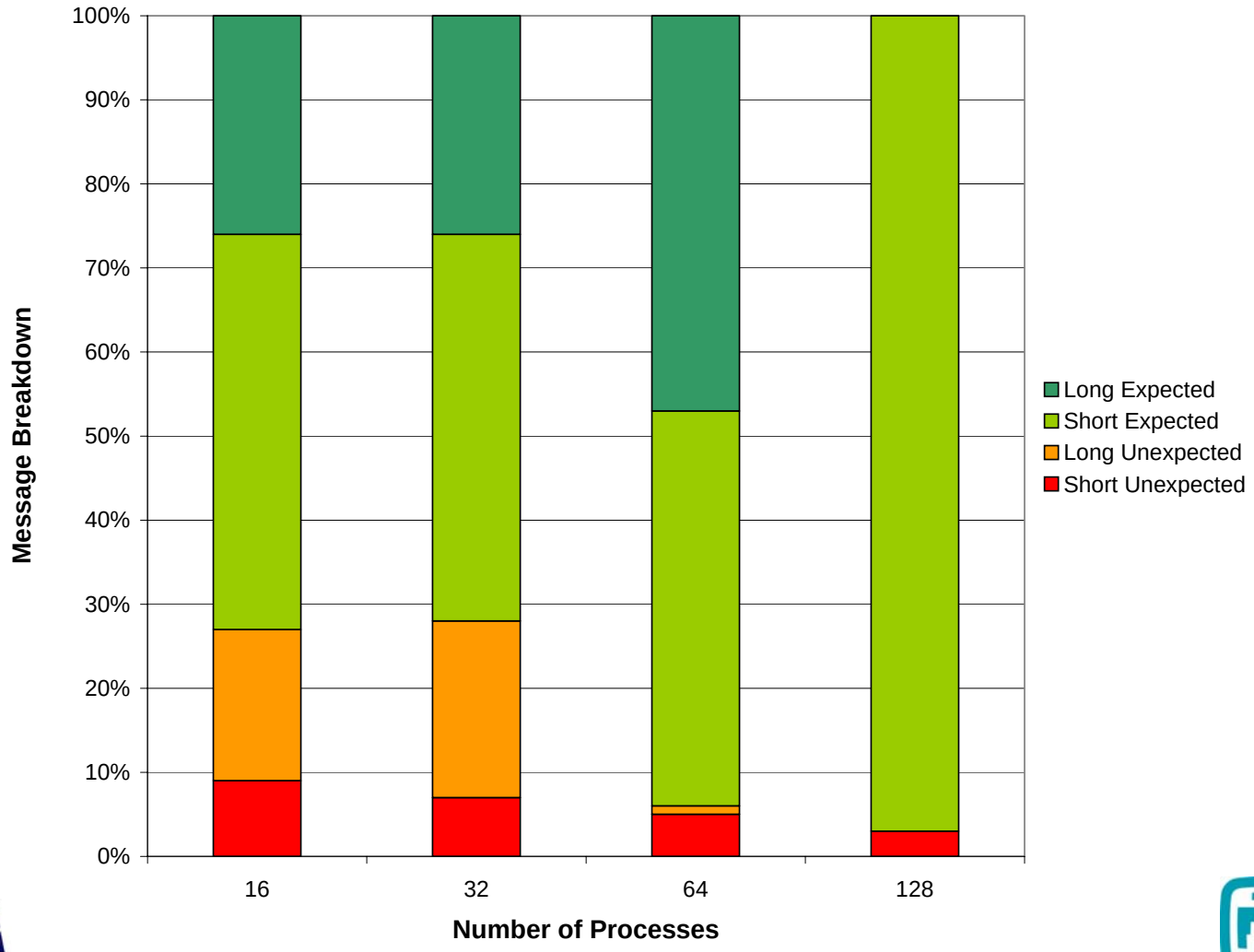
IS – Elan3


8
7
6
5
4
3
2
1
0
Runtime (seconds)
7.18
5.87
4.09
3.29
2.07
1.73
4
8
16
Number of Processes
SHMEM
Tports

IS - Red


25
20
15
10
5
0
Runtime (seconds)
21.45
21.58
20.05
20.10
12.35
12.47
11.61
11.47
7.96
7.97
7.37
7.37
5.39
5.36
4.75
4.80
2.99
3.07
2.82
3.05
4
8
16
32
64
Number of Nodes
Eager (P0)
Eager (P1)
Rendezvous (P0)
Rendezvous (P1)

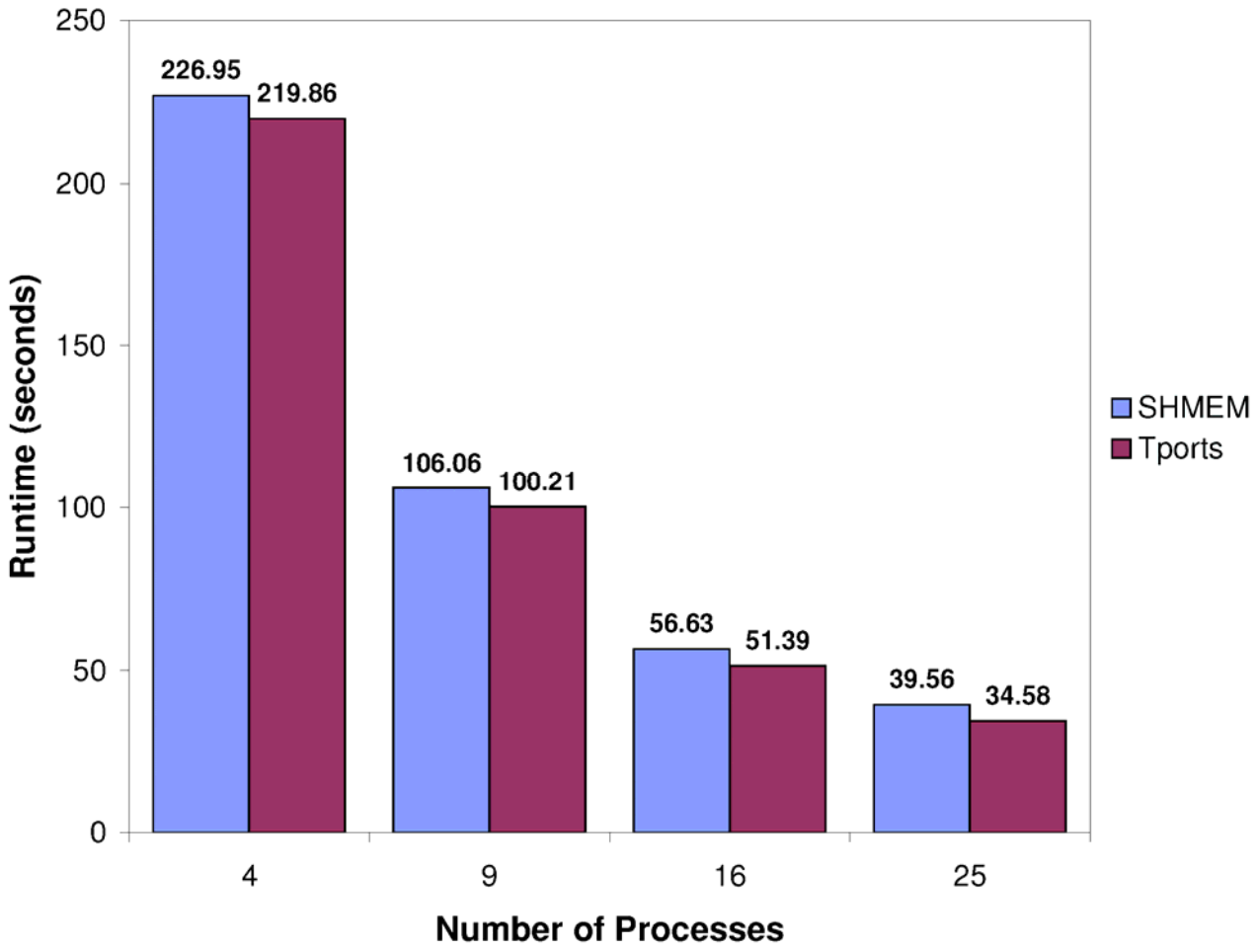
IS – Message Breakdown



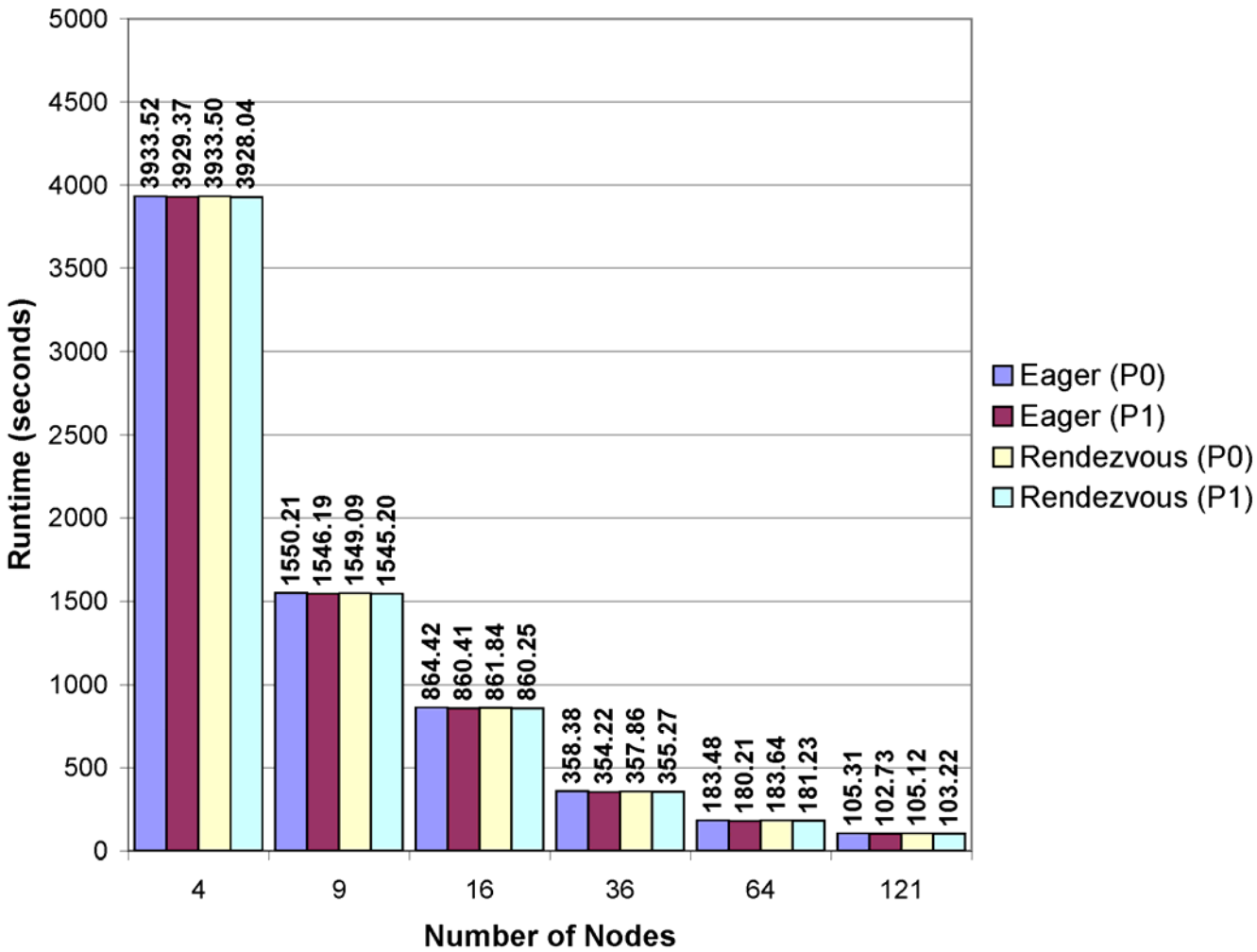
IS

- Performance on Elan-3 can be explained by partially by MPI_Alltoallv() performance
- PO and P1 performance on Red is nearly identical
- Price of independent progress on Red is high for large number of unexpected messages
- Performance is determined by independent progress

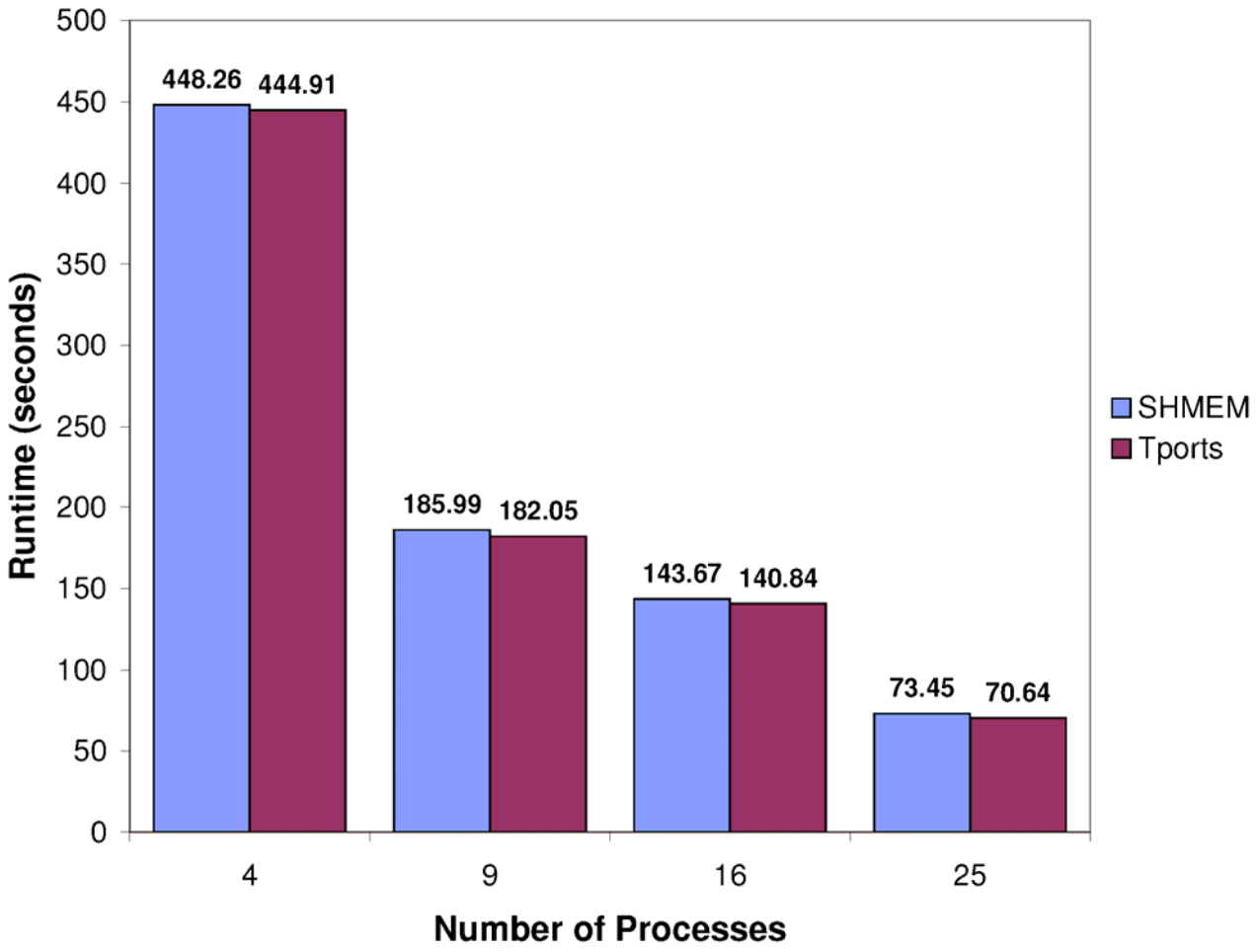
SP – Elan3


250
200
150
100
50
0
Runtime (seconds)
226.95
219.86
106.06
100.21
56.63
51.39
39.56
34.58
4
9
16
25
Number of Processes
SHMEM
Tports

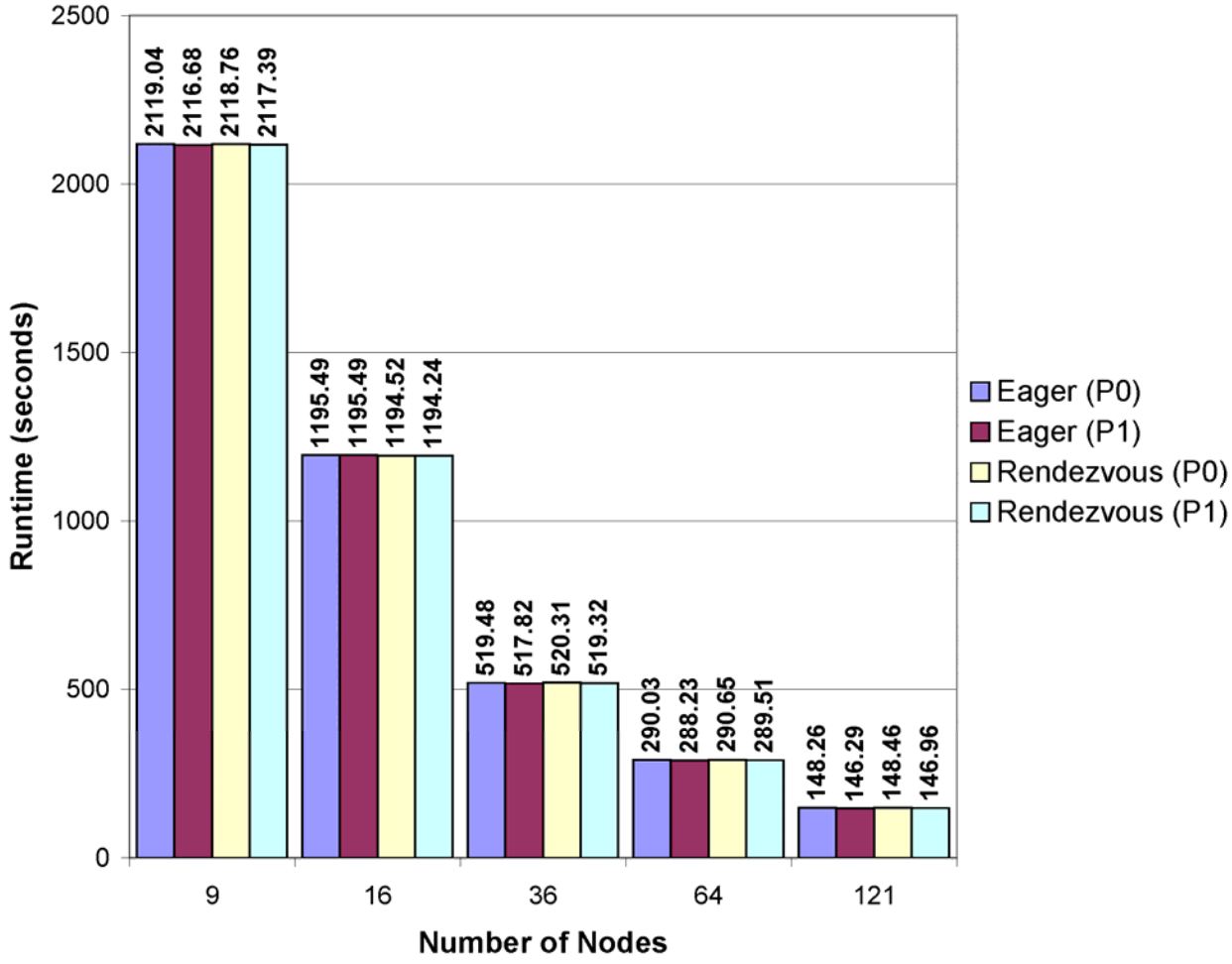
SP – Red


Runtime (seconds)
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
3933.52
3929.37
3933.50
3928.04
1550.21
1546.19
1549.09
1545.20
864.42
860.41
861.84
860.25
358.38
354.22
357.86
355.27
183.48
180.21
183.64
181.23
105.31
102.73
105.12
103.22
4
9
16
36
64
121
Number of Nodes
Eager (P0)
Eager (P1)
Rendezvous (P0)
Rendezvous (P1)

BT – Elan3


500
450
400
350
300
250
200
150
100
50
0
Runtime (seconds)
448.26
444.91
185.99
182.05
143.67
140.84
73.45
70.64
4
9
16
25
Number of Processes
SHMEM
Tports

BT - Red


2500
2000
1500
1000
500
0
Runtime (seconds)
2119.04
2116.68
2118.76
2117.39
1195.49
1195.49
1194.52
1194.24
519.48
517.82
520.31
519.32
290.03
288.23
290.65
289.51
148.26
146.29
148.46
146.96
9
16
36
64
121
Number of Nodes
Eager (P0)
Eager (P1)
Rendezvous (P0)
Rendezvous (P1)

Important Points

- Overlap, independent progress, and offload must be combined to achieve the greatest performance benefit
  - The whole is greater than the sum of the parts
- Independent progress is an important factor in performance
  - The way it is implemented is important
  - Interrupts results in a loss in performance for IS on Red
  - Offload results in a significant performance gain for Elan-3

Future Work

- Explore the reasons why independent progress, and overlap improve performance
- Develop further techniques to isolate the benefits of each
  - Be able to identify applications that would benefit from each
- Extend this analysis to applications in Sandia's workload