Performance Impact of the Red Storm Upgrade

Ron Brightwell **Keith Underwood** **Courtenay Vaughan**

Center for Computation, Computers, Information and Math

Sandia National Laboratories

Cray Users Group Meeting

May 8, 2007

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.

Outline

- Hardware upgrade
- Software upgrade
- Micro-benchmark results
- Application results
- Conclusions
- Future work

Red Storm

Before Upgrade

- 10,880 2.0 GHz single-core AMD Opteron CPUs
  - 43.52 TF/s peak
- SeaStar 1.2 network
  - 1.1 GB/s one-way bandwidth
- 2 GB DDR 3333 memory
- #9 on June 2006 Top 500 list
- Catamount LWK

After Upgrade

- 13,272 2.4 GHz dual-core AMD Opteron CPUs
  - 127.41 TF/s peak
- SeaStar 2.1 network
  - 2.1 GB/s one-way bandwidth
- 2 GB DDR 333 memory
- #2 on current Top 500 list
- Catamount LWK with virtual node mode support

Red Storm is still a Highly Balanced System

Machine	Peak Node (GFLOPS)	Peak BW (GB/s)	Ratio
IBM BG/L	5.6	0.35	0.0625
Cray Red Storm'07	9.6	4.8	0.5000
IBM Purple	48	8	0.1700
SGI Columbia	24	6.4	0.2700
Dell Thunderbird	14.4	2	0.1300
Cray Red Storm'04	4	4.8	1.2000
NEC Earth Simulator	64	12.3	0.1920
Mare Nostrum	17.6	0.5	0.0280
Thunder	22.4	1	0.0400

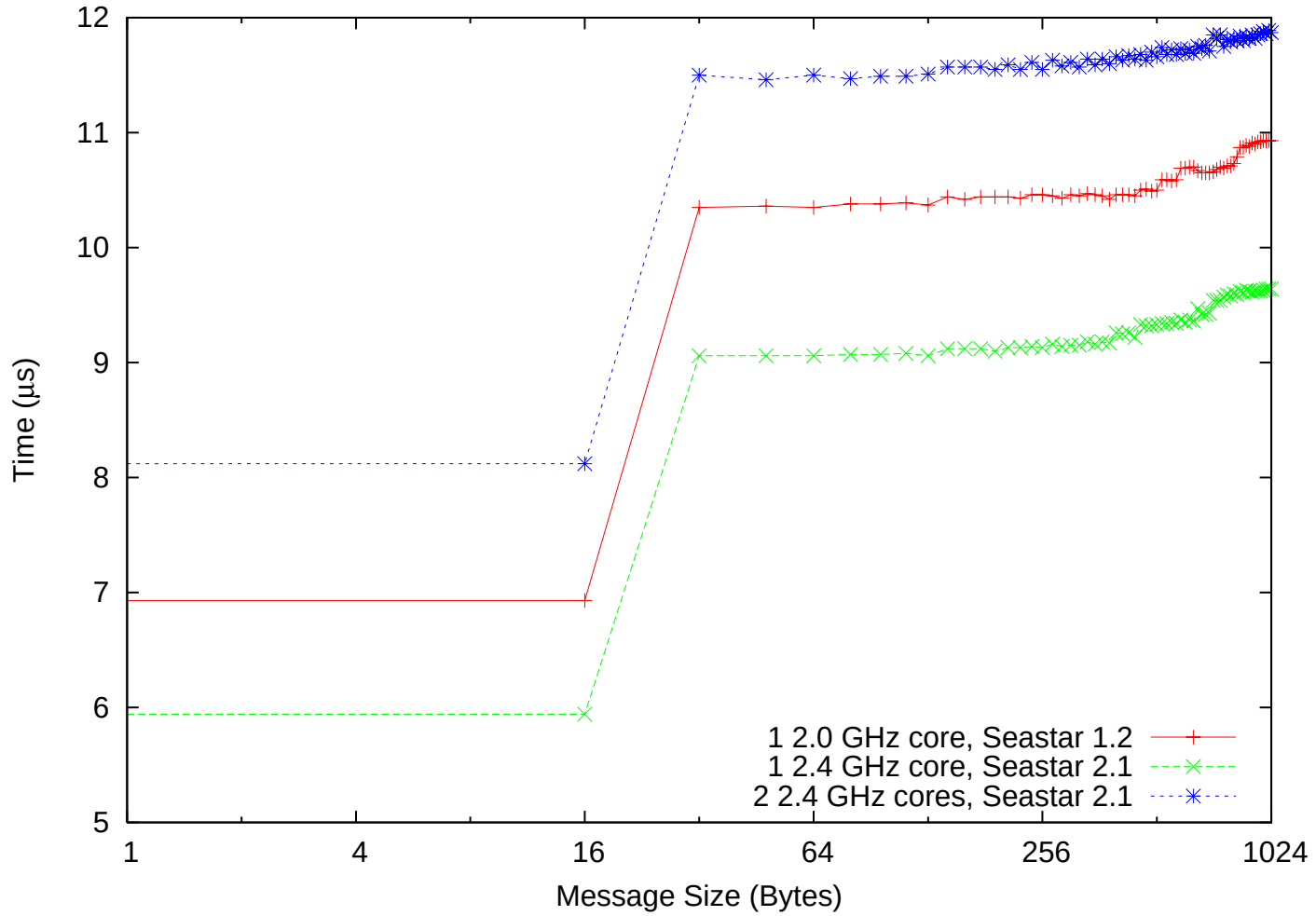
Software Upgrade

- Virtual Node Mode
  - No shared memory communication
  - Generic Portals uses memory copies for intra-node messages (up to a certain size)
  - Second processor essentially waits for work
  - Shared text between processes
- Hang around for John VanDyke's talk ☺

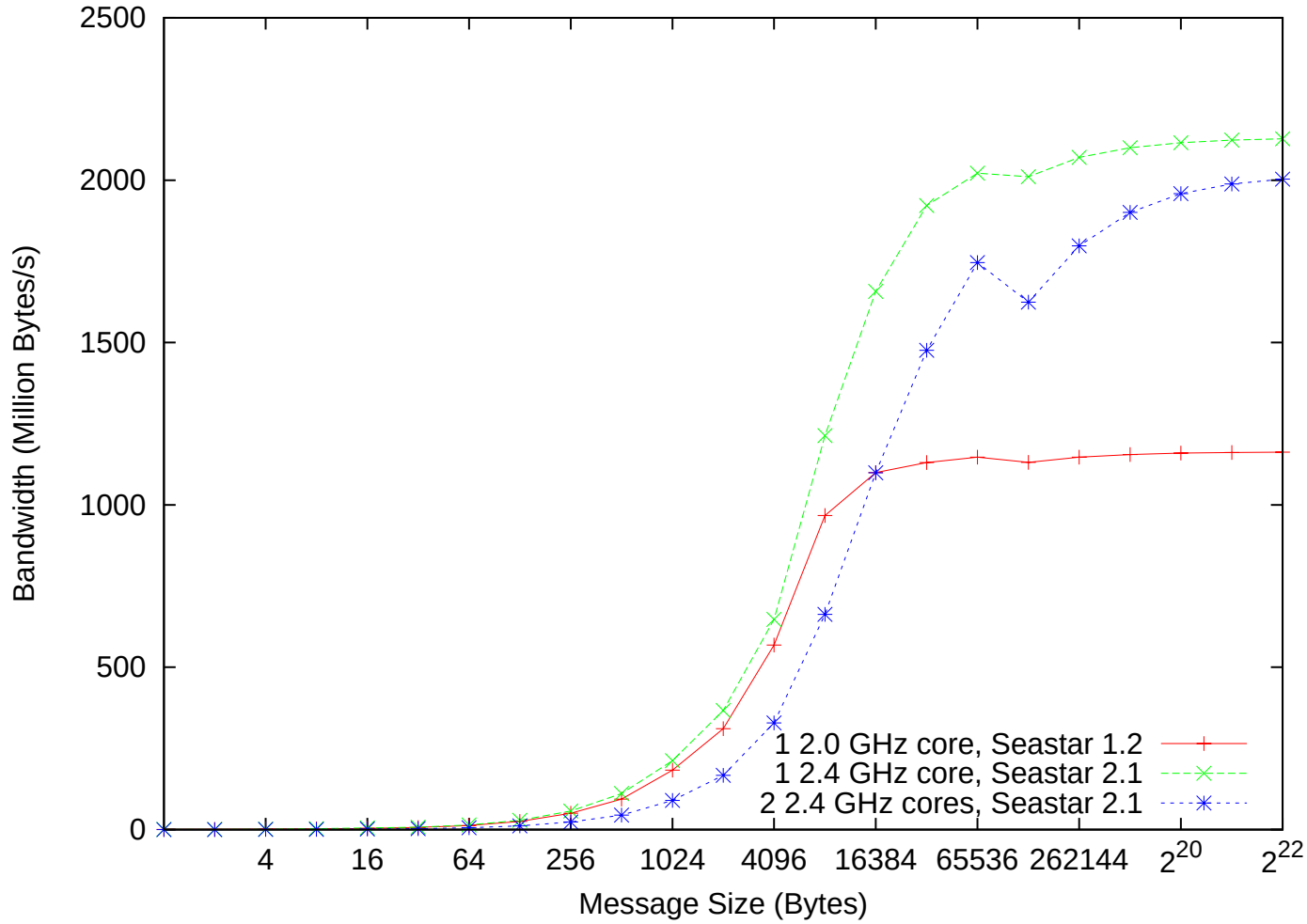
Communication Micro-Benchmarks

- MPI ping-pong latency benchmark
- Ohio State streaming bandwidth benchmark
  - 64 messages at a time
- Analysis of intra-node message passing
  - Processes on separate nodes
  - Processes on the same node

MPI Latency


12
11
10
9
8
7
6
5
Time (µs)
1
4
16
64
256
1024
Message Size (Bytes)
1 2.0 GHz core, Seastar 1.2
1 2.4 GHz core, Seastar 2.1
2 2.4 GHz cores, Seastar 2.1

MPI Streaming Bandwidth



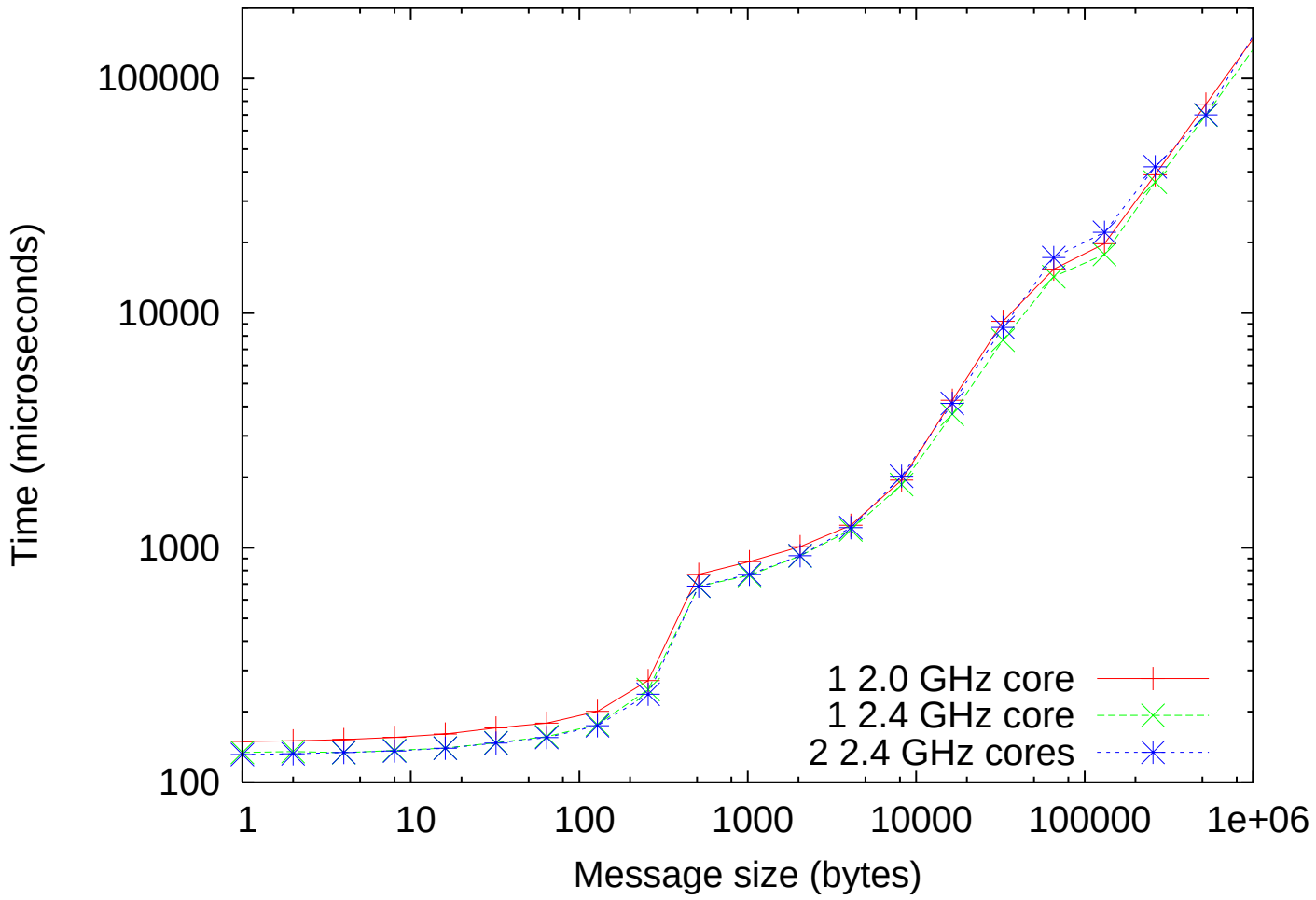
Analysis

- Faster processor reduces latency by 1 µs for SN
- Serialized Portals processing for VN increases latency by more than 2 µs
- Cray has since addressed some of the issues for Generic Portals and is moving to fine-grained locking
  - See Mark Pagel's talk on Thursday
- Accelerated Portals doesn't have these issues
  - See my talk tomorrow
- 20% increase in CPU speed led to 20% increase in streaming bandwidth for SN
- Reduction in peak bandwidth for VN is due to competition for HyperTransport bandwidth

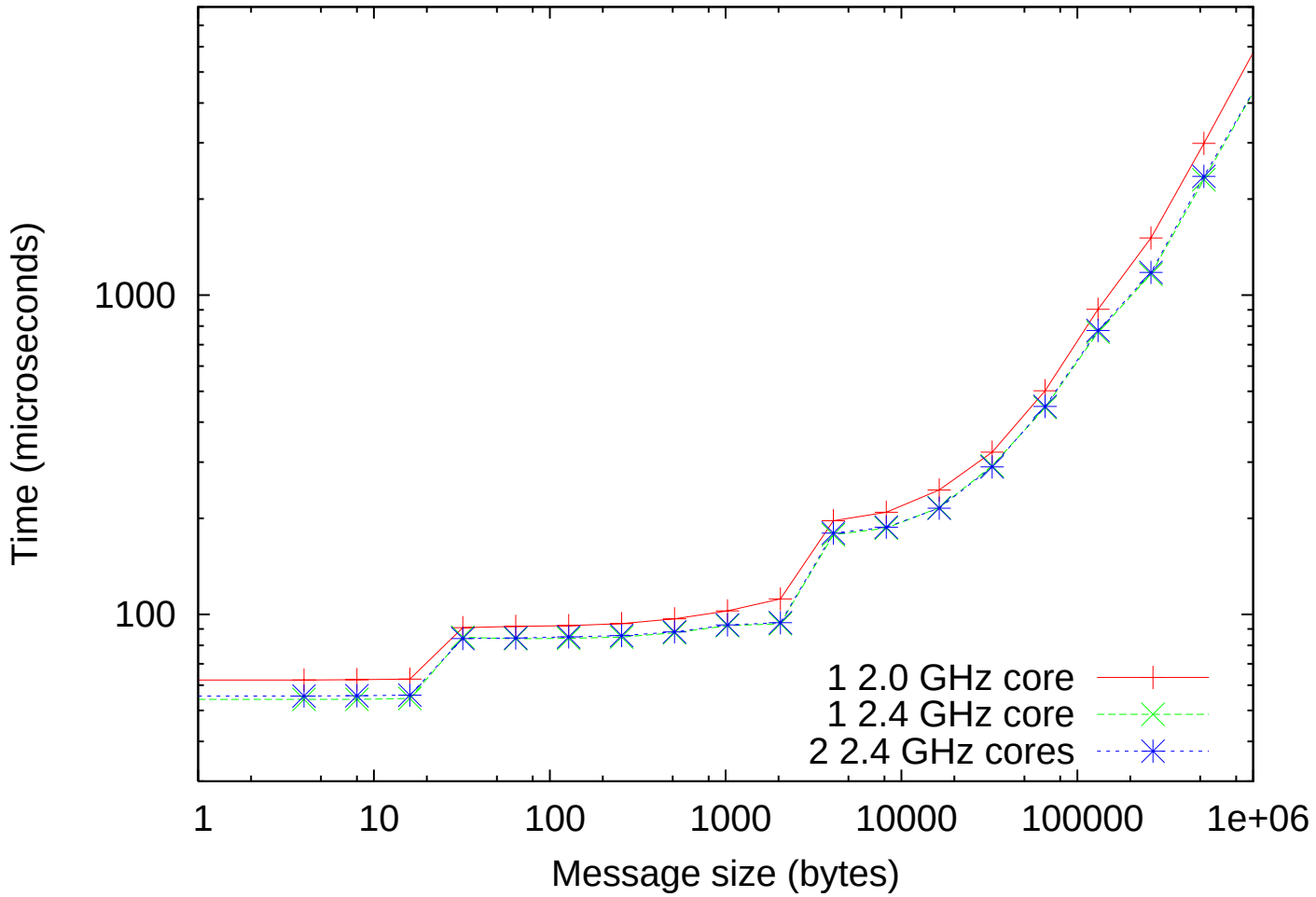
Higher-Level Benchmarks

- Pallas MPI Benchmarks (PMB)
  - Alltoall
  - Allreduce
  - Reduce
- HPC Challenge Benchmarks
  - HPL
  - STREAMS
  - PTRANS
  - FFT

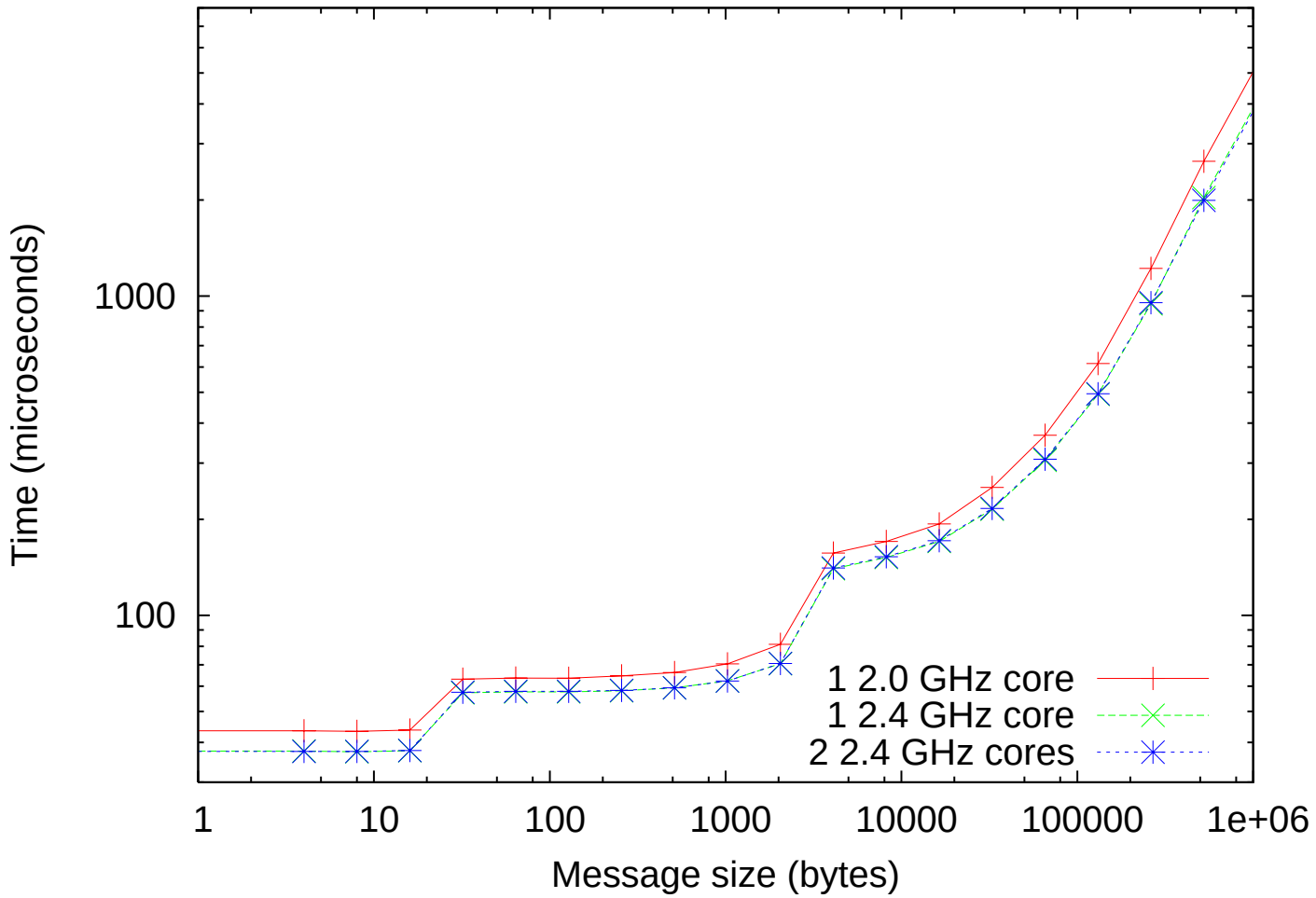
64-Node PMB Alltoall



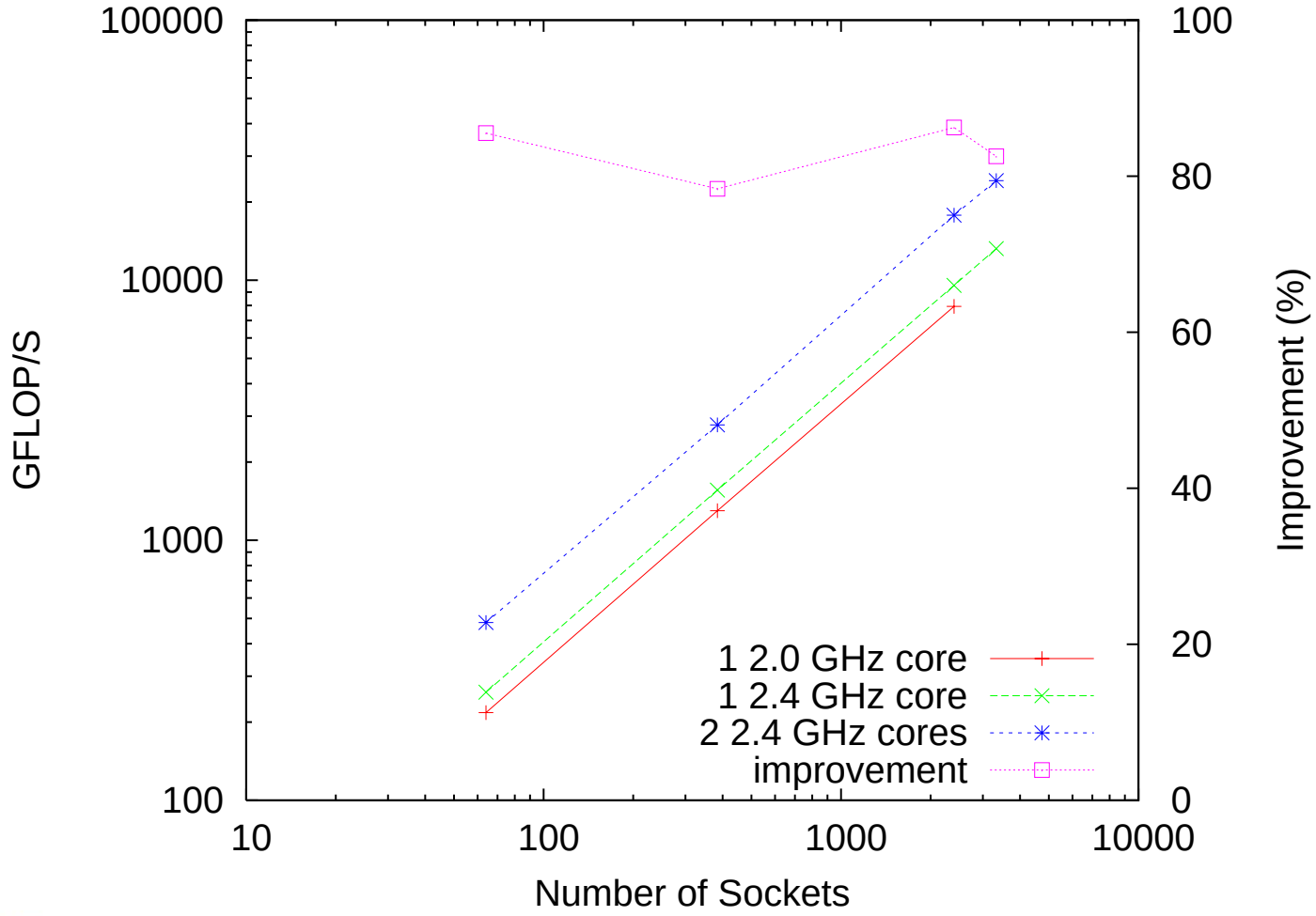
64-Node PMB Allreduce



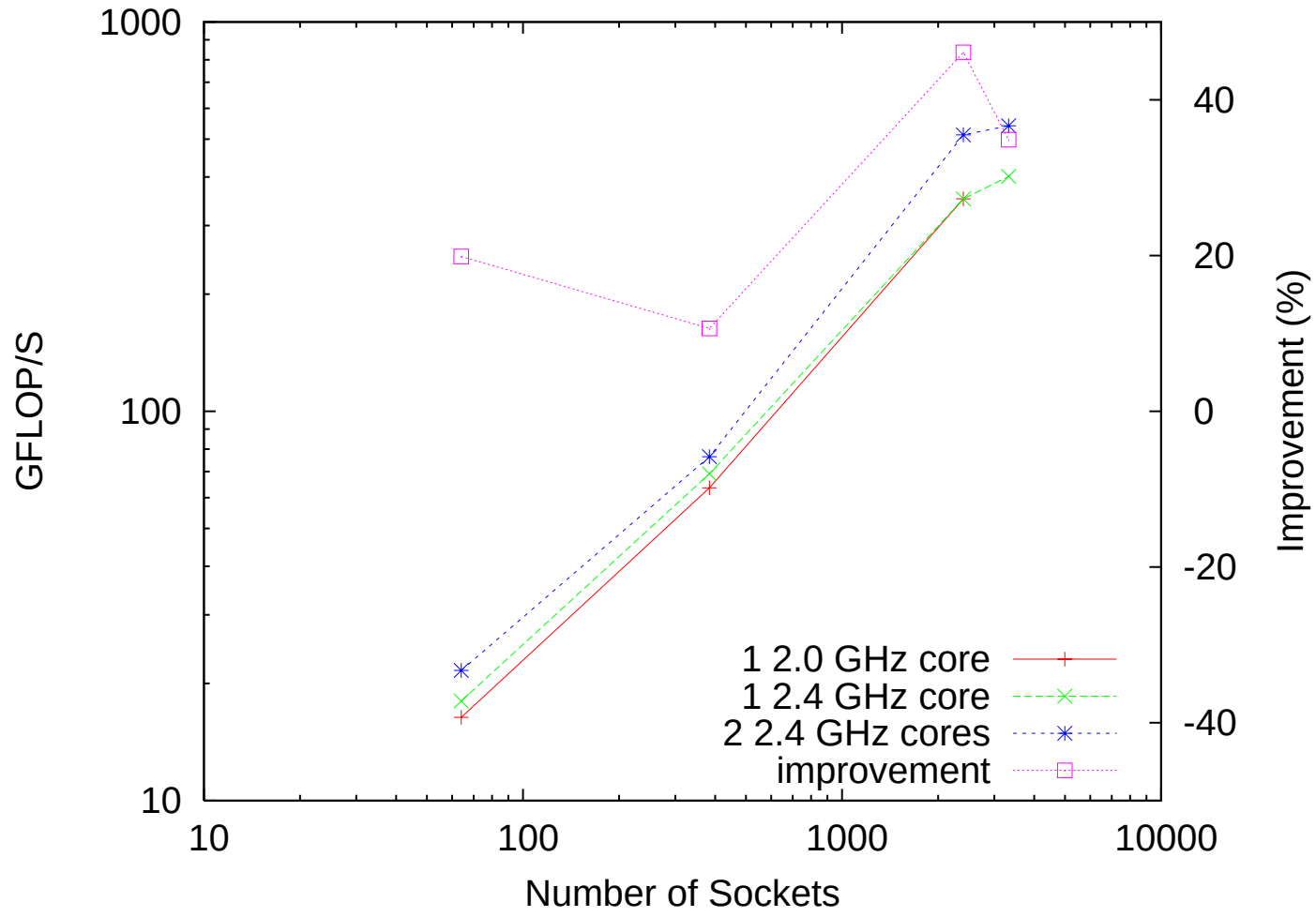
64-Node PMB Reduce



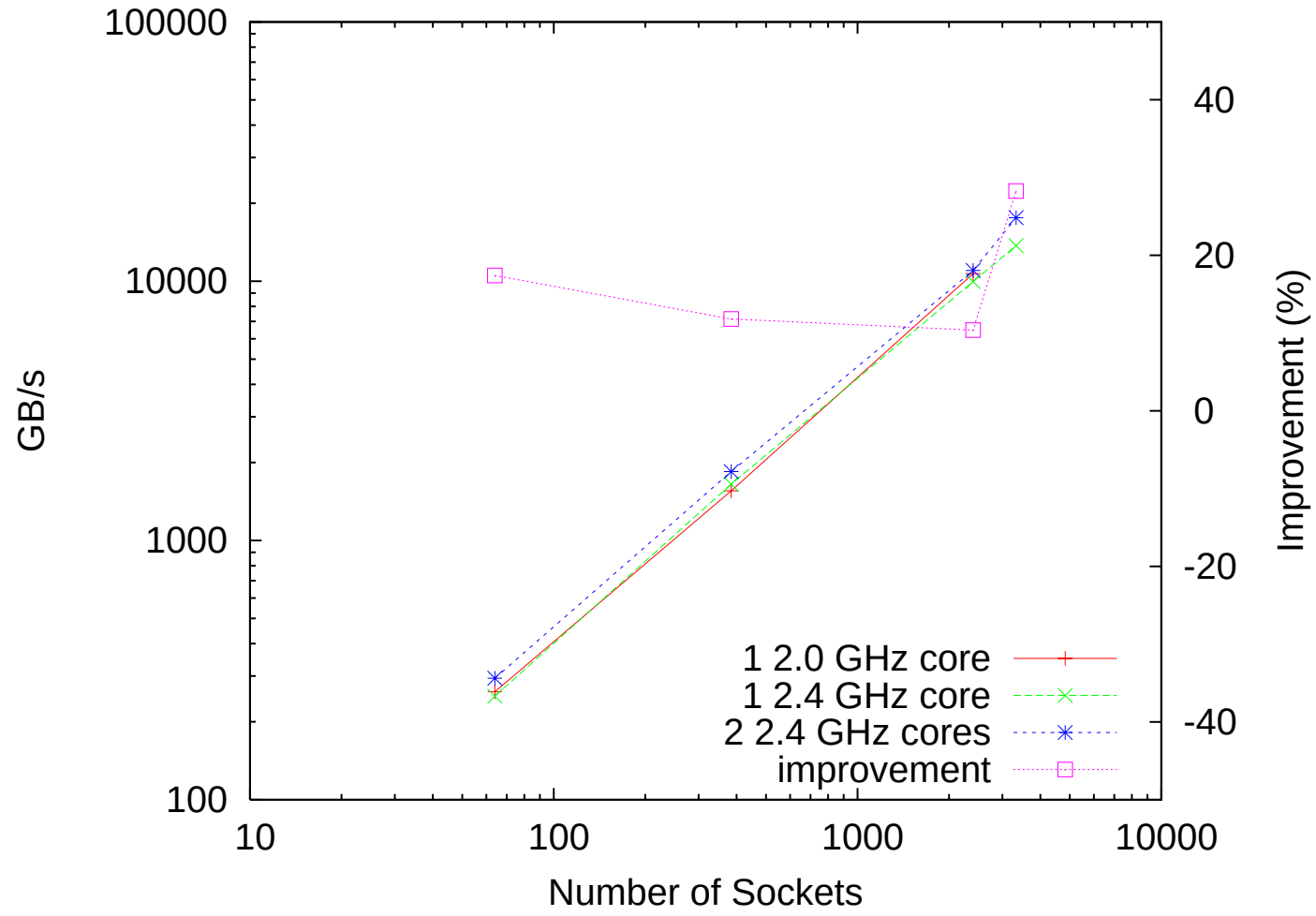
HPL


100000
10000
1000
100
GFLOP/S
100
80
60
40
20
0
Improvement (%)
10
100
1000
10000
Number of Sockets
1 2.0 GHz core
1 2.4 GHz core
2 2.4 GHz cores
improvement

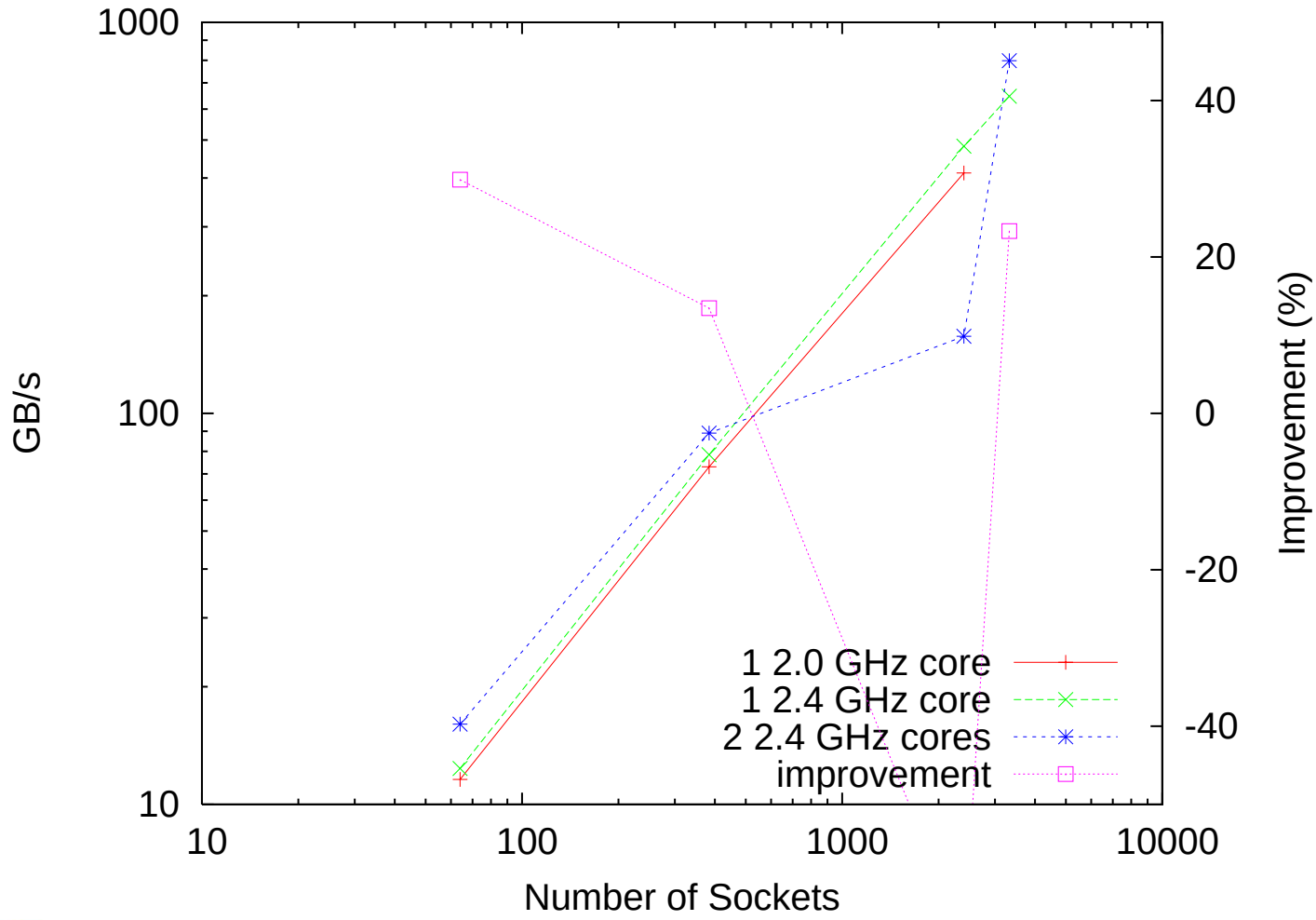
FFT


1000
100
10
GFLOP/S
40
20
0
-20
-40
Improvement (%)
10
100
1000
10000
Number of Sockets
1 2.0 GHz core
1 2.4 GHz core
2 2.4 GHz cores
improvement

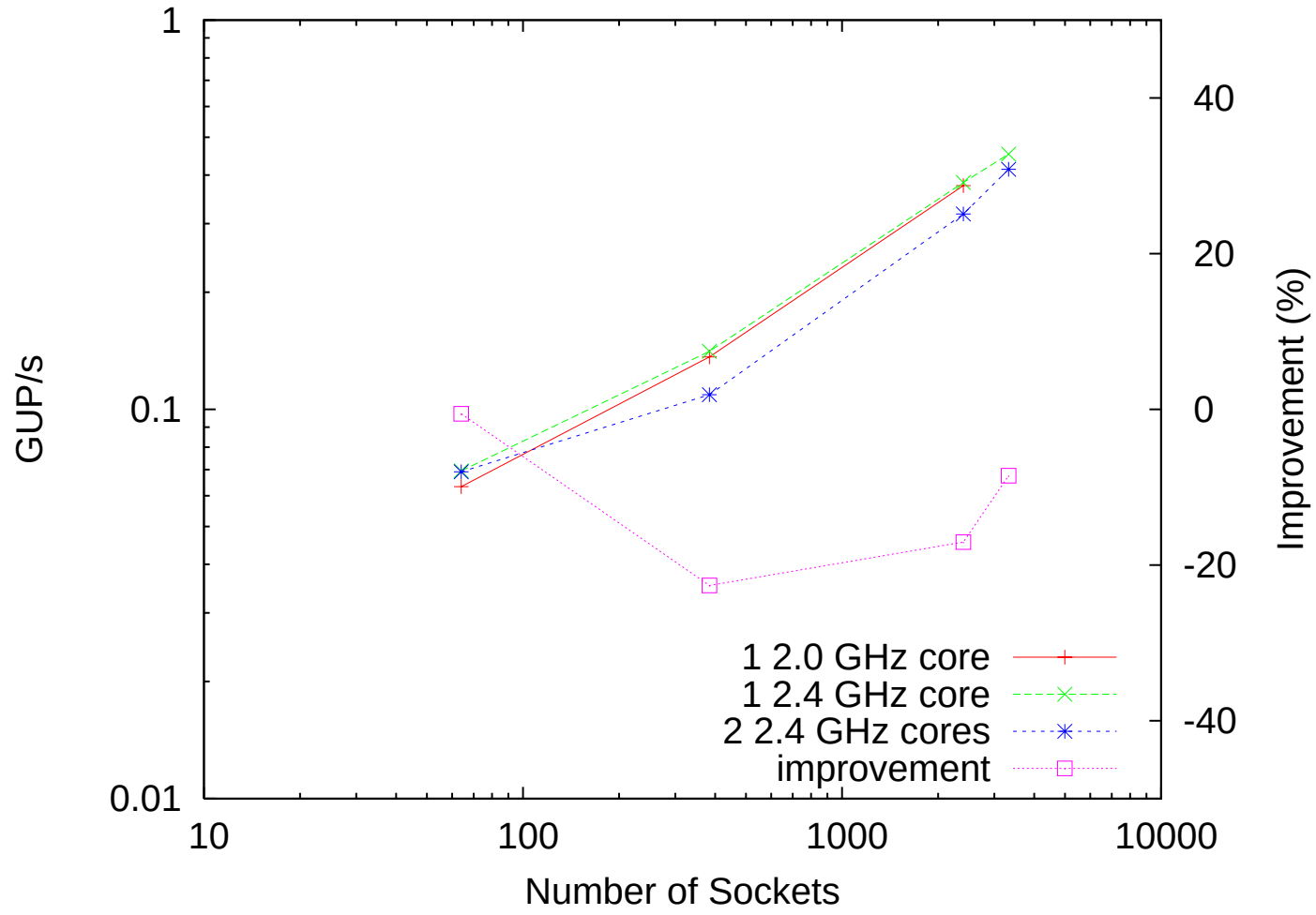
STREAMS


100000
10000
1000
100
GB/s
40
20
0
-20
-40
Improvement (%)
10
100
1000
10000
Number of Sockets
1 2.0 GHz core
1 2.4 GHz core
2 2.4 GHz cores
improvement

PTRANS


1000
100
10
GB/s
40
20
0
-20
-40
Improvement (%)
10
100
1000
10000
Number of Sockets
1 2.0 GHz core
1 2.4 GHz core
2 2.4 GHz cores
improvement

RandomAccess



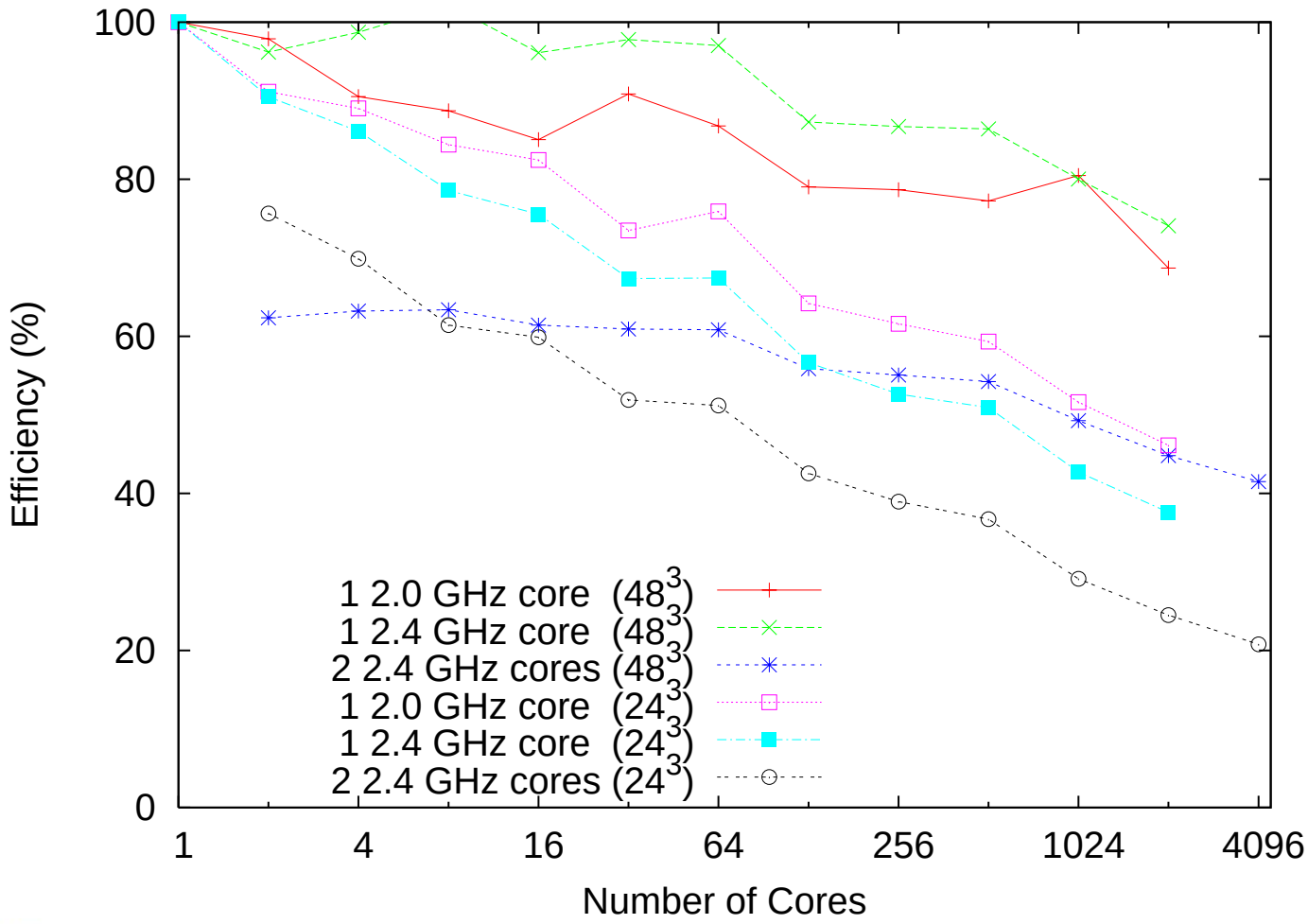
Applications

- SAIC's Adaptive Grid Eulerian (SAGE)
  - Multi-dimensional, multi-material Eulerian hydrodynamics code
  - Large class of production appications at LANL
- Parallel, Time-dependent SN (PARTISN)
  - Designed to solve th time-independent or dependent multigroup discrete ordinates form of the Boltzmann transport equation in several different geometries
- CTH
  - Multi-material, large deformation, strong shock wave, solid mechanics code

Analysis of Dual-Core Impacts

- Scaling efficiency
  - Number of cores
  - Fixed size per core
- Overall improvement
  - Number of sockets
  - Fixed size per socket

PARTISN - Diffusion

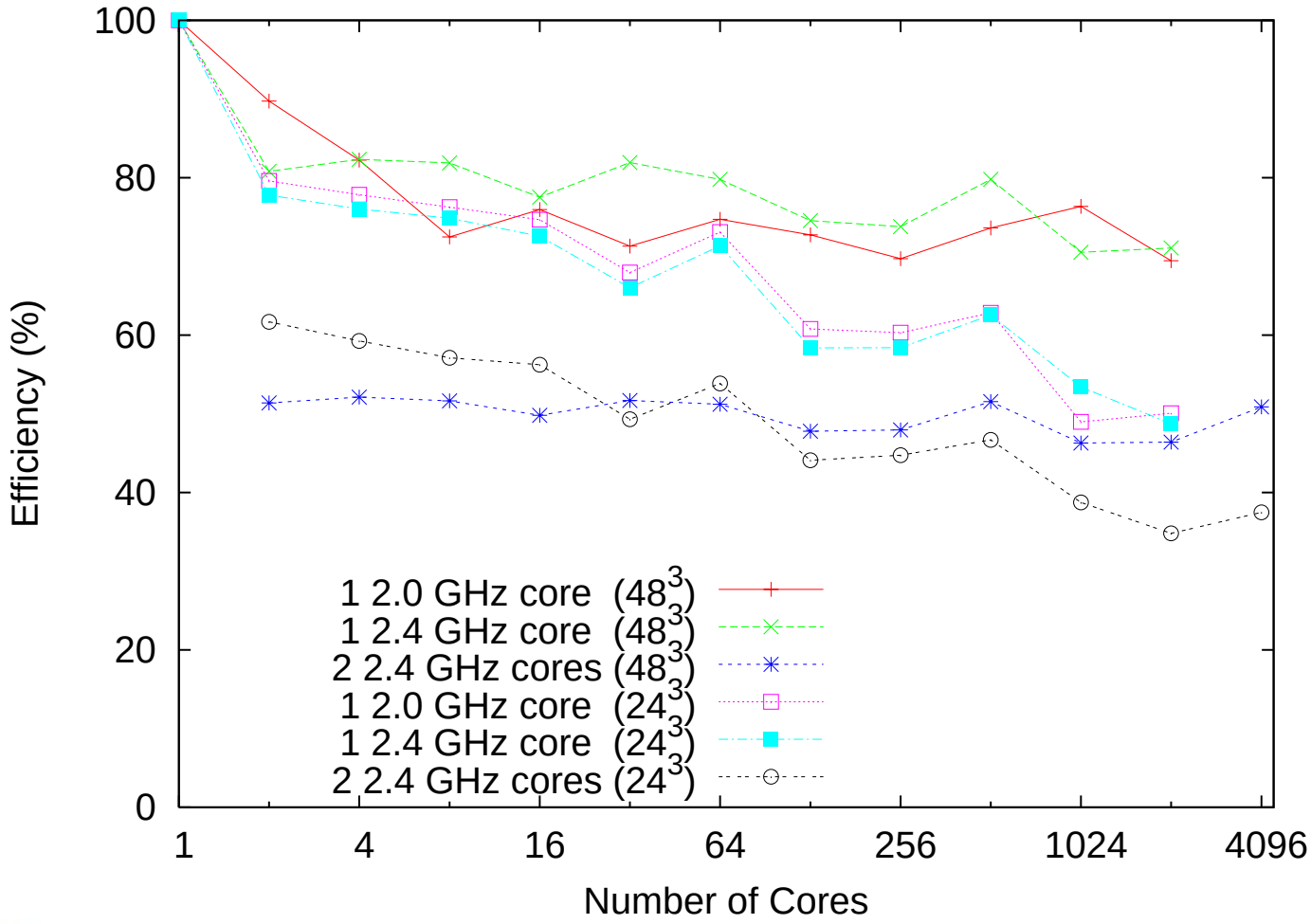


CCIM

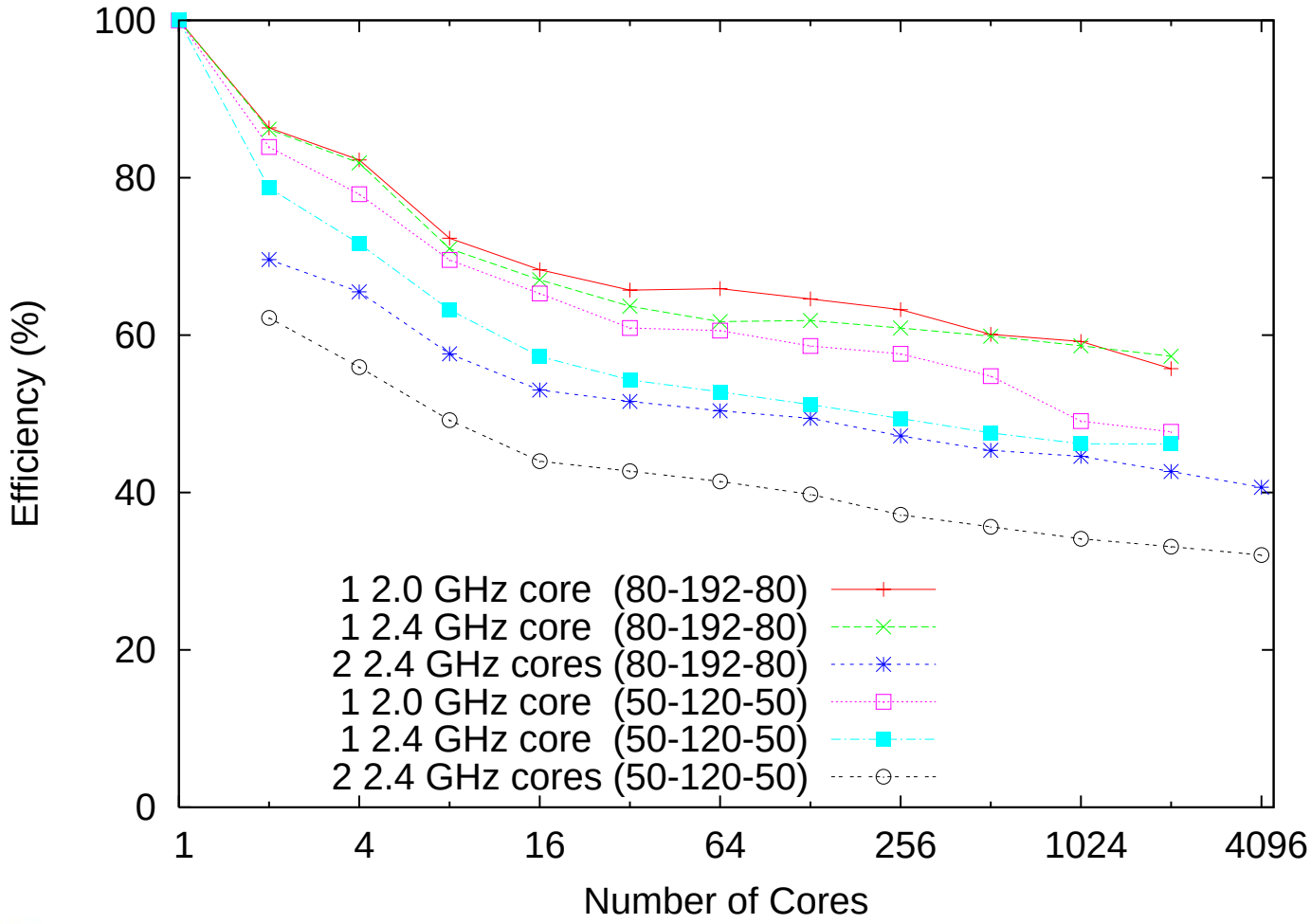
NNSA National Nuclear Security Administration

Sandia National Laboratories

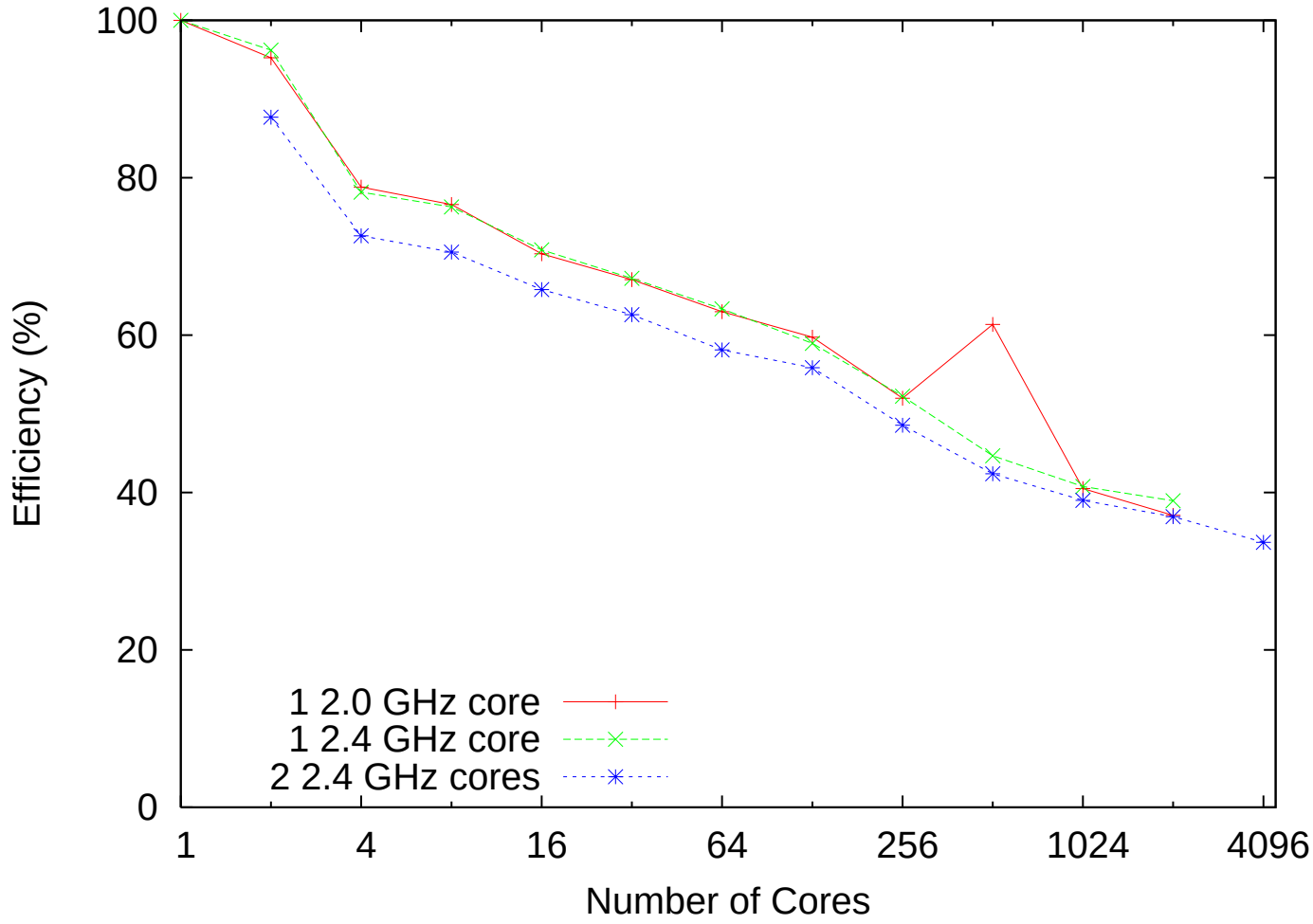
PARTISN - Transport



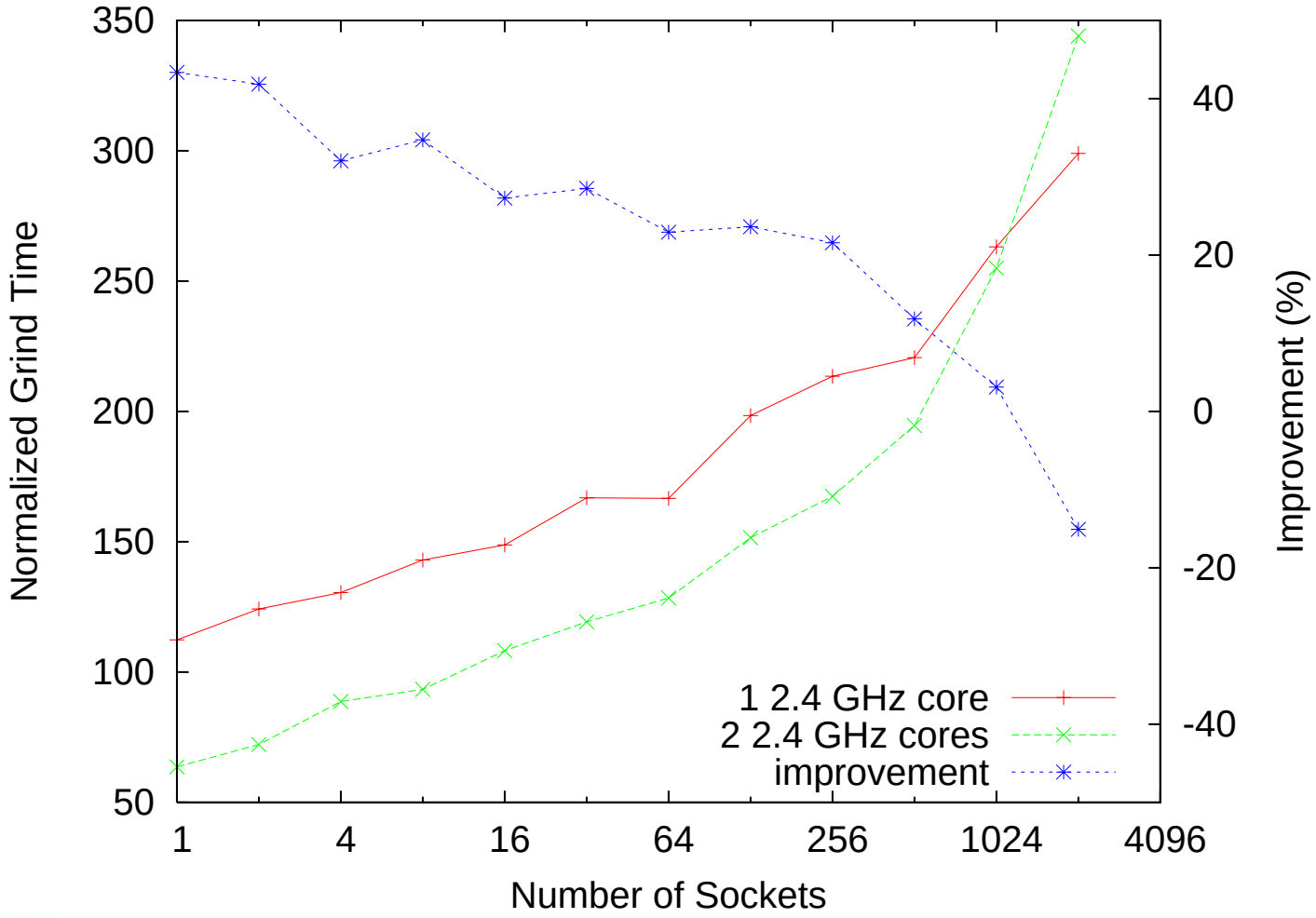
CTH


100
80
60
40
20
0
Efficiency (%)
1
4
16
64
256
1024
4096
Number of Cores
1 2.0 GHz core (80-192-80)
1 2.4 GHz core (80-192-80)
2 2.4 GHz cores (80-192-80)
1 2.0 GHz core (50-120-50)
1 2.4 GHz core (50-120-50)
2 2.4 GHz cores (50-120-50)

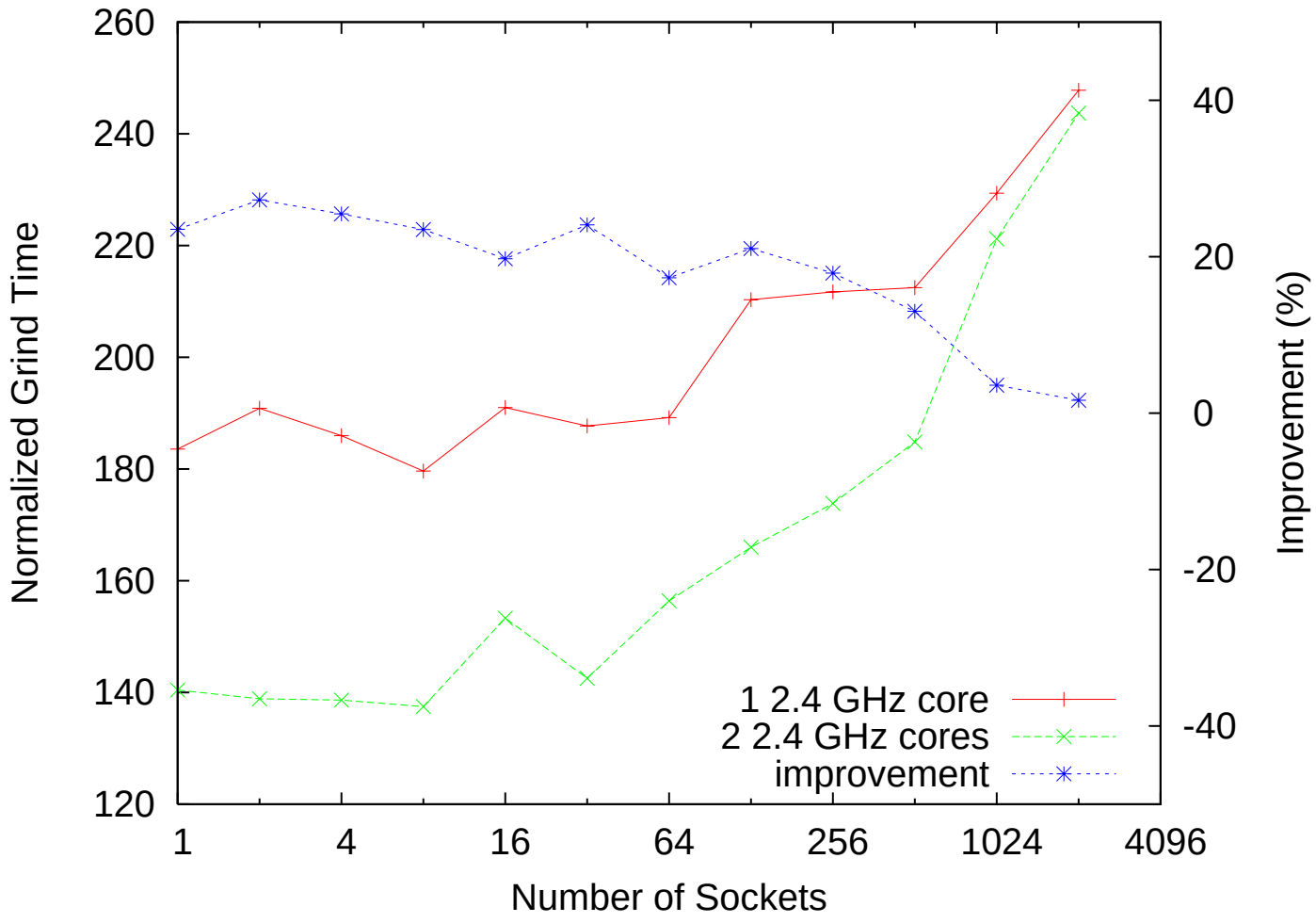
SAGE


100
80
60
40
20
0
Efficiency (%)
1
4
16
64
256
1024
4096
Number of Cores
1 2.0 GHz core
1 2.4 GHz core
2 2.4 GHz cores

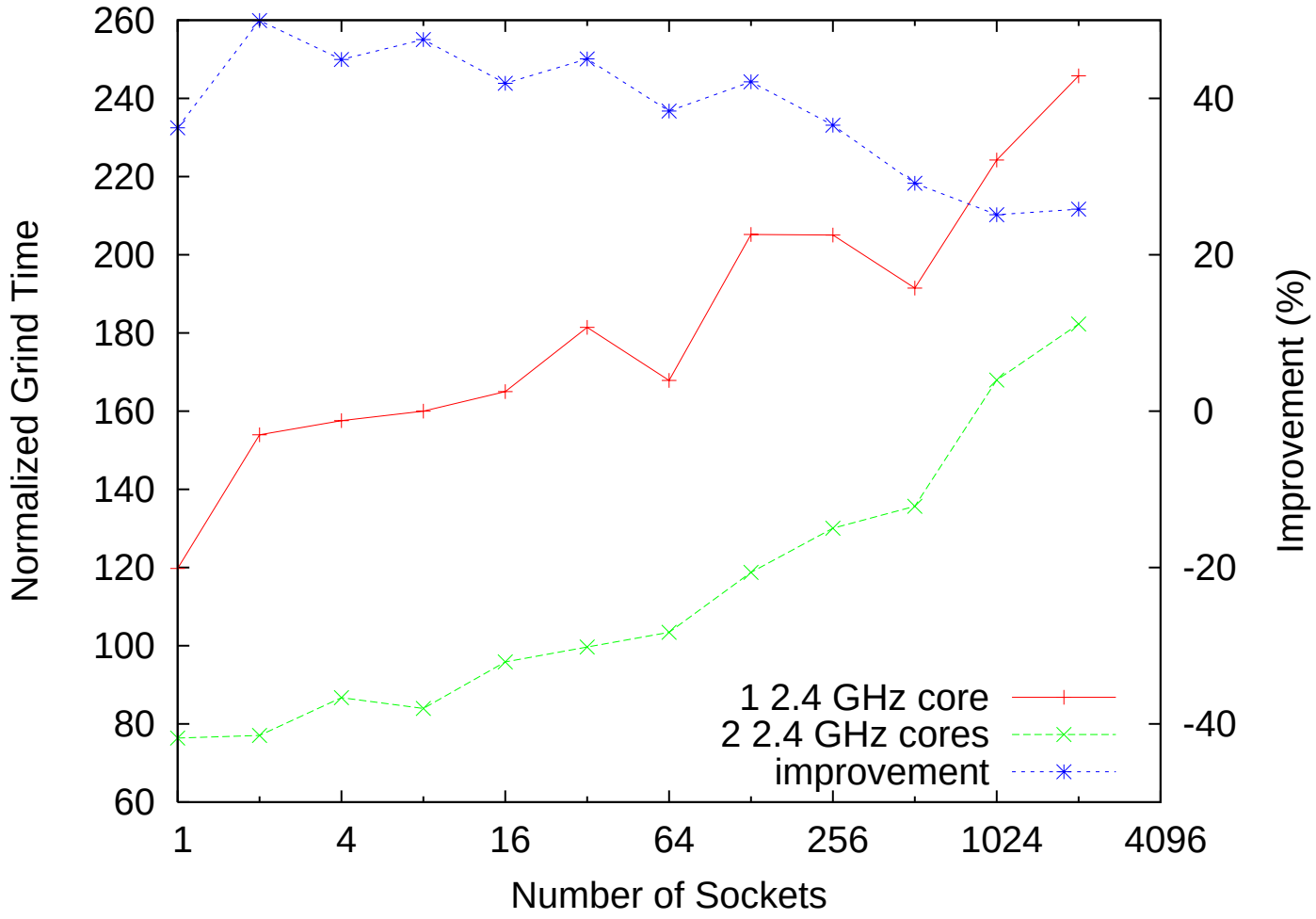
PARTISN - Diffusion - 24^3



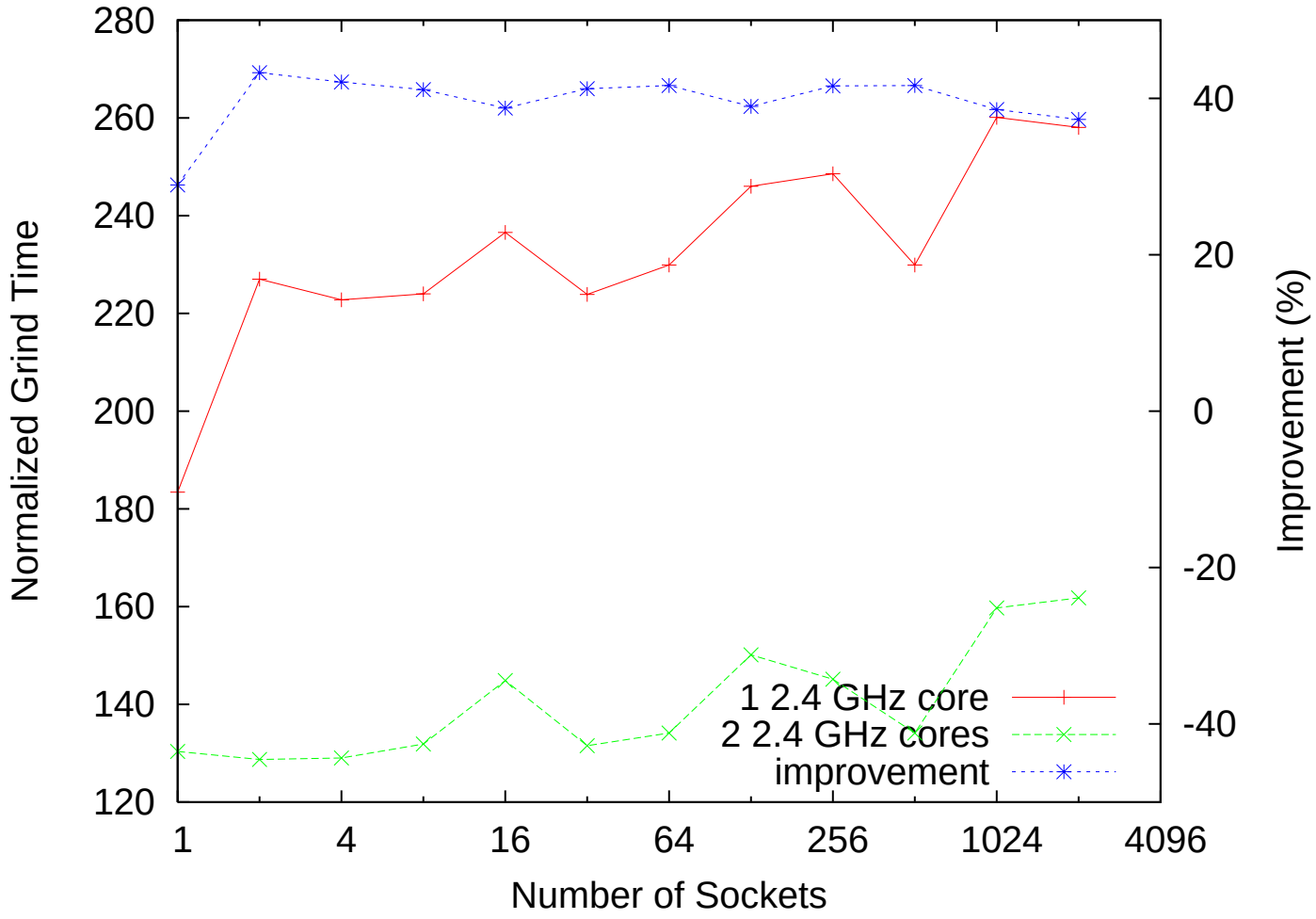
PARTISN – Diffusion - 48^3



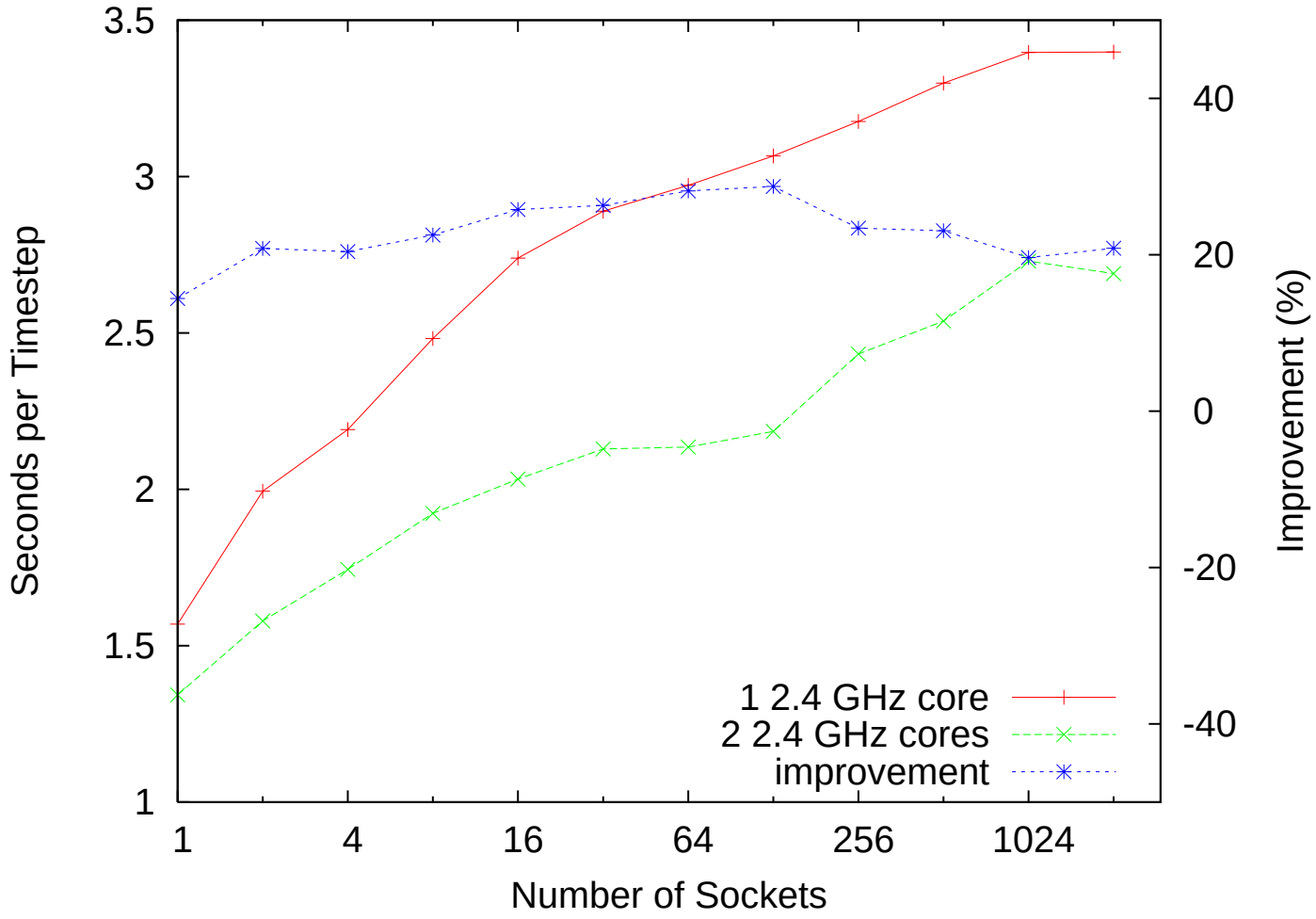
PARTISN – Transport - 24^3



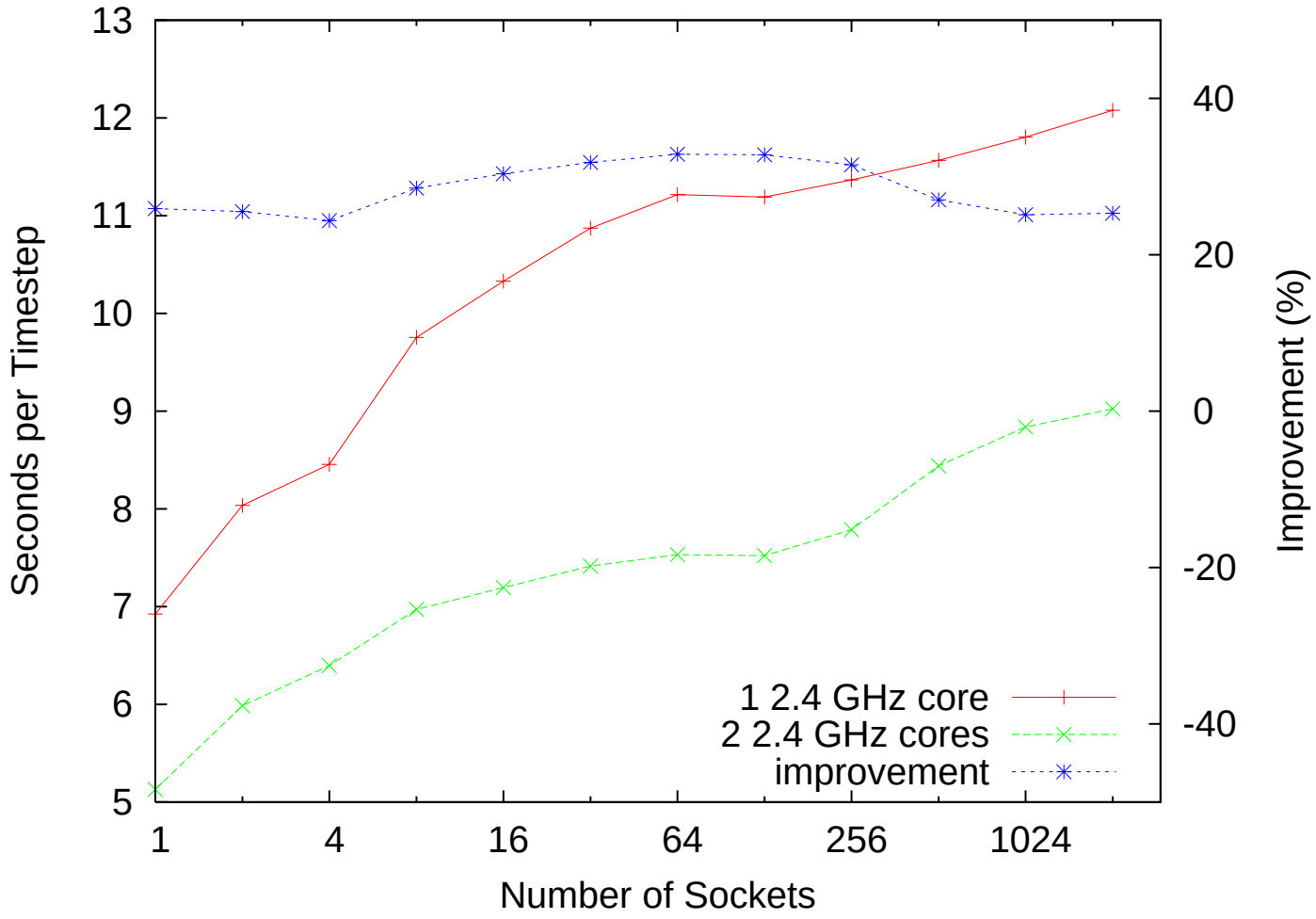
PARTISN – Transport - 48^3



CTH – 50x120x50



CTH – 80x190x80



Conclusions

- Half round-trip latency decreases by 15%
- Peak uni-directional bandwidth doubled
- Peak bi-directional bandwidth increased ~80%
- Small message throughput increased by over 20%
- Adding a second core provides 20-50% performance increase to real applications
- Impact on scalability is minor
- Scaling degradation visible at very high node counts

Acknowledgments

- Sue Kelly
- Kevin Pedretti
- John VanDyke