SAND2011-3341C

Investigating the Impact of the Cielo Cray XE6 Architecture on Scientific Application Codes

Courtenay Vaughan, Mahesh Rajan, Richard Barrett, Douglas Doerfler, and Kevin Pedretti

Sandia National Laboratories

LSPP workshop at IPDPS
May 2011

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.

Cielo

- Cray XE6 just upgraded to 8944 compute nodes
- dual-socket oct-core AMD Magny-Cours nodes
- clocked at 2.4 GHz
- 32 GB of 1.333 GHz DDR3 memory per node
- 3D torus with Gemini interconnect
- have large machine and smaller machines

- smaller systems configured briefly as XT6 with same nodes and SeaStar interconnect

XT5

- Cray XT5 with 160 compute nodes
- dual socket with 6 core AMD Istanbul processors
- 2.4 GHz processors
- 32 GB of 800 MHz DDR2 Memory per node
- 6 x 4 x 8 3D torus with SeaStar 2.2

Sandia National Laboratories

Red Storm

- Started as 10368 single-core nodes and SeaStar 1.2 network
  - 2.0 GHz nodes with 333 MHz DDR1 memory
- Upgraded to 5 rows (12960) nodes, dual-core nodes, and SeaStar 2.1 network
  - 2.4 GHz nodes with 400 MHz DDR1 memory
- Center section then upgraded to quad-core nodes
  - Quad-cores are 2.2 GHz with 800 MHz DDR2 memory

Cray XT Evolution to XE6

Arch	Year	# nodes	cores/ soc	soc/ node	# cores	GHz	ops/ clock	DDR	GHz
XT3	2005	10368	1	1	10368	2.0	2	1	0.33
XT4	2006	12960	2	1	25920	2.4	2	1	0.4
XT4	2007	6720 6240	2 4	1	38400	2.4 2.2	2 4	1 2	0.4 0.8
XT5	2010	160	6	2	1920	2.4	4	2	0.8
XT6	2010	20	8	2	320	2.4	4	3	1.333
XE6	2011	8944	8	2	143104	2.4	4	3	1.333

Sandia National Laboratories

XE6 node

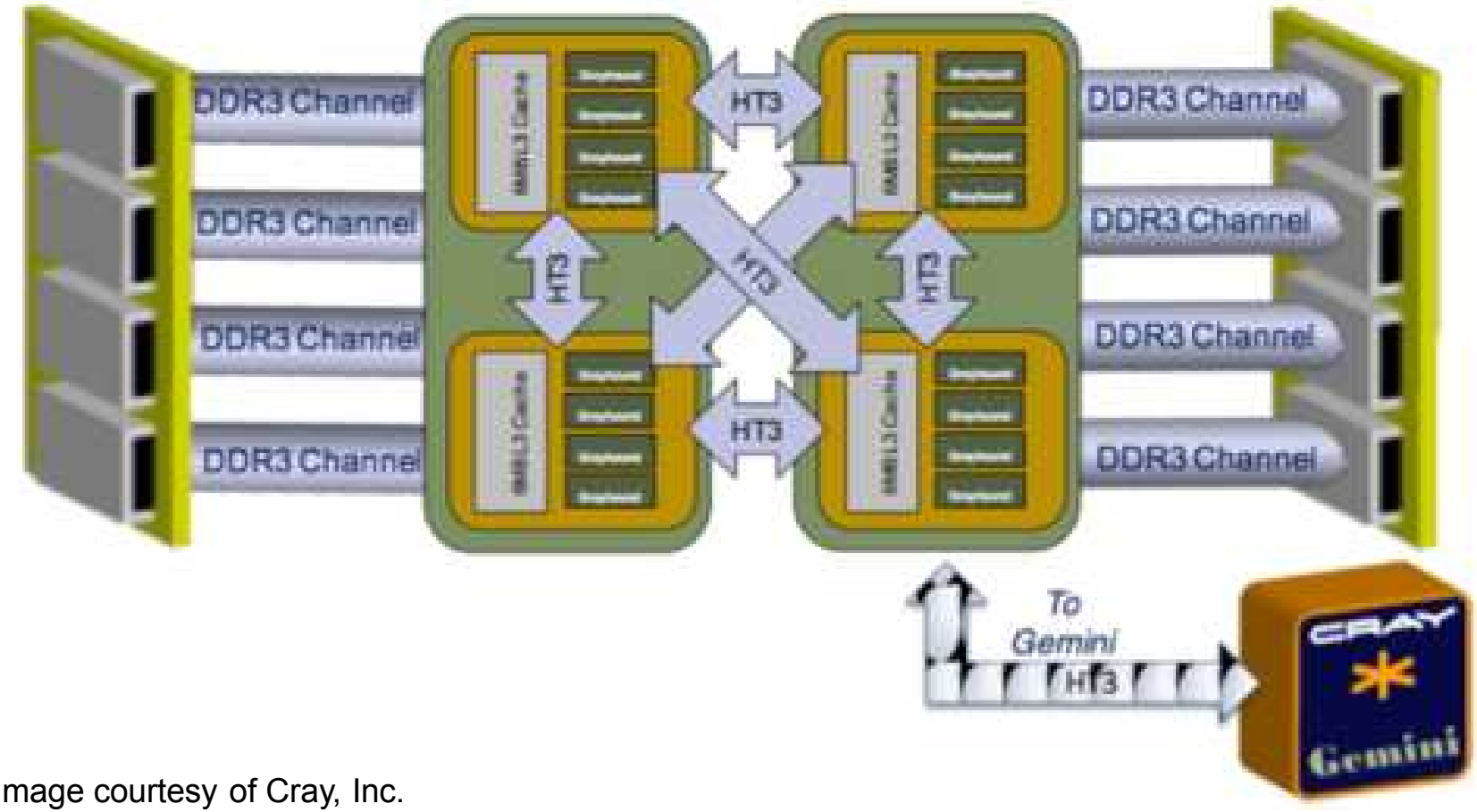


Image courtesy of Cray, Inc.

SeaStar and Gemini Interconnects

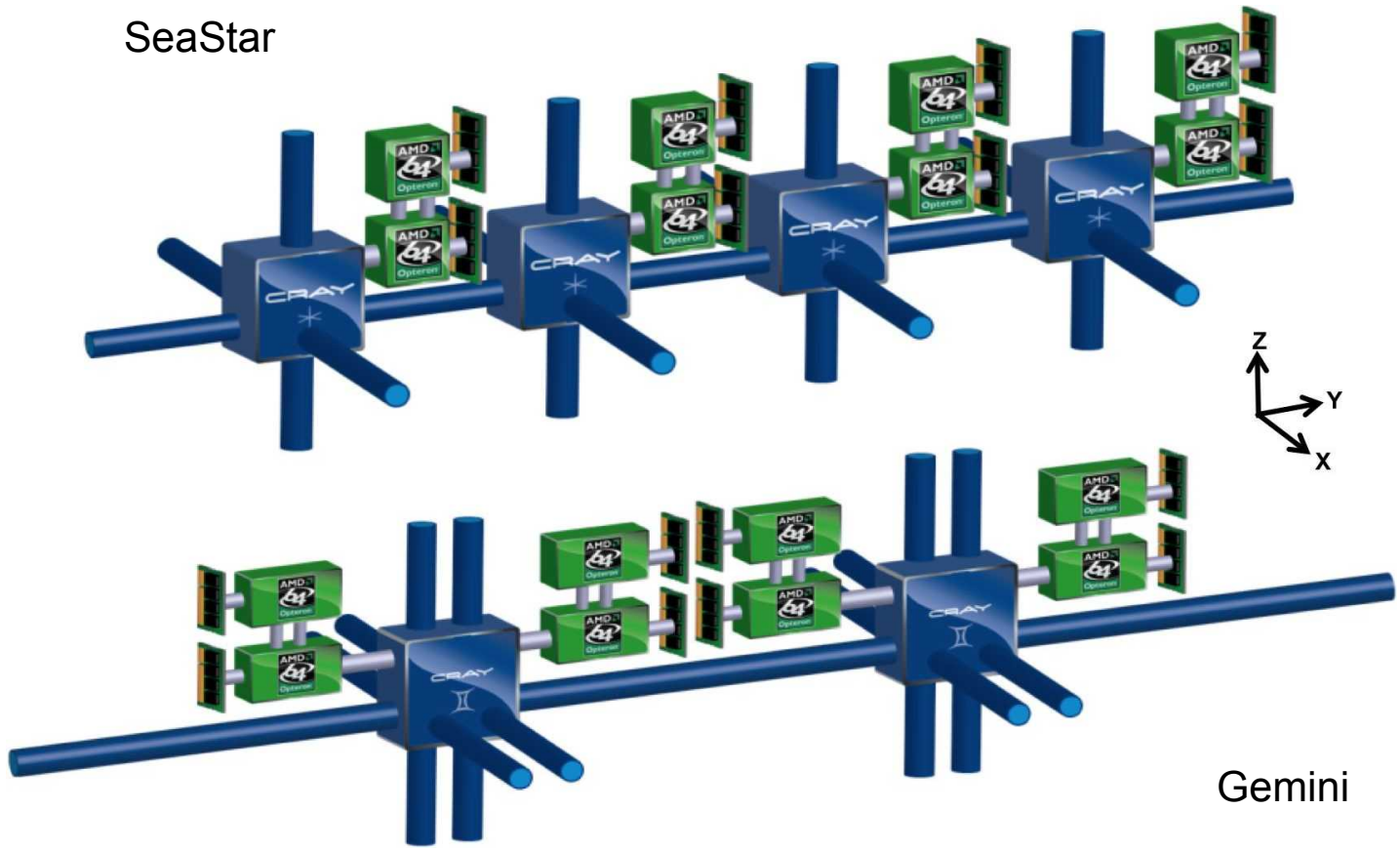
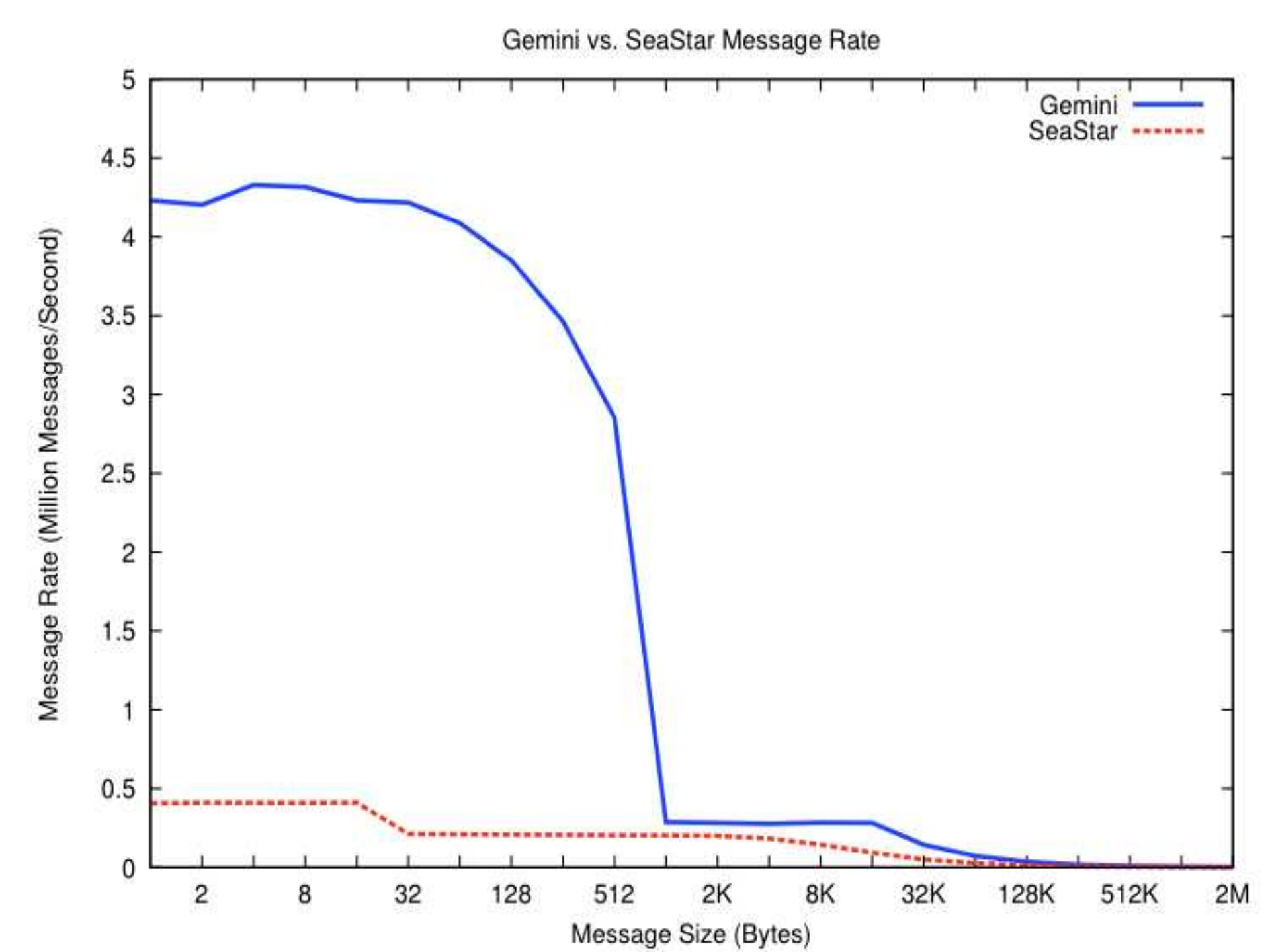
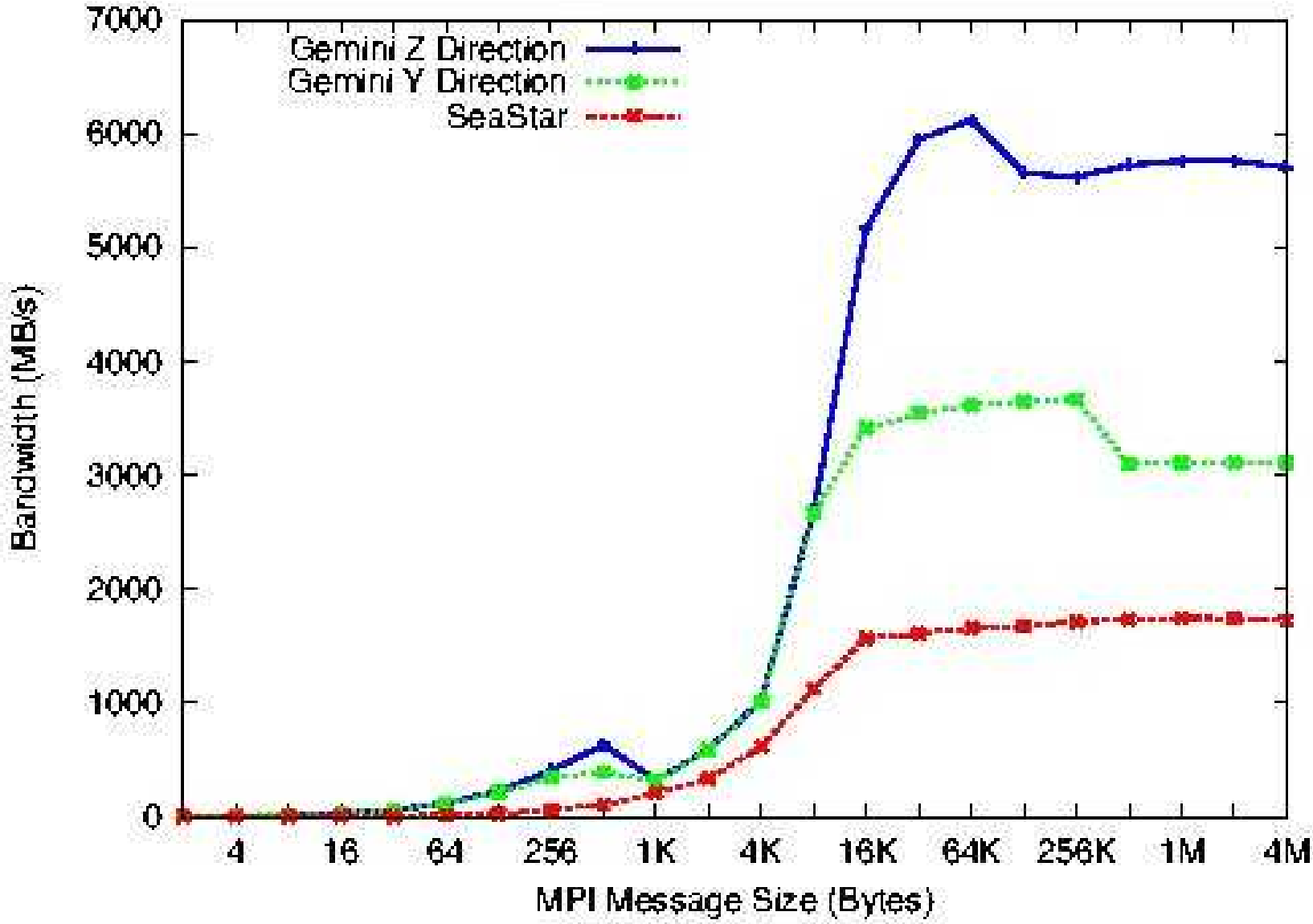


Image courtesy of Cray, Inc.


Gemini vs. SeaStar Message Rate
Gemini
SeaStar
Message Rate (Million Messages/Second)
Message Size (Bytes)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
2
8
32
128
512
2K
8K
32K
128K
512K
2M

Point to Point BW: Gemini vs. Seastar



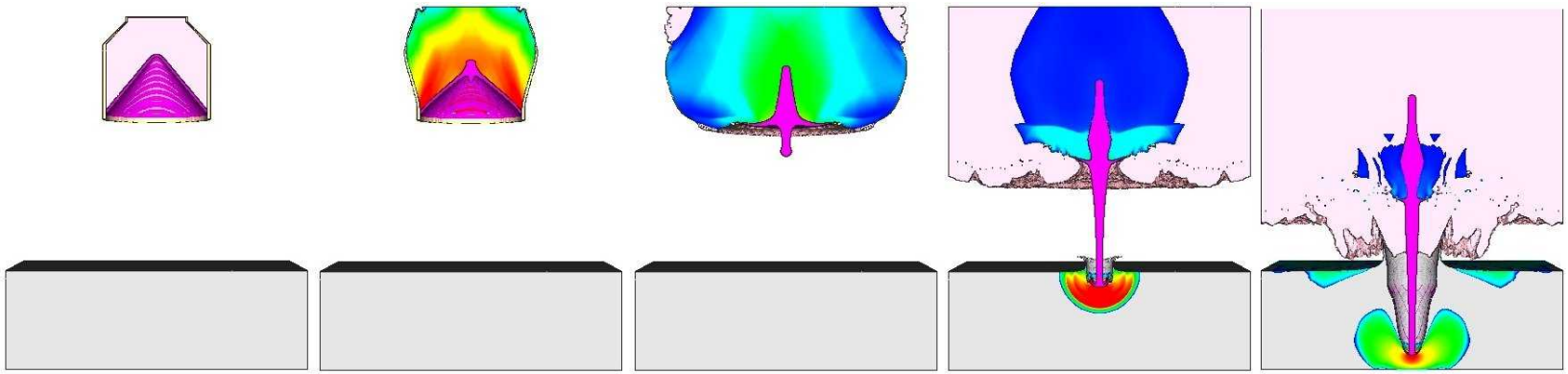
CTH

- Three-dimensional Eulerian shock hydrodynamics code
- Ran in flat mesh mode - no AMR (Automatic Mesh Refinement)
- Several points in each timestep where each processor sends a few large messages to up to six neighbors
- Messages are aggregated from several variables per cell

CTH Problem

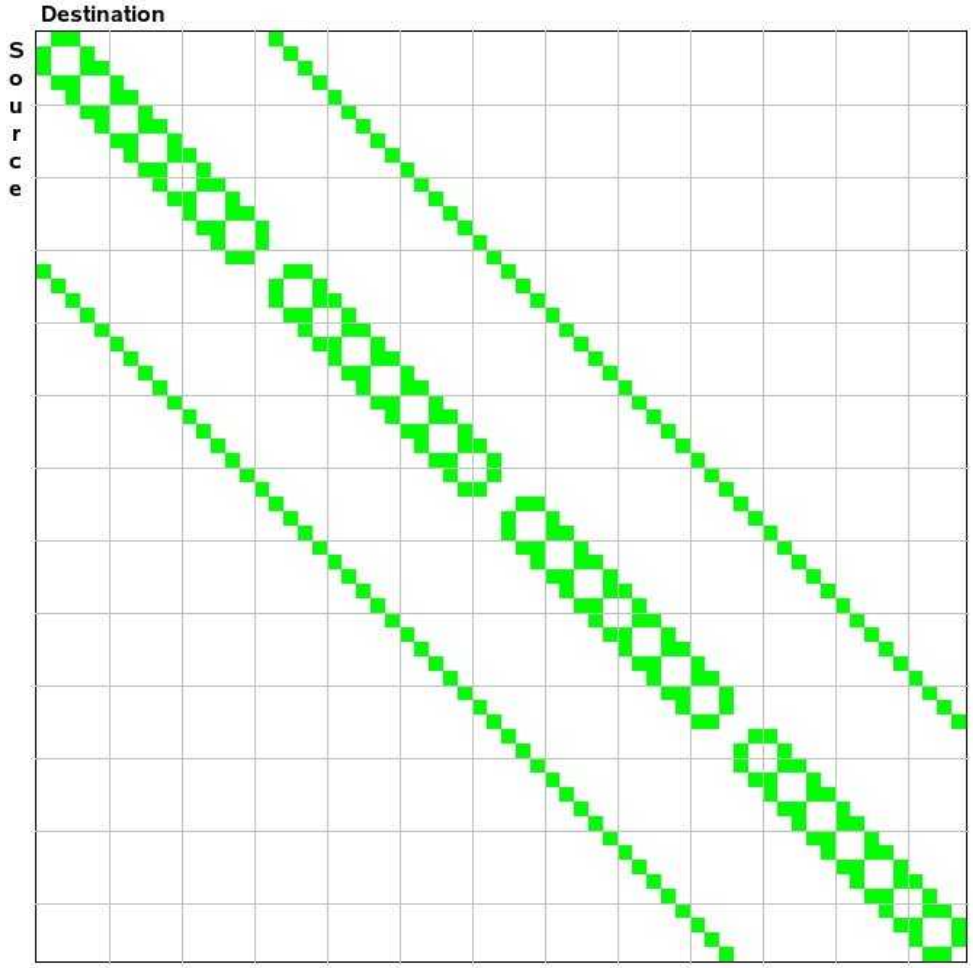
- explosively formed Shaped-Charge problem with 4 materials, high explosives, and 80 x 192 x 80 cells/processor in weak scaling mode
  - Messages aggregate 40 variables per cell and average 4.1 MB

Shaped Charge Problem



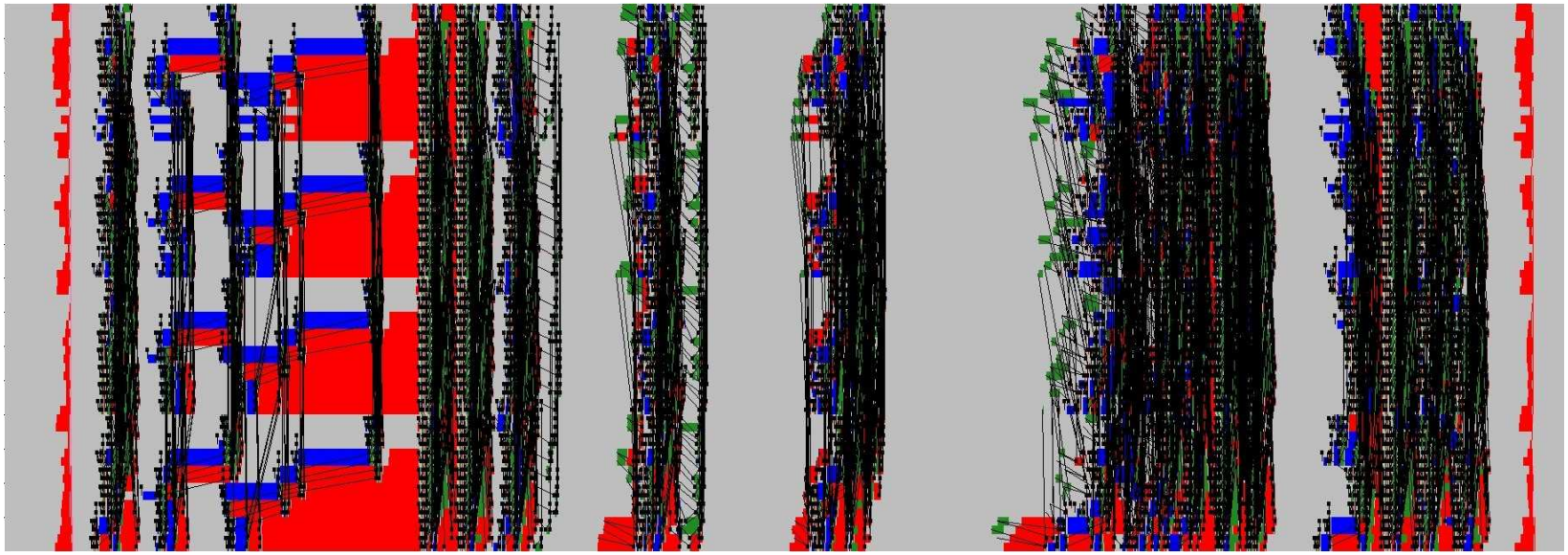
Sandia National Laboratories

CTH communication Matrix on 64 Cores



Sandia National Laboratories

CTH Communication Trace on 64 Cores

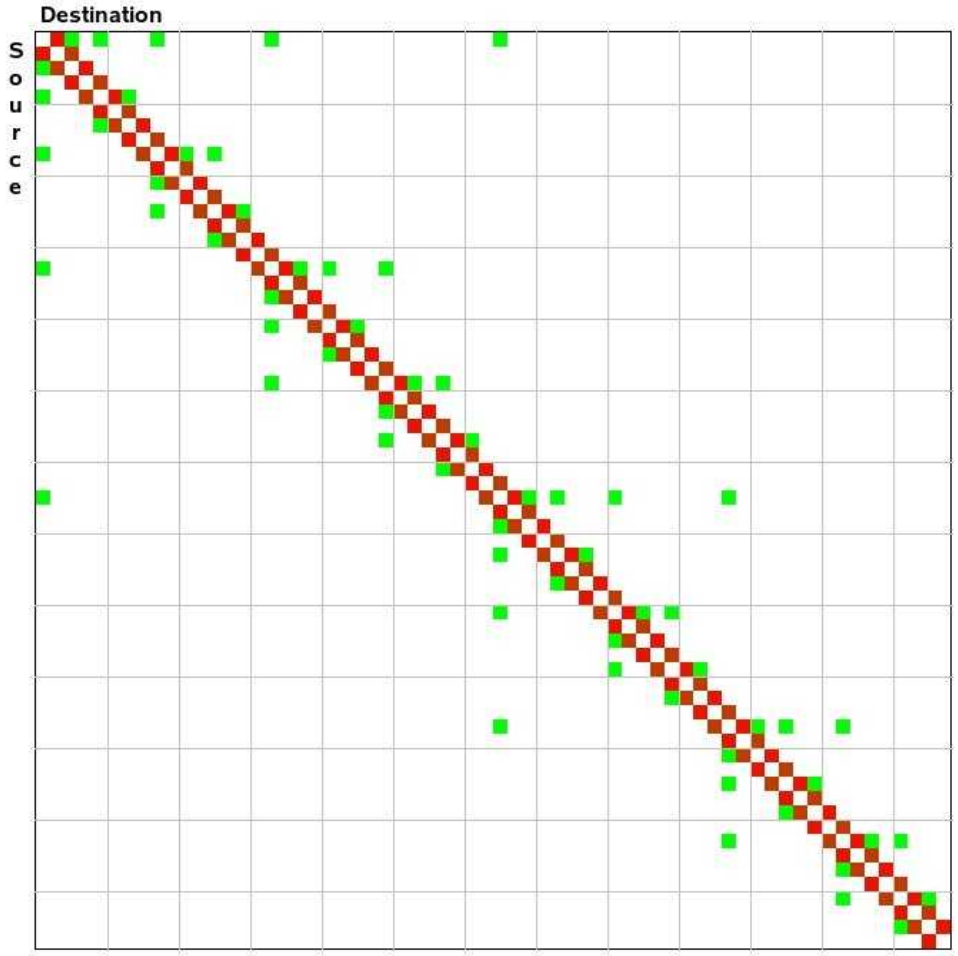


SAGE

- SAIC Adaptive Grid Eulerian
- multi-dimensional multi-material Eulerian hydrodynamics with AMR
- Uses modified 1D decomposition to partition the 3D mesh

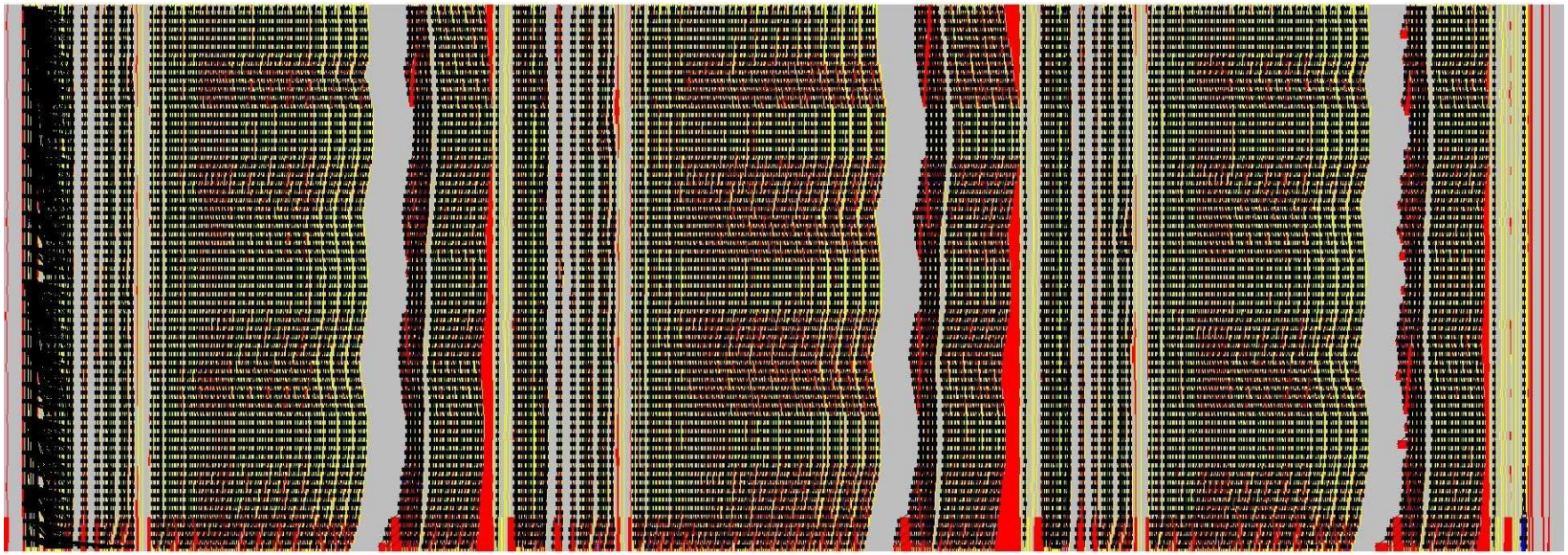
- Problem used is: timing
  - Does not make use of AMR

SAGE Communication Matrix on 64 Cores



Sandia National Laboratories

SAGE Communication Trace on 64 Cores

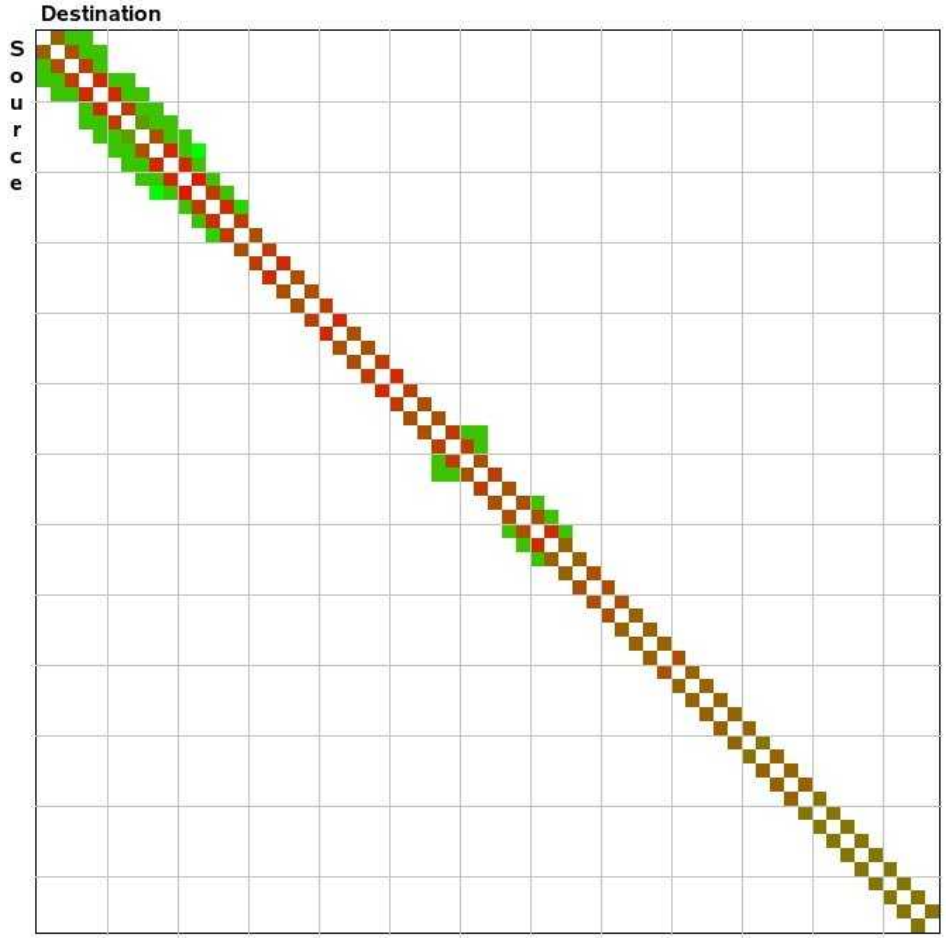


xNOBEL

- One, two, or three dimensional multi-material Eulerian hydrodynamics code
  - has support for high explosives
- Uses Continuous Adaptive Mesh Refinement
- Communicaton intensive with many small messages

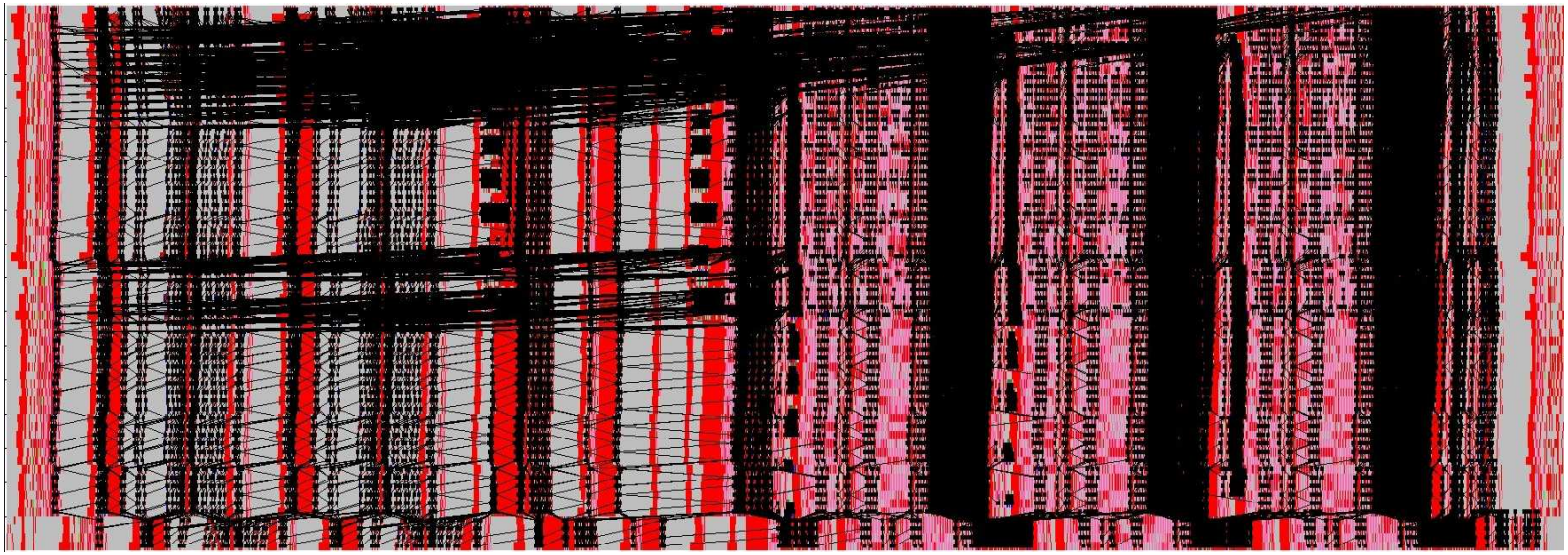
- Problem is a 2D shaped-charge problem

xNOBEL Communication Matrix on 64 Cores



Sandia National Laboratories

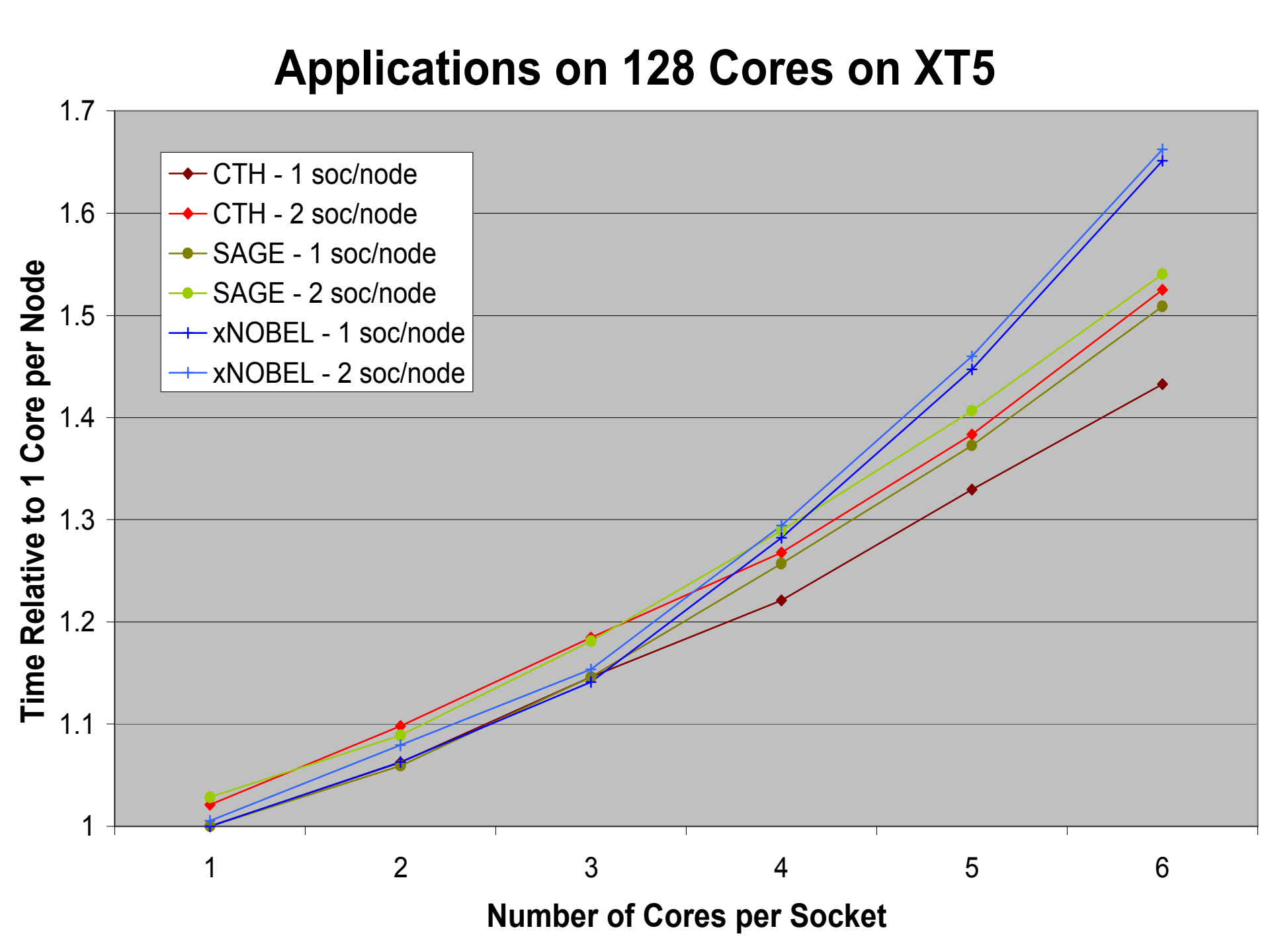
xNOBEL Communication Trace on 64 Cores

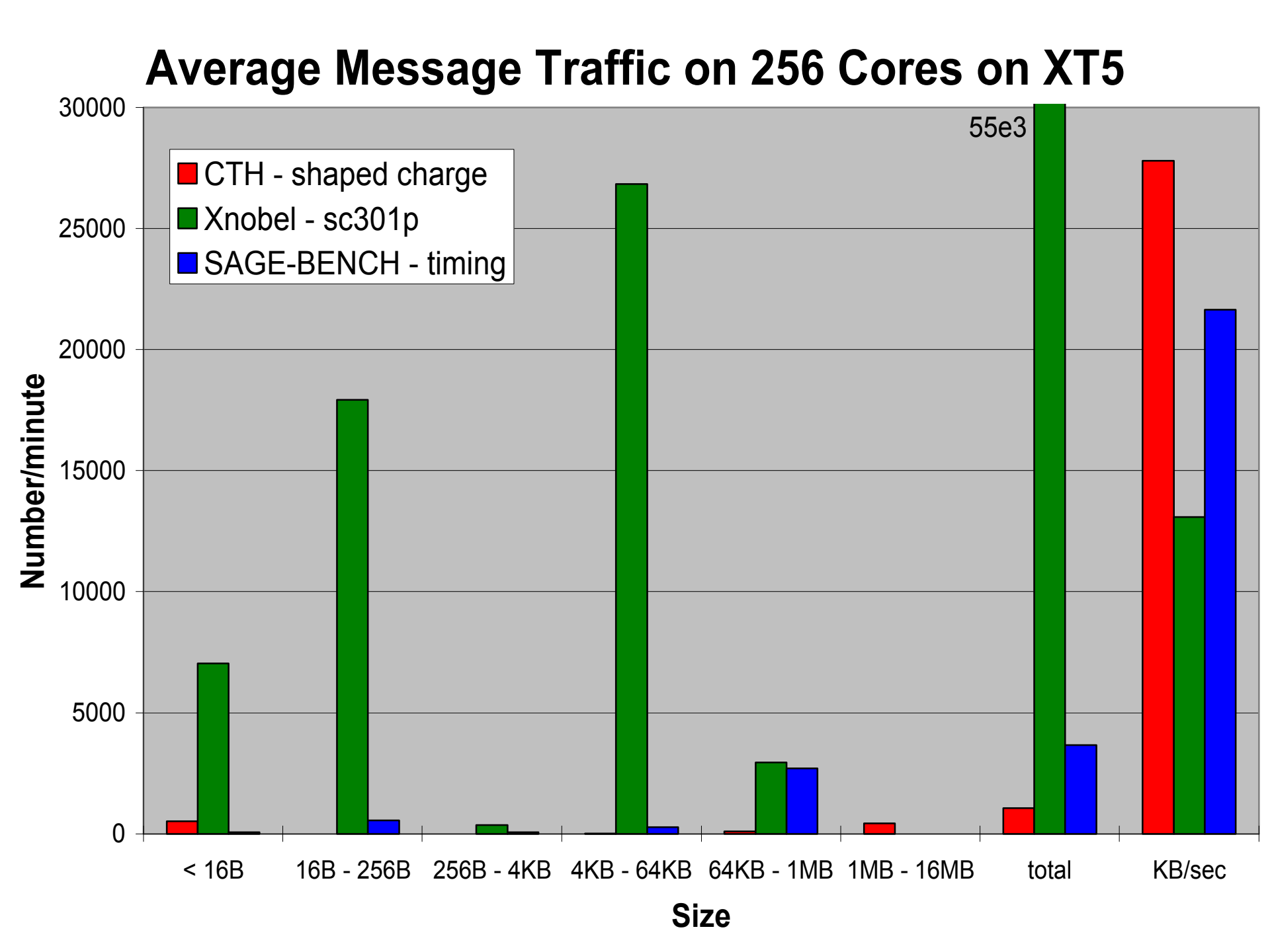


Sandia National Laboratories

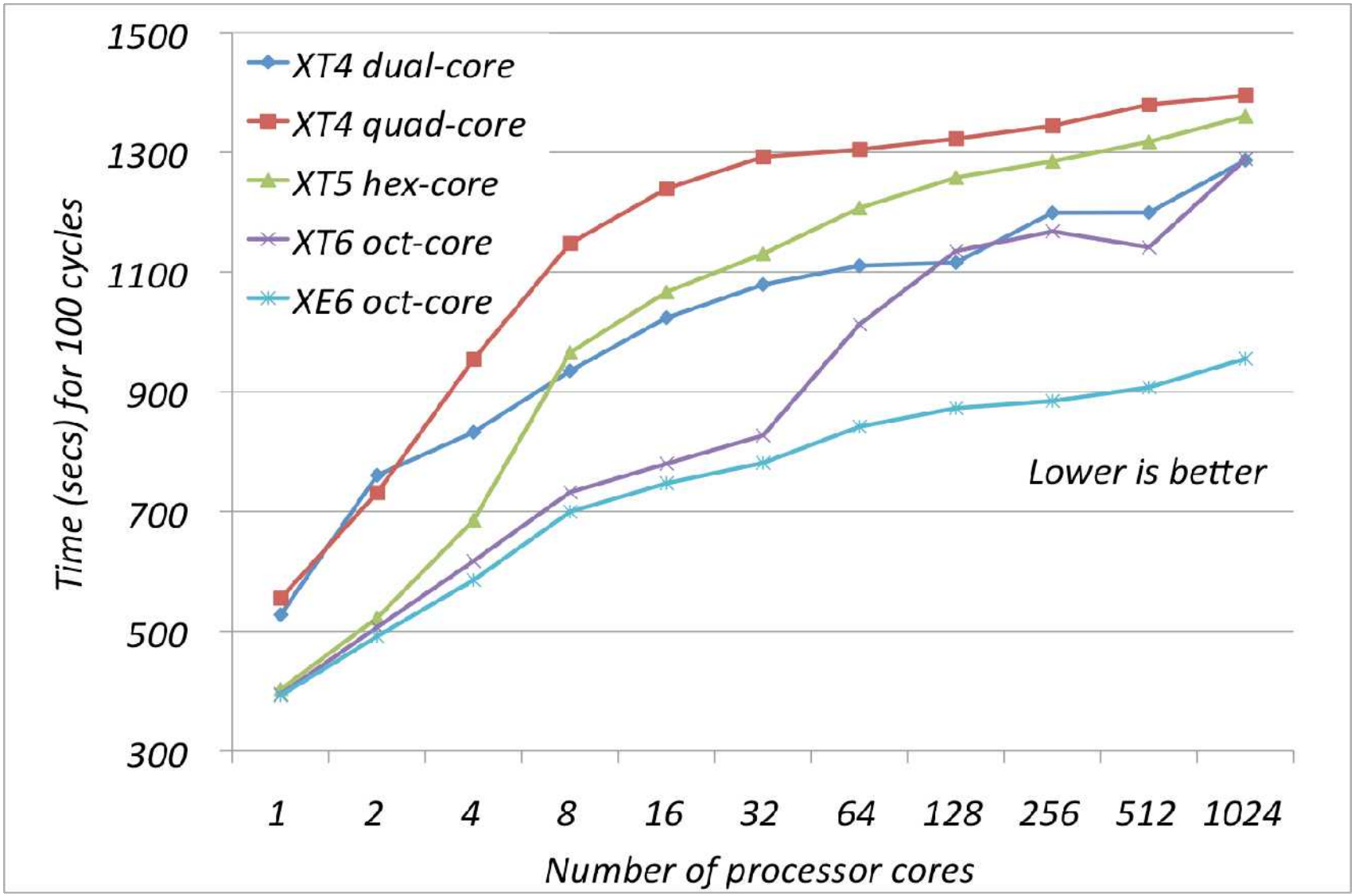
Profile on 256 Cores on XT5

Code	% time in compute	% time in MPI	% time in MPI_SYNC
CTH	52.2	38.4	9.5
SAGE	84.3	11.8	3.9
xNOBEL	33.3	22.8	44.0

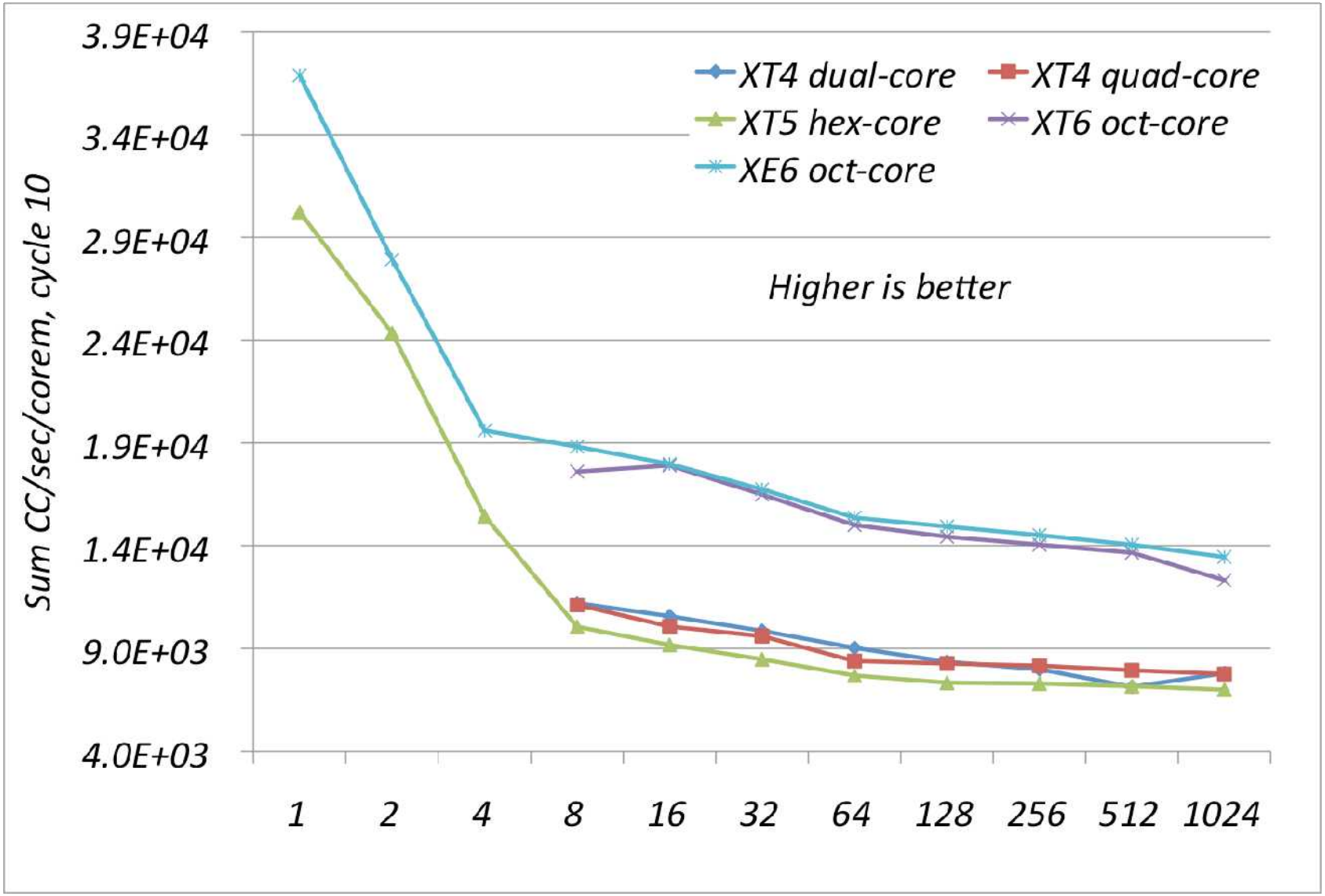

Applications on 128 Cores on XT5
Time Relative to 1 Core per Node
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1
1
2
3
4
5
6
Number of Cores per Socket
CTH - 1 soc/node
CTH - 2 soc/node
SAGE - 1 soc/node
SAGE - 2 soc/node
xNOBEL - 1 soc/node
xNOBEL - 2 soc/node


Average Message Traffic on 256 Cores on XT5
30000
25000
20000
15000
10000
5000
0
Number/minute
55e3
CTH - shaped charge
Xnobel - sc301p
SAGE-BENCH - timing
< 16B
16B - 256B
256B - 4KB
4KB - 64KB
64KB - 1MB
1MB - 16MB
total
KB/sec
Size

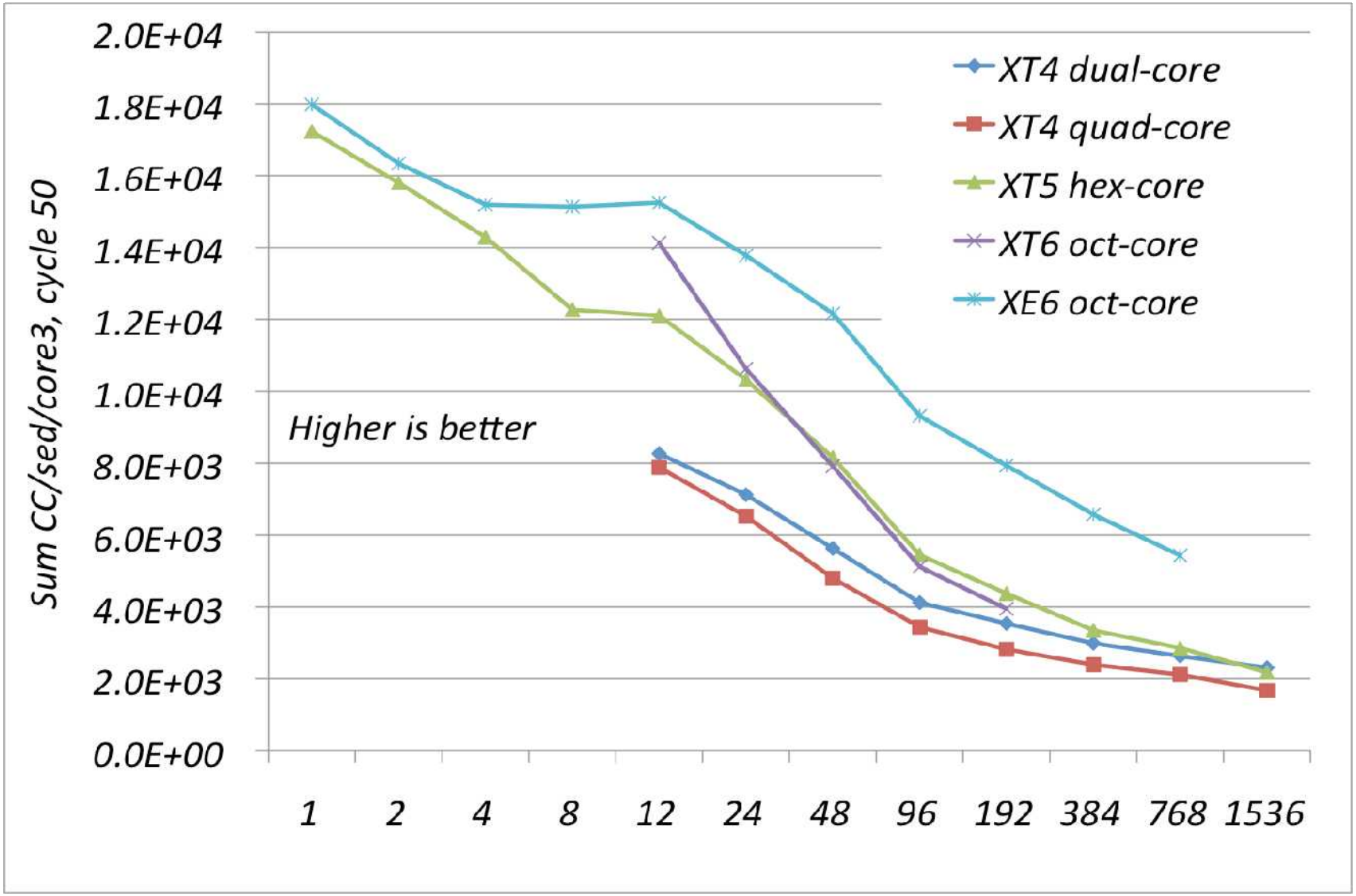
CTH Performance



SAGE Performance



xNOBEL Performance



Summary of Results

- Three codes that do similar computations differently and are helped differently moving through the machine evolution
- CTH has large network bandwidth requirements and shows some evolutionary performance improvement moving to the XE6
- SAGE has large memory bandwidth needs and is helped by the XT6/XE6 node
- xNOBEL sends lots of small messages and is helped by the Gemini network

Future Work

- Would like to run these codes on XE6 with 6 cores per NUMA region
- Working on mini-apps to be able to dive deeper into performance issues and to experiment with different communication schemes without the burden of a large code