SAND2011-2763P

Using LANL Applications for Performance Testing of Cielo, Red Sky, …

Douglas Doerfler
Sandia National Laboratories
Scalable Computer Architectures Dept., 1422
EAP Colloquium, LANL, April 19th, 2011

SAND 2011-xxxxP
Unlimited Release
Printed April, 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.

Sandia National Laboratories

CIELO 6X ACCEPTANCE TESTING

Sandia National Laboratories

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

Acknowledgements

- **Cray:** Cindy Nuss, Mike Davis, Steve Whalen, Ting-Ting Zhu, Ron Pfaff, Stephan Behling, Kevin McMahon, Frank Kampe, David Whitaker
- **LANL:** Scott Pakin, Mike Lang, Craig Idler
- **LLNL:** Scott Futral, Tom Spelce
- **SNL:** Paul Lin, Courtenay Vaughan

Sandia National Laboratories

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

6x Performance Applications

Lab	Code	Fortran	Python	C	C++	MPI	OpenMP	Description
SNL	RAMSES/ Charon			X	X	X		A transport reaction code to simulate the performance of semiconductor devices under irradiation
SNL	CTH	X		X		X		Explicit, multi-material shock hydrodynamics code
LANL	xNOBEL	X		X		X		Continuous Adaptive Mesh Refinement (CAMR) code: Hydrodynamics with adaption and high-explosive burn modeling
LANL	SAGE	X		X		X		Multi-dimensional multi-material Eulerian hydrodynamics code with adaptive mesh refinement.
LLNL	AMG2006			X		X	X	Algebraic Multi-Grid linear system solver for unstructured mesh physics packages
LLNL	UMT2006	X	X	X	X	X	X	Single physics package code. Unstructured-Mesh deterministic radiation Transport.

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

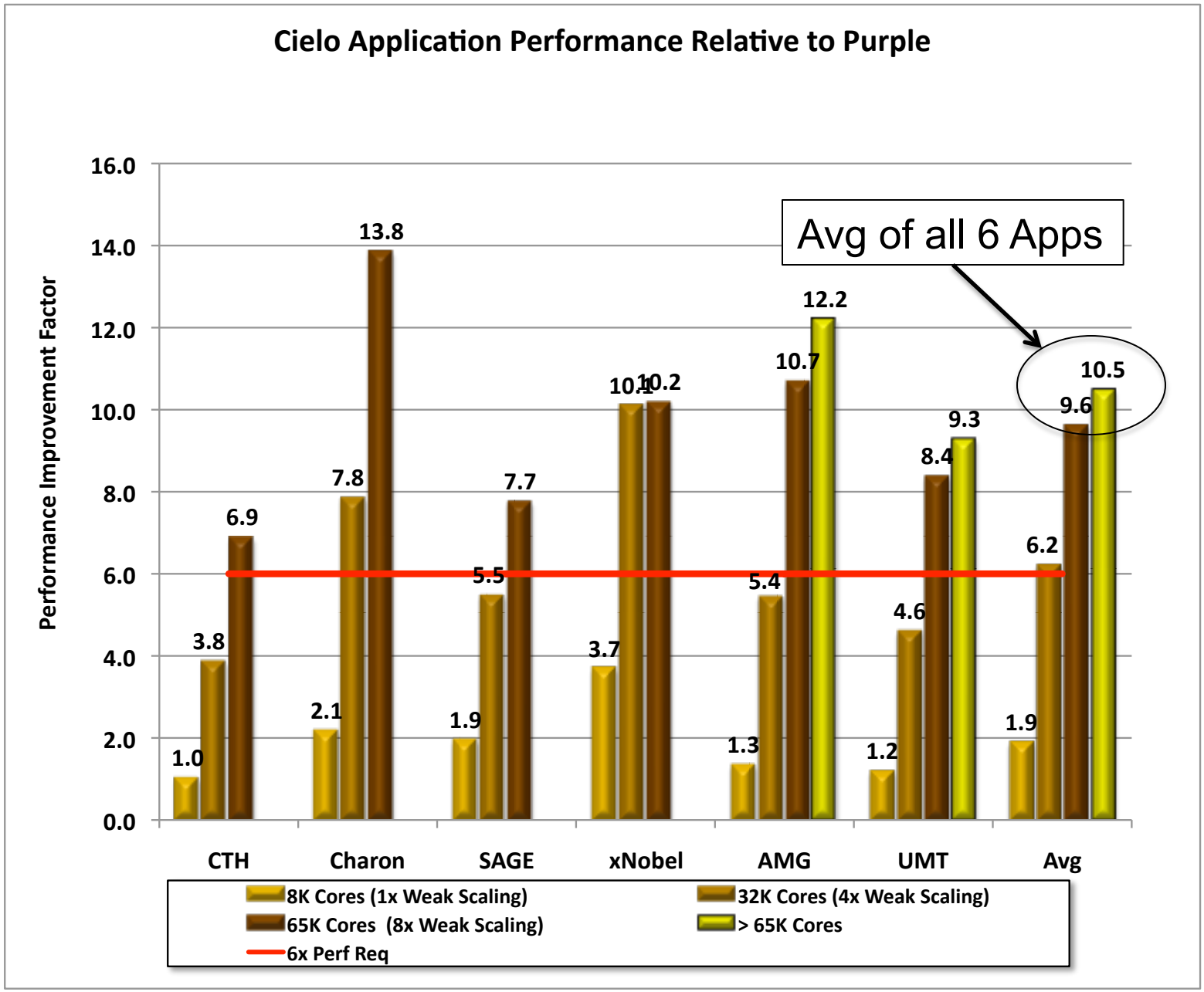
6x Application Improvement Relative to Purple (WEAK SCALING)

- **Improvement = (speedup in runtime) * (increased problem size)**
- **Runtime metric will be chosen to measure platform performance, not algorithmic performance**
- **E.g., a problem run at 8x the problem size as the Purple baseline problem, and the runtime speedup is 1.33x (i.e. ¾ the time) that of Purple**
  - **Speedup = 8x * 1.33x = 10.6x**
- **Application must use at least 4,469 (2/3) of Cielo's compute nodes**

Performance Metric	Purple	Cielo	Ratio
Application Performance	1x	> 6x Purple	TBD
Number of nodes used	1,024 (of 1,336)	up to 5,138 (of 6,704)	5.02x
Number of cores used	8,192	up to 82,208	10.0x
Peak FP	62.3 TF	789 TF	12.7x
Peak Memory BW	102 TB/s	438 TB/s	4.29x
Total Memory Capacity	32 TB	160 TB	5.0x
Memory per node	32 GB	32 GB	1.0x
Memory per core	4 GB	2 GB	0.5x

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

Executive Summary

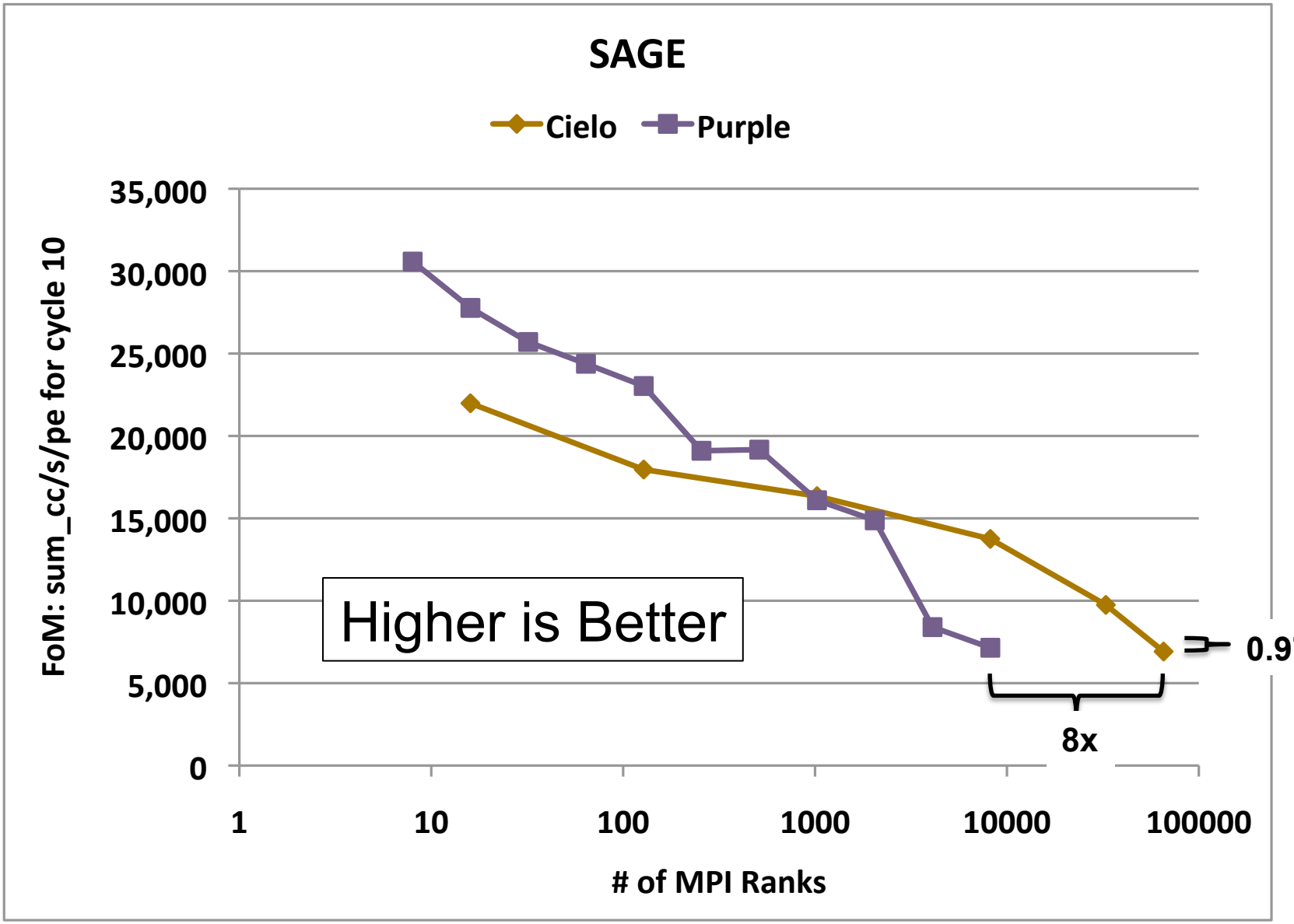


Sandia National Laboratories

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

SAGE: 8x * 0.97x = 7.7x

Weak scaling: timing_h input, 17,500 cells/PE, 10 cycles

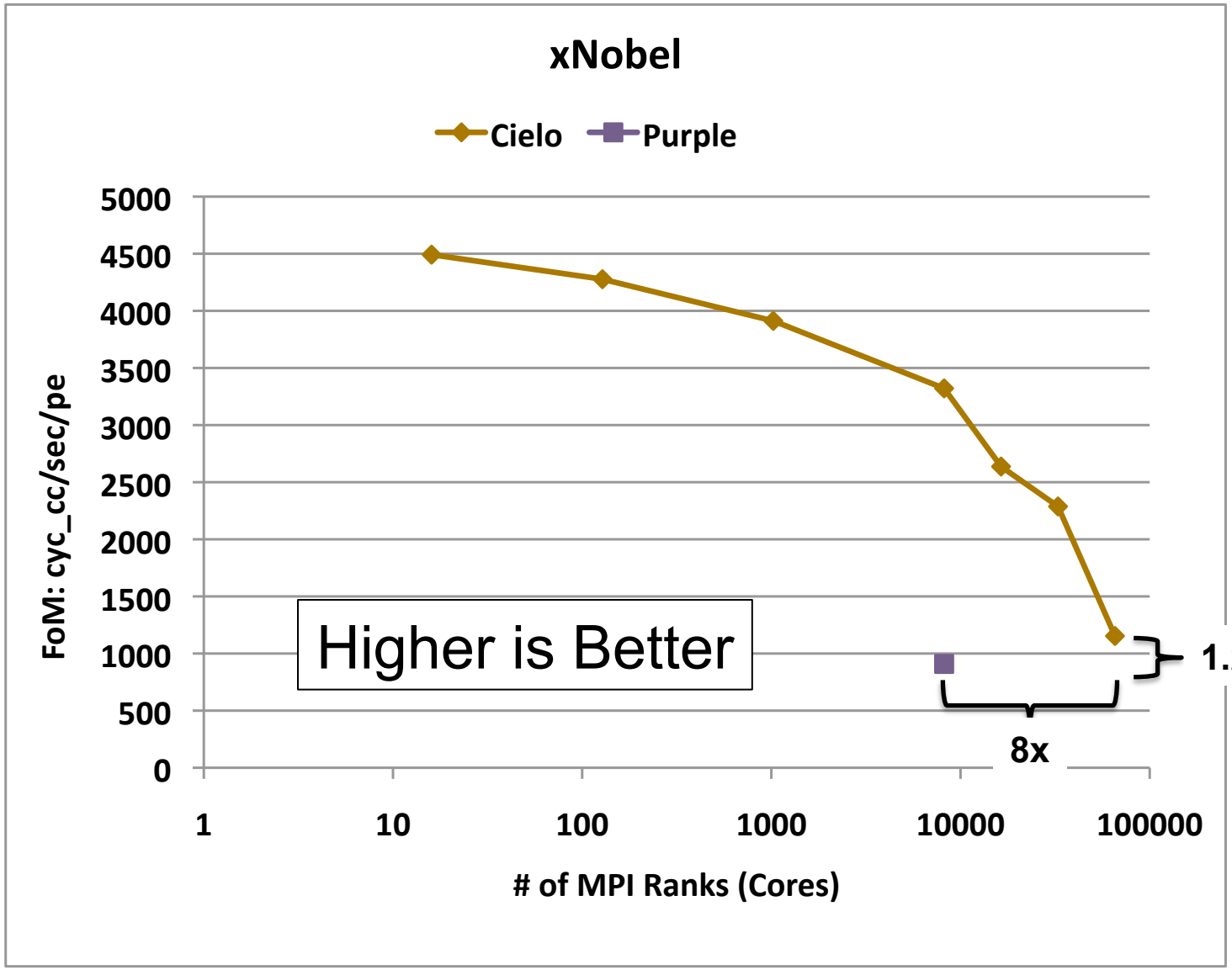


Sandia National Laboratories

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

xNOBEL: 8x * 1.27x = 10.2x

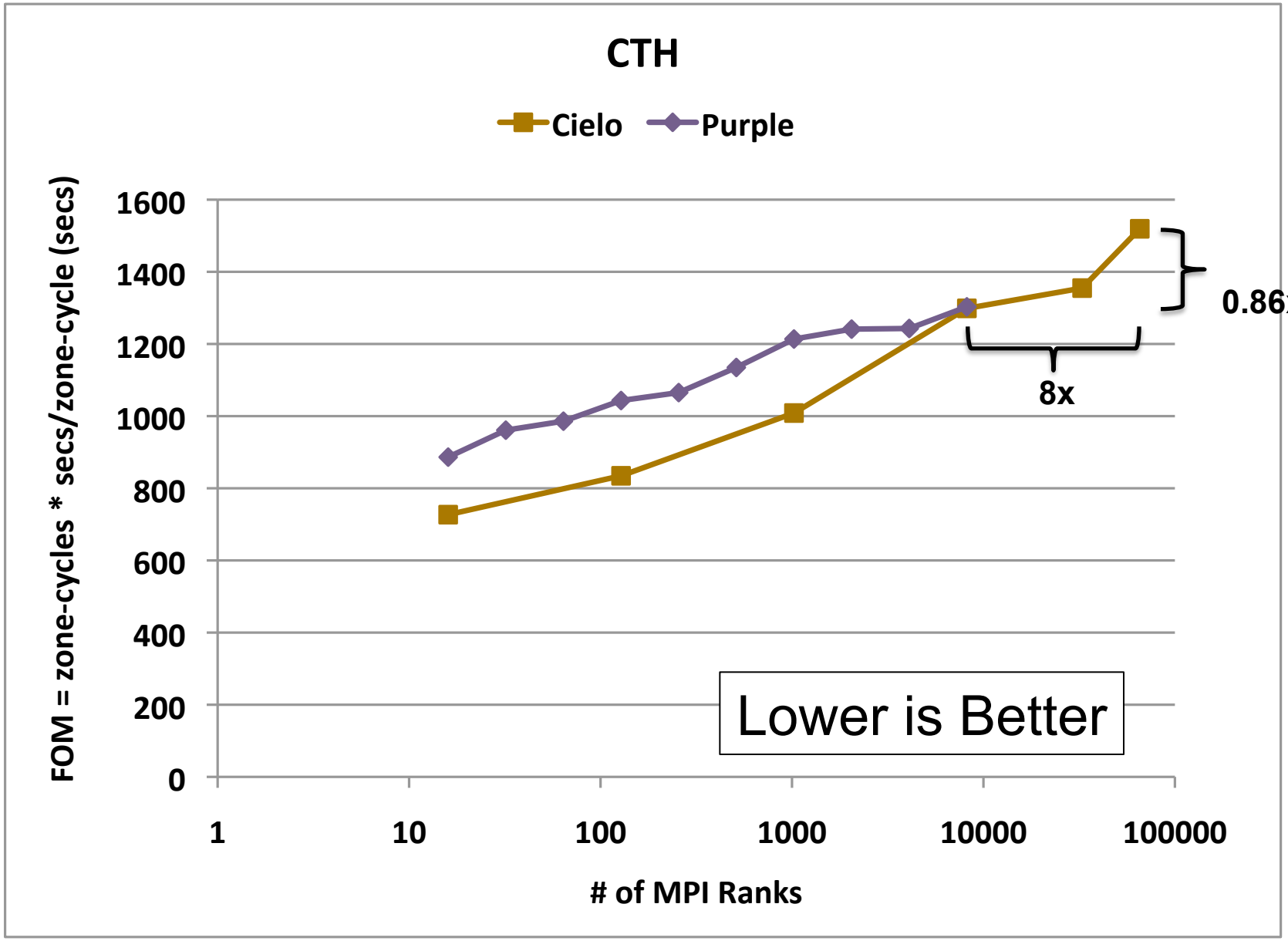
Weak scaling: sc301p input, ~16K cells/PE, 50 iterations



Sandia National Laboratories

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

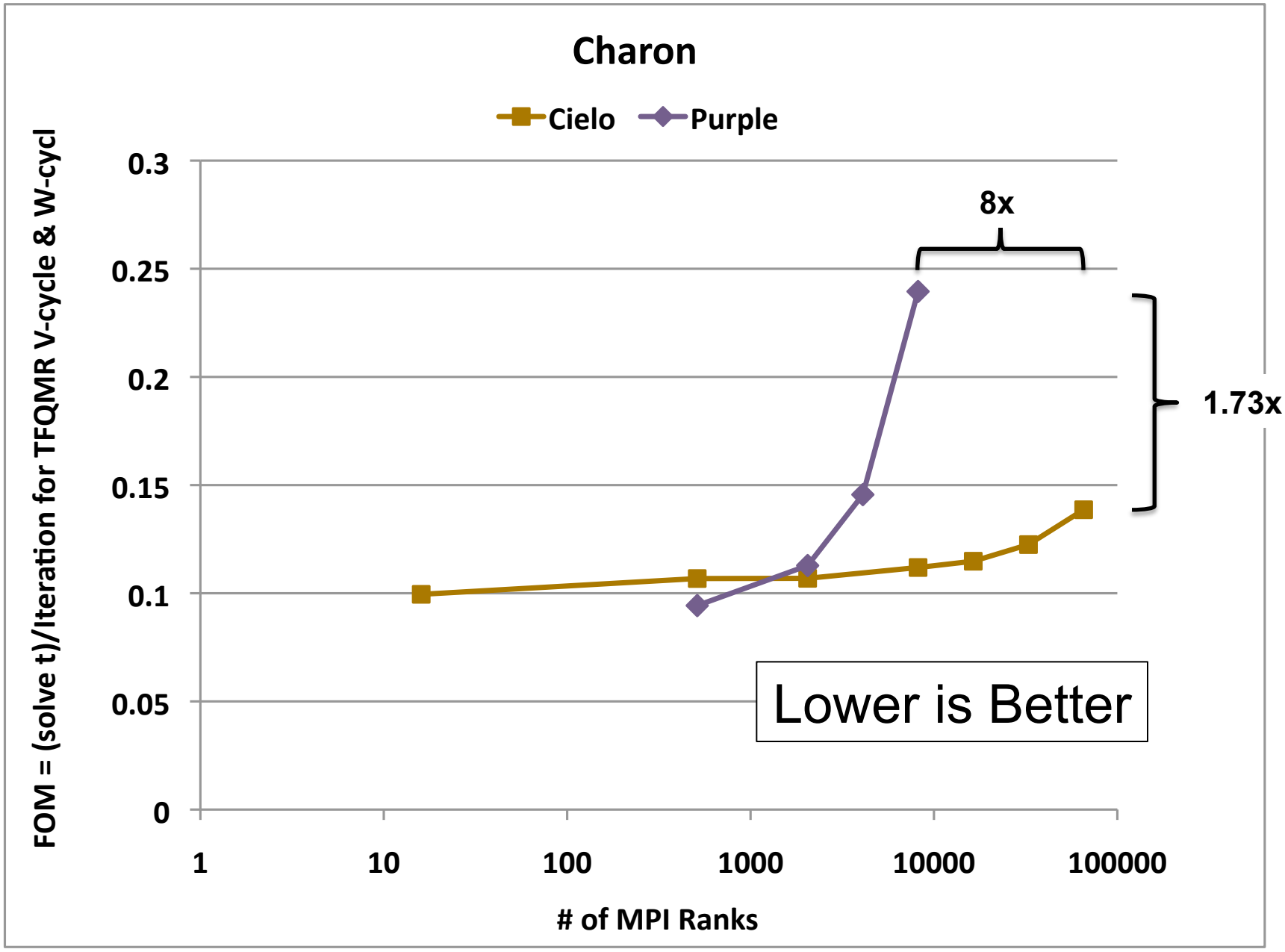
CTH: 8x * 0.86x = 6.9x



Sandia National Laboratories

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

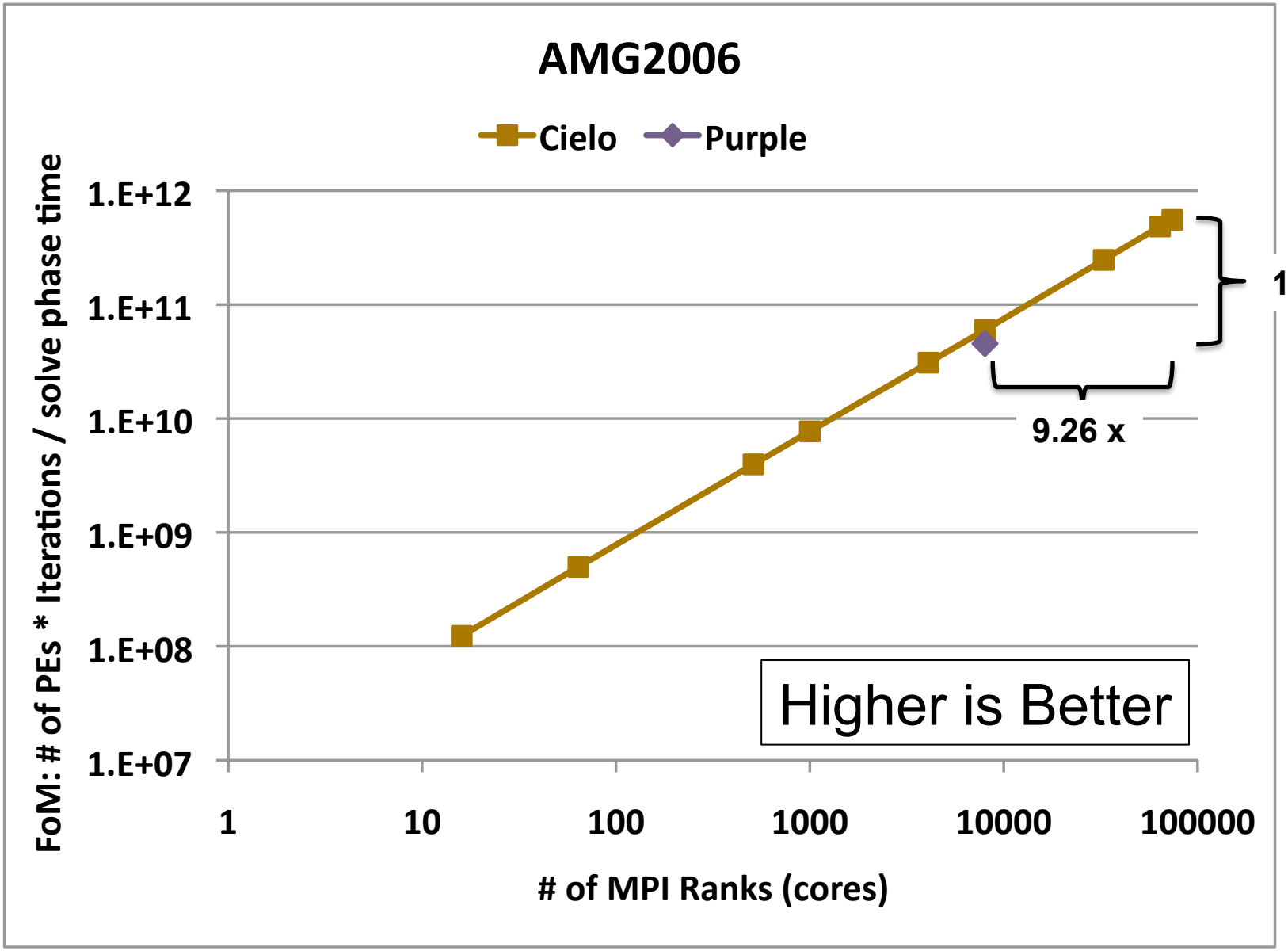
Charon: 8x * 1.73x = 13.8x



Sandia National Laboratories

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

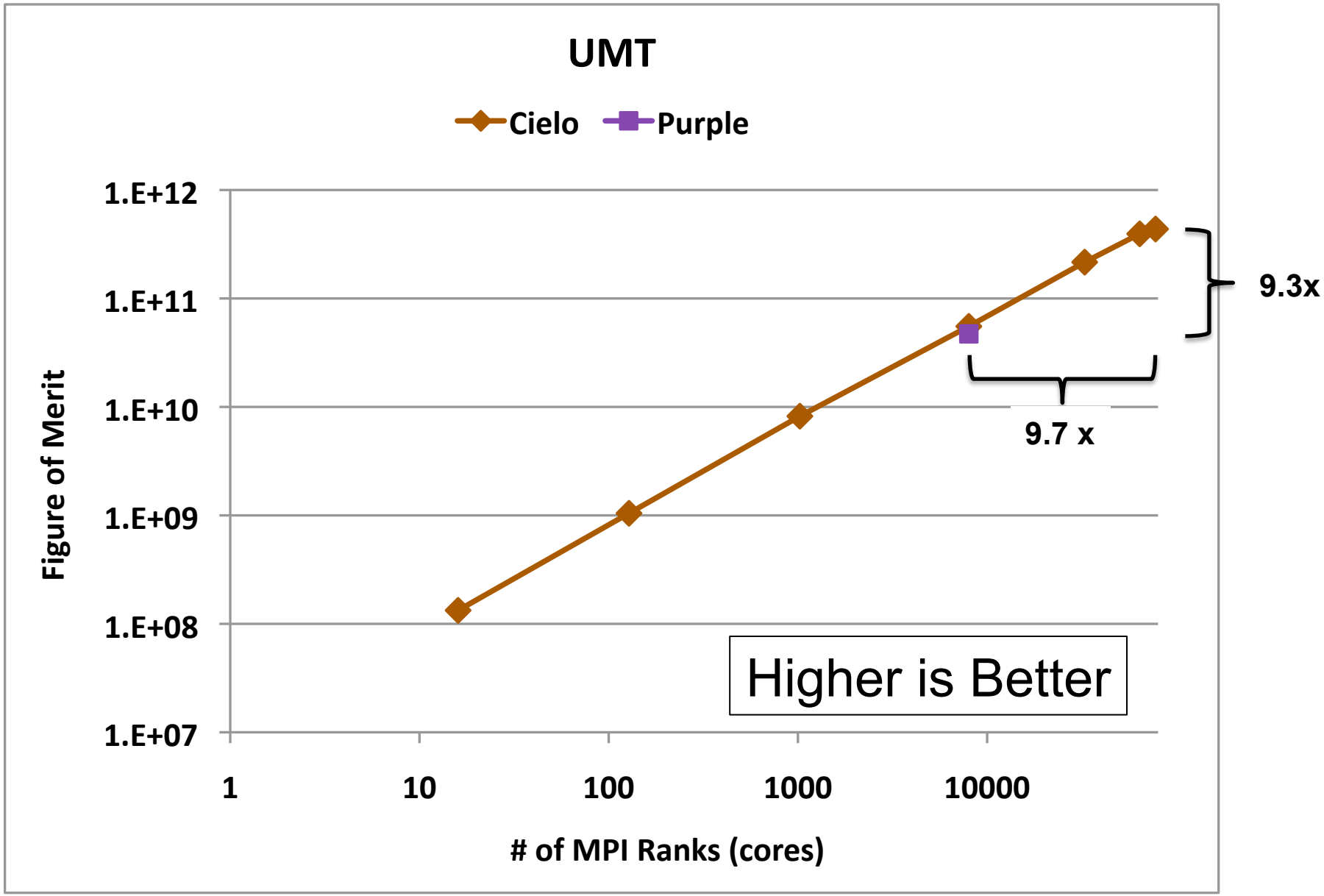
AMG: 12.2x (speedup already factored into FOM)



Sandia National Laboratories

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

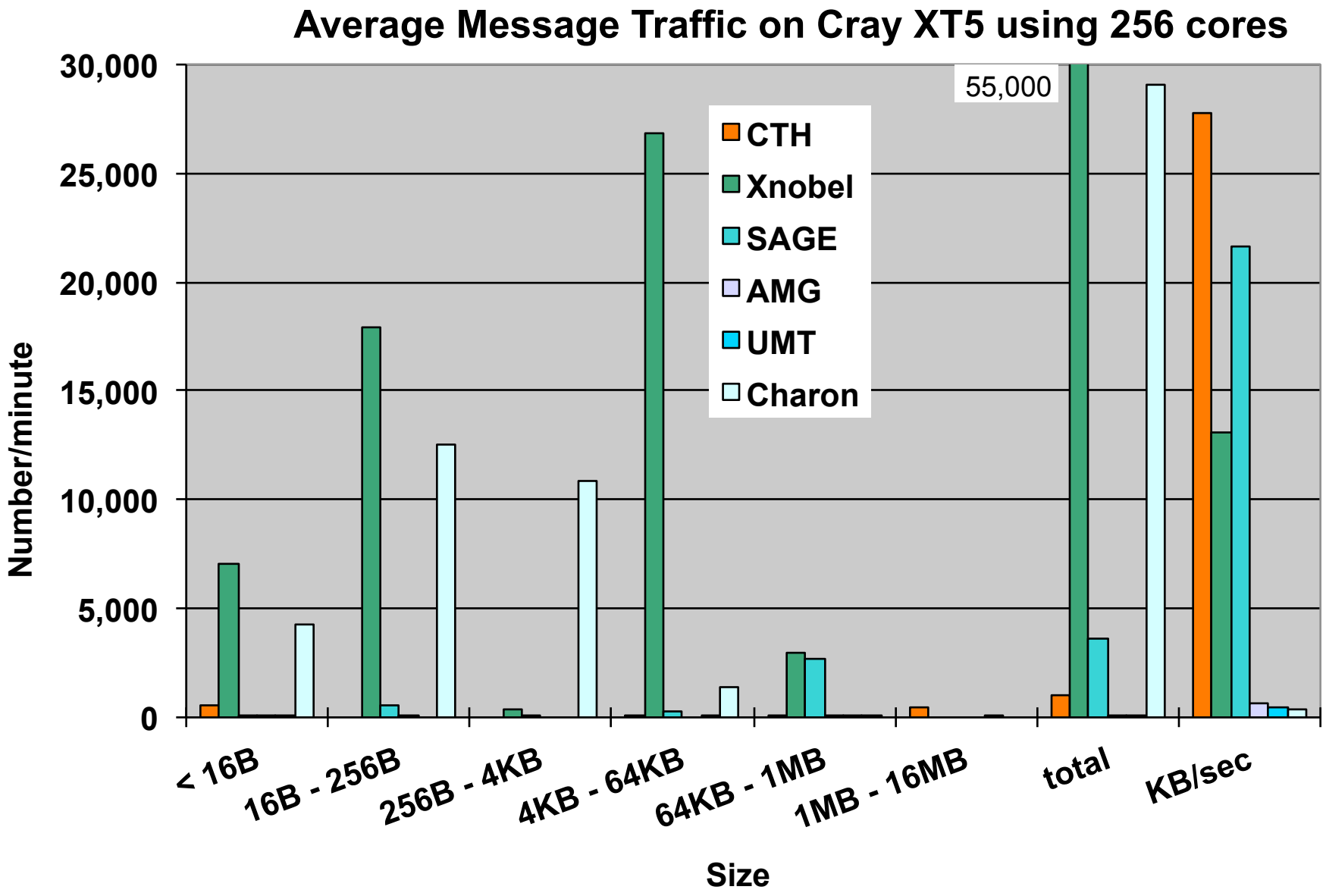
UMT: 9.3x (speedup already factored into FOM)



Sandia National Laboratories

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

6x Application Messaging Characteristics

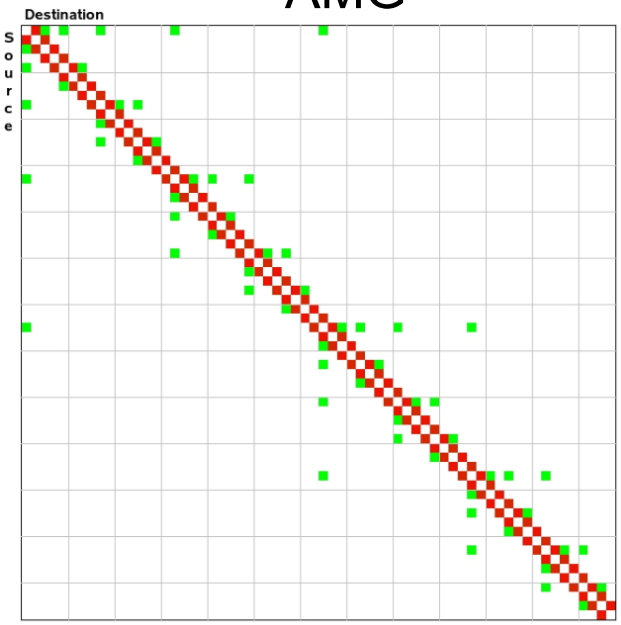


Sandia National Laboratories

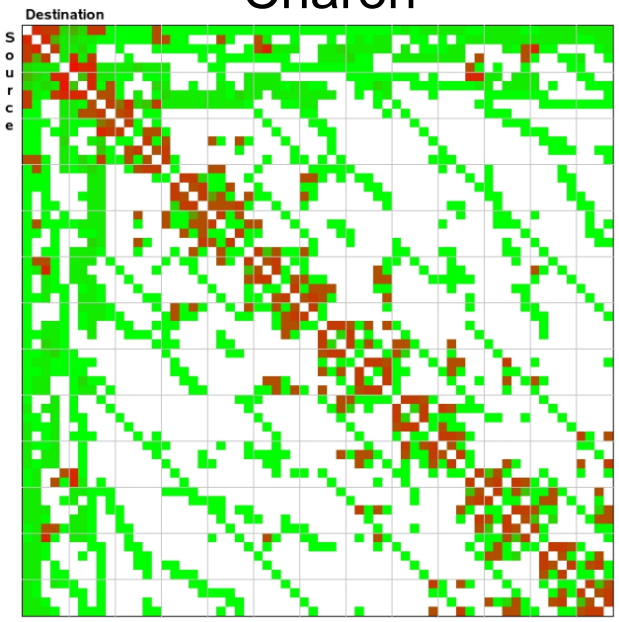
cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

6x Applications Messaging Patterns

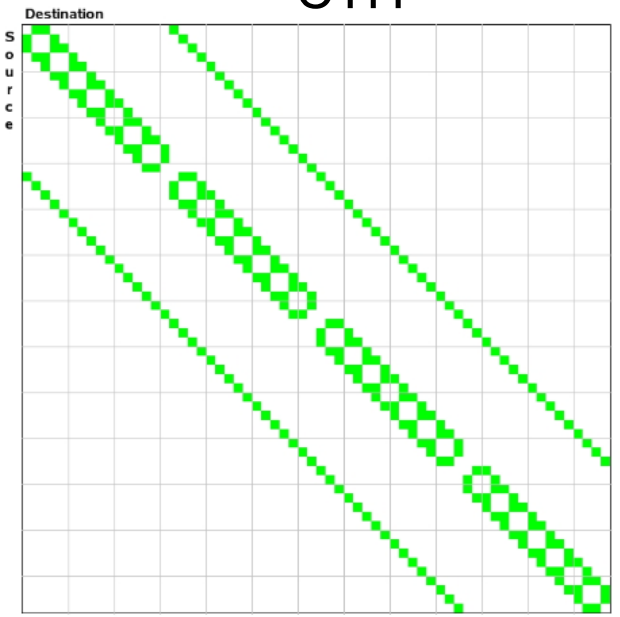
AMG



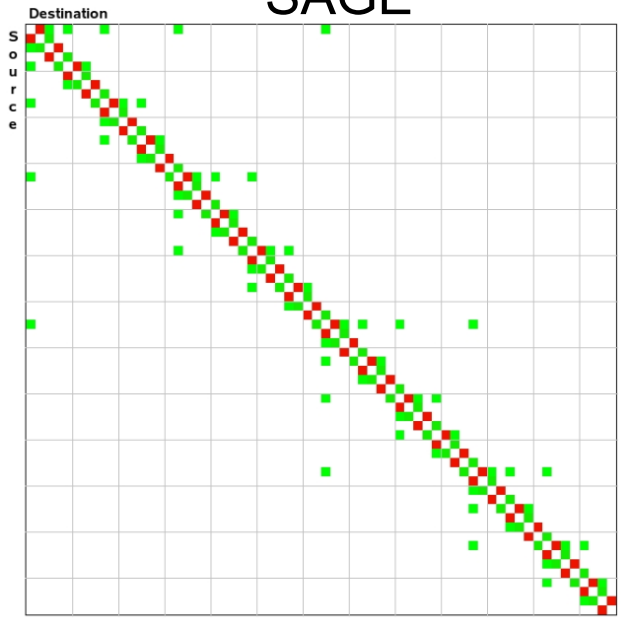
Charon



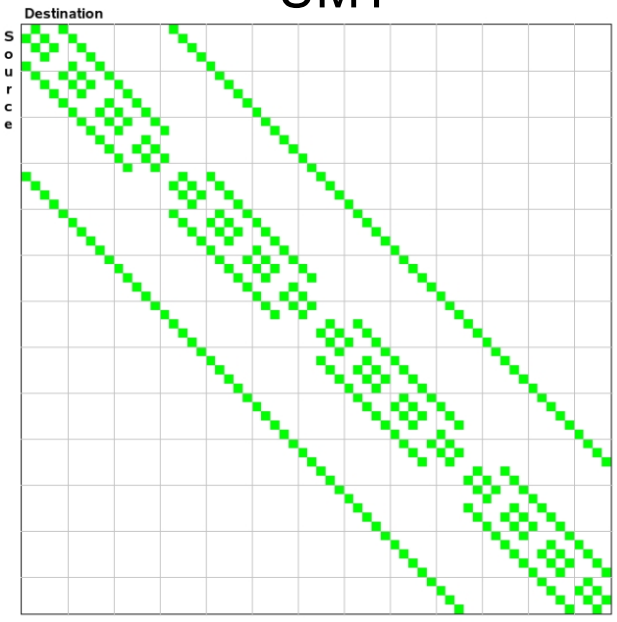
CTH



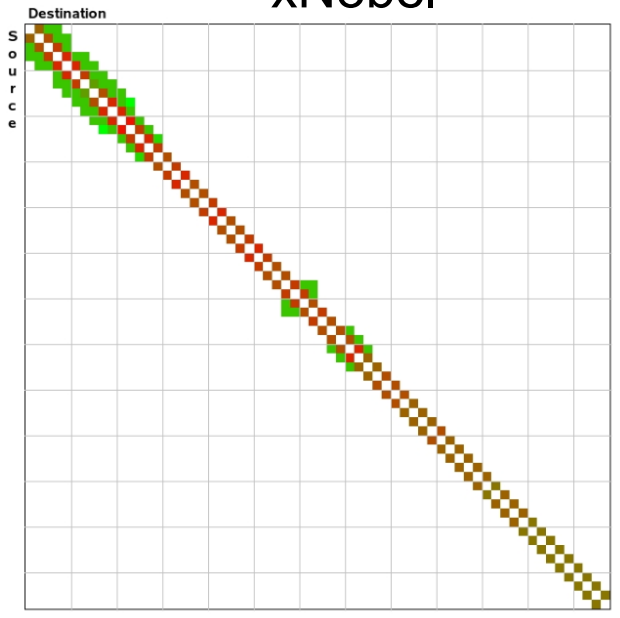
SAGE



UMT



xNobel



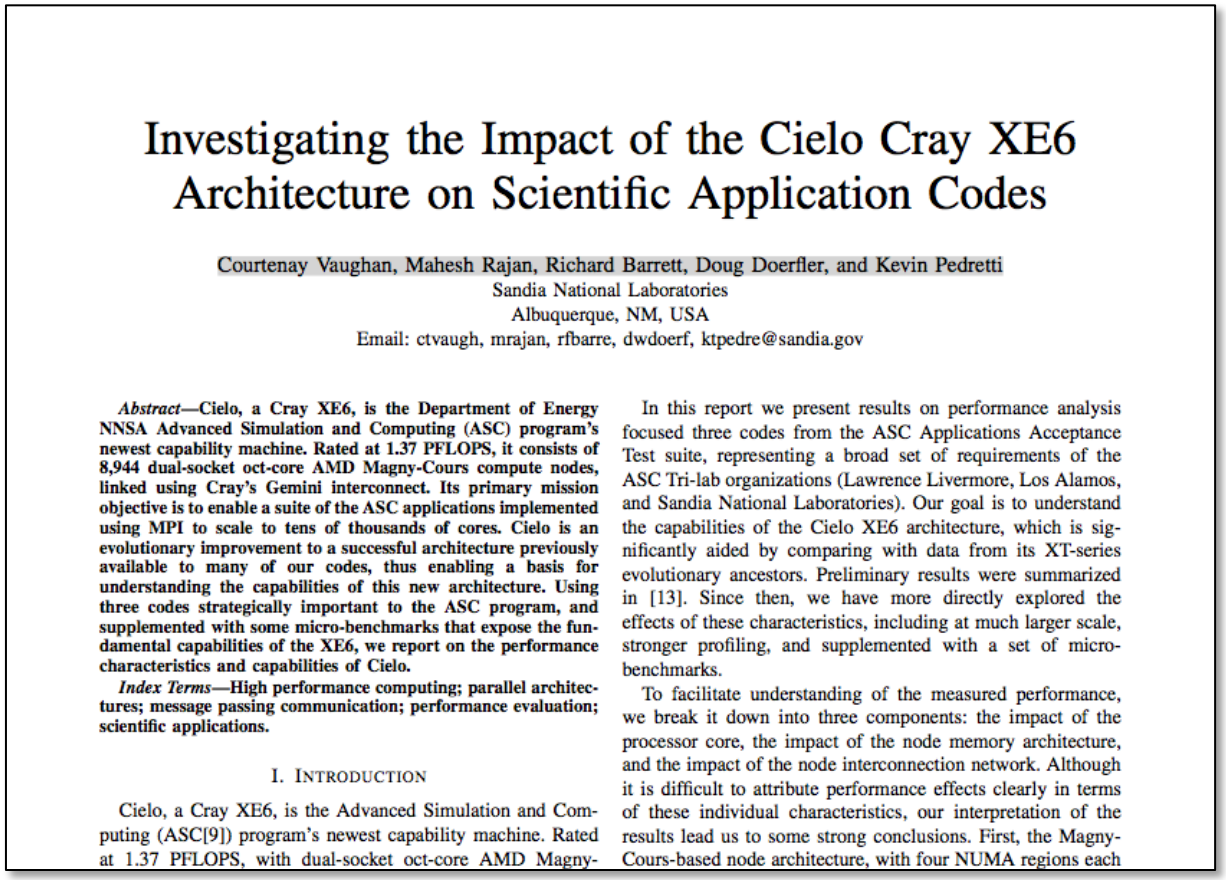
National Laboratories

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

Takeaways: Food for Thought

- Charon, Sage and xNOBEL have limitations in the total number of "cells" that can be processed in a single job due to the use of 32-bit signed integers, i.e. 2 Billion "cells" total
  - I/O needed to be controlled if not totally eliminated in some circumstances
- What applications/methods do we use for 140,000+ core Cielo Phase 2 system?
  - Mini-Apps?
- Phase 2 acceptance time period will be short

Sandia National Laboratories



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

Courtenay Vaughan, Mahesh Rajan, Richard Barrett, Doug Doerfler, and Kevin Pedretti
Sandia National Laboratories
Albuquerque, NM, USA
Email: ctvaugh, mrajan, rfbarre, dwdoerf, ktpedre@sandia.gov

***Abstract*—Cielo, a Cray XE6, is the Department of Energy NNSA Advanced Simulation and Computing (ASC) program's newest capability machine. Rated at 1.37 PFLOPS, it consists of 8,944 dual-socket oct-core AMD Magny-Cours compute nodes, linked using Cray's Gemini interconnect. Its primary mission objective is to enable a suite of the ASC applications implemented using MPI to scale to tens of thousands of cores. Cielo is an evolutionary improvement to a successful architecture previously available to many of our codes, thus enabling a basis for understanding the capabilities of this new architecture. Using three codes strategically important to the ASC program, and supplemented with some micro-benchmarks that expose the fundamental capabilities of the XE6, we report on the performance characteristics and capabilities of Cielo.**

***Index Terms*—High performance computing; parallel architectures; message passing communication; performance evaluation; scientific applications.**

I. Introduction

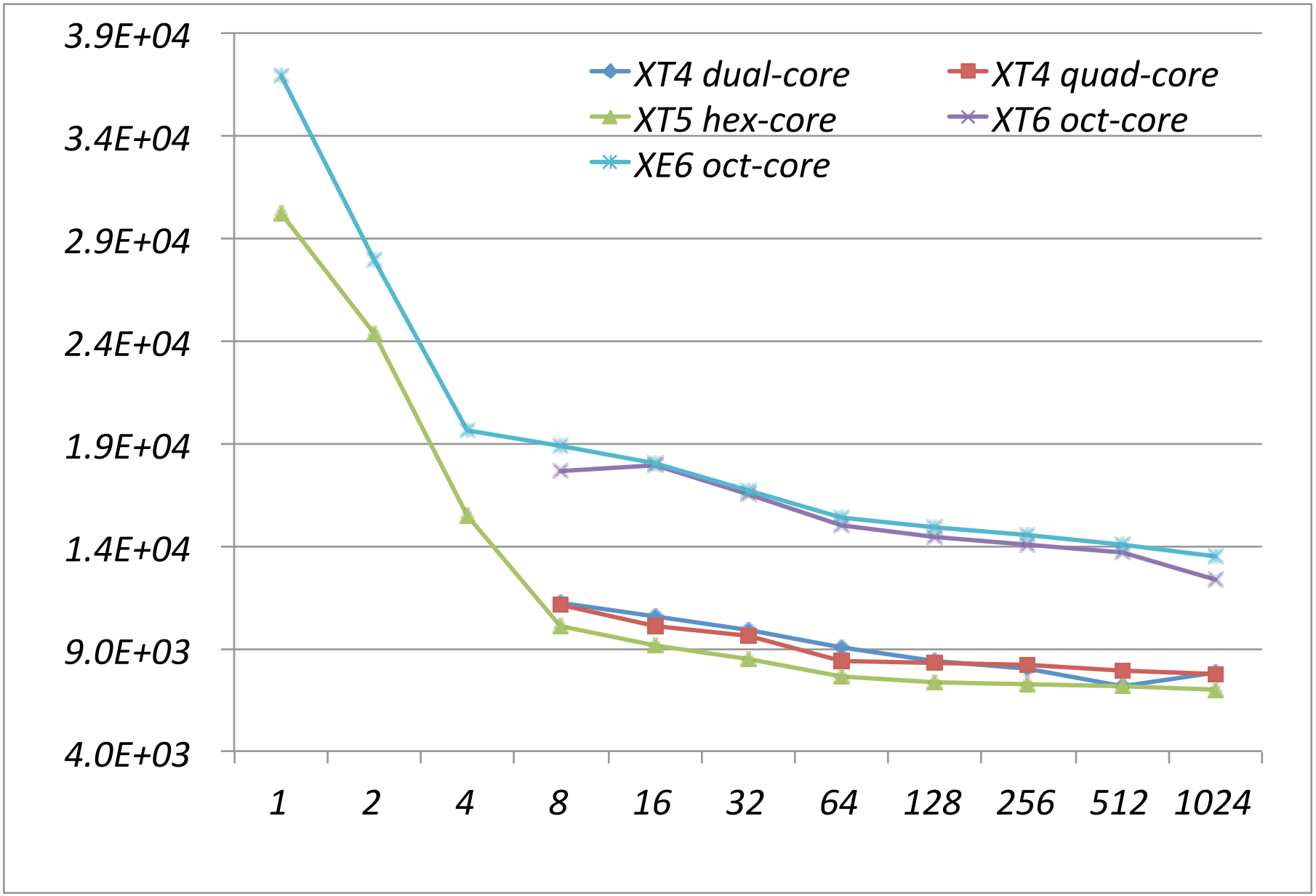
Cielo, a Cray XE6, is the Advanced Simulation and Computing (ASC[9]) program's newest capability machine. Rated at 1.37 PFLOPS, with dual-socket oct-core AMD Magny-

In this report we present results on performance analysis focused three codes from the ASC Applications Acceptance Test suite, representing a broad set of requirements of the ASC Tri-lab organizations (Lawrence Livermore, Los Alamos, and Sandia National Laboratories). Our goal is to understand the capabilities of the Cielo XE6 architecture, which is significantly aided by comparing with data from its XT-series evolutionary ancestors. Preliminary results were summarized in [13]. Since then, we have more directly explored the effects of these characteristics, including at much larger scale, stronger profiling, and supplemented with a set of micro-benchmarks.

To facilitate understanding of the measured performance, we break it down into three components: the impact of the processor core, the impact of the node memory architecture, and the impact of the node interconnection network. Although it is difficult to attribute performance effects clearly in terms of these individual characteristics, our interpretation of the results lead us to some strong conclusions. First, the Magny-Cours-based node architecture, with four NUMA regions each

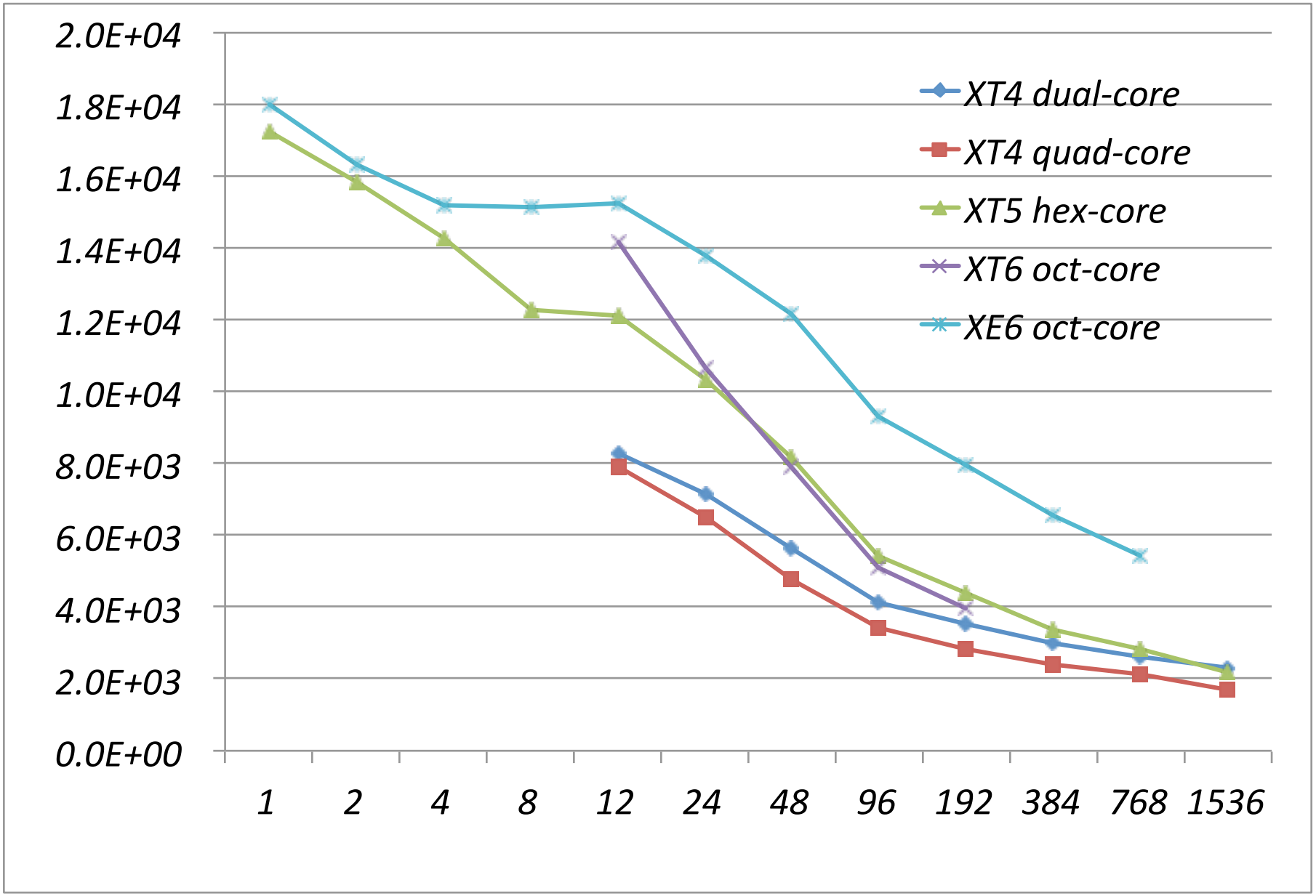
IPDPS/LSPP PAPER: ACCEPTED

Sage

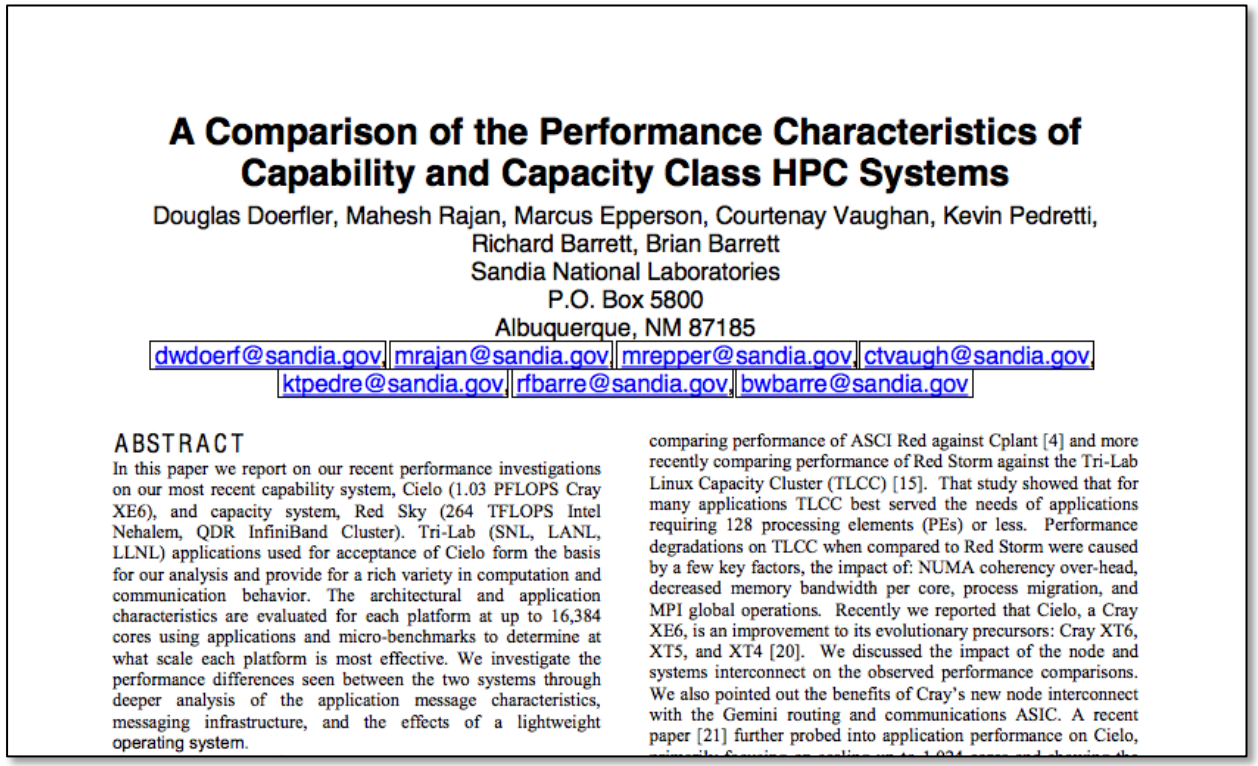


Sandia National Laboratories

xNobel



Sandia National Laboratories



A Comparison of the Performance Characteristics of Capability and Capacity Class HPC Systems

Douglas Doerfler, Mahesh Rajan, Marcus Epperson, Courtenay Vaughan, Kevin Pedretti,
Richard Barrett, Brian Barrett
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185
dwdoerf@sandia.gov, mrajan@sandia.gov, mrepper@sandia.gov, ctvaugh@sandia.gov,
ktpedre@sandia.gov, rfbarre@sandia.gov, bwbarre@sandia.gov

ABSTRACT

In this paper we report on our recent performance investigations on our most recent capability system, Cielo (1.03 PFLOPS Cray XE6), and capacity system, Red Sky (264 TFLOPS Intel Nehalem, QDR InfiniBand Cluster). Tri-Lab (SNL, LANL, LLNL) applications used for acceptance of Cielo form the basis for our analysis and provide for a rich variety in computation and communication behavior. The architectural and application characteristics are evaluated for each platform at up to 16,384 cores using applications and micro-benchmarks to determine at what scale each platform is most effective. We investigate the performance differences seen between the two systems through deeper analysis of the application message characteristics, messaging infrastructure, and the effects of a lightweight operating system.

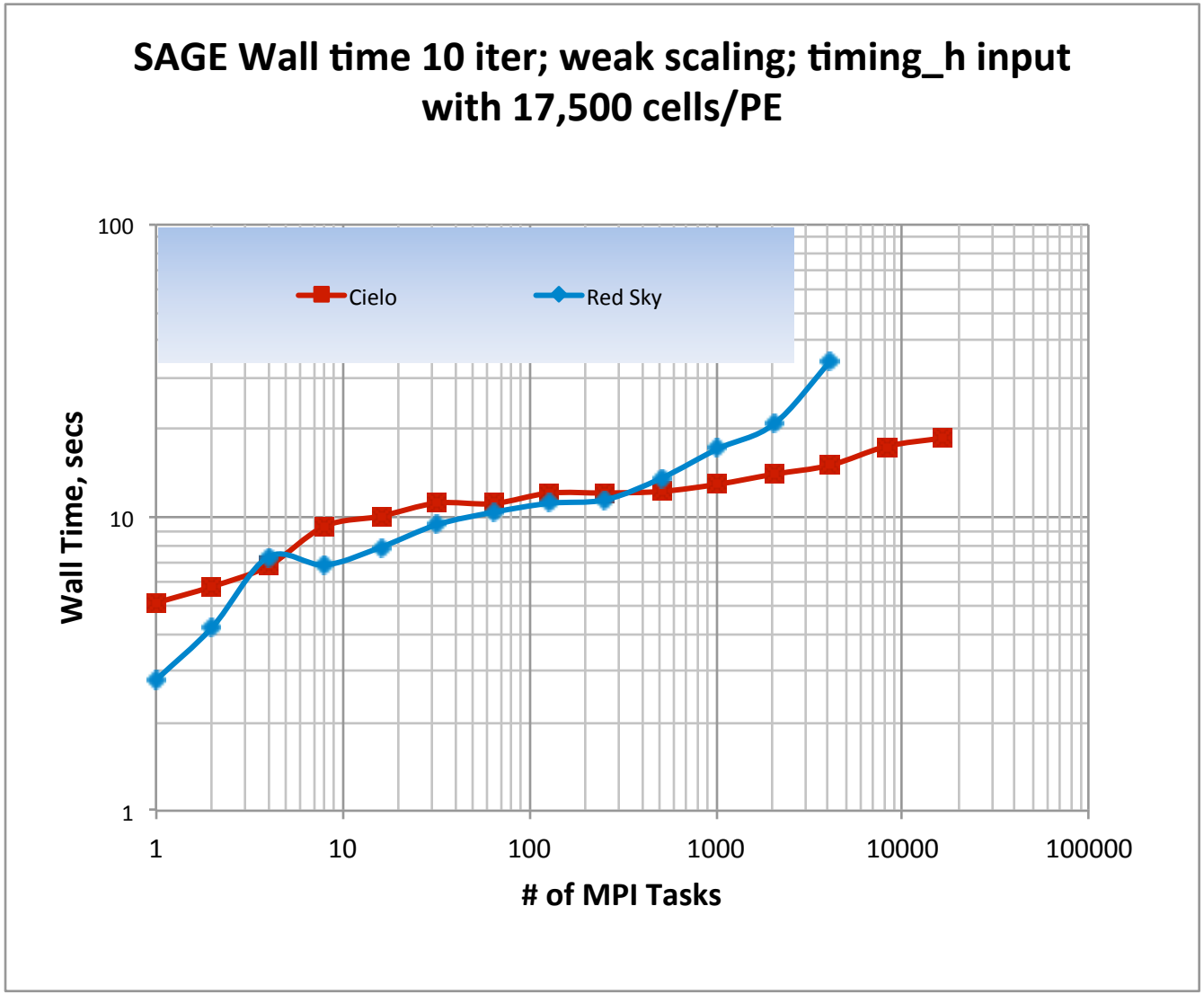
comparing performance of ASCI Red against Cplant [4] and more recently comparing performance of Red Storm against the Tri-Lab Linux Capacity Cluster (TLCC) [15]. That study showed that for many applications TLCC best served the needs of applications requiring 128 processing elements (PEs) or less. Performance degradations on TLCC when compared to Red Storm were caused by a few key factors, the impact of: NUMA coherency over-head, decreased memory bandwidth per core, process migration, and MPI global operations. Recently we reported that Cielo, a Cray XE6, is an improvement to its evolutionary precursors: Cray XT6, XT5, and XT4 [20]. We discussed the impact of the node and systems interconnect on the observed performance comparisons. We also pointed out the benefits of Cray's new node interconnect with the Gemini routing and communications ASIC. A recent paper [21] further probed into application performance on Cielo,

USING THE 6X APPLICATIONS FOR ANALYZING CAPACITY COMPUTING AT SNL

SC11 PAPER: SUBMISSION

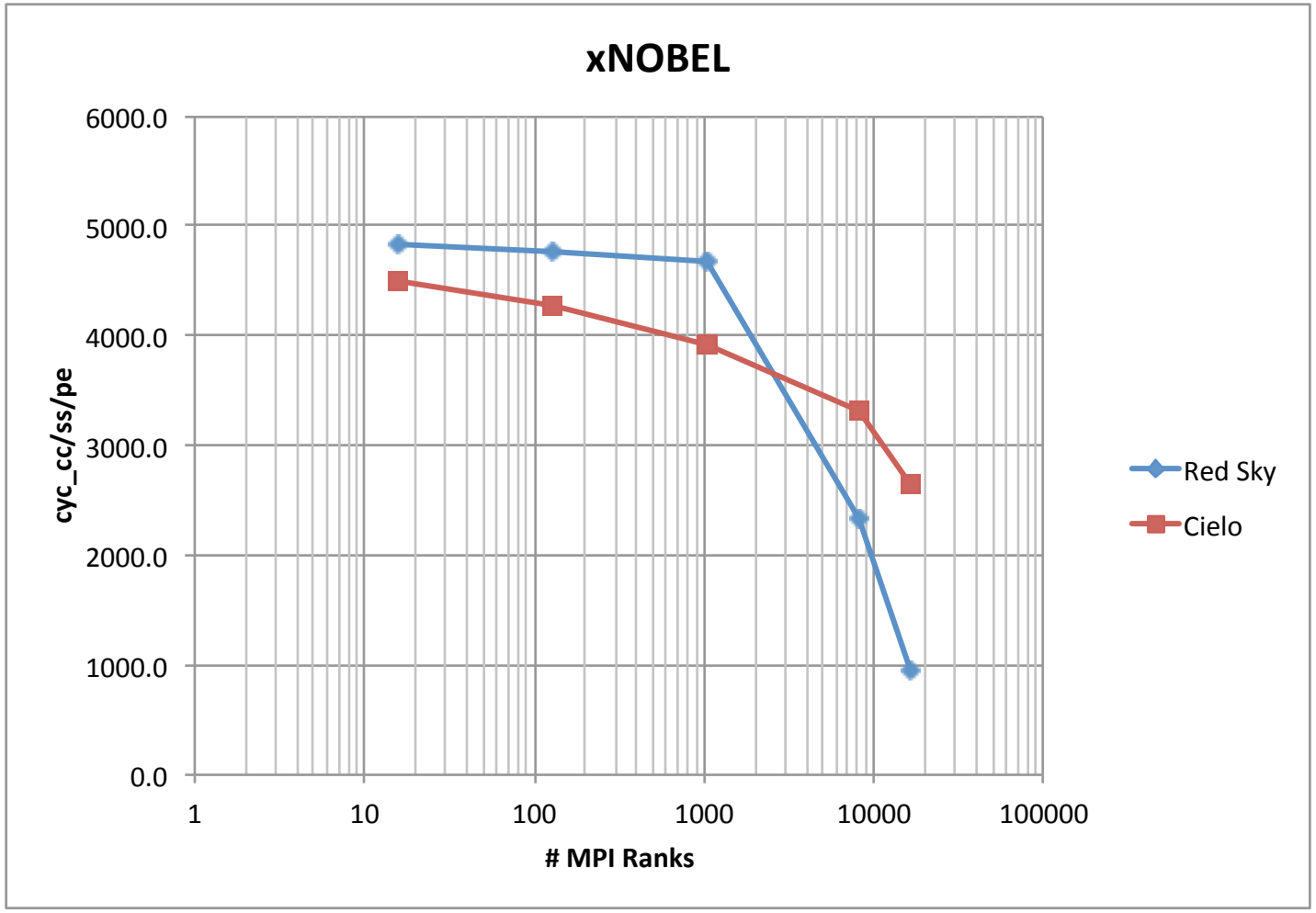
Sandia National Laboratories

Sage on Red Sky



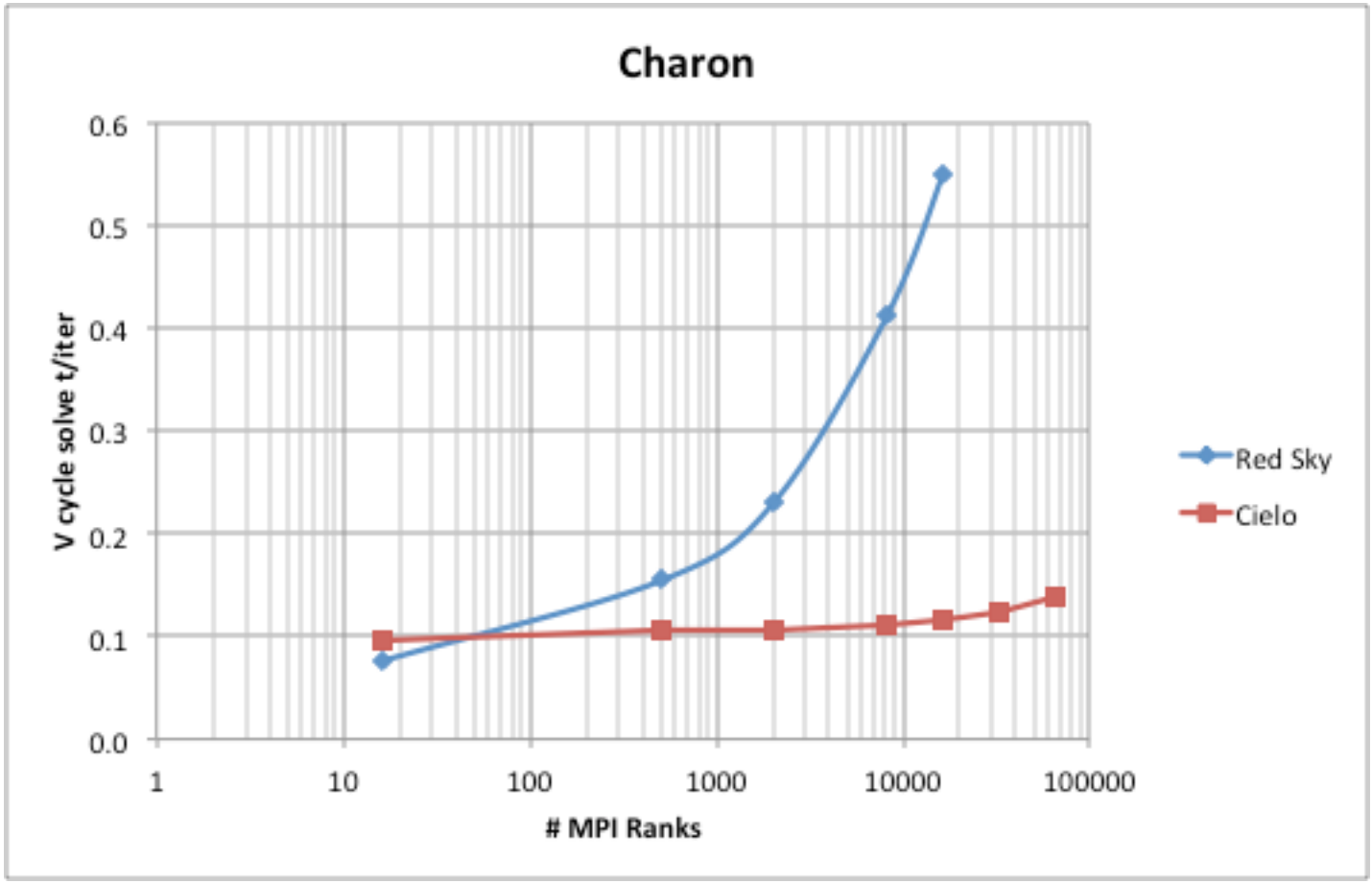
Sandia National Laboratories

xNobel on Red Sky



Sandia National Laboratories

Charon on Red Sky



DWDOERF@SANDIA.GOV

QUESTIONS, COMMENTS, DISCUSSION

BACKUP SLIDES

Cielo Platform Overview

Douglas Doerfler
Cielo System Architect

Cielo Roadshow
October 26th, 27th and 28th 2010

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

Design Philosophy & Goals

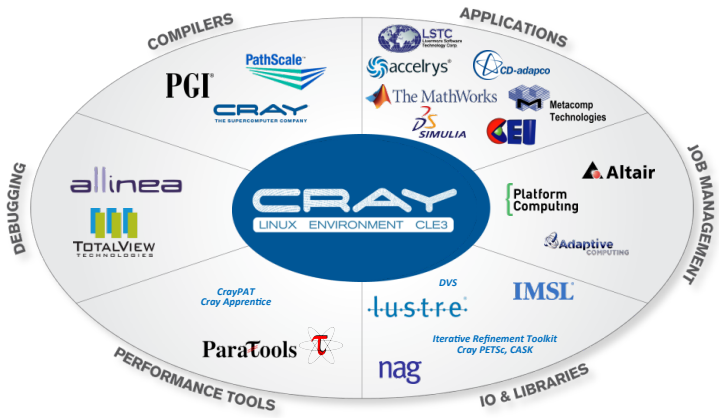
- Petascale production capability to be deployed in Q1FY11
  - Take over the role Purple currently plays
  - Usage Model will follow the Capability Computing Campaign (CCC) process
  - Capability: Capable of running a single application across the entire machine
- Easy migration of existing integrated weapons codes
  - MPI Everywhere is the nominal programming model
  - 2GB memory per core (minimum) to support current application requirements
- Productivity goal is to achieve a 6x to 10x improvement over Purple on representative CCC applications
  - Memory subsystem performance will be the major contributor to node performance
  - Interconnect performance will be major contributor to scaling performance
  - Reliability will be major contributor to CCC total time to solution
- Upgrade path to allow increased capability in out years
- Key challenges: Reliability, Power, HW and SW Scalability, Algorithmic Scaling to 80K to 100K MPI ranks

Sandia National Laboratories

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

Cielo at a Glance

- Platforms
  - Cielo – Petascale Capability Computing Platform
  - Cielito – Small application development testbed
- Cray XE6 Architecture
  - 3D Torus Topology using Cray Gemini high-speed interconnect
  - AMD Magny-Cours based nodes
- Cray Linux Environment (CLE) System Software
  - ALPS Runtime with Moab batch scheduling
  - CrayPat & Apprentice2 performance analysis tools
  - TotalView debugger
  - PGI, Cray and GNU compiler suites
  - Etc.



- Integrated Visualization & Analysis Partition
  - 64 GB memory per node partition
- Integrated into LANL's Parallel Scalable Backbone Network (PaScalBB)
  - 10 PB of user available storage
  - 200 GB/s of network bandwidth
  - 160 GB/sec of parallel file system bandwidth

Sandia National Laboratories

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

Cielo Hardware Architecture

- AMD Magny-Cours Node
  - Dual-socket AMD 6136 Processors
  - 2 x 8 = 16 total cores
  - 2.4 GHz core frequency
  - 32 GB of 1333 DDR3 memory
    - 64 GB for Visualization Nodes
  - 153.6 peak DP GFLOPs
  - 85.3 peak GB/s memory BW
- Gemini High-Speed Interconnect
  - 3D Torus topology
  - Phase 1: 18x8x24

    X bisection: > 4.38 TB/s

    Y bisection: > 4.92 TB/s

    Z bisection: > 3.92 TB/s
  - Phase 2: 16x12x24

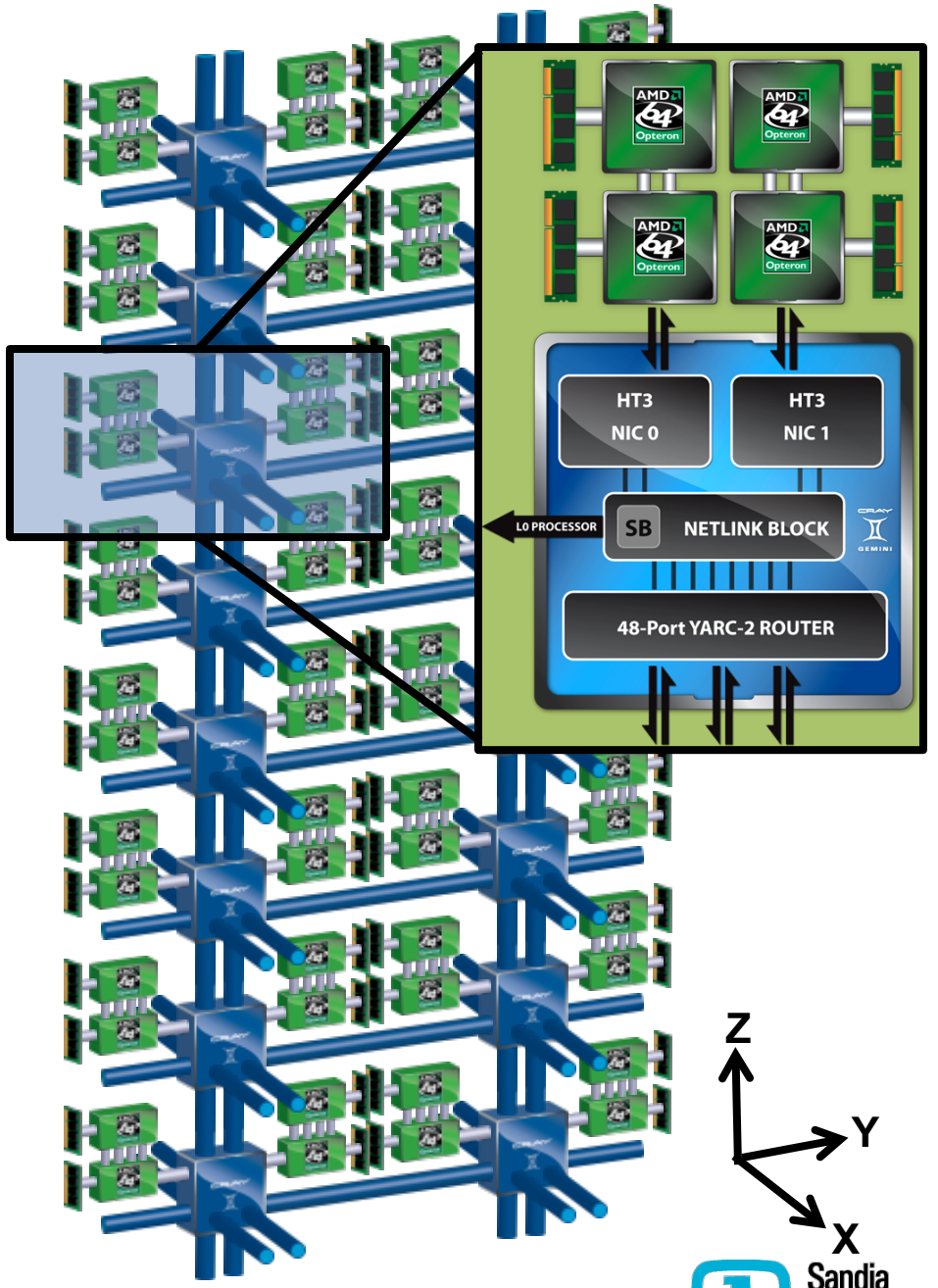
    X bisection: > 6.57 TB/s

    Y bisection: > 4.38 TB/s

    Z bisection: > 4.38 TB/s
  - Node Injection

    > 6 GB/s/dir sustained BW

    > 8 MMsgs/sec sustained



Sandia National Laboratories

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

Cielo By Numbers

	Phase 1	Phase 2	Cielito
# of Cabinets	72	96	1
# of Service Nodes	208	272	14
# of Compute Nodes	6,704*	8,944*	68
# of Visualization Nodes	(376)	(376)	(4)
# of Compute Cores	107,264	143,104	1,088
Peak Memory BW	572 TB/s	763 TB/s	5.8 TB/s
Memory Capacity per Core	2 GB (4 GB)	2 GB (4 GB)	2 GB (4 GB)
Compute Memory Capacity	226.6 TB	298.2 TB	2.3 TB
Peak Compute FLOPS	1.03 PF	1.37 PF	10.4 TF
Sustained PFS BW	> 160 GB/s		TBD
System Power	< 3.9 MW	< 4.4 MW	
Full System Job MTBI	> 25 hours		
System MTBI	> 200 hours		

* Total compute nodes including Viz nodes and nodes allocated for other services

Sandia National Laboratories

Capability vs. Capacity; HPC systems application performance comparisons: Cielo vs. Red Sky

Doug Doerfler and Mahesh Rajan

Sandia National Laboratories

2/9/2011

System Comparison

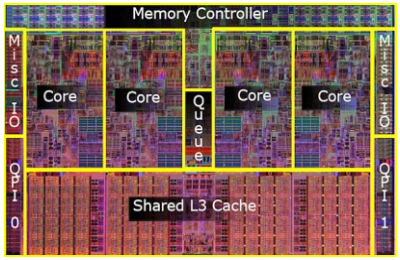
SYSTEM	Red Sky	Cielo
Num Compute Nodes	2318	6704
Num Compute Cores	18,544	107,264
Processor	Dual Intel Nehalem 2.93 GHz	Dual AMD Magny-Cours, 2.4 GHz
Cores / node	8	16
Memory / Core	1.5 GB	2 GB
Peak Node GFLOPS	93.76	153.6
Memory	3 channels/socket, DDR3, 1333 MHz	4 channels/socket, DDR3,1333 MHz
Cache	L1=4x32KB I,D L2=4x512KB L3=8MB	L1=8x64 KB, I,D L2=8x512KB L3=12MB (10MB)
Interconnect / Topology	QDR IB, Torus	Gemini, Torus
Compute Node OS	TOSS	CNL
MPI	OpenMPI 1.4.1	MPT 5.1.4
Compilers	Intel 11.1	PGI 10.x; Cray CCE 7.x

Red Sky Software

- OS: CentOS Red Hat Linux based with patches
- MPI: OpenMPI & MVAPICH (IB OFED stack)
- Scheduler: Slurm and Moab
- Compilers: Intel and GNU
- Debugger: TotalView
- Math Libraries: BLACS, FFTW, MKL
- Performance Tools: OpenSpeedShop, TAU, mpiP
- User Control: via Modules

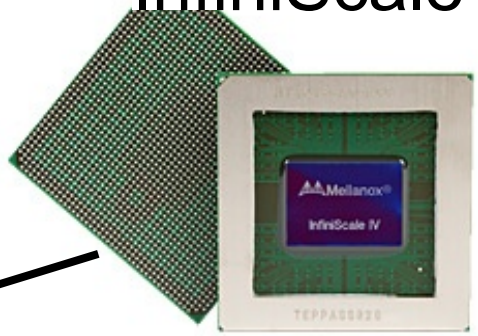
Red Sky Node & Chassis Architecture

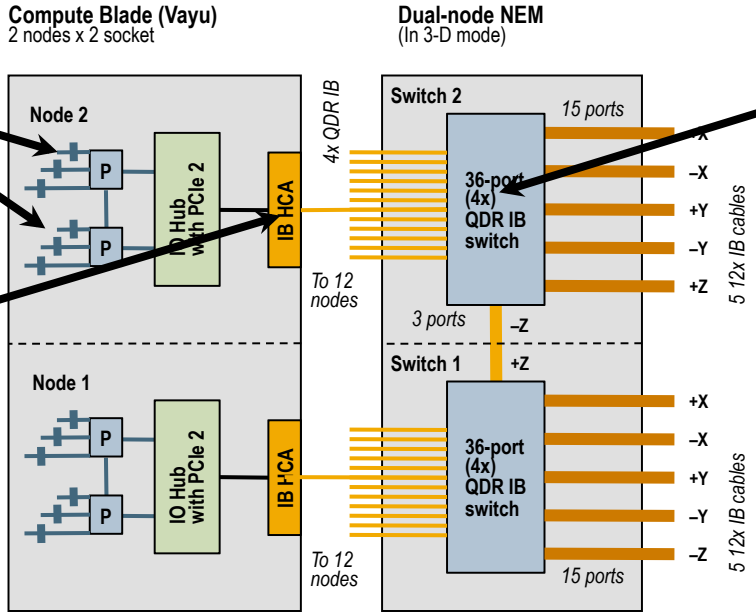
Intel Nehalem

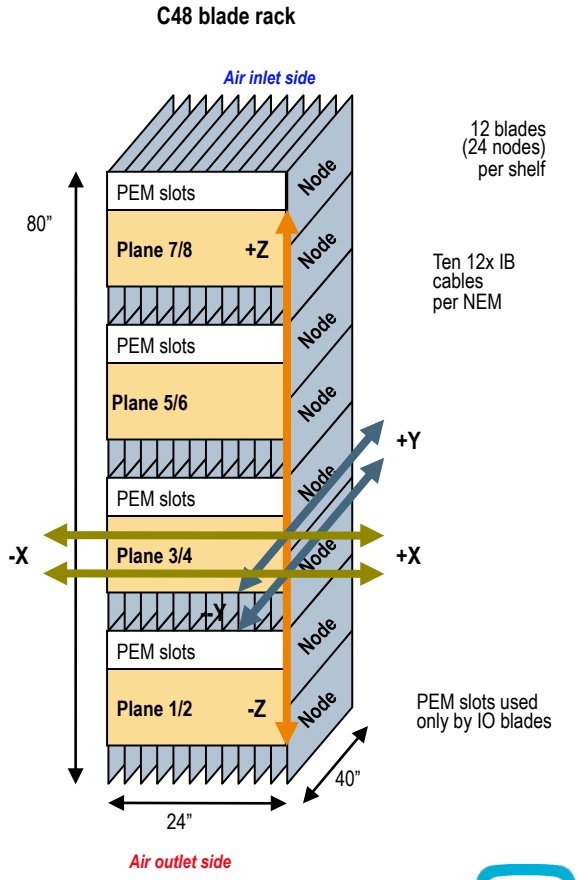


Mellanox ConnectX

Mellanox InfiniScale IV







Sun microsystems

Sandia National Laboratories