SAND2011-3362C

Application-Driven Acceptance of Cielo, an XE6 Petascale Capability Platform

Douglas Doerfler, **Mahesh Rajan**, *Sandia National Laboratories,* **Cindy Nuss**, *Cray Inc.,* **Cornell Wright**, *Los Alamos National Laboratory and* **Tom Spelce**, *Lawrence Livermore National Laboratory*

ABSTRACT: *Cielo is one of the first instantiations of Cray's new XE6 architecture and will provide capability computing for the NNSA's Advanced Simulation and Computing (ASC) Campaign. A primary acceptance criteria for the initial phase of Cielo was to demonstrate a six times (6x) performance improvement for a suite of ASC codes relative to its predecessor, the ASC Purple platform. This paper describes the 6x performance acceptance criteria and discusses the applications and the results. Performance up to tens of thousands of cores are presented with analysis to relate the architectural characteristics of the XE6 that enabled the platform to exceed the acceptance criteria.*

KEYWORDS: Application Performance, XE6

1. Introduction[1]

Cielo is the current capability computing platform for the NNSA's Advanced Simulation and Computing (ASC) Campaign. Its programmatic predecessor is the Purple platform, operated by Lawrence Livermore National Laboratory, which was retired in November 2010. Cielo is the initial project of the Alliance for Computing at the Extreme Scale (ACES), a collaboration between Los Alamos National Laboratory and Sandia National Laboratories to create a New Mexico center for high performance computing [1].

The initial deployment of Cielo was completed in December 2010. The primary metrics for the acceptance of Cielo were availability, reliability and application performance. All key criteria for ensuring a productive and successful platform. In this paper we describe the criteria and analyse the results of the application performance testing that was performed as a part of the acceptance of the platform. The application performance requirement is to demonstrate a six times (6x) improvement in capability using a suite of ASC codes. Improvement is relative to the Purple system. Using a suite of codes for acceptance criteria is not new, a similar approach was used in the acceptance of the Red Storm platform at Sandia National Laboratory [2]. Using real applications for evaluating performance was an extremely effective method for Red Storm and the Cielo design team chose to add a similar acceptance test for Cielo.

In section 2 we describe the high-level architecture of Cielo and Purple. Section 3 provides a high level description of each of the 6x applications. In section 4 the capability improvement factor is defined, the key performance characteristics that were used to establish the 6x requirement is described and results are presented and analysed. Sections 5 and 6 address potential future work and concluding remarks.

2. Architecture Descriptions

Cielo

Cielo is the latest ASC Tri-Lab capability computing system and is one of the first instantiations of the Cray XE6 architecture [3]. At the time of this study the Cielo system was composed of 6,704 compute nodes, each configured with Advanced Micro Devices 2.4 GHz, eight-

[1] This work was supported by the U.S. Department of Energy. Sandia is a multi program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States National Nuclear Security Administration and the Department of Energy under contract DE-AC04-94AL85000.

core (model 6136) Magny-Cours processor for a total of 107,264 compute cores and a peak performance of 1.03 PFLOPS. The system will grow in May 2011 to 8,894 compute nodes, for a total of 142,304 cores and 1.37 PFLOPS peak performance.

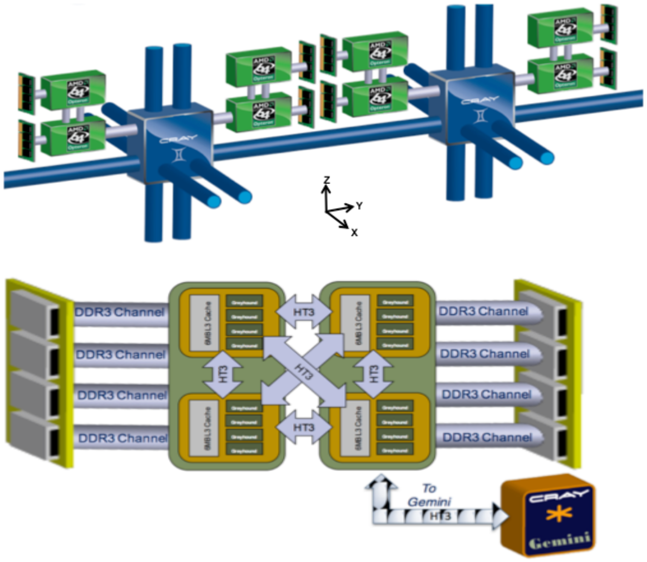

Figure 1. Cielo Node and the Gemini Interconnect (Courtesy, Cray Inc.)

Each compute node has two processors, with each processor consisting of two four-core dies for a total of sixteen cores per node, arranged as four separate NUMA regions. HyperTransport™ links connect the dies as shown in figure 1. As one would expect the NUMA nature of the node needs to be considered when optimizing node performance. Note the arrangement of the four DDR3 memory channels (two per die) providing direct access to 4 GB DIMMS for a total of 32 GB per node.

For the XE6 architecture the Gemini high-speed interconnect replaces the SeaStar interconnect used in the XT. Gemini was designed to support multicore processors and scale up to millions of cores in a single system.

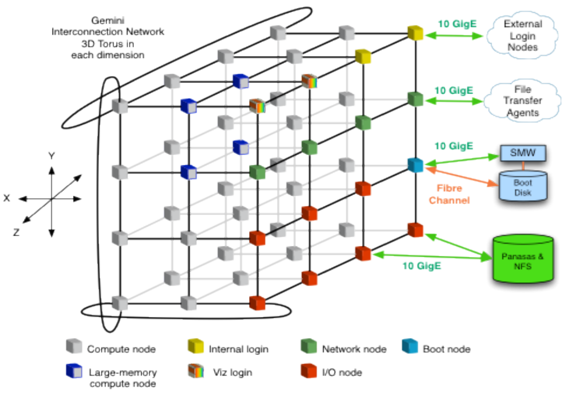

Figure 2. Cielo Configuration Diagram

The Cielo system is configured as an 18x8x24 3D-torus network. A pictorial representation of Cielo identifying the principal components is shown in figure 2.

Purple

The ASC Purple platform is Cielo's predecessor as the production capability computer for the ASC program. Sited at and operated by Lawrence Livermore National Laboratory, Purple was initially deployed in 2005 and retired in November 2010. An instantiation of IBM's POWER Architecture, Purple consisted of 1,336 IBM p5 575 compute nodes connected by the Federation high-speed interconnect, with an aggregate peak performance of 81.2 TFLOPS.

The Purple compute node architecture consisted of eight IBM Power5-based Dual Chip Modules (DCM) that together operated as a single SMP system. Each DCM contained a Power5 processor chip coupled with a separate 36 MB L3 cache chip. However, in Purple only one core per Power5 processor was enabled, leaving the full L3 cache and memory capacity available to the single active core on each chip [4].

The Federation high-speed interconnect consists of switch network interfaces (SNI) on each node and High Performance Switches (HPS) connecting the nodes. Each Purple node has a single two port SNI, where each port is capable of 4 GB/second peak bi-directional bandwidth. The high performance switches are connected in a fat tree topology with 3 levels of switches.

A comparison of key Cielo and Purple specifications is shown in table 1.

Table 1: Cielo and Purple Specifications

	Cielo	Purple
# of Compute Nodes	6,704	1,336
# of Processors/Node	2	8
# of Cores/Processor	8	1
Total # of Compute Cores	107,264	10,688
Processor	AMD Magny-Cours	IBM Power5
Frequency	2.4 GHz	1.9
FLOPS/Clock	4	4
GFLOPS/Node	153.6	60.8
Memory Type	1333 MHz DDR3	533 MHz DDR1
Memory/Node	32 GB	32 GB
Peak Memory BW/Node	85.3 GB/s	99.2 GB/s
Network Interface	Cray Gemini	IBM Federation
Network Topology	18x8x24 3D Torus	Fat-Tree, 3 Level
Ping-Pong Latency	~1.3 uS	~4.4 uS
Bidirectional Link BW	18.8 GB/s X&Z 9.4 GB/s Y	8 GB/s

3. Application Descriptions

Applications were chosen to be representative of the type of workloads expected to run on Cielo in production. Each of the three ASC laboratories was allowed to choose two applications, for a total of six. A taxonomy of languages used in provided in table 2

Charon (SNL)

Charon is a semiconductor device simulation code. Charon uses a drift-diffusion model, which is a coupled system of nonlinear partial differential equations that relate the electric potential to the electron and hole concentrations. The problem used is an example of a 2D steady-state drift-diffusion solution for a bipolar junction transistor, applied in a weak-scaled method. There are approximately 31,000 degrees of freedom per MPI rank. Inter-process communication involves 100's of Bytes to 10's of KB message transfers and small message reduction operations. At scale, Charon becomes communication bound and is sensitive to small message MPI_SEND rates, MPI_ALLREDUCE collective performance and OS Noise.

CTH (SNL)

CTH is a multi-material, large deformation, strong shock wave, solid mechanics code. CTH has models for multi-phase, elastic, viscoplastic, porous and explosive materials, using second-order accurate numerical methods to reduce dispersion and dissipation and produce accurate, efficient results. The problem used for this study is the shaped charge problem, in three dimensions on a rectangular mesh, in a weak-scaled method. Inter-process communication aggregates cell data into MB size MPI messages. CTH uses MPI Send calls with matching MPI Recv calls, communicates in a relatively small localized region, is limited by peak interconnect bandwidth and can be sensitive to node placement.

SAGE (LANL)

SAGE is a multidimensional, multi-material Eulerian hydrodynamics code. The timing_h problem in a weak-scaled method is used. Inter-process communication is through a bulk-synchronous gather/scatter abstraction, which collects off-process data and inserts it into doubly indexed arrays; the receiver unpacks the message, also using a doubly indexed array. MPI message sizes are in the 100's of KB to 1 MB range. At scale, this problem can be communication bound if MPI performance for these byte ranges is low. Sage can also be sensitive to MPI_ALLGATHER collective performance.

xNobel (LANL)

xNobel is a one, two, or three dimensional, multi-material Eulerian hydrodynamics code. It was developed for solving a variety of high deformation flow of materials problems, with the ability to model high explosives. The problem used for this study is the sc301p shape charge problem in three dimensions, in a weak-scaled method. Interprocess communication consists of relatively small messages in the 10's of bytes to 100's of KB size. At scale, this problem becomes communication bound and is sensitive to small message MPI_ISSEND transfer rates and latency.

AMG2006 (LLNL)

AMG2006 is a parallel algebraic multigrid solver of linear systems arising from problems on unstructured grids. Configured for weak scaling on a logical three-dimensional processor grid (px*py*pz), AMG solves the Laplace equations on a global grid of dimension px*220 x py*220 x pz*220. The figure of merit is related to the solve phase time for the preconditioned conjugate gradient solver for 100 iterations (higher is better) as defined in table 4. Runtime is dominated by the memory bandwidth requirements of the sparse matrix-vector product at small core counts and by MPI_ALLREDUCE performance at large core counts with a message size of about 2 KB. The other MPI routines, mostly non-blocking point-to-point communication, consume a negligible fraction of the communication cost.

UMT2006 (LLNL)

The UMT benchmark is a 3D, deterministic, multigroup, photon transport code for unstructured meshes. The deterministic transport code solves the first-order form of the steady-state Boltzmann transport equation. The equation's energy dependence is modeled using multiple photon energy groups. The angular dependence is modeled using a collocation of discrete directions, or "ordinates." The spatial variable is modeled with an "upstream corner balance" finite volume differencing technique. The solution proceeds by tracking through the mesh in the direction of each ordinate. For each ordinate direction all energy groups are transported, accumulating the desired solution on each zone in the mesh. The MPI messaging demands of UMT are low.

Table 2: Taxonomy of application languages

Lab	Code	Fortran	Python	C	C++	MPI	OpenMP
SNL	Charon			X	X	X	
SNL	CTH	X		X		X	
LANL	xNobel	X		X		X	
LANL	SAGE	X		X		X	
LLNL	AMG2006			X		X	X
LLNL	UMT2006	X	X	X	X	X	X

4. Results

Method

The purpose of the application acceptance test is to demonstrate the increased capability of the Cielo platform relative to its programmatic predecessor, the ASC Purple platform. The requirement is to demonstrate at least a six times improvement (6x) in capability, defined to be the product of increased problem size and runtime performance speedup relative to Purple. For example, if the problem size executed on Cielo is eight times larger then the one executed on Purple (i.e. 8x weak scaling) and the runtime metric of interest demonstrates a speedup of 1.25 relative to Purple, then the capability improvement becomes 8x * 1.25 = 10x.

The factor of 6x is somewhat arbitrary. But it is a factor that the ACES design team felt was achievable for the state of technology and budget available at the time of procurement. And from a programmatic point of view it’s roughly correlated to Moore’s Law. Cielo is being deployed approximately 5 years after Purple. If you use the interpretation of Moore’s Law that performance doubles every two years, you expect your next generation capability platform to provide roughly 6 times improvement in capability for your applications. So 6x feels about right and provides a challenging target for the vendor without being unrealistically too high.

The Purple baseline data was collected at a nominal scale of 1024 Purple nodes (8192 processors). This scale was chosen because most scaling studies are easily sized to fit a power of 2 and it has been shown that Purple works very well at this job size. For Cielo, it is desirable to use as much of the platform as possible, but to be fair no more than 5,138 nodes could be used, the same ratio of 1024 out of the 1,336 total Purple compute nodes.

Some of the key characteristics that translate to performance are captured in table 3 and a ratio of Cielo to Purple is provided. Although the peak double precision floating-point capability of Cielo is more than 12 times that of Purple, aggregate memory bandwidth is a more appropriate measure of computational performance for the ASC codes and Cielo has a little more than 4 times increase in this metric. That is, from a processor-to-processor perspective a Purple node is a very capable platform and achieving a 6x improvement in capability requires Cielo to demonstrate excellent scalability at the larger scales. The 6x requirement was not viewed as an easy metric to meet by the Cielo design team or Cray.

Table 3: Comparison of key Cielo and Purple characteristics

	Purple	Cielo	Ratio
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

The runtime figure of merit (FOM) speedup is not necessarily wall clock time and is application dependent. The intent was to pick a metric which is a measure of the platforms scaling characteristics, NOT that of the application. For many applications total time to solution is not a good performance metric from an algorithm perspective. For example, many ASC codes use iterative solvers and as the problem scales to a larger number of MPI processes the number of iterations to solution increases and hence time to solution increases. In this case, a more appropriate metric would be the average time per iteration step. Application FOMs are captured in table 4.

Table 4: Application figure of merit

	Figure of merit	Direction
Charon	Seconds (Solve time per iteration)	Lower is better
CTH	Seconds (Total Zone-cycles * Seconds per Zone-cycle)	Lower is better
SAGE	Sum_cell_cycles/second/PE	Higher is better
xNobel	Sum_cell_cycles/second/PE	Higher is better
AMG2006	# of PEs * # of Iterations / Solve Phase time	Higher is better
UMT2006	FOM as reported by code	Higher is better

Application Scaling

Scaling studies were performed for each application on Cielo, using increasingly larger number of PEs, figures 3 through 8. Although it was desirable to collect scaling study results on Purple, due to time and man power constraints some results were only collected for the 8,192 baseline data point (for AMG and UMT the baseline was 8,000 PEs). All of the studies utilized weak scaling, so for Charon, CTH, SAGE and xNobel a horizontal line would represent perfect scaling. The FOM for AMG and UMT increase linearly with scale for perfect scaling.

It was not the goal of the testing to analyse the scaling characteristics of each application, for Cielo or Purple. But interesting results include: Charon's poor scaling on Purple which resulted in a high improvement factor for that application; CTH scaled well on Purple up to 8K PEs, where Cielo scaling gets progressively worse, thus contributing to a low improvement factor for Cielo at 64K PEs; and the rapid roll off of xNobel scaling on Cielo. The applications Charon, SAGE and xNobel were limited in the number of cells per PE that could be used to do the fact that these applications use signed 32-bit integers to store the total number cells for the problem. This limitation forced testers to use a smaller overall problem size than was desired for these applications. This is partly responsible for relatively poor scaling demonstrated by SAGE and xNobel as the computational work was not sufficient to offset the increasing communication costs as scale increases.

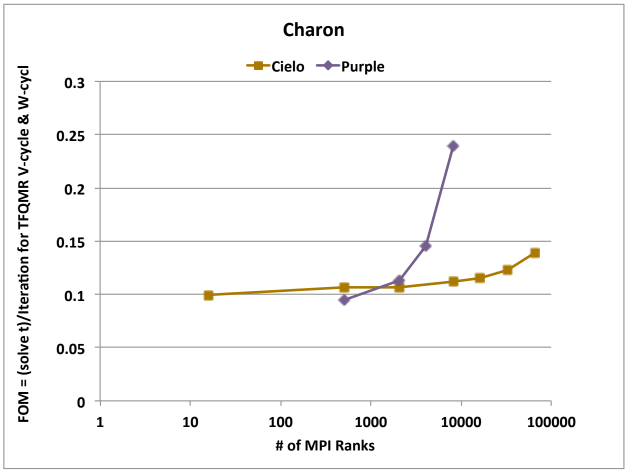

Figure 3: Charon scaling. Lower is better.

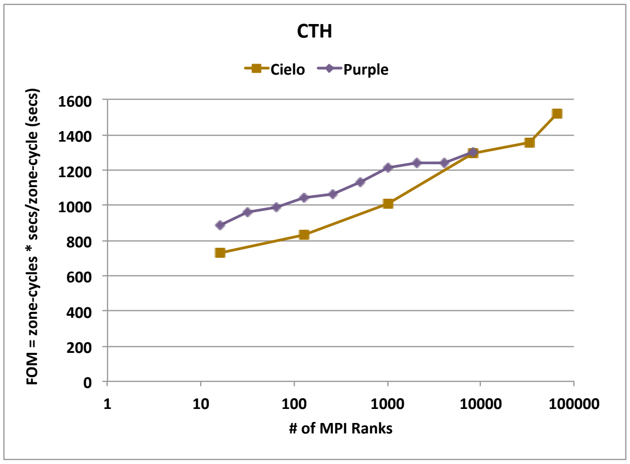

Figure 4: CTH scaling. Lower is better.

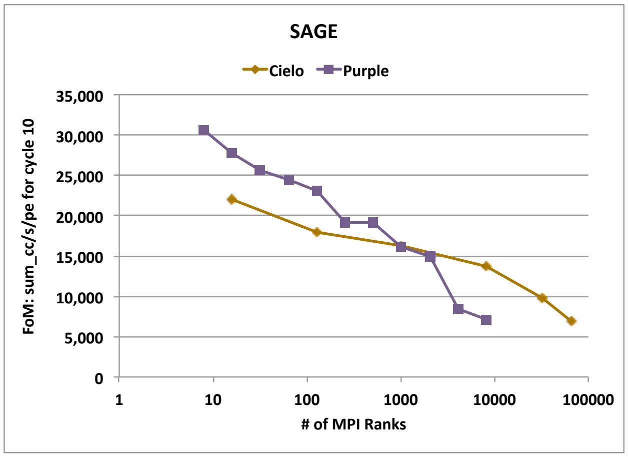

Figure 5: SAGE scaling. Higher is better.

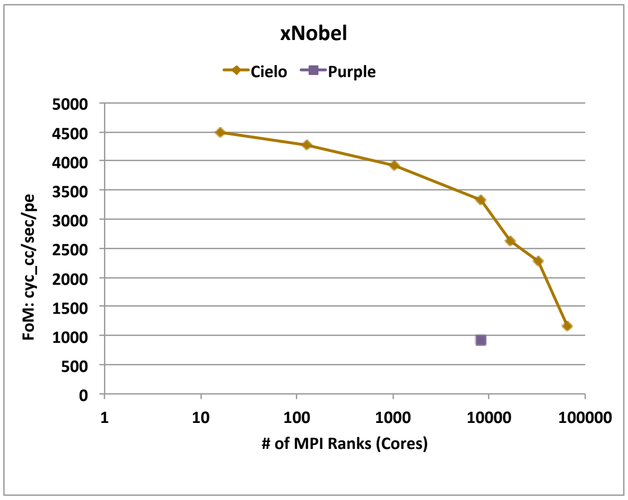

Figure 6: xNobel scaling. Higher is better.

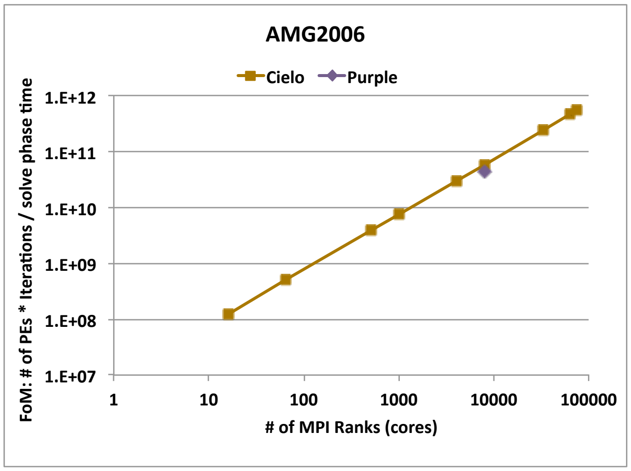

Figure 7: AMG2006 scaling. Higher is better.

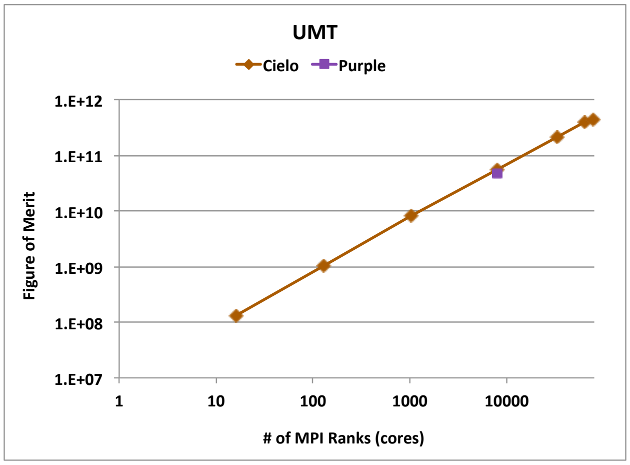

Figure 8: UMT scaling. Higher is better.

Capability Improvement

As defined above, capability improvement of Cielo relative to Purple is the product of the increase in problem size and the FOM speedup observed by the application at a given scale. The improvement factor was calculated for varying degrees of weak scaling: 1x in which results for 8K PEs are directly compared for both platforms, 4x in which Cielo results at 32K PEs are compared to 8K PEs of Purple, 8x in which 64K PEs are compared to 8K PEs of Purple, and >8x in which the application results for AMG at 74,088 PEs and UMT at 77,616 PEs were used.

At 64K PEs, the demonstrated Cielo capability improvement is 9.6x as shown in table 5 and figure 9. Using the AMG and UMT results that were >64K PEs, the demonstrated improvement is 10.5x. Thus exceeding the requirement of 6x and passing the acceptance test for application performance.

Table 5: Capability Improvement Factors

	8K Cores (1x Weak Scaling)	32K Cores (4x Weak Scaling)	64K Cores (8x Weak Scaling)	> 64K Cores
CTH	1.0	3.8	6.9	
Charon	2.1	7.8	13.8	
SAGE	1.9	5.5	7.7	
xNobel	3.7	10.1	10.2	
AMG	1.3	5.4	10.7	12.2
UMT	1.2	4.6	8.4	9.3
Average	1.9	6.2	9.6	10.5

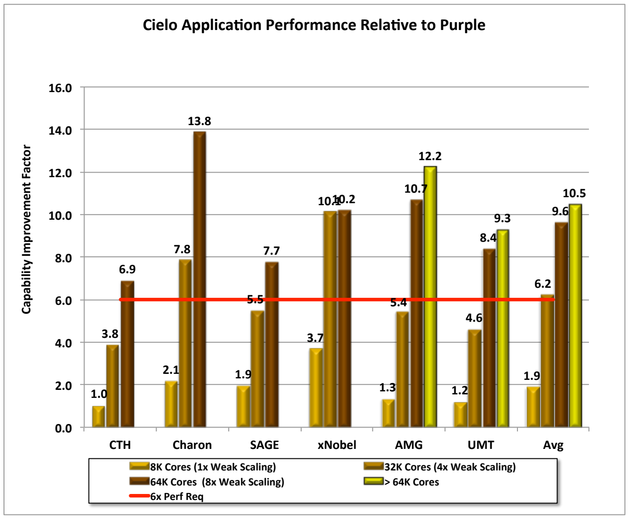

Figure 9: Capability improvement factors for the six applications at varying levels of increasing problem size, 1x, 4x, 8x, and >8x (AMG ran up to 74,088 PEs and UMT ran up to 77,616 PEs.

5. Future Work

This testing was only for phase 1 of the Cielo deployment. Phase 2 takes place in May 2011 and the current schedule, which is being driven by the desire to return the platform to production work as soon as possible, does not allow enough time to repeat a study as extensive as these tests. As was seen in phase 1 testing three of the six applications have limitations on the total number of cells in a problem. As such for phase 2 application acceptance testing will be limited to dedicated system time for large scale testing and a reduced number of applications. In addition, the Purple platform will no longer be used as a point of reference. For phase 2 capability improvement will be measured against phase 1 results.

Analysing the scaling characteristics of the platform does not end with acceptance testing and will continue throughout the life the platform with the goal of improving application productivity of Cielo and Cray's XE6 product line. Specifically for the phase 1 acceptance applications CTH and xNobel, it will be desirable to obtain a better understanding of issues that are limiting the scaling characteristics of the applications.

6. Conclusion

For the Cielo design team, the key metrics of success for the platform are availability, reliability and application performance. In this paper we described the criteria that was used to judge application performance of the Cielo platform for phase 1 acceptance testing. Detailed scaling results were presented in addition to capability improvement factors, relative to the ASC Purple platform, at varying scales. In general Cielo exceeded design requirements and demonstrated up to 10.5x improvement factors. Very good scaling characteristics were demonstrated for the test applications. Although issues were observed for a few, such as reduced scaling characteristics of CTH and xNobel at large scale. The Cielo team will continue to evaluate application performance of the platform and acceptance testing is only the first steps towards evaluating the platform, with the goal of improving platform productivity over its lifetime.

7. Acknowledgments

This effort involved many people and the authors would like to thank them for their contributions. Cray contributors included Mike Davis, Steve Whalen, Ting-Ting Zhu, Ron Pfaff, Stephan Behling, Kevin McMahon, Frank Kampe and David Whitaker. LANL contributors include Scott Pakin, Mike Lang and Craig Idler. From LLNL the authors would like to thank Scott Futral, and from SNL Paul Lin and Courtenay Vaughan.

8. About the Authors

Douglas Doerfler is a Principle Member of Technical Staff at Sandia National Laboratories. Doug is the ACES Cielo Architect and his research interests include high-performance computer architectures and performance analysis.

Mahesh Rajan is a Distinguished Member of Technical Staff at Sandia National Laboratories. Mahesh supported the Cielo application acceptance tests and his research interests include high-performance computer architectures and performance analysis.

Cindy Nuss is the manager of the Benchmarking and Performance Analysis group at Cray Inc. Cindy was a member of the ACES/Cray Application Benchmarking Team working on the Cielo 6x application acceptance.

Cornell Wright is a Member of Technical Staff at Los Alamos National Laboratory. Cornell leads the Application Readiness team at LANL and is co-lead for Cielo Application Readiness.

Tom Spelce is a Member of the Technical Staff at Lawrence Livermore National Lab. Tom's interests include performance analysis, development tools and emerging hardware architectures.

References

[1] James Ang, et al., "The Alliance for Computing at the Extreme Scale", Cray User Group (CUG) 2010, University of Edinburgh, England, May 2010.
[2] Ron Brightwell, et al., Architectural specification for massively parallel computers: an experience and measurement-based approach, *Concurrency and Computation: Practice and Experience*, August 2005
[3] Cray XE6, http://www.cray.com/Products/XE/Systems/XE6.aspx/
[4] ASC Purple, https://asc.llnl.gov/computing_resources/purple/
[5] Blaise Barney, Using ASC Purple - A Tutorial, https://computing.llnl.gov/tutorials/purple/.
[6] Ron Kalla, Balaram Sinharoy, and Joel M. Tendler. IBM Power5 Chip: a Dual-Core Multithreaded Processor. IEEE Micro, 24(2):40–47, Mar-Apr 2004.