Performance on Trinity (a Cray XC40) with Acceptance-Applications and Benchmarks

Nathan Wichmann, Cindy Nuss, Pierre Carrier, Ryan Olson, Sarah Anderson and Mike Davis
Cray Inc.,
wichmann, cnuss, pcarrier, ryan, saraha, u3186@cray.com

Randal Baker, Erik W. Draeger, Stefan Domino, Anthony Agelastos, Mahesh Rajan
rsb@lanl.gov, draeger1@llnl.gov, spdomin@sandia.gov, amagela@sandia.gov, mrajan@sandia.gov

Cray User Group Meeting, May 8-11, 2016, London, UK

This work was supported in part 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.

National Nuclear Security Administration

Sandia National Laboratories

NNSA's First Advanced Technology System(ATS-1)
Previous Capability Computing Systems: Cielo, Sequoia

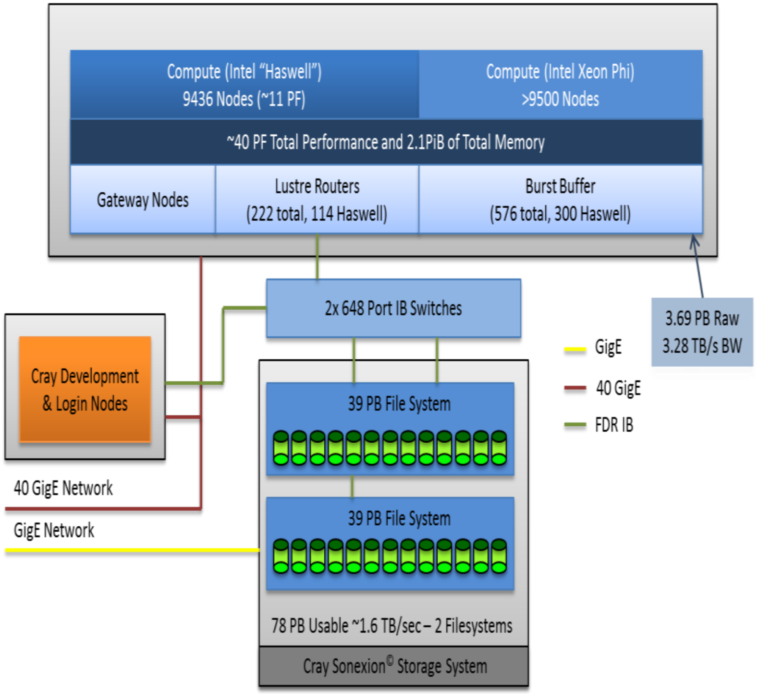
- Trinity (ATS-1) deployed by ACES (New Mexico Alliance for Computing at Extreme Scale) and sited at Los Alamos.
- ATS-2 will be led by LLNL, ATS-3 by ACES, …

Cielo
- Cray XE6
- Nodes =8944
- Memory > 291.5TB
- Peak Performance =1.37 PF
- AMD MagnyCours(16 cores/node)

Sequoia
- IBM BG/Q
- Nodes = 98,304
- Memory = 1.6PB
- Peak Performance = 20PF
- IBM PowerPC A2 (16 cores/node)

Trinity
- Cray XC40
- Nodes > 19000
- Memory > 2PB
- Peak Performance > 40PF
- Intel Haswell (32 cores/node) & Knights Landing(72 cores/node)

Trinity Architecture: Phase-1 Haswell Nodes(installed 2015); Phase-2 KNL nodes (in 2016)



- Peak Node Performance: 32cores*16FLOPs/cycle*2.3GHz = 1,177.6 GFLOPS/node
- Intel® Hyper-Threads and Intel® Turbo Boost

Trinity Phase-1 Acceptance

Completed December 2015

Focus here on application performance measures

Primary Focus

Acceptance Tests and Criteria

1) **Capability Improvement(CI) metric**
   - CI Metric = problem-size-increase x run-time-speedup
   - **4X over a baseline performance measured on 2/3rd of the nodes on Cielo**
   - **runs at near full scale of Trinity**
   - **May use appropriately scaled inputs**
   - **Applications representative of planned Tri-lab productions apps**
2) NERSC's Sustained System Performance (SSP) target of 400; specified input: "large"
3) Microbenchamrks: Stream, OMB, SMB, mpimemu, psnap, pynamic
4) Run at full scale SSP benchmarks: miniFE, miniGhost, AMG, UMT and SNAP

Cielo, Trinity Architectural Parameter Comparisons

System	Cielo (XE6)	Trinity(XC40)
Total Nodes	8,894	9,436
Total Cores	142,304	301,952
Processor	AMD MagnyCours	Intel Haswell
Processor ISA	SSE4a	AVX2
Clock Speed(GHz)	2.40	2.30
Cores/node	16	32
Memory-per-core(GB)	2	4
Memory	DDR3 1,333 MHz	DDR4 2,133 MHz
Peak node GFLOPS	153.6	1,177.6
Channels/socket	4	4
Processor Cache: L1(KB) L2(KB) L3(MB)	 8 x 64 8 x 512 10	 16 x 32 16 x 256 40
Interconnect Topology	Gemini 3D Torus 18x12x24	Aries Dragonfly

CI Metric and Applications

SNL App: SIERRA/Nalu:

- Low Mach CFD code for incompressible flows; unstructured mesh; LES/Turbulence Models
- Test Problem:
  - Turbulent open jet (Reynolds number of ~6,000)
  - Weak scaling meshes (R1:268k elements, R2:2.15M elements, ….. R6: 9 billion elements)
- Figure of Merit: Solve time/Linear iteration (66%)& Assemble time/non-linear step(34%)

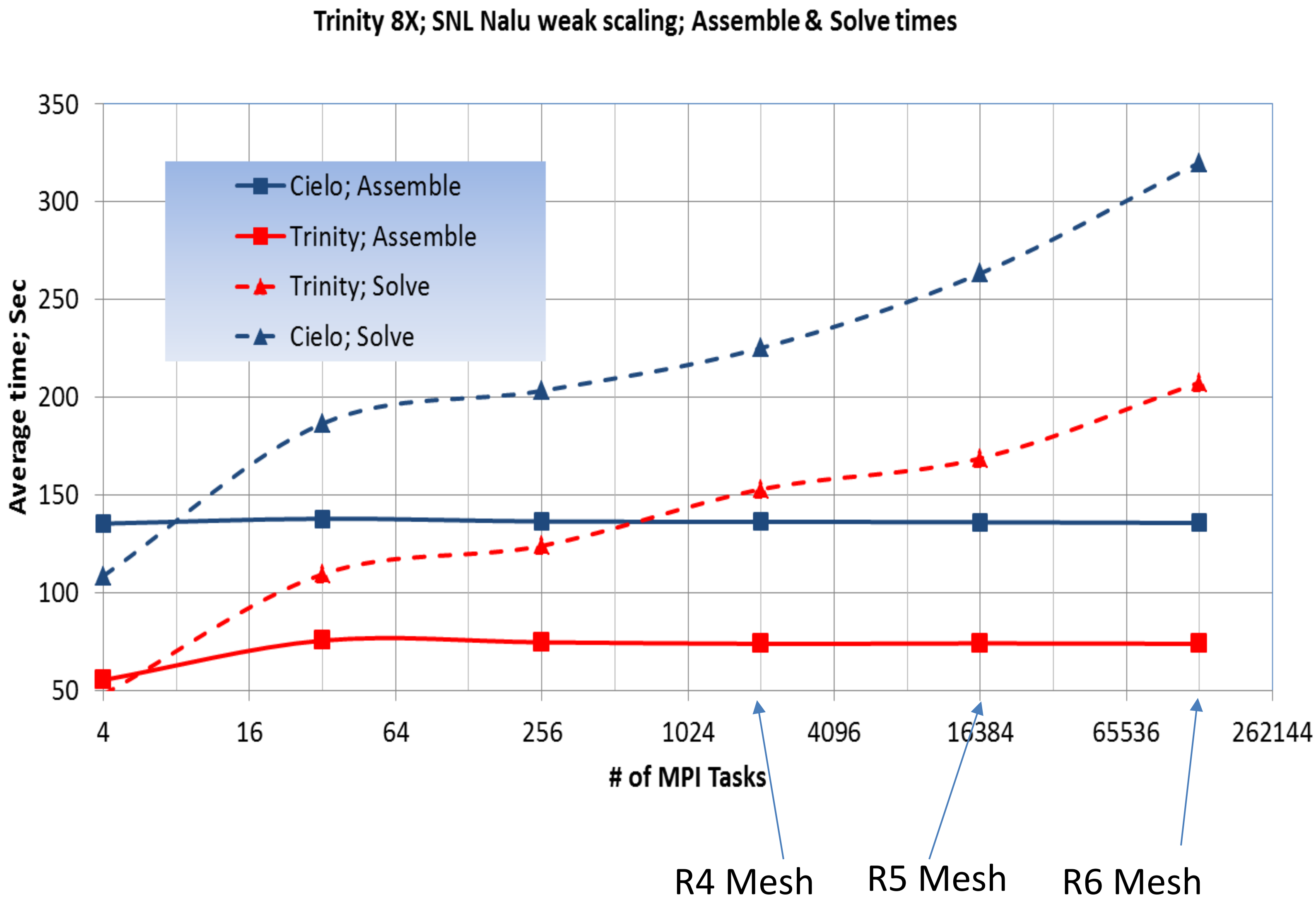
LANL App: PARTISN:

- Particle transport code provides neutron transport solutions on orthogonal meshes in one, two, and three dimensions
- Test Problem: MIC_SN (MIC with group-dependent Sn quadrature).
- Figure of Merit: *Solver Iteration Time (should stay constant for weak scaling)*

LLNL App: Qbox:

- first-principles molecular dynamics code used to compute the properties of materials at the atomistic scale
- Test Problem: benchmark problem is the initial self-consistent wave function convergence of a large crystalline gold system (FCC, a0 = 7.71 a.u).
- Figure of Merit: maximum total wall time to run a single *self-consistent iteration* with three non-self-consistent inner iterations)

SIERRA/Nalu Weak Scaling



SIERRA/Nalu CI Performance

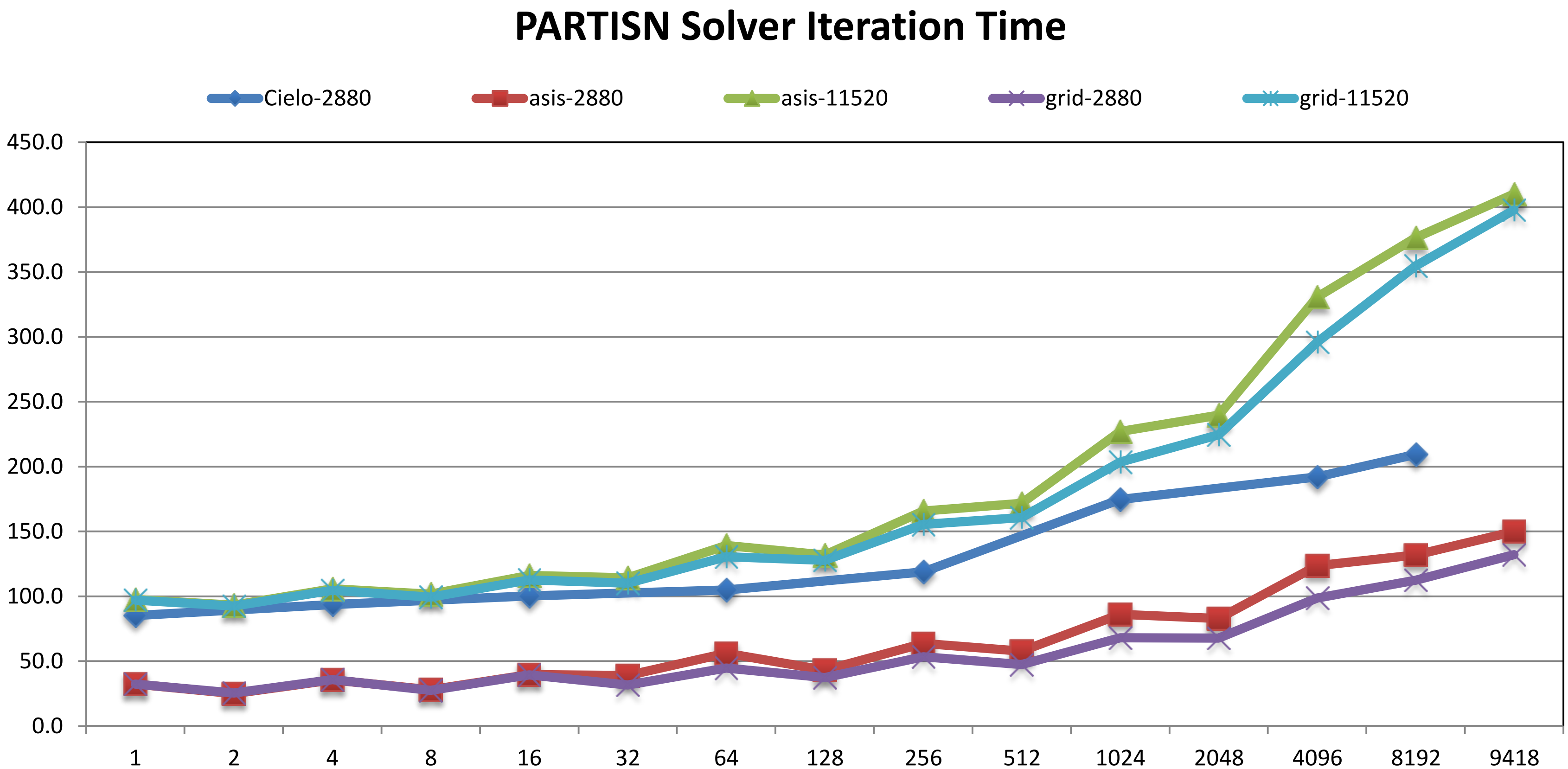
BASELINE on Cielo				Trinity Phase -1 CI Results			
Nodes	MPI Tasks	Problem Size Complexity Measure	RunTime FOM	Nodes	MPI Tasks	Problem Size Complexity Measure	RunTime FOM
8,192	131,072	*R6*: 9B elements	1.15	9,420	301,240	*R6*: 9B elements	0.286

Running the same problem (9 Billion element mesh) on 2.3 times the number of PEs resulted in a Capability Improvement Metric of =1.15/0.286 = **4.02**

PARTISN Weak Scaling

(2,880 & 11,520 zones/core)

('asis': default MPI mapping; 'grid' mapping with *grid_order)*



Nodes: XC40 32 ranks/node, one thread per rank- CIELO: 4 Ranks/node 4 threads/rank

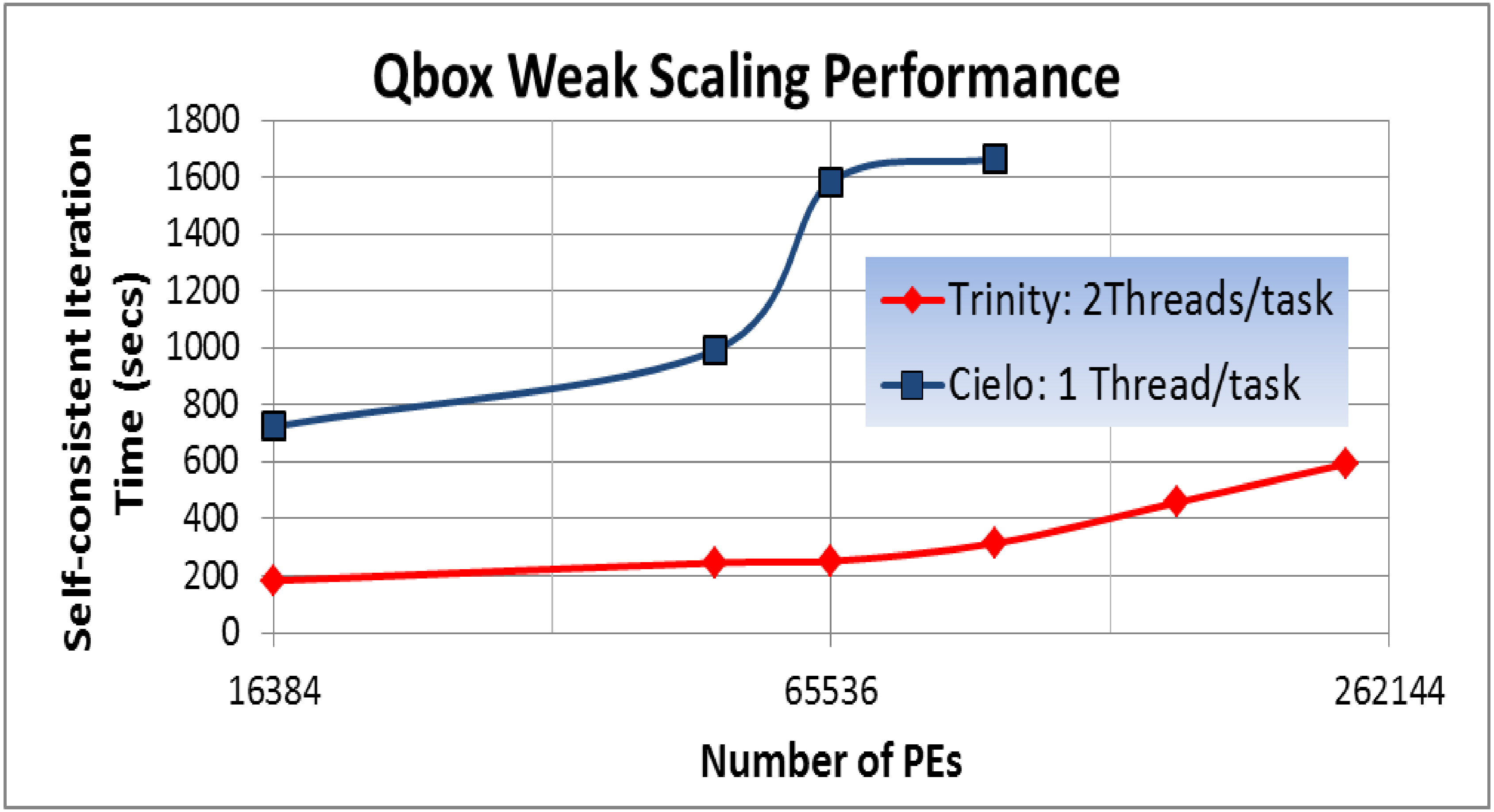
PARTISN CI Performance

BASELINE on Cielo				Trinity Phase -1 CI Results			
Nodes	MPI Tasks (4 Threads/task)	Problem Size Complexity Measure	RunTime FOM	Nodes	MPI Tasks	Problem Size Complexity Measure	RunTime FOM
8,192	32,768	*2,880 zones/core*	209.4 secs	9,418	301,376	*11,520 zones/core*	397.71 secs

Running a (11,520/2880)*2.3 = 9.19, larger problem on 2.3 times the number of PEs, took 1.899 times longer *solver iteration time* leading to a Capability Improvement Metric of = 9.19 / 1.899 = **4.84**

Qbox Weak Scaling

(1600 Atoms on 98,304 PEs; Trinity is 5.3X faster)



Qbox CI Performance

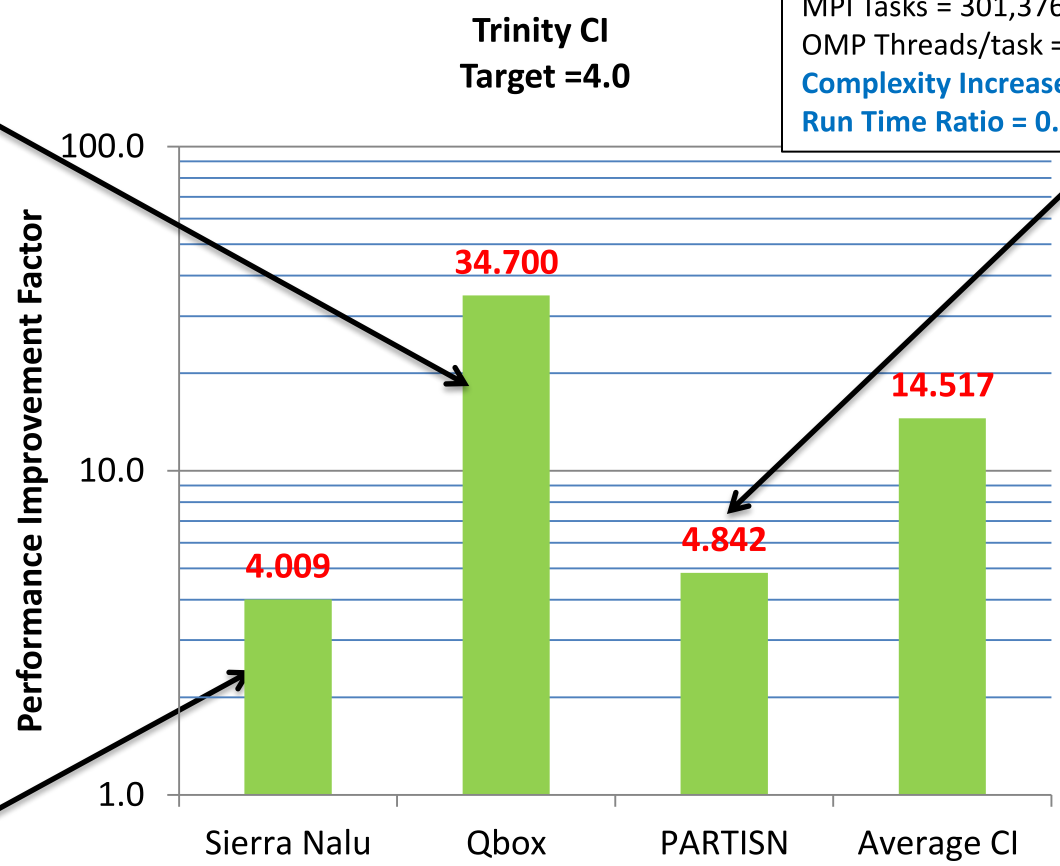
BASELINE on Cielo				Trinity Phase -1 CI Results			
Nodes	MPI Tasks 1 thread/task	Problem Size Complexity Measure	RunTime FOM	Nodes	MPI Tasks 8 threads/Task Hyperthreads	Problem Size Complexity Measure	RunTime FOM
6,144	98,304	*1,600 Atoms*	1663 secs	9,418	75,344	*8,800 Atoms*	7974 secs

Running a **(8800/1600)³** = 166.375 times larger problem on 3.065 times the number of PEs took 4.79 times longer leading to a Capability Improvement Metric of = 166.375 / 4.79 = **34.7**

Capability Improvement Summary

Qbox:
Nodes, = 9418
MPI Tasks= 75,344
OMP Threads/task=8
Hyperthreading used
Complexity Increase = 166.3
Run Time Ratio = 0.208

PARTISN:
Nodes, = 9418
MPI Tasks = 301,376
OMP Threads/task = None
Complexity Increase = 9.2
Run Time Ratio = 0.526



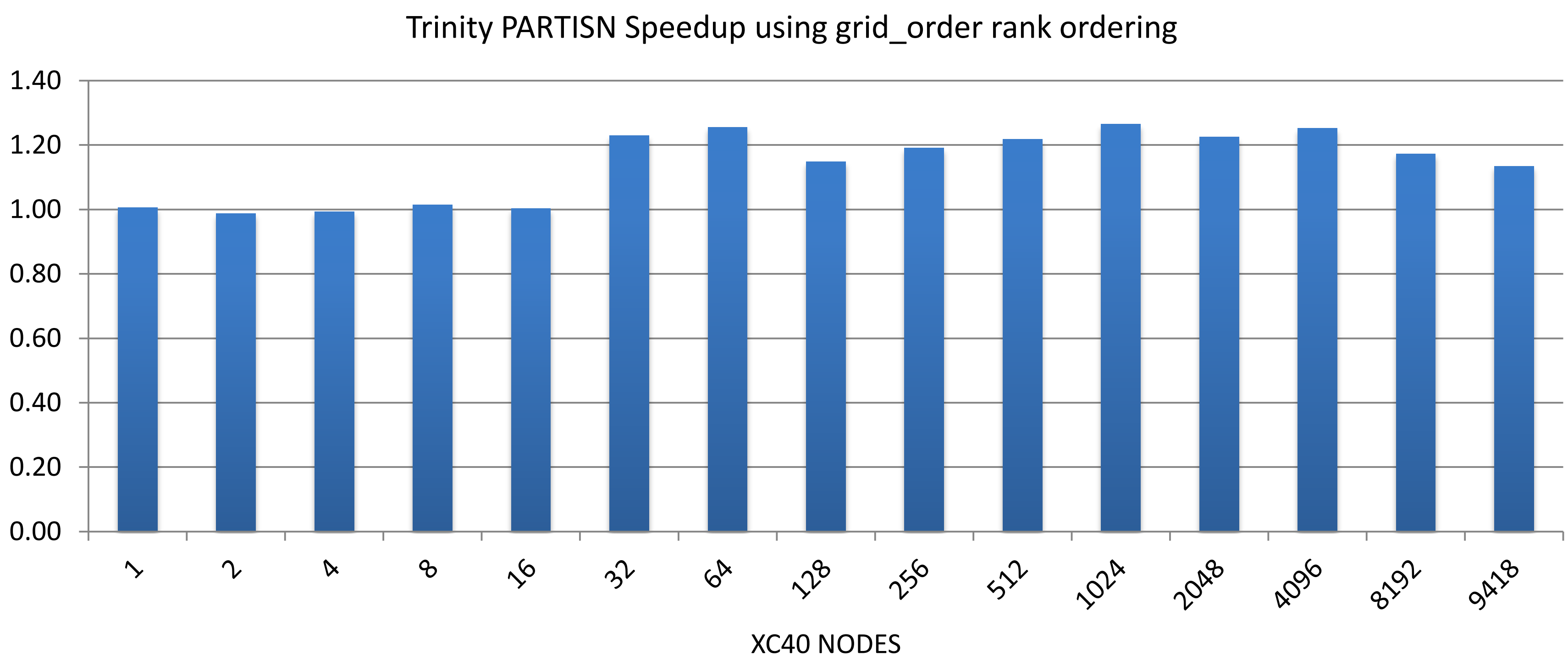
Nalu:
Nodes, = 9420
MPI Tasks = 301,240
OMP Threads/task=None
Complexity Increase = 1
Run Time Ratio = 4.009

PARTISN: Performance tuning

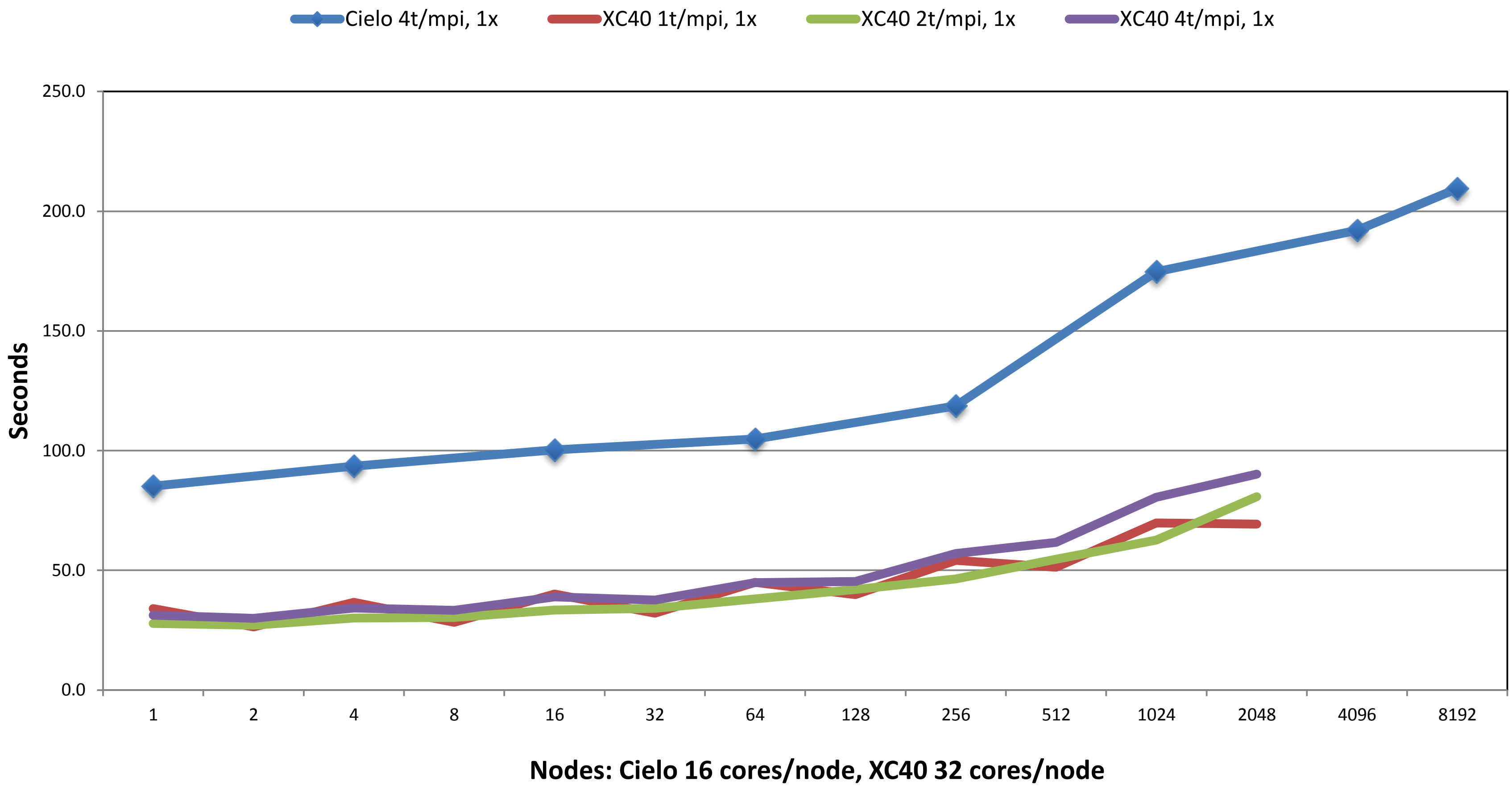
- opt_sweep3d(), which actually performs the KBA sweep that comprises the wave-front algorithm, took 85% of time; optimized for vectorization by Randy Baker and team at LANL.
- MPI Isend/MPI_Recv communications were frequent on the 2D processor mesh. Cray utility *grid_order* which “repacks” MPI ranks so that Cartesian mesh communication neighbors are more often on node was used to minimize communication overhead.
  - For example, a 16384 PE 128x128 mesh problem, use of *grid_order* improved MPI time by 42% and overall time by18%
- Example MPICH_RANK_ORDER file:

```
# grid_order -R -Z -c 2,8 -g 554,544 -m 301376 -n 32
# Region 0: 0,0 (0..301375)
0,1,2,3,4,5,6,7,544,545,546,547,548,549,550,551,8,9,10,11,12,1
3,14,15,552,553,554,555,556,557,558,559
16,17,18,19,20,21,22,23,560,561,562,563,564,565,566,567,24,25,
26,27,28,29,30,31,568,569,570,571,572,573,574,575 ...
```

grid_order impact studied at all scales up to 25% speedup!!



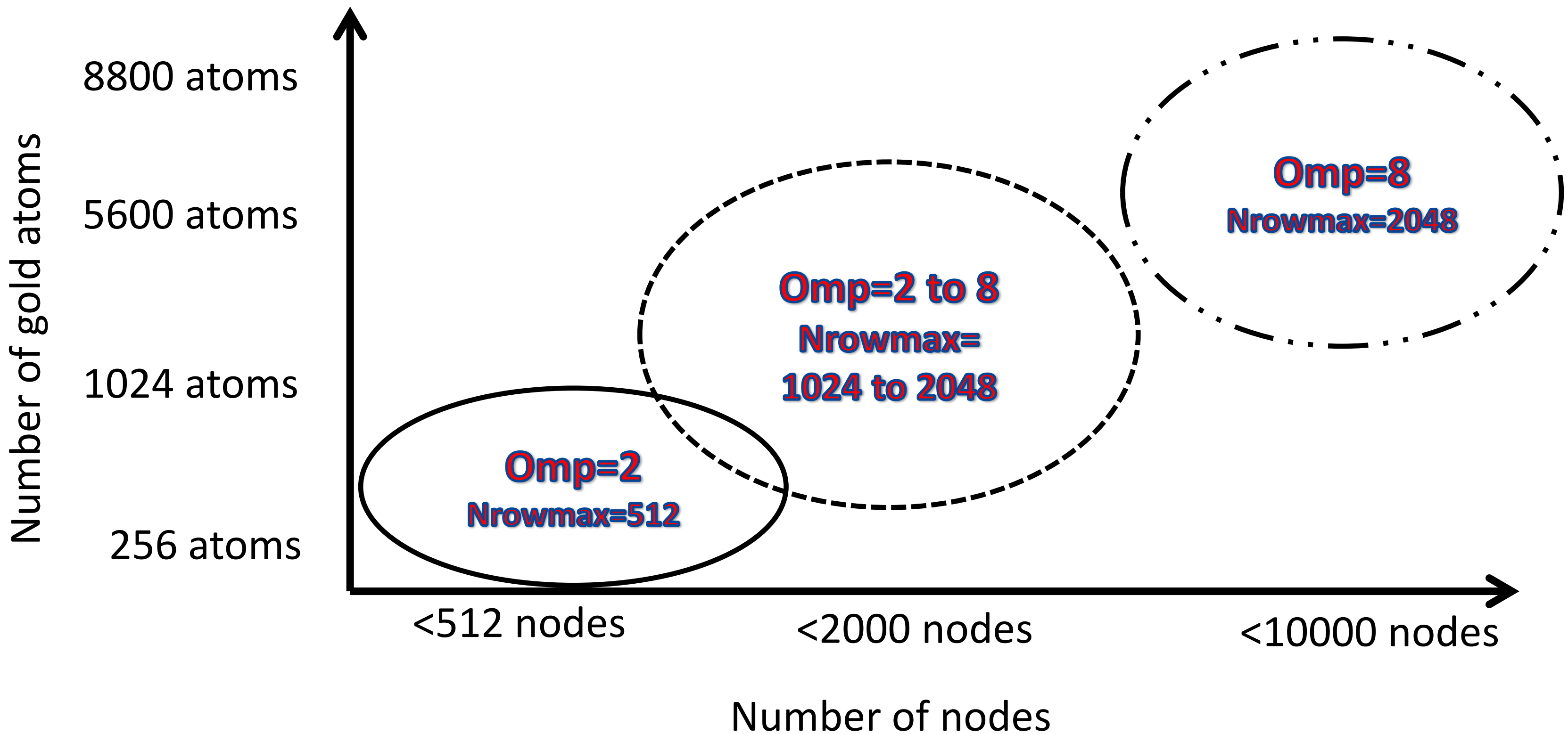
Trinity's best performance was with 1 thread/MPI rank; Cielo's 4 threads/rank



Qbox: performance tuning

- Compute time dominated by parallel dense linear algebra and parallel 3D complex-to-complex Fast Fourier Transforms (FFT)
- Efficient single-node kernels necessary to achieve good peak performance
- The communication patterns are complex, with nonlocal communication occurring both within the parallel linear algebra library (ScaLAPACK) and in sub-communicator collectives within Qbox, which are primarily MPI_Allreduce and MPI_Alltoallv operations.
- Threading implemented as a mix of OpenMP and threaded single-node linear algebra kernels
- 5,600 atom Qbox simulation showed 2X performance gain for 8 OpenMP threads/task when compared to 2

Qbox optimal performance: Impact of *nrowmax* and threads/task



Qbox performance: *grid_order* gave 6% to 30% performance gain

Qbox *iteration time* using MPI Grid ordering for 9418 node run
"grid_order -R -P -c 2,2 -g 34,1108" (without hyperthreading)
"grid_order -R -P -c 4,2 -g 68,1108" (with hyperthreading)

	2400 gold atoms	8800 gold atoms
Without grid order	456	9383
With grid order	315	8834
With grid order and hyperthreading	---	7974

NERSC’s Sustained System Performance (SSP) Metric

- A set of benchmark programs that represent a workload
- Computed as a geometric mean of the performance of eight Tri-Lab and NERSC benchmarks
  - *miniFE, miniGhost, AMG, UMT, SNAP, miniDFT, GTC and MILC*
- Problem size (“large”) specified
- Baseline data collected on NERSC’s Hopper (XE6); Baseline node-count suggested by production use
- Trinity run node-count not specified; For a few benchmarks, use of fewer nodes than the baseline, skewed the metric
- SSP being revised by NERSC, ACES for use with ATS-3

Trinity Phase-1 SSP target was 400: Achieved 500

Baseline SSP performance on NERSC's Hopper (Cray XE6)

Hopper Nodes	6384					
Hopper SSP						
Application Name	MPI Tasks	Threads	Nodes Used	Reference Tflops	Time (seconds)	Pi
miniFE	49152	1	2048	1065.151	92.4299	0.0056
miniGhost	49152	1	2048	3350.20032	95.97	0.0170
AMG	49152	1	2048	1364.51	151.187	0.0044
UMT	49152	1	2048	18409.4	1514.28	0.0059
SNAP	49152	1	2048	4729.66	1013.1	0.0023
miniDFT	10000	1	417	9180.11	906.24	0.0243
GTC	19200	1	800	19911.348	2286.822	0.0109
MILC	24576	1	1024	15036.5	1124.802	0.0131
					Geom. Mean=	0.0082
					SSP=	52.1212

SSP performance on Trinity

Trinity Nodes	9436					
Trinity SSP						pi: Rate(TF/s per Node)
Application Name	MPI Tasks	Threads	Nodes Used	Reference Tflops	Time (seconds)	Pi
miniFE	49152	1	1536	1065.151	49.5116	0.0140
miniGhost	49152	1	1536	3350.20032	1.77E+01	0.1229
AMG	49152	1	1536	1364.51	66.233779	0.0134
UMT	49184	1	1537	18409.4	454.057	0.0264
SNAP	12288	2	768	4729.66	1.77E+02	0.0348
miniDFT	2016	1	63	9180.11	377.77	0.3857
GTC	19200	1	300	19911.348	868.439	0.0764
MILC	12288	1	384	15036.5	393.597	0.0995
					Geom. Mean=	0.0530
					SSP=	500.0177

Micro-benchmark Results

STREAM Performance at a Trinity node

Function	Copy	Scale	Add	Triad
Rate (MB/s)	108,014	108,653	118,850	119,077

PSNAP OS jitter/Noise Benchmark Results

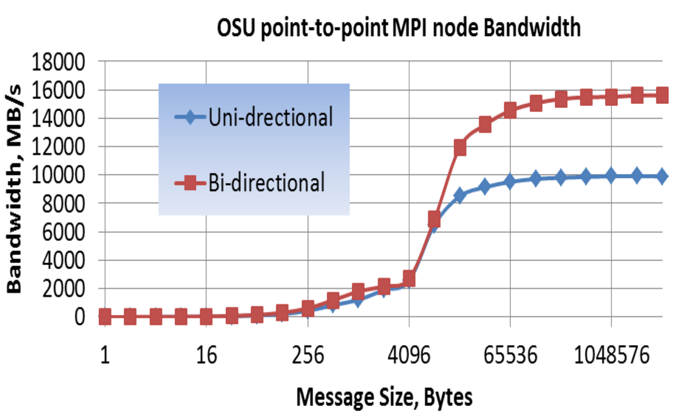
NR: 9436

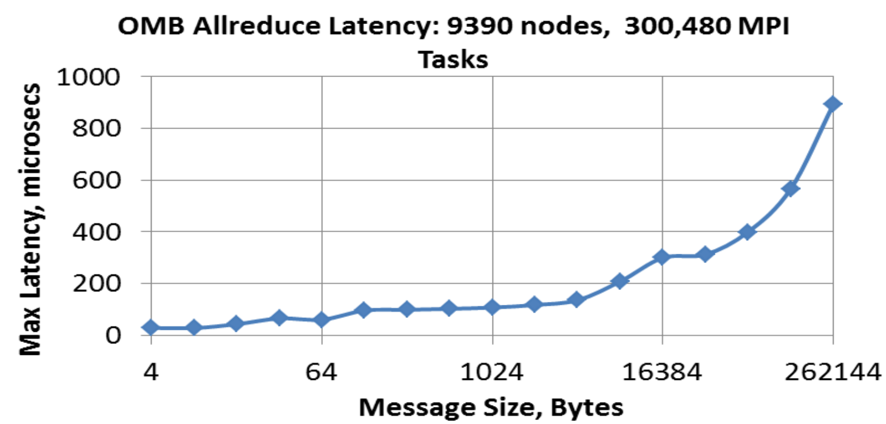
Average Slowdown: 0.15%

Min Slowdown: 0.13%

Max Slowdown: 0.18%

Maximum percentage slowdown at a core was measured to be 0.18%.





Conclusions

- Several months of effort by the Cray and Tri-Lab teams resulted in exceeding acceptance requirements
- Based on benchmark results, we anticipate production Trinity apps, will see a gain of 2x-6x over Cielo (with same number of PEs)
- The benefit of a hybrid code (MPI+threads) clearly seen with Qbox
  - ✓ 2x the performance with 8 threads over 2 threads per task
- use of *grid_order* resulted in good performance gains