TRINITY

SAND2017-9335PE

A Presentation on Trinity Performance

Dennis C. Dinge

Sandia National Laboratories Organization 9326

Los Alamos NATIONAL LABORATORY EST. 1943

Sandia National Laboratories

NNSA

Alliance for Computing at Extreme Scale

This presentation is almost completely lifted from the two presentations below.

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

Performance on Trinity Phase 2
(a Cray XC40 utilizing Intel Xeon Phi processors) with Acceptance-Applications and Benchmarks

A. Agelastos*, M. Rajan*, N. Wichmann#, P. Lin*, R. Baker+, S. Domino*, E. Draeger@,
S. Anderson#, J. Balma#, S. Behling#, M. Berry#, P. Carrier#, M. Davis#,
K. McMahon#, D. Sandness#, K. Thomas#, S. Warren#, and T. Zhu#
* Sandia National Laboratories, Albuquerque, NM
+ Los Alamos National Laboratory, Los Alamos, NM
@ Lawrence Livermore National Laboratory, Livermore, CA
# Cray, Inc., St. Paul, MN

JOWOG-34; June 26-29, 2017; Los Alamos, NM, USA

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

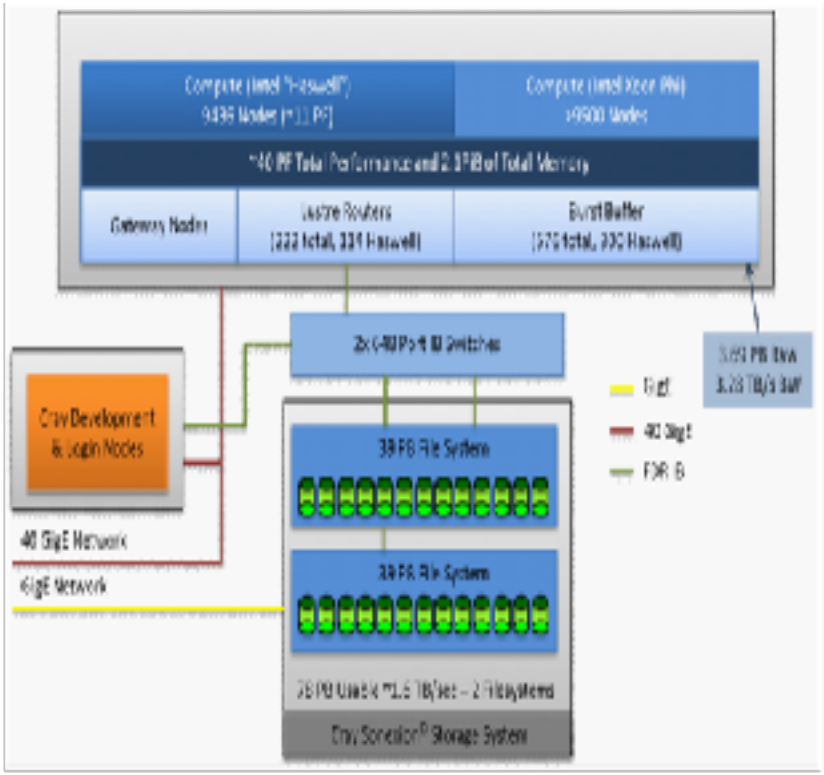
Trinity (ATS-1) deployed by ACES (New Mexico Alliance for Computing at Extreme Scale, i.e. Los Alamos & Sandia) and sited at Los Alamos. ATS–2 will be led by LLNL, ATS-3 by ACES

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)

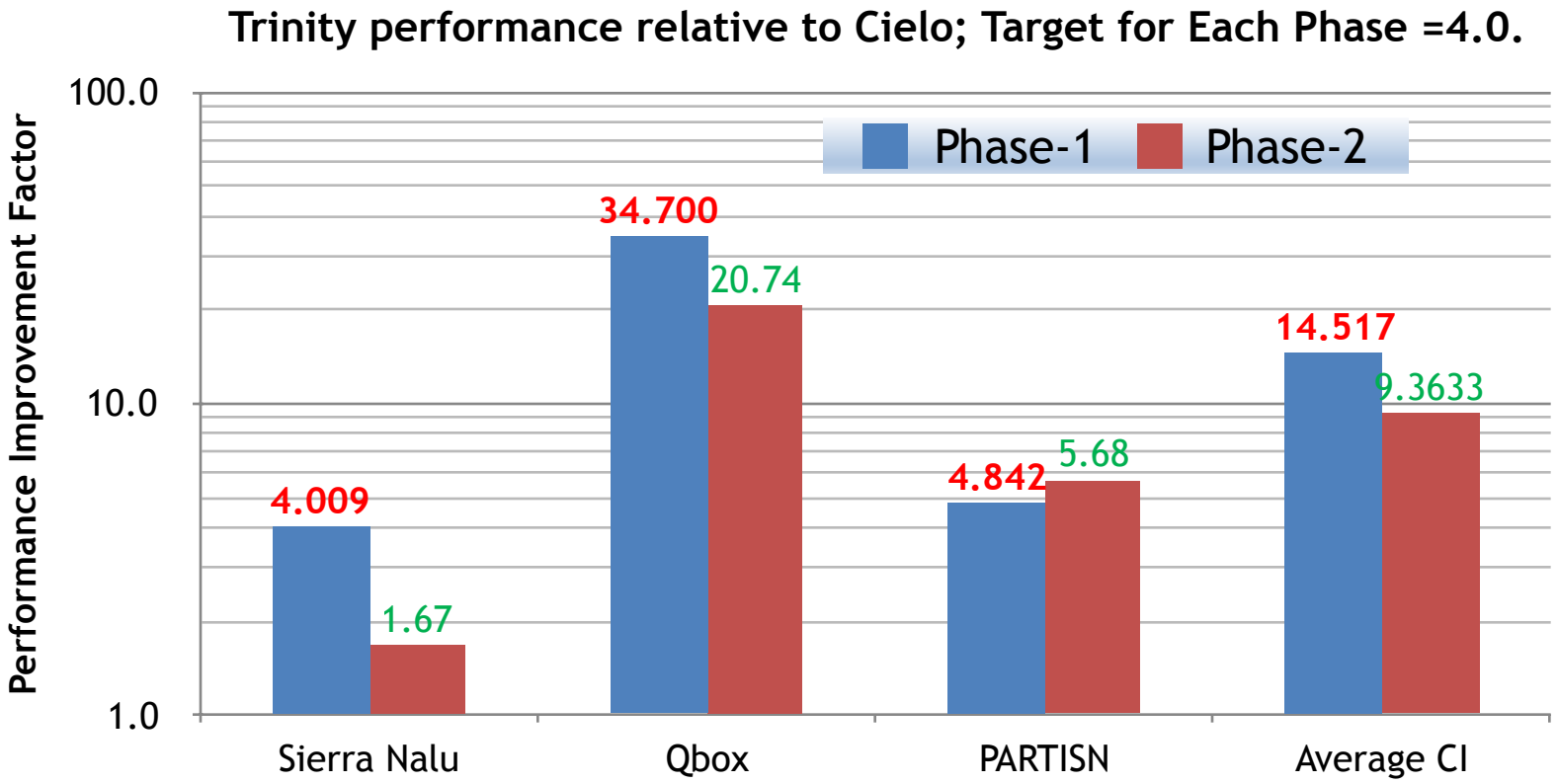
Trinity
- Cray XC40
- Nodes > 19,420
- Memory > 2PB
- Peak Performance > 40PF
- Intel Haswell (32 cores/node) & Knights Landing(68 cores/node)

Trinity Architecture: Phase-1 with Haswell Nodes accepted December 2015; Phase-2 KNL nodes accepted December2016



- Peak Haswell Node Performance: 32cores*16FLOPs/cycle*2.3GHz = 1,177.6 GFLOPS/node
- Peak KNL Node Performance: 68cores*32FLOPs/cycle*1.4GHz = 3, 046.4 GFLOPS/node
- Intel Turbo Mode Boost enabled
- Intel Hyper-Threads enabled

Capability Improvement Summary



With Qbox, KNL is overall about 1.75 times slower than Xeon, per node

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%)

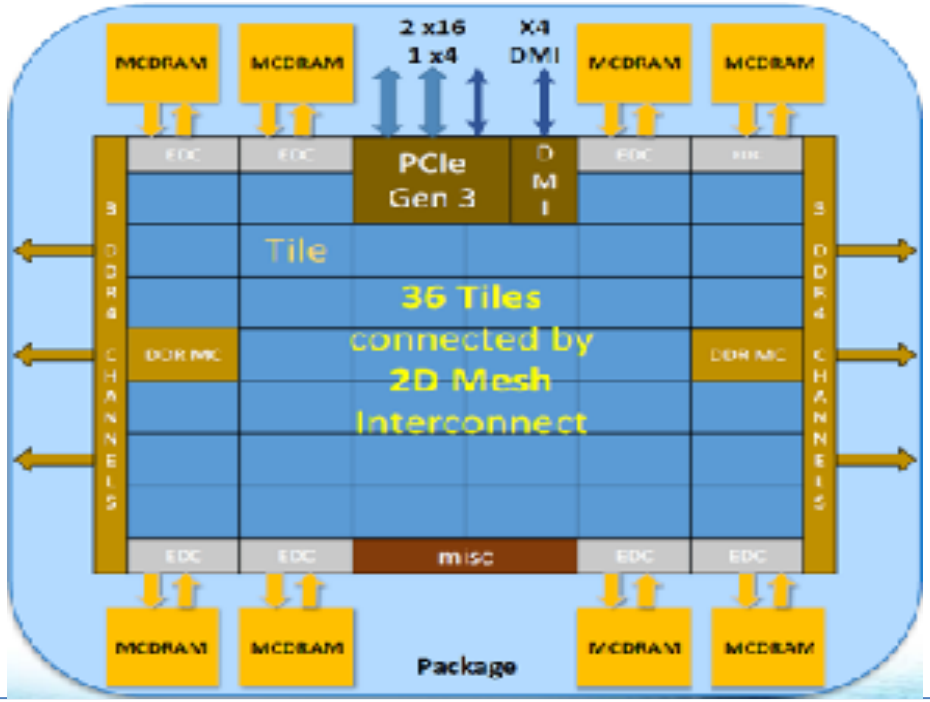
LLNL App: Qbox:

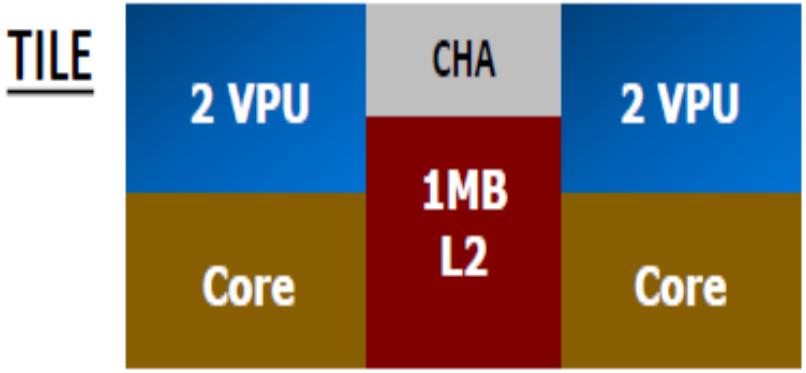
- first-principles molecular dynamics code used to compute the properties of materials at the atomistic scale
- Test Problem: benchmark problem is the initial self-consistent wave function convergence of a large crystalline gold system (FCC, a0 = 7.71 a.u).
- Figure of Merit: maximum total wall time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

LANL App: PARTISN:

- PARTISN particle transport code [6] 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)*

Trinity KNL Processor Architecture





- Three Cluster Modes: All-to-All, Quadrant, Sub-NUMA Clustering
- MCDRAM Modes
  - Flat: Exposed as a separate NUMA node
  - Cache: Direct mapped Cache, 64B lines
  - Hybrid: Part Cache, Part Memory; 25% or 50% cache

- Intel Turbo Mode Boost enabled
- 4 Threads per core. Simultaneous Multithreading (SMT)
- 2 VPU per core: 2x AVX 512 units; 32 SP/16DP per unit
- 2D-Mesh connections for tile
- DDR4: 6 channels @ 2.4 GHz

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

Trinity Phase-2 Acceptance Completed December 2016

1) Capability Improvement(CI) metric
   - ➢ 4X over a baseline performance measured on 2/3rd of the nodes on Cielo
   - ➢ runs at near full scale
   - ➢ May use appropriately scaled inputs
   - ➢ Three applications representative of Tri-lab productions apps ; **Nalu, PARTISN, QBOX**
2) NERSC's Sustained System Performance (**SSP**) target of 489; specified input: "large"
3) Microbenchmarks: 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)	Phase-1	Phase-2
Total Nodes	8,894	9,436	9,984
Total Cores	142,304	301,952	678,300
Processor	AMD MagnyCours	Intel Haswell	Intel Xeon Phi (KNL)
Processor ISA	SSE4a	AVX2	AVX-512
Clock Speed(GHz)	2.40	2.30	1.4
Cores/node	16	32	68
Memory-per-core(GB)	2	4	1.41 (DDR4) 0.235 (MCDRAM)
Memory	DDR3 1,333 MHz	DDR4 2,133 MHz	DDR4 2,400 MHz
Peak node GFLOPS	153.6	1,177.6	3,046.4
DDR Channels/socket	4	4	6
Cache L1(KB) L2(KB) L3(MB)	8 x 64 8 x 512 10	16 x 32 16 x 256 40	68x32 34 x 1,024 16 GB MCDRAM (if in Cache
Interconnect Topology	Gemini 3D Torus 18x12x24	Aries Dragonfly	

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**

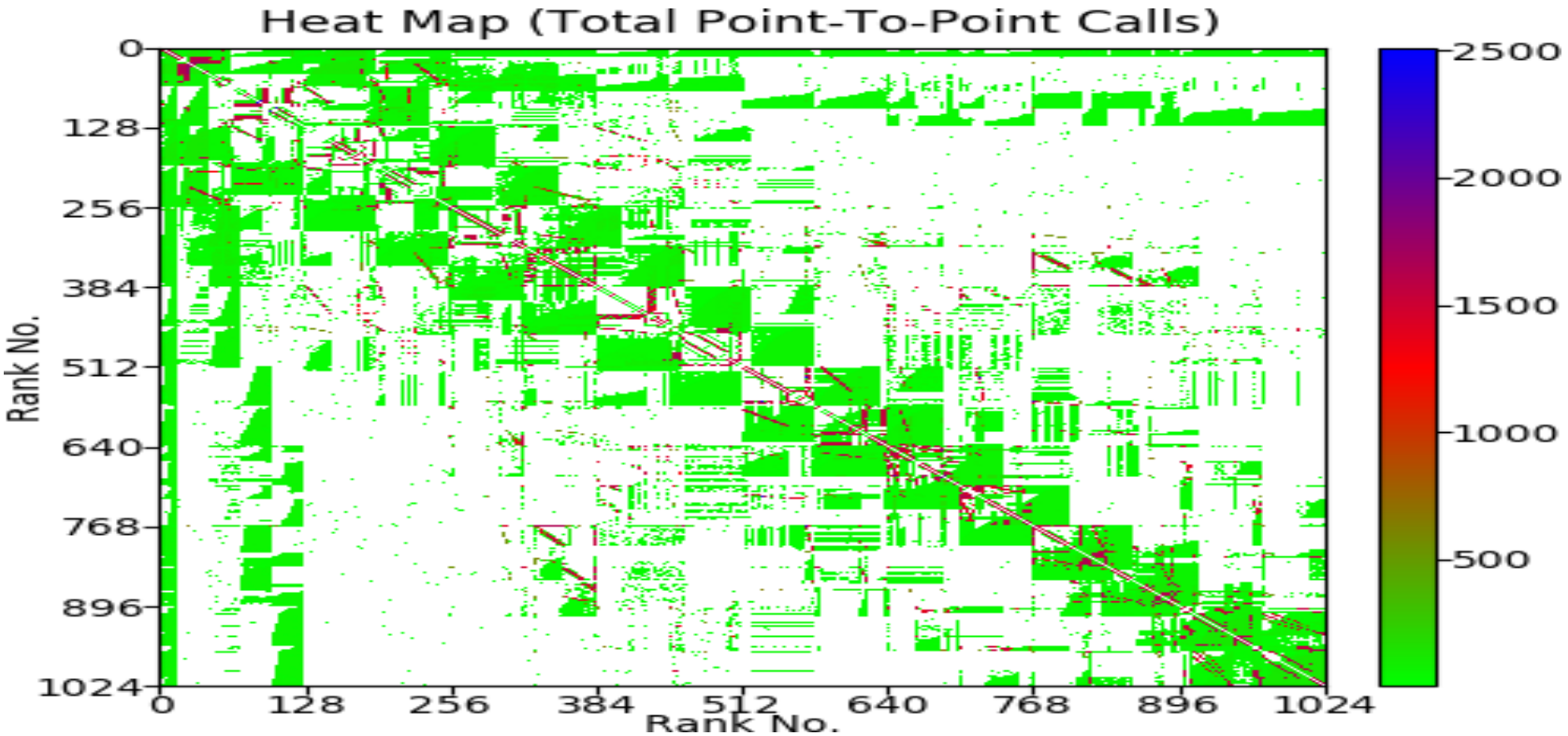
SIERRA/Nalu CI Performance

BASELINE on Cielo				Trinity Phase -2 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	4,096	262,144	*R6*: 9B elements	0.689

Running the same problem (9 Billion element mesh) on 2 times the number of PEs in quad-cache mode resulted in a Capability Improvement Metric of =1.15/0.689 = **1.67**

CI for Nalu fell short of the target 4.0 because out-of-memory (OOM) errors prevented runs at 8192 KNL nodes in quad-cache mode

KNL MPI memory usage Investigation; Nalu p2p communications



Heat map of total number of point-to-point (p2p) calls for "R3 Mesh" 1,024 MPI ranks
Rank 14 for the "R3 Mesh" and Rank 13 for the "R4 Mesh" are connected to 1,023 and 7,671 ranks, respectively (horizontal block of green at the top of the Heat Map). This heavy p2p connectivity on few MPI tasks was an initial "theory" for what could be causing OOM due to in MPI buffer growth.

SIERRA/Nalu Trinity Phase 2 acceptance test

- Root cause of Nalu Trinity Phase 2 acceptance test failing for 8192 KNLs (524,288 MPI processes) was due to memory explosion for >= 131072 MPI processes due to poor scaling of Cray MPI_Reduce_scatter
- Tiny driver code that performs Reduce_scatter and collects memory information
- "memory used" is the memory usage for the Reduce_scatter only

# KNLs	MPI	Mem used/ node(GB)	time(s)
1	64	0.01	0.004
2	128	0.07	0.02
16	1024	0.08	0.02
128	8192	0.15	0.08
256	16384	0.10	0.05
512	32768	0.14	0.07
1024	65536	0.18	0.15
2048	131072	22.3	465
4096	262144	45.0	1478
8192	524288	No time to measure	~5000

Half the node memory!

- For 262,144 MPI processes, the separate Reduce+Scatter calls took a total of ~0.3 seconds and max memory per KNL increased by ~0.5 GB (contrast that with 465 second and 22 GB/NKL respectively).

PARTISN CI Performance

BASELINE on Cielo				Trinity Phase -1 CI Results			
Nodes	MPI Tasks (4 Threads/task)	Problem Size Complexity Measure	RunTimeFOM	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**

PARTISN CI Performance

BASELINE on Cielo				Trinity Phase -2 CI Results			
Nodes	MPI Tasks	Problem Size Complexity Measure	RunTimeFOM	Nodes	MPI Tasks	Problem Size Complexity Measure	RunTime FOM
8,192	32,768 (4 OMP threads/ task)	*2,880 zones/core*	209.4 secs	8,192	262,144 (2 OMP threads/ task)	*6,480 zones/core*	332.047 secs

Running a (6480*262144*2)/(2880*32768*4) = 9 times larger problem on 4 times the number of PEs took 1.585 times longer *solver iteration time* leading to a Capability Improvement Metric of = 9/ 1.585 = **5.68**

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**

Qbox CI Performance

BASELINE on Cielo				Trinity Phase -2 CI Results			
Nodes	MPI Tasks 1 thread/task	Problem Size Complexity Measure	RunTime FOM	Nodes	MPI Tasks 4 threads/Task Hyperthreads	Problem Size Complexity Measure	RunTime FOM
6,144	98,304	*1,600 Atoms*	1663 secs	8,504	136,064	*6,000 Atoms*	4,227.45 secs

Running a (6000/1600) ** 3 = 52.73 time larger problem on 5.536 times the number of PEs took 2.54 times longer *self-consistent iteration time* leading to a Capability Improvement Metric of = 52.73 / 2.54 = **20.74**

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

Trinity Phase-1 SSP target was 400: Achieved 500

Baseline SSP performance on NERSC's Hopper, Cray XE6 (Top) and Trinity Phase I SSP performance (Bottom)

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

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

Trinity Phase-2 SSP target was 489: Achieved 581

Baseline SSP performance on NERSC's Hopper, Cray XE6 (Top) and Trinity Phase II SSP performance (Bottom)

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

Application Name	Nodes Used	NERSC Elapsed Time Proposed	Proposed Pi	Elapsed Time	December times	KNL Pi December
miniFE	3840	7.7	0.03602	7.22	7.20	0.0385
miniGhost	768	29.6	0.14737	33.86	33.95	0.1285
AMG	768	165	0.01077	140.70	160.36	0.0111
UMT	769	552	0.04337	449.94	467.26	0.0512
SNAP	768	216	0.02851	216.07	212.00	0.0290
miniDFT	47	1020	0.19149	478.87	471.62	0.4142
GTC	150	2396	0.05540	2195.05	2118.73	0.0627
MILC	384	882	0.04440	612.93	631.27	0.0520
		Geom. Mean=	0.04901	Geom. Mean=		0.0582
				Trinity SSP =		580.92
				Target Trinity SSP =		489

Conclusions-I

Several months of effort by the Cray and Tri-Lab teams resulted in exceeding performance acceptance requirements

Based on benchmark results we anticipate production Trinity apps will see a gain of 2x-6x over Cielo

Benefit of hybrid MPI + Threads clearly seen with Qbox

use of grid_order resulted in good performance gains for CI and many SSP apps

Conclusions-II

KNL's quad/cache mode is a good-performing, general purpose mode for applications which have not yet directly mapped selected data structures to MCDRAM.

- *highly susceptible to thrashing if important data aliases to the same location*
- *This becomes more and more likely as node counts increase (seen with SSP apps: GTC, miniFE)*

Both Intel and Cray compilers were used depending upon the application.

Static linking, more often than not, achieves better performance than dynamic linking.

Huge pages typically provides a performance increase and should be investigated for each application.

Grid ordering improves performance for many applications and should also be investigated for each application.

Hybrid parallelism can improve performance over MPI everywhere on KNL, however Cray's MPICH implementation on Trinity scales remarkably well and can be used in the interim while developing hybrid parallelism within an application.

It is important to closely monitor application communication patterns at larger scales.