BLUE WATERS

SUSTAINED PETASCALE COMPUTING

May 10, 2017

Updating the SPP Benchmark Suite for Extreme-Scale Systems

Gregory Bauer, Victor Anisimov, Galen Arnold, Brett Bode, Robert Brunner, Tom Cortese, Roland Haas, Andriy Kot, William Kramer, JaeHyuk Kwack, Jing Li, Celso Mendes, Ryan Mokos, Craig Steffen

NCSA

NCSA NSF GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION CRAY

Outline

I. Introduction
II. Background
III. New SPP Benchmark Suite
IV. Website for SPP Distribution
V. Conclusion

Introduction

- Blue Waters System
  - Funded by NSF, installed at NCSA in Fall’2012
  - One of the major open-science systems world-wide
  - Largest machine produced by Cray
- Balanced Architecture
  - 13+ PF peak performance, 1+ PF sustained
  - 1.5 PB memory – 64 GB per node
  - 26 PB of useable disk space, 1.0+ TB/s write rate
  - 4 tape libraries (HPSS), 200+ PB of useable space

➢ Can serve well a wide range of science applications!

Introduction (cont.)

- Blue Waters Acceptance Tests
  - Traditional benchmarks (HPCChallenge, etc)
  - Sustained petascale problems on full system
  - SPP suite of applications, on CPU and GPU
    - Based on expected NSF workload for Track-1
  - Many other functionality tests
  - Some tests continue to run periodically, via Jenkins
- Acceptance Test Results
  - Mendes, C. et al - *Deployment and testing of the sustained petascale Blue Waters system*, *Journal of Computational Science*, v. 10, p. 327-337, Sep. 2015. DOI: 10.1016/j.jocs.2015.03.007

Outline

I. Introduction
II. Background
III. New SPP Benchmark Suite
IV. Website for SPP Distribution
V. Conclusion

Background

- Traditional Benchmarks
  - Good for quick assessment of systems
  - Typically focus on one aspect of the systems
    - e.g. Linpack, Stream, IOR, etc
  - Different benchmarks rank systems differently
  - No single benchmark is universally “perfect”
    - Top500 → Green500, Graph500
    - Linpack → HPCG, etc
  - Merging of HPC+BigData can exacerbate this!

Background (cont.)

- Sustained Petascale Performance (SPP) Benchmarks
  - Based on *Sustained System Performance* method: SSP (Berkeley, 2008)
  - Focuses on time-to-solution of real programs
  - Represents the behavior of full applications, including pre/post-processing, I/O, checkpointing, etc
  - Allows combining results into a single index reflecting how much real work a system can produce
    - Different contributions from distinct system partitions in heterogeneous systems

Background (cont.)

- Original SPP Benchmarks (2012):
  - NAMD*: molecular dynamics
  - MILC, CHROMA*: particle physics
  - VPIC, SPECFEM3D: geophysics
  - WRF: weather forecast
  - PPM: astrophysics
  - NWCHEM, GAMESS*: quantum chemistry
  - QMCPACK*: materials science

  * CPU- and GPU-based versions employed

  Code selection based on expected system workload

Background (cont.)

- Observed SPP values on XE nodes (2012):

Application	Flop Count	Number of Nodes	Time (s)	Node Rate (GF/s)	BW Rate (PF/s)
VPIC	1.83×10^{18}	4,608	5,811.0	68.26	1.65
QMCPACK	4.71×10^{17}	4,800	1,852.0	52.98	1.28
NAMD	8.29×10^{17}	5,000	5,432.4	30.50	0.74
WRF	1.05×10^{18}	4,560	8,931.0	25.81	0.62
MILC	4.73×10^{17}	4,116	5,099.5	22.52	0.54
PPM	2.57×10^{19}	8,256	46,848.0	66.45	1.61
SPECFEM3D	6.30×10^{18}	5,419	16,918.4	68.70	1.66
NWCHEM	5.95×10^{18}	5,000	24,852.2	47.87	1.16

Background (cont.)

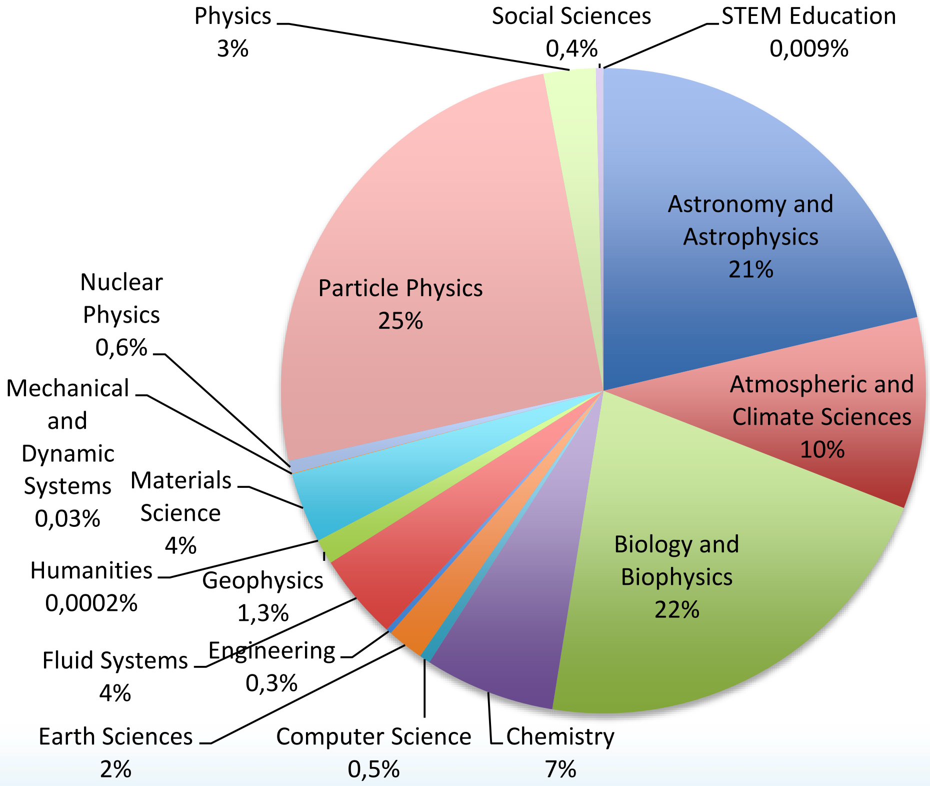
- Observed SPP values on XE nodes (2012):

Application	Flop Count	Number of Nodes	Time (s)	Node Rate (GF/s)	BW Rate (PF/s)
VPIC	1.83×10^{18}	4,608	5,811.0	68.26	1.65
QMCPACK	4.71×10^{17}	4,800	1,852.0	52.98	1.28
NAMD	8.29×10^{17}	5,000	5,432.4	30.50	0.74
WRF	1.05×10^{18}	4,560	8,931.0	25.81	0.62
MILC	4.73×10^{17}	4,116	5,099.5	22.52	0.54
PPM	2.57×10^{19}	8,256	46,848.0	66.45	1.61
SPECFEM3D	6.30×10^{18}	5,419	16,918.4	68.70	1.66
NWCHEM	5.95×10^{18}	5,000	24,852.2	47.87	1.16

Geometric Mean: 1.063 PF/s = SPP_{CPU}

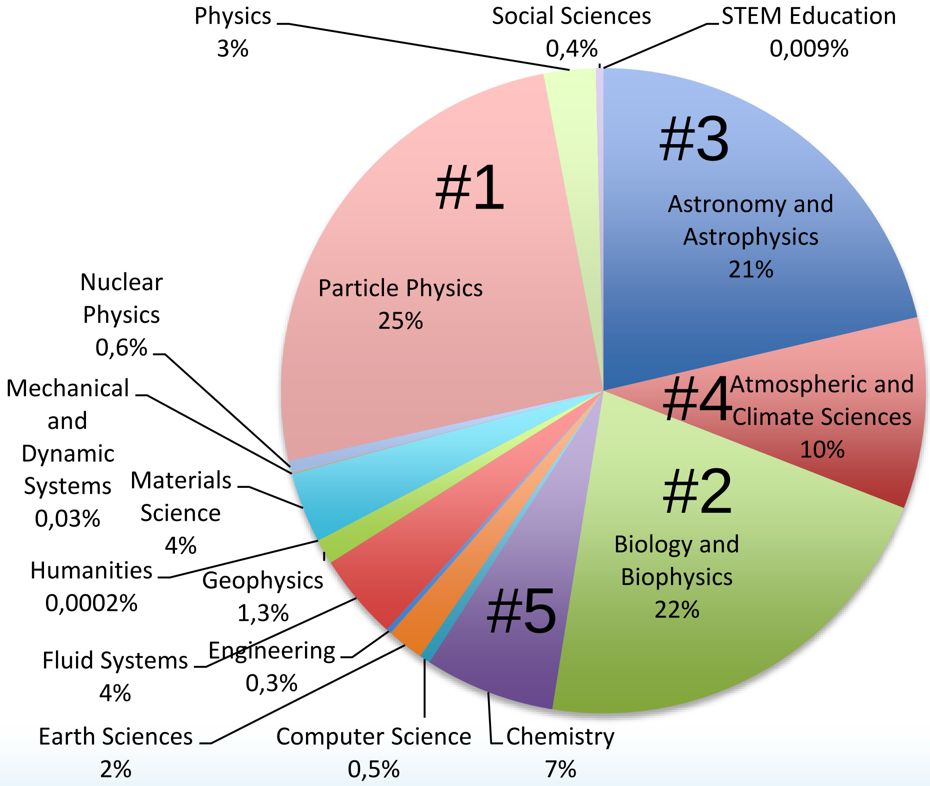
Background (cont.)

- Observed Blue Waters Workload – Jul’2013/Aug’2014:



Background (cont.)

- Observed Blue Waters Workload – Jul’2013/Aug’2014:



#1,2,3,4,5: >85%

Outline

I. Introduction
II. Background
III. New SPP Benchmark Suite
IV. Website for SPP Distribution
V. Conclusion

New SPP Benchmark Suite

- Motivations for SPP Suite Update
  - Track most used science applications
  - Keep suite relevant for assessment of new systems
- Guidelines for code selection
  - Provided by Blue Waters workload study
    - Conducted by SUNY-Buffalo, sponsored by NSF
    - Analyzed 3.5 years of Blue Waters jobs
    - Report available at Blue Waters Portal
- New selection of SPP codes
  - Codes expected to run on Blue Waters and on other systems

New SPP Benchmark Suite

- Selected Codes (CPU versions only, for now):
  - AWP-ODC: geophysics
  - CACTUS: astrophysics
  - MILC: quantum chromodynamics
  - NAMD: molecular dynamics
  - NWCHEM: computational chemistry
  - PPM: astrophysics
  - PSDNS: turbulence
  - QMCPACK: materials science
  - RMG: electronic structure of materials
  - VPIC: movement of charged particles
  - WRF: weather forecast

New SPP Benchmark Suite

- Characteristics of the codes:

Application	Language	Parallelization
AWP-ODC	Fortran, C++	MPI
CACTUS	Fortran, C, C++	MPI+OpenMP
MILC	C, C++	MPI+OpenMP
NAMD	C++	Charm++
NWCHEM	Fortran, C, C++	Global Arrays
PPM	Fortran	MPI+OpenMP
PSDNS	Fortran	MPI+OpenMP+CAF
QMCPACK	C, C++	MPI+OpenMP
RMG	Fortran, C, C++	MPI+pthreads
VPIC	C++	MPI+OpenMP
WRF	Fortran	MPI+OpenMP

Programming Language - C++: 8/11; Fortran: 7/11; C: 5/11

Parallelization Paradigm - MPI: 9/11; OpenMP: 8/11

New SPP Benchmark Suite

- Two problem sizes defined:
  - Compact problem:
    - Intended to allow quick build/run process
    - Can be executed on up to a few hundred nodes
  - Large problem:
    - Reflects production runs of each code
    - Can be executed on thousands of nodes
    - May have significant resource demands (e.g. I/O)

New SPP Benchmark Suite

- Problem sizes:

Application	Compact Input	Large Input
AWP-ODC	128^3 mesh	5600x2800x1024 mesh
CACTUS	$6.7\text{x}10^7$ grid points	$2.8\text{x}10^9$ grid points
MILC	36x36x36x72 lattice	72x72x72x144 lattice
NAMD	2 ps simulation of 100 M atoms	20 ps simulation of 100 M atoms
NWCHEM	32 atoms, 4 ccsd iterations	32 atoms, 20 ccsd iterations
PPM	$1{,}280^3$ zone mesh	$5{,}120^3$ zone mesh
PSDNS	$2{,}048^3$ grid points	$8{,}192^3$ grid points
QMCPACK	$5.12\text{x}10^4$ Monte Carlo samples	$2.56\text{x}10^6$ Monte Carlo samples
RMG	302 water molecules	4,096 atoms
VPIC	$1{,}200^3$ grid, $8.64\text{x}10^{10}$ particles	$1{,}536^3$ grid, $1.16\text{x}10^{12}$ particles
WRF	9120x9216x48 grid, 900 T-steps	9120x9216x48 grid, 9,000 T-steps

New SPP Benchmark Suite

- Recent executions on Blue Waters:
  - Goal: illustrate code utilization on a real system
  - Both compact and large cases were tested
  - Arbitrary selection of compilers
  - Codes without Blue Waters-specific optimizations
    - Should hopefully run on other systems too
  - Optimization efforts ongoing in some cases
    - Done jointly with science teams

New SPP Benchmark Suite

- Recent executions on Blue Waters (large input):

Application	Number of Nodes	Time (s)
AWP-ODC	2,048	855
CACTUS	4,096	4,800
MILC	1,296	7,916
NAMD	4,500	242.2
NWCHEM	5,000	22,201
PPM	8,448	7,790
PSDNS	8,192	1,538
QMCPACK	5,000	1,765
RMG	3,456	7,310
VPIC	4,608	4,218
WRF	4,560	10,260

Times:

Min. ≈ 4 minutes

Max. ≈ 6 hours

Number of Nodes:

Mean = 4,655

≈ 49 Blue Waters cabinets

≈ 20% of XE cabinets

New SPP Benchmark Suite

- Potential for improvements in SPP codes:

  CACTUS:

  - I/O parts: parallel I/O
  - Communication intensive: topology-awareness

  MILC:

  - Numerical parts: fused multiply-add (FMA) instructions
  - Non-uniform domain decompositions (but watching for numerical instabilities)
  - Rank-reordering for improved communication

New SPP Benchmark Suite

- Potential for improvements in SPP codes (cont.):

NAMD:

- GPU offloading (ORNL’s Summit)
- AVX-512 vectorization on KNL (ANL’s Aurora)
- Specialized Charm++ communication layers

NWCHEM:

- Network-specialized layer (for GlobalArrays)
- Vectorization via SIMD support in numerical parts
- Other algorithmic optimizations in CCSD part

New SPP Benchmark Suite

- Potential for improvements in SPP codes (cont.):

PPM:

- Optimized placement of team members/servers, to avoid network congestions

  e.g. via node selection and/or rank reordering

- Vectorization by compiler in numerical/math functions

QMCPACK:

- Use of fused multiply-add (FMA) instructions
- Adoption of mixed-precision in some numerical parts

New SPP Benchmark Suite

- Potential for improvements in SPP codes (cont.):

VPIC:

- Compiler-generated vectorization, FMA instructions
- Dynamic load balance for particles (monitor mem. use)
- Use of parallel I/O techniques for output of data

WRF:

- Vectorization for processing of vertical columns
- Optimized placement for near-neighbor communication
- Input data with multiple ranks/files

Outline

I. Introduction
II. Background
III. New SPP Benchmark Suite
IV. Website for SPP Distribution
V. Conclusion

Website for SPP Distribution

- Motivation
  - Promote SPP usage widely
  - Share experiences with the community
- Initial site implementation
  - *https://bluewaters.ncsa.illinois.edu/benchmarks*
  - Sources, build/run scripts for Blue Waters, inputs
    - Parts of input data require GlobusOnline
  - Links to SPP and other regular benchmarks

Outline

I. Introduction

II. Background

III. New SPP Benchmark Suite

IV. Website for SPP Distribution

V. Conclusion

Conclusion

- Blue Waters System
  - Tremendous asset for open-science community
  - Blue Waters workload enables many scientific discoveries
- Updated SPP Benchmark Suite
  - 11 apps, representing codes using Blue Waters today
  - Provides time-to-solution measurements
  - Openly available to the community
  - Potential for performance improvements typically includes:
    - Vectorization, FMA instructions
    - Task placement, rank-reordering – for better communication

Acknowledgments

- Funding: NSF OCI-0725070/ACI-1238993, state of Illinois
- Personnel: **NCSA** Blue Waters team, Cray site team