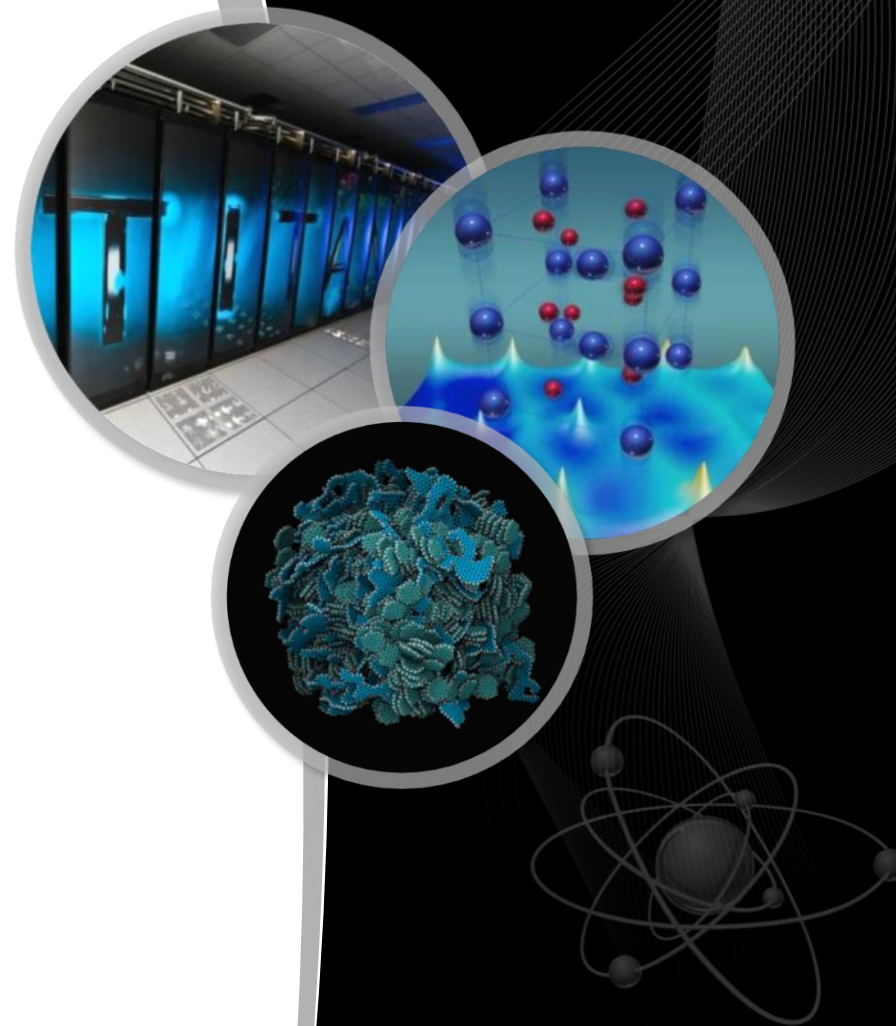


Early Application Results on Summit

T. P. Straatsma

**National Center for Computational Sciences
Oak Ridge National Laboratory**



OLCF Program to Ready Application Developers and Users

- We are preparing users through:
 - Application readiness through Center for Accelerated Application Readiness (CAAR)
 - Early science projects for CAAR and selected other ESP
 - Training and web-based documentation
 - Access on Titan, SummitDev, Summit Phase I, Summit
 - Access for broader user base on final system after acceptance
- Goals:
 - Early science achievements
 - Transferable experience
 - Demonstrate application readiness
 - Prepare INCITE & ALCC proposals
 - Harden Summit for full-user operations

CAAR Projects



ACME/E3SM: Climate Research: Advancing Earth System Models
PI: Dr. David Bader, Lawrence Livermore National Laboratory
Science Domain: Climate Science
CAAR Liaison: Dr. Matt Norman
CSEEN Postdoc: Dr. Anikesh Pal
NESAP



DIRAC: CAAR Oak Ridge Proposal for getting the Relativistic Quantum Chemistry Program Package DIRAC ready for SUMMIT
PI: Prof. Dr. Lucas Visscher, Free University Amsterdam, the Netherlands
Science Domain: Relativistic Quantum Chemistry
CAAR Liaisons: Dr. Dmitry Liakh, Dr. Tjerk Straatsma
CSEEN Postdoc: TBD (backfill Dr. Amelia Fitzsimmons)



FLASH: Using FLASH for Astrophysics Simulations at an Unprecedented Scale
PI: Dr. Bronson Messer, Oak Ridge National Laboratory
Science Domain: Astrophysics
CAAR Liaisons: Dr. Bronson Messer
CSEEN Postdoc: Dr. Austin Harris (backfill Dr. Thom Papatheodore)



GTC: Particle Turbulence Simulations for Sustainable Fusion Reactions in ITER
PI: Prof. Dr. Zhihong Lin, University of California - Irvine
Science Domain: Plasma Physics
CAAR Liaison: Dr. Wayne Joubert
NESAP



HACC: Cosmological Simulations for Large-scale Sky Surveys
PI: Dr. Salman Habib, Argonne National Laboratory
Science Domain: Cosmology
CAAR Liaison: Dr. Bronson Messer
NESAP, ESP



LS-DALTON: Large-scale Coupled-cluster Calculations of Supramolecular Wires
PI: Prof. Dr. Poul Jørgensen, Aarhus University
Science Domain: Quantum Chemistry
CAAR Liaison: Dr. Dmitry Liakh
INCITE

NAMD: Molecular Machinery of the Brain
PI: Dr. James Phillips, University of Illinois at Urbana-Champaign
Science Domain: Biophysics
CAAR Liaison: Dr. Tjerk Straatsma
NESAP

NUCCOR: Nuclear Structure and Nuclear Reactions
PI: Dr. Gaute Hagen, Oak Ridge National Laboratory
Science Domain: Nuclear Physics
CAAR Liaison: Dr. Gustav Jansen
CSEEN Postdoc: TBD (backfill Dr. Micah Schuster)

NWCHEM: Developing Coupled Cluster Methodologies for GPUs
PI: Dr. Karol Kowalski, Pacific Northwest National Laboratory
Science Domain: Computational Chemistry
CAAR Liaison: Dr. Dmitry Liakh
IBM Postdoc: Dr. David Appelhans
NESAP

QMCPACK: Materials Science Research for High-Temperature Superconductors
PI: Dr. Paul R. C. Kent, Oak Ridge National Laboratory
Science Domain: Materials Science
CAAR Liaison: Dr. Ying Wai Li
CSEEN Postdoc: Dr. Andreas Tillack

RAPTOR: Fluid Dynamics Research to Accelerate Combustion Science
PI: Dr. Joseph Oefelein, Sandia National Laboratory, Livermore
Science Domain: Engineering/Combustion
CAAR Liaison: Dr. Ramanan Sankaran
CSEEN Postdoc: TBD (backfill Dr. Kalyana Gottiparthi)

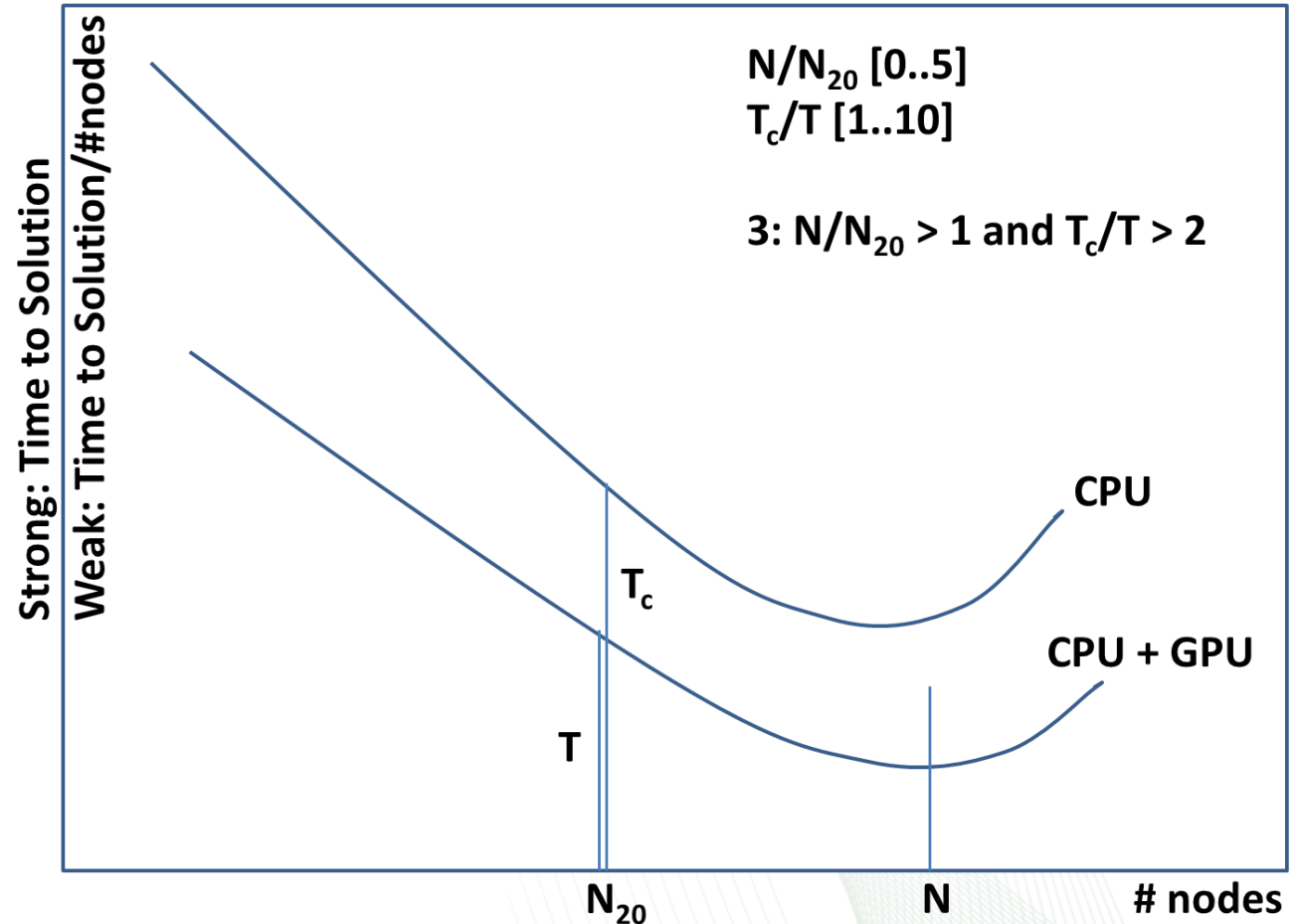
SPECFEM: Mapping the Earth's Interior Using Big Data
PI: Dr. Jeroen Tromp, Princeton University, Princeton University
Science Domain: Seismology
CAAR Liaison: Dr. Judy Hill
CSEEN Postdoc: TBD (backfill Dr. Yangkang Chen)

XGC: Multiphysics Magnetic Fusion Reactor Simulator, from Hot Core to Cold Wall
PI: Dr. CS Chang, Princeton Plasma Physics Laboratory, Princeton University
Science Domain: Plasma Physics
CAAR Liaison: Dr. Ed D'Azevedo
CSEEN Postdoc: TBD (backfill Dr. Stephen Abbott)
NESAP



Center for Accelerated Application Readiness – Success Metrics

- **Scalability:** Applications should demonstrate reduced time to solution (for strong scaling benchmarks) or time to solution divided by the number of nodes used (for weak scaling benchmarks) to 20% or more of the full Summit machine, N_{20} . This is also known as the capability metric.
- **Accelerated Performance:** Applications should demonstrate a performance improvement of a factor of two or better by using all six GPUs compared to using both CPUs only, with a job that runs on 20% of the full Summit machine.



GTC

Prof. Dr. Zhihong Lin, University of California, Irvine
Prof. Dr. William Tang, Princeton University
Dr. Ihor Holod, University of California, Irvine
Dr. Animesh Kuley, University of California, Irvine
De. Bei Wang , Princeton University

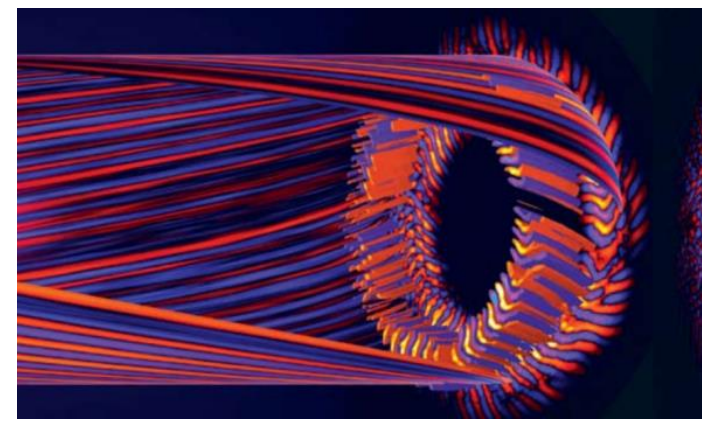
Domain Area: Plasma Physics / Fusion



Zhihong Lin



Wayne Joubert



Plasma simulations supporting the ITER project are a key DOE/FES focus and are required to understand the complex kinetic dynamics governing magnetic confinement properties of fusion-grade plasmas.

The Gyrokinetic Toroidal Code (GTC) is a massively parallel particle-in-cell code for first-principles, integrated simulations of burning plasma experiments such as the International Thermonuclear Experimental Reactor (ITER), the crucial next step in the quest for the fusion energy. GTC solves the five-dimensional (5D) gyrokinetic equation in full, global torus geometry to address kinetic turbulence issues in magnetically-confined fusion tokamaks.

The proposed CAAR project will develop the kinetic capability for first-principles-based direct numerical simulations of key instabilities that limit the burning plasma performance and threaten device integrity in magnetically-confined fusion systems. Of critical mission importance for Fusion Energy Research & Development is the timely achievement of the capability to understand, predict, control, and mitigate performance-limiting and integrity-threatening instabilities in burning plasmas.

Targeted for CAAR:

The GTC particle-in-cell (PIC) algorithm is the most computationally dominant component of the GTC code. As a sequel to previous work, a large part of the project's performance optimization work will thus focus on efficient multithreading of this computation for Summit. The particle PUSH and SHIFT operations are the two most dominant operations of the PIC computation. These two operations will be targeted for acceleration on Summit.

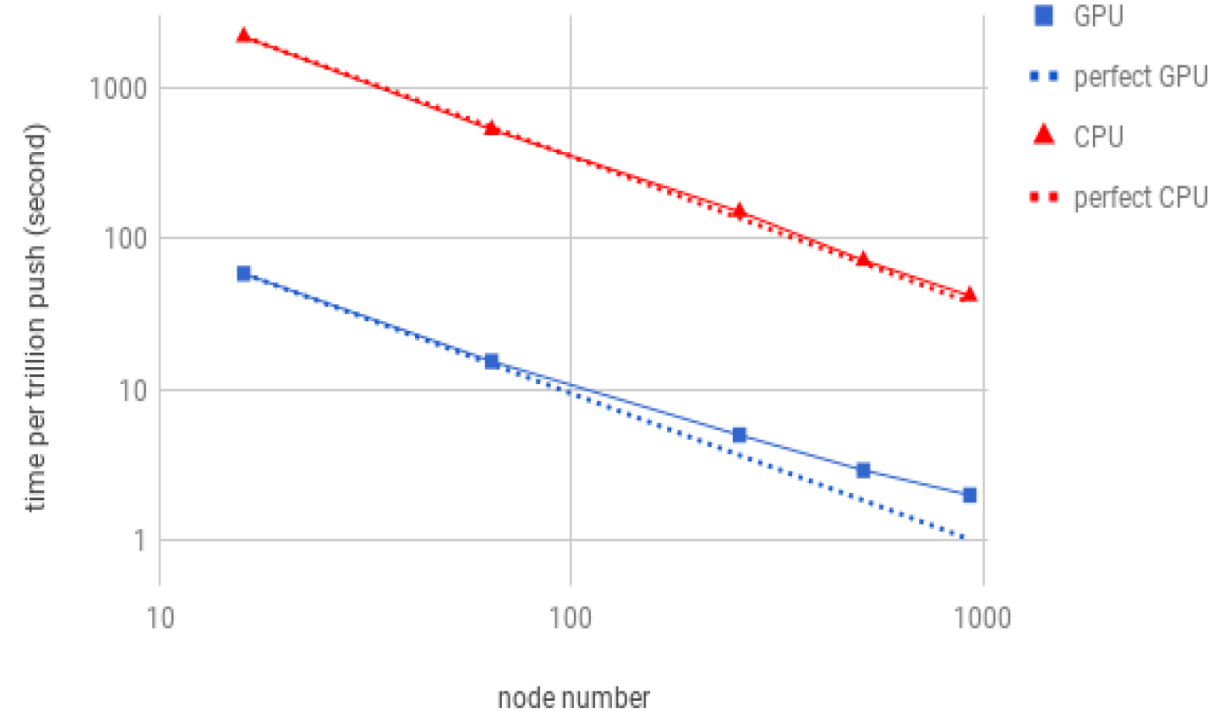
http://phoenix.ps.uci.edu/gtc_group/

GTC Application Readiness Status

- GTC is a gyrokinetic toroidal fusion code for modeling fusion reactors
- Much of the code already used GPUs effectively on Titan, e.g., particle push, using OpenACC
- Additional development work was done to optimize for Summit
- Code uses the NVIDIA AmgX solver to improve performance over the previous PETSc field solver

Early Results on Summit

Wall-clock time for one trillion particle pushes in the GTC weak scaling test on Summit

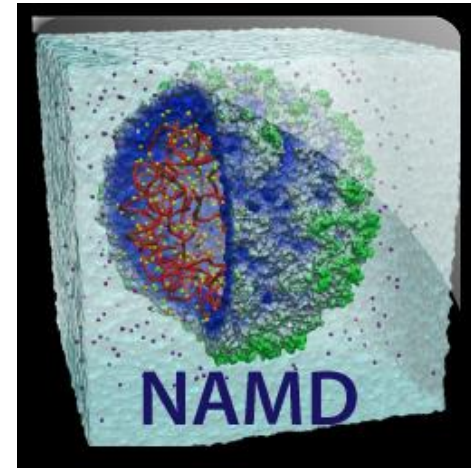


NAMD

Dr. James Phillips, University of Illinois at Urbana-Champaign (UIUC)
Dr. Laxmikant Kalé (CoPI), Professor of Computer Science, UIUC
Eric Bohm, Kirby Vandivort, and John Stone - Senior Research Programmers, UIUC
David Hardy, Research Programmer, UIUC
Ronak Buch, Bilge Acun, and Michael Robson, Research Assistants, UIUC



Jim Phillips



Domain Area: Computational Biophysics

DOE/BER funded programs are using molecular dynamics simulations to acquire a molecular level understanding of the relationships between biomass structure and recalcitrance.

NAMD is a high performance molecular dynamics code that is capable of massively parallel simulations, and it will be used to study the molecular level neural mechanisms of cellular and neural signaling. The proposed research requires the computational power of Summit and significant modifications of NAMD will be required in order to exploit Summit's computational capabilities. NAMD has a large worldwide user base and these modifications will also greatly benefit the larger computational biophysical community.

The BRAIN (Brain Research through Advancing Innovative Neurotechnologies) initiative is one of the Administration's "Grand Challenges". This research will elucidate molecular details of neural dynamics, synapse dynamics, and neural to synapse dynamics.

Targeted for CAAR:

1. Replace Particle Mesh Ewald (PME) with the Multilevel Summation Method (MSM) for solving the long ranged electrostatics. PME is the primary performance bottleneck in large molecular biomolecular simulations. MSM provides better performance, and flexibility.
2. Convert NAMD from a thread-centric to process centric design. This enables (i) better aggregation of work in order to effectively utilize multiple GPU's and (ii) reductions in latency and maximizing throughput.
3. Modify NAMD data structures to allow as much reasonable code re-use between Xeon Phi, CPU's, and GPU's.
4. Explore writing OpenMP 4.0 SIMD kernels for addressing cross-platform vector instructions.

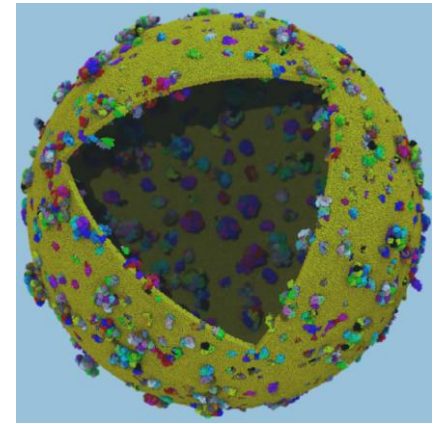
<http://www.ks.uiuc.edu/Research/namd/>

NAMD Application Readiness Status

1. Major challenge is time to solution on small problems (strong scaling)
2. Multiple levels of parallelism:
 - Within GPU work is parallelized to thousands of threads and to overlapping GPU kernels
 - Within node, work is split to among GPUs and CPUs
 - System is divided spatially into groups of atoms that are distributed to the nodes
3. Atoms are split on to nodes and only atom coordinates and forces that are needed are communicated from other nodes
4. C++ using Charm++ parallel library for thread and node parallelism. CUDA C is used for implementing the GPU kernels
5. Current GPU implementation does not prevent other hardware implementations
6. Direct GPU-GPU communication (both NVLINK and over IB)

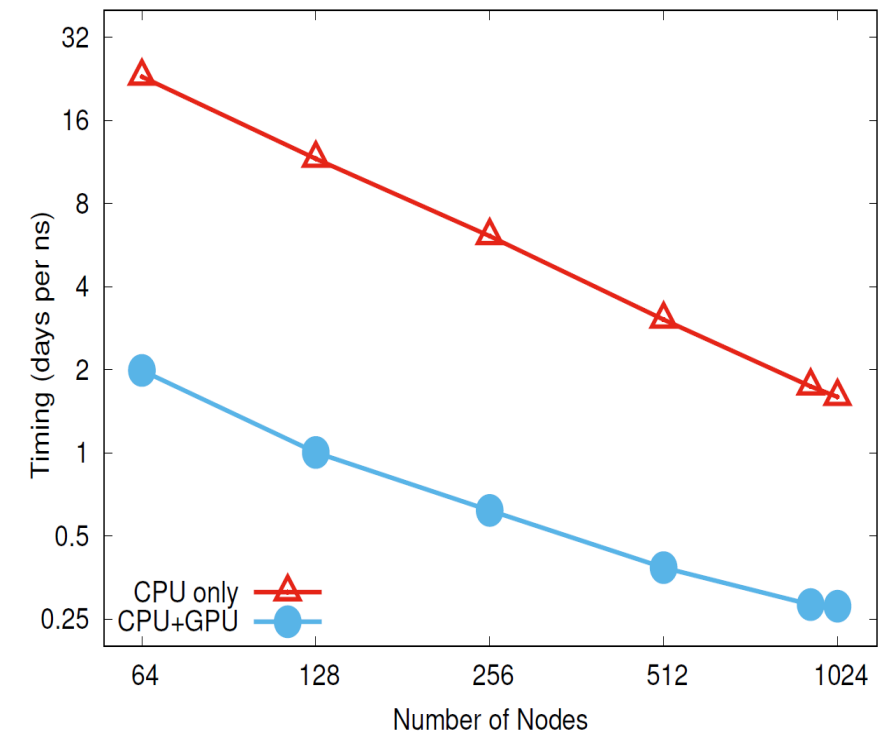
CAAR Accomplishments

1. New non-bonded force CUDA kernels
2. New PME reciprocal force CUDA kernels
3. Explicit solvent simulations are 1.3 – 2.6 times faster on GPUs
4. Implicit solvent (Generalized Born) simulations are 1.8 – 5.5 times faster
5. Faster simulations on systems with multiple GPUs per node



Early Results on Summit

NAMD strong scaling performance for a one billion atom HIV capsid proto-cell simulation on Summit.



NUCCOR

Dr. Gaute Hagen, Oak Ridge National Laboratory, Oak Ridge, TN
Prof. Dr. Thomas Papenbrock and Dr. Gustav Jansen, University of Tennessee, Knoxville, TN



Gaute Hagen Gustav Jansen



Nuclear Landscape mapped

Nuclear density functional theory, coupled with start-of-the-art computational tools, was used to estimate the borders of the nuclear landscape & explore its properties

Domain Area: Nuclear Physics

Nuclear physics theory and computation is central to the DOE/NP mission of improving our understanding of the building blocks of matter, discovering the origins of nuclei, and identifying the forces that transform matter.

NUCCOR is an application for the computation of the structure and reactions of atomic nuclei, implementing a set of algorithms that solve the quantum mechanical nuclear many-body problem using state-of-the-art nuclear interactions and currents. These include Hartree Fock, Coupled Cluster, and Equation of Motion methods.

An highly optimized NUCCOR through CAAR will impact the field of low-energy nuclear physics through enabling benchmarks and quality standards for neutrinoless double-beta decay, nuclear structure calculations of experimentally relevant nuclei for guiding, interpreting and predicting experimental research, and enabling nuclear structure and reaction of nuclei and their behaviors with previously unattainable detail.

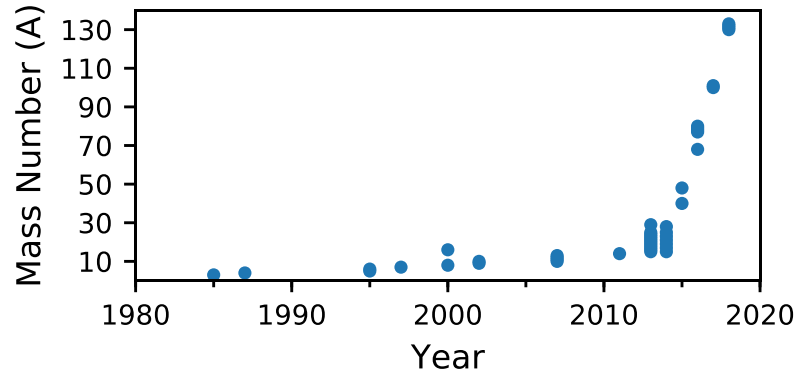
Work targeted for CAAR:

1. Optimization of reordering and expansion of packed tensors
2. Use task parallelism in mapping matrix multiplications onto accelerators using BLAS
3. Packing and aggregation of resulting tensor components

<https://www.olcf.ornl.gov/caar/nuccor/>

NUCCOR Application Readiness Status

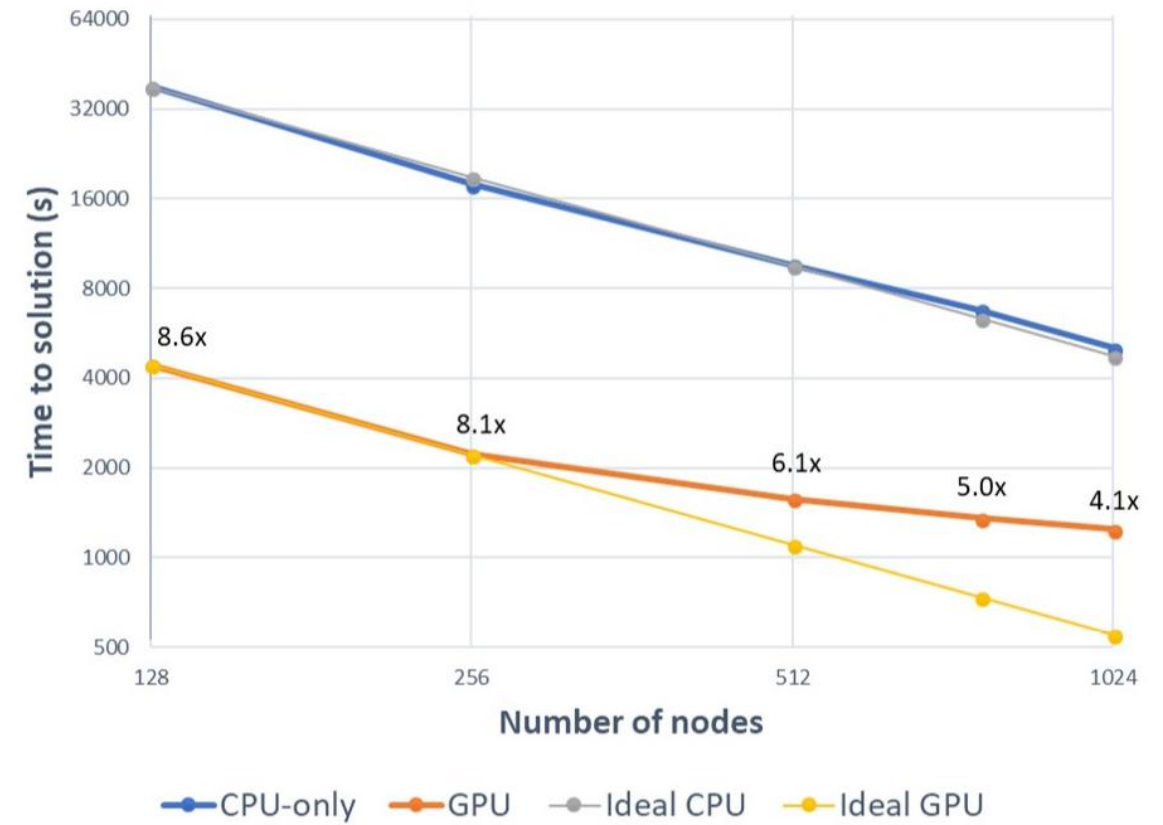
1. It is now possible to describe larger nuclei to a higher degree of precision from first principles



2. The spherical tensor contractions at the heart of NUCCOR has been pushed to external libraries to allow multiple versions for different programming models
 - MPI + OpenMP for CPU-only usage
 - MPI + OpenMP + CUDA for GPU accelerated kernels
3. Object-oriented design to allow selecting of computational kernels at run-time for easy testing and instrumentation, while retaining the possibility of selecting kernels for optimal performance
4. Testing framework to allow unit, integration and regression testing currently covering over 70% of the new libraries

Early Results on Summit

Strong scaling of a full application run to compute a state in an $A = 48$ system, like ^{48}Ca or ^{48}Ti needed in this project, using up to 1024 nodes on Summit for both the CPU-only version and the GPU version.



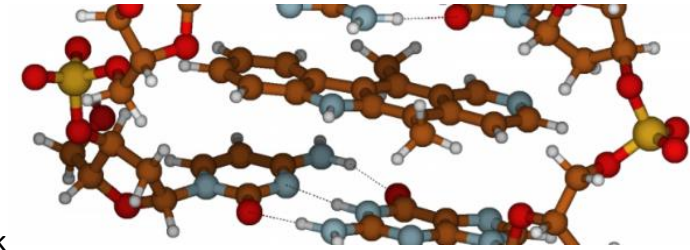
QMCPACK

Dr. Paul R. C. Kent, Oak Ridge National Laboratory, Oak Ridge, TN
Dr. F. A. Reboredo and Dr. J. T. Krogel, ORNL
Dr. Ying Wai Li and Dr. Andreas Tillack, ORNL

Domain Area: Materials Science



Paul Kent Ying Wai Li Andreas Tillack



Materials science research is a DOE/BES Grand Challenge area that aligns with the White House “Materials Genome Initiative” for accelerating new materials prediction, discovery and design.

QMCPACK is an open source, ab initio quantum Monte Carlo code for the study and prediction of materials properties by stochastically solving quantum many-body problems using variational MC and more accurate but computationally demanding diffusion MC. It allows for highly accurate, predictive ab initio calculations of larger and more complex systems that are limited or inaccessible by conventional methods such as density functional theory (DFT).

Algorithmic improvements and enhanced predictive capacity targeted by CAAR will accelerate the understanding of a wide range of materials, e.g., strongly correlated systems and transition elements, which are essential for materials design for energy capture, storage and conversion, as well as high temperature superconductors.

Targeted for CAAR:

1. Orbital evaluation and memory usage
 - reduce memory usage by different approaches for wavefunctions representation
 - analyze and balance storage vs re-computation of trial wavefunctions
2. Slater determinant evaluation
 - examine and improve numerical stability of current (less computationally intensive) updating scheme
3. QMC energy evaluation
 - optimize and explore extra level(s) of parallelism to improve time-to-solution ratio
4. Load balancing
 - reduce synchronizations and global collection operations to maximize performance

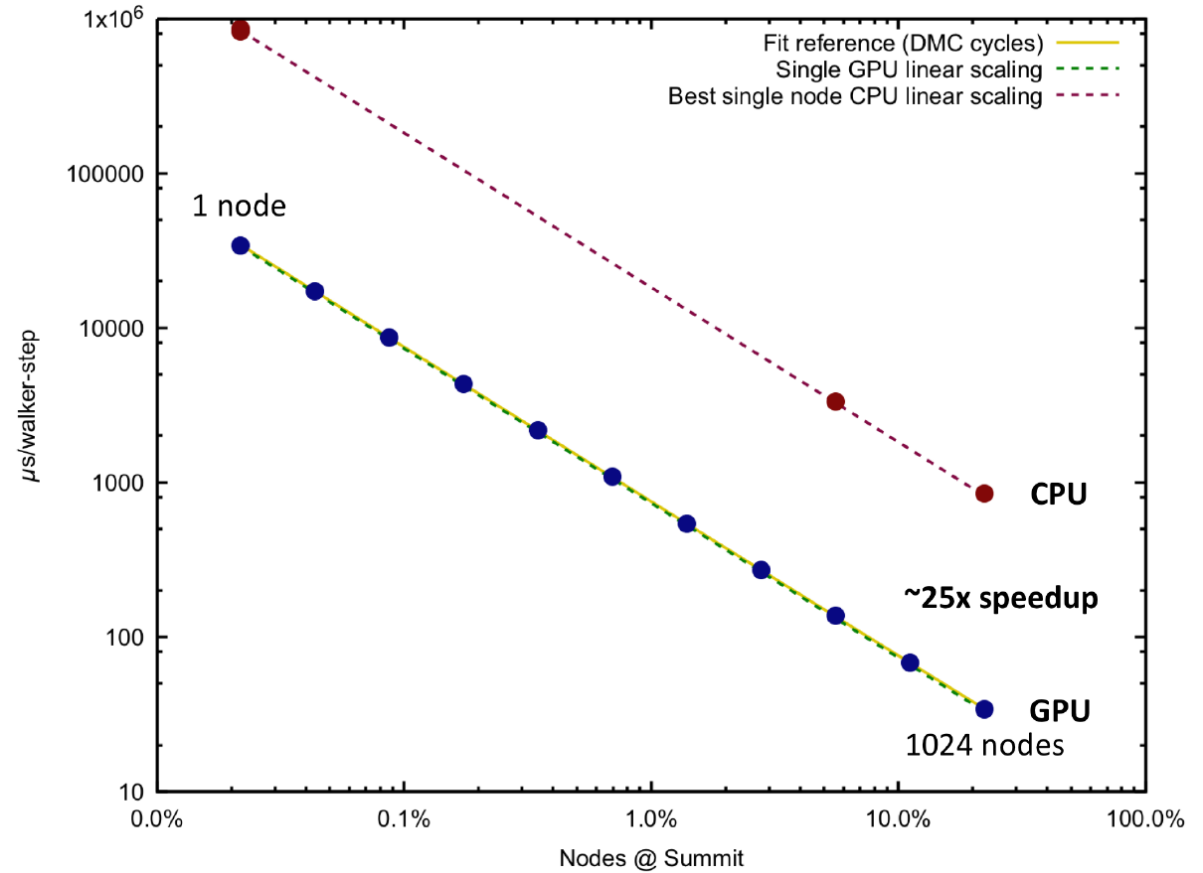
<http://qmcpack.org/>

QMCPACK Application Readiness Status

1. Implementation of complex wavefunctions on GPUs to handle general twist boundary conditions. This added functionality enables most standard QMC calculations to use GPUs for acceleration.
2. Development and implementation of a new Monte Carlo update scheme, the rank-k delayed updates algorithm, to increase compute intensity. The multiple, sequential BLAS-2 matrix operations are fused into a single BLAS-3 operation.
3. Implementation that enables a more general combination of the number of MPI ranks and GPUs on a node.
4. Investigation and implementation of distributed read-only data (“spline table”) over multiple GPUs, as well as mutual access of GPU memory among MPI ranks and GPUs, within a node. This lifted the on-chip memory limitations of a GPU, enabling more memory intensive calculations.
5. Investigation of using task-based programming techniques to improve parallelism on GPUs.

Early Results on Summit

Scaling plot of NiO 256 atom cell runs on up to 1024 nodes of Summit.



RAPTOR

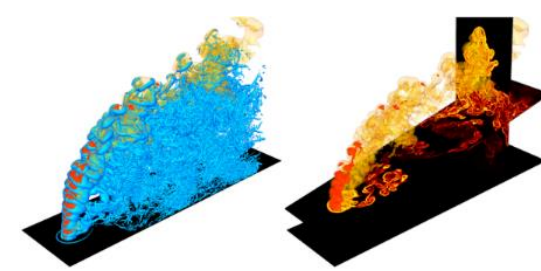
Dr. Joseph Oefelein, Sandia National Laboratory, Livermore, CA
Dr. Ramanan Sankaran, Oak Ridge National Laboratory, Oak Ridge, TN
Dr. Guilhem Lacaze, Dr. Rainer Dahms, and Dr. Anthony Ruiz, SNL



Joseph C. Oefelein



Ramanan Sankaran



Domain Area: Engineering/Combustion

Large Eddy Simulation (LES) of turbulent combustion is a research program that spans DOE-SC/BES and DOE-EERE/VTO with the objective of enabling predictive simulations of engine combustion

RAPTOR is a Computational Fluid Dynamics (CFD) solver designed for LES of a wide variety of turbulent combustion problems in complex geometries. It is designed to handle high-Reynolds-number, high-pressure, real-gas and/or liquid conditions over a wide Mach operating range, including liquid injection and sprays. It accounts for detailed thermodynamics and transport processes at the molecular level, and is sophisticated in its ability to handle a generalized sub-filter model framework in both the Eulerian and Lagrangian frames.

LES with RAPTOR will allow better insights to be gained from complementary experiments and thus provide insights into the key processes that must be accounted for in engineering models. It will enable predictive simulations of advanced concepts will accelerate the design cycle of internal combustion engines and gas turbines.

CAAR targets

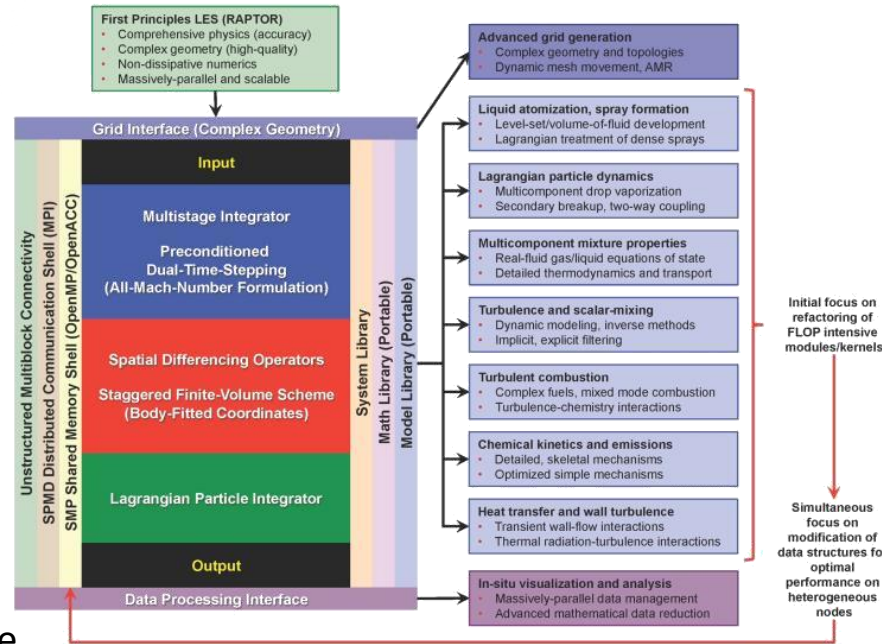
1. MPI+Directives(OpenMP/OpenACC) approach to accelerate the flow solver
2. Physics kernels will be implemented as portable libraries using Kokkos framework for device abstraction
3. Load balancing through task based model with guided task placement

<http://crf.sandia.gov/tag/raptor/>

RAPTOR Current Status

1. Physics models that are computationally intensive are externalized as libraries

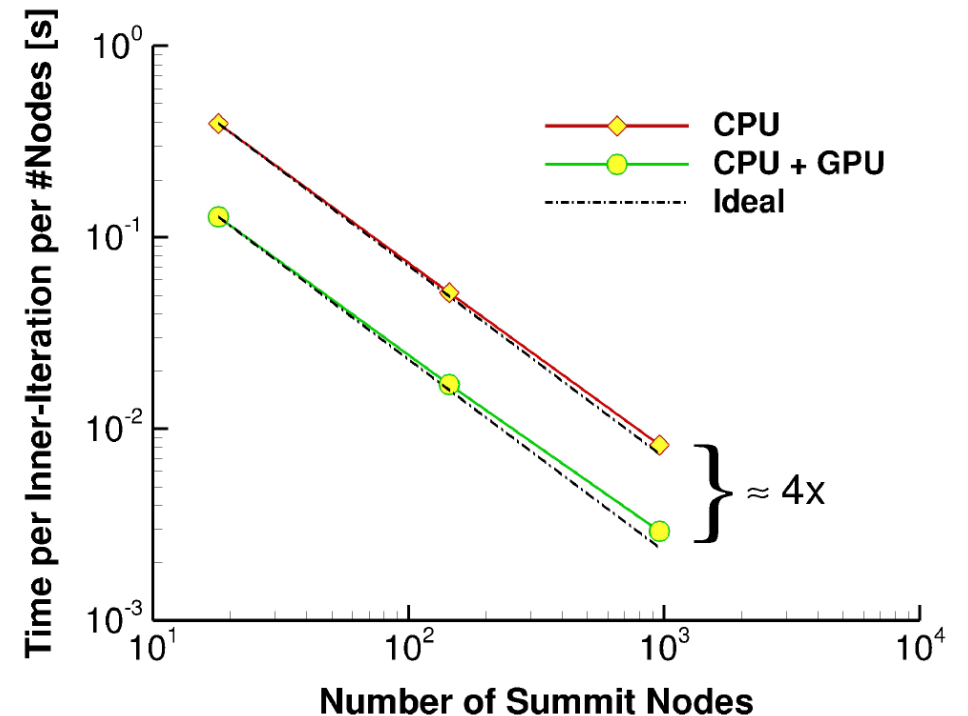
- New accelerated version of the libraries implemented in templated C++ using Kokkos programming model
- Performance portable through use of multiple backends (Cuda, OpenMP etc.)
- Tests developed to verify correctness against the original implementation
- Interfaces developed to invoke C++ library from original code and exchange data



2. Main flow solver and other physics models, that are not rewritten, are accelerated through a hybrid MPI+OpenMP programming model
3. Performance portability is emphasized in both Kokkos and the MPI+OpenMP developments
4. RAPTOR uses the GPU accelerators on Summit for a significant fraction of the computation

Early Results on Summit

Weak scaling attributes of the hybrid-MPI+OpenMP version of RAPTOR on Summit



CAAR Current Status (August 2018)

Preliminary status of the 13 CAAR projects in meeting the stated goals:

1. Scalability to 20% of the number of Summit nodes
2. Performance improvement with factor 2 or better from utilizing the GPUs.

Currently:

- 2 Applications meet performance metrics
- 6 Applications exceed performance metrics

<u>Application</u>	<input type="checkbox"/>	<u>Scaling</u>	<u>Perf</u>
ACME/E3SM	<input type="checkbox"/>		
DIRAC	<input type="checkbox"/>		
FLASH	<input type="checkbox"/>	512	2.3
GTC	✓	928	20
HACC	✓	1024	30
LS-DALTON		512	2.6
NAMD	✓	1024	5.7
NUCCOR	✓	1024	4.1
NWChem			
QMCPACK	✓	1024	24.7
RAPTOR	✓	960	4
SPECFEM	✓	960	4.9
XGC	✓	1024	13.2

Summit Early Science Program (ESP)

1. Call for Early Science Proposals

- a. Issued December 2017, resulting in 62 Letters of Intent (LOI) submitted by year's end
- b. 62 LOI teams were provided access in "waves" to Summit Phase I.
 - CAAR-related ESP projects plus 2 ORNL deep-learning projects form the first wave
 - Prior to final acceptance, all activity is porting, scaling, application readiness
 - Access to Summit was provided to demonstrate scalability and performance of their codes
- c. 48 ESP proposals were submitted by the deadline of June 2018
 - About 30 proposals meet computational readiness and scientific relevance criteria

2. Gordon Bell

Five teams with work on Summit were selected as Gordon Bell finalists

- Teams had access to Summit to demonstrate scalability and performance of their codes
- Opportunity to achieve updated numbers on Summit, as allowed by acceptance work

3. Goals: Early Science achievements, Demonstrate CAAR Work, Prepare for INCITE & ALCC, Harden Summit for Full-User Operations.

Five Gordon Bell Finalists used Summit

A Fast Scalable Implicit Solver for Nonlinear Time-Evolution Earthquake City Problem on Low-Ordered Unstructured Finite Elements with Artificial Intelligence and Transprecision Computing

Tsuyoshi Ichimura^{1,2,3}, Kohei Fujita^{1,3}, Takuma Yamaguchi¹, Akira Naruse⁴, Jack C. Wells⁵, Thomas C. Schulthess⁶, Tjerk P. Straatsma⁵, Christopher J. Zimmer⁵, Maxime Martinasso⁶, Kengo Nakajima^{7,3}, Muneco Hori^{1,3}, Lalith Maddegadara^{1,3}

¹Earthquake Research Institute & Department of Civil Engineering, The University of Tokyo
²Center for Advanced Intelligence Project, RIKEN, ³Center for Computational Science, RIKEN
⁴NVIDIA Corporation, ⁵Oak Ridge National Laboratory
⁶Swiss National Supercomputing Centre, ⁷Information Technology Center, The University of Tokyo

Attacking the Opioid Epidemic: Determining the Epistatic and Pleiotropic Genetic Architectures for Chronic Pain and Opioid Addiction

Wayne Joubert¹, Deborah Weighill^{1,4}, David Kainer¹, Sharlee Climer², Amy Justice³, Kjersten Fagnan⁵, Daniel Jacobson¹

¹Oak Ridge National Laboratory, ²University of Missouri-St. Louis,
³Yale University/Department of Veterans Affairs, ⁴University of Tennessee, ⁵DOE Joint Genome Institute

167-PFlops Deep Learning for Electron Microscopy: From Learning Physics to Atomic Manipulation

Robert M. Patton¹, J. Travis Johnston¹, Steven R. Young¹, Catherine D. Schuman¹, Don D. March², Thomas E. Potok¹, Derek C. Rose³, Seung-Hwan Lim¹, Thomas P. Karnowski³, Maxim A. Ziatdinov^{4,5}, and Sergei V. Kalinin^{4,5}
Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6085

Simulating the *weak* death of the neutron in a femtoscale universe with near-Exascale computing

Evan Berkowitz*, M.A. Clark[†], Arjun Gambhir^{‡§¶}, Ken McElvain^{¶§}, Amy Nicholson^{||}, Enrico Rinaldi^{**§}, Pavlos Vranas^{‡§} André Walker-Loud^{§‡¶}, and Chia Cheng Chang[§], Balint Joo^{††}, Thorsten Kurth^{‡‡§}, Kostas Orginos^{xxi},

* Institut für Kernphysik and Institute for Advanced Simulation, Forschungszentrum Jülich, 52425 Jülich Germany
[†] NVIDIA Corporation, Santa Clara, California, 95051, USA
[‡] Physical Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California 94550, USA
[§] Nuclear Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[¶] Department of Physics, University of California, Berkeley, CA 94720, USA
^{||} Department of Physics and Astronomy, University of North Carolina, Chapel Hill, NC 27516-3255, USA
^{**} RIKEN-BNL Research Center, Brookhaven National Laboratory, Upton, NY 11973, USA
^{††} Scientific Computing Group, Thomas Jefferson National Accelerator Facility, Newport News, VA 23606, USA
^{‡‡} NERSC, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
^x Department of Physics, The College of William & Mary, Williamsburg, VA 23187, USA
^{xxi} Theory Center, Thomas Jefferson National Accelerator Facility, Newport News, VA 23606, USA

Exascale Deep Learning for Climate Analytics

Thorsten Kurth
Lawrence Berkeley National Laboratory
Berkeley, CA 94720, USA
tkurth@lbl.gov

Sean Treichler
NVIDIA
Santa Clara, CA 95051, USA
sean@nvidia.com

Joshua Romero
NVIDIA
Santa Clara, CA 95051, USA
josh@nvidia.com

Mayur Mudigonda
Lawrence Berkeley National Laboratory
Berkeley, CA 94720, USA
mudigonda@berkeley.edu

Nathan Luehr
NVIDIA
Santa Clara, CA 95051, USA
nluehr@nvidia.com

Everett Phillips
NVIDIA
Santa Clara, CA 95051, USA
ephillips@nvidia.com

Ankur Mahesh
Lawrence Berkeley National Laboratory
Berkeley, CA, 94720, USA
amahesh@lbl.gov

Michael Matheson
Oak Ridge National Laboratory
Oak Ridge, TN 37831, USA
mathesonma@ornl.gov

Jack Deslippe
Lawrence Berkeley National Laboratory
Berkeley, CA 94720, USA
jrdeslippe@lbl.gov

Massimiliano Fatica
NVIDIA
Santa Clara, CA 95051, USA
mfatica@nvidia.com

Prabhat
Lawrence Berkeley National Laboratory
Berkeley, CA 94720, USA
prabhat@lbl.gov

Michael Houston
NVIDIA
Santa Clara, CA 95051, USA
mhouston@nvidia.com

CoMet

Dr. Dan Jacobson, Dr. Wayne Joubert, Oak Ridge National Laboratory



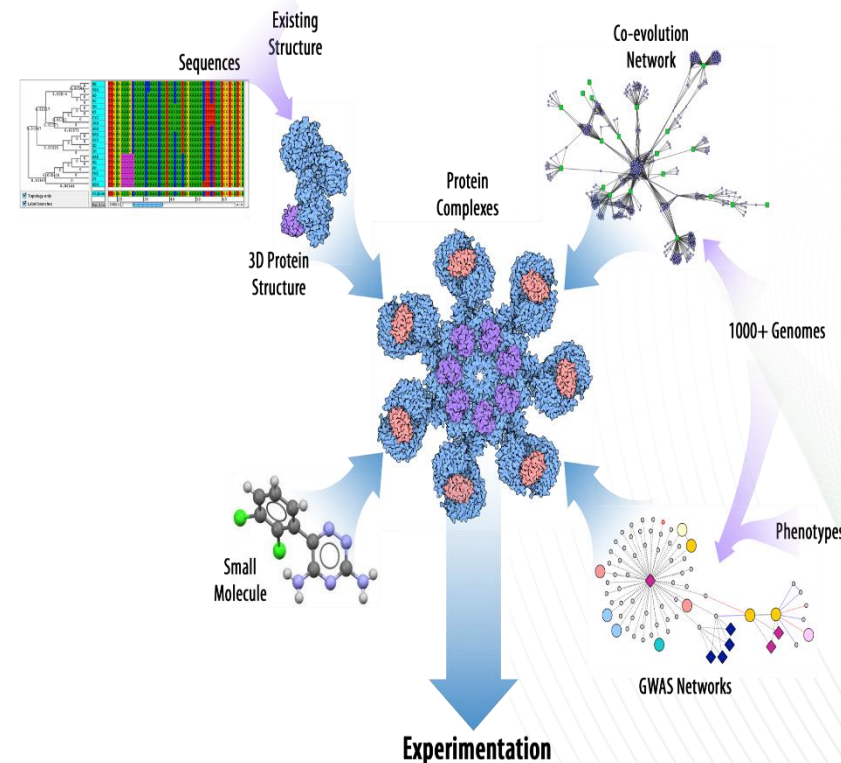
Dan Jacobson Wayne Joubert

Domain Area: Bioinformatics

CoMet is a new data analytics application for comparative genomics studies targeting the discovery of co-occurrences of specific genetic features within a population. It is neither a traditional modeling and simulation or a machine learning application, but provides an integrated component in an AI workflow.

Implementation approach:

- Operates on allele data that is encoded directly into long vectors of 2-bit values
- Performs an all-to-all comparison of vectors – this requires aggressive overlap of communication, transfers, GPU computation, CPU computation
- Original implementation does calculations on the GPU using bitwise operations (AND, OR, NOT, mask, __popc, etc.) in a GEMM-like computational framework
- The new method maps the vector elements to FP16 data values and exploits the Volta tensor cores using cuBLAS GEMM calls

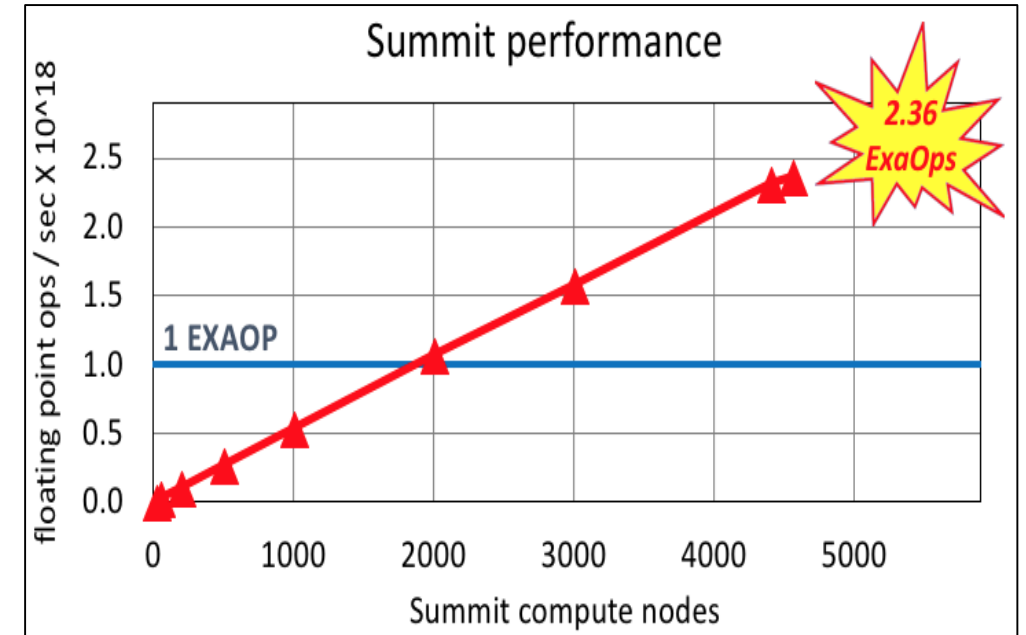


CoMet Application Readiness Status

- Each vector containing 2-bit values is converted into two columns representing the number of 0s and 1s in each element, forming new matrix V'
- Applying dense matrix-matrix product to calculate $V'^T V'$ generates all vector-vector correlation tables
- Use cuBlasGemmEx
- Input values are FP16
- Results are computed and stored as FP32

Performance

- Achieved **2.36 ExaOps** (mixed precision ExaFlops) at 4,560 nodes (99% of Summit) using the Tensor Cores – first reported application to reach ExaOp
- Equivalent to **86.4 TF** per GPU for the whole computation (including communications and transfers) at 4,560 nodes
- Excellent scaling made possible by Summit fat tree network with adaptive routing
- **> 4X faster** than original bitwise (non-flop) implementation on GPUs (= 4X more science possible)



GronOR

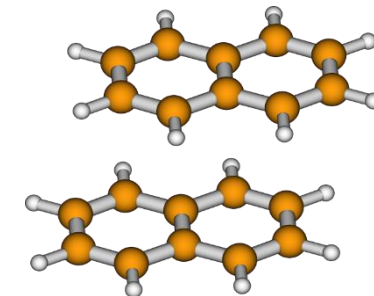
Dr. Remco W. A. Havenith, University of Groningen
Dr. Tjerk Straatsma, Oak Ridge National Laboratory



Remco Havenith



Tjerk Straatsma



Domain Area: Computational Chemistry

GronOR is a non-orthogonal configuration interaction application based on the factorization method in the General Non-Orthogonal Matrix Element (GNOME) code. COIN is a substantially refactored, massively MPI-parallelized code base that can take advantage of GPU acceleration. Scalability and load balancing is achieved through use of a task based algorithm. The algorithm is implemented in a fault tolerant way.

The intended application of GronOR is for small clusters of molecules, with special interest in molecular systems relevant for photovoltaic applications.

Implementation approach:

- OpenACC for GPU off-loading
- Master-slave model with task based load balancing
- MPI parallelization with point-to-point non-blocking communication
- Avoid global synchronization and global reduction operations
- Fault resilient implementation

Targeted for ESP:

- Acceleration of additional computational kernels for GPUs on Summit.
- Demonstration of fault tolerance.
- Application to molecular clusters relevant to photovoltaic systems.

<http://gronor.org>

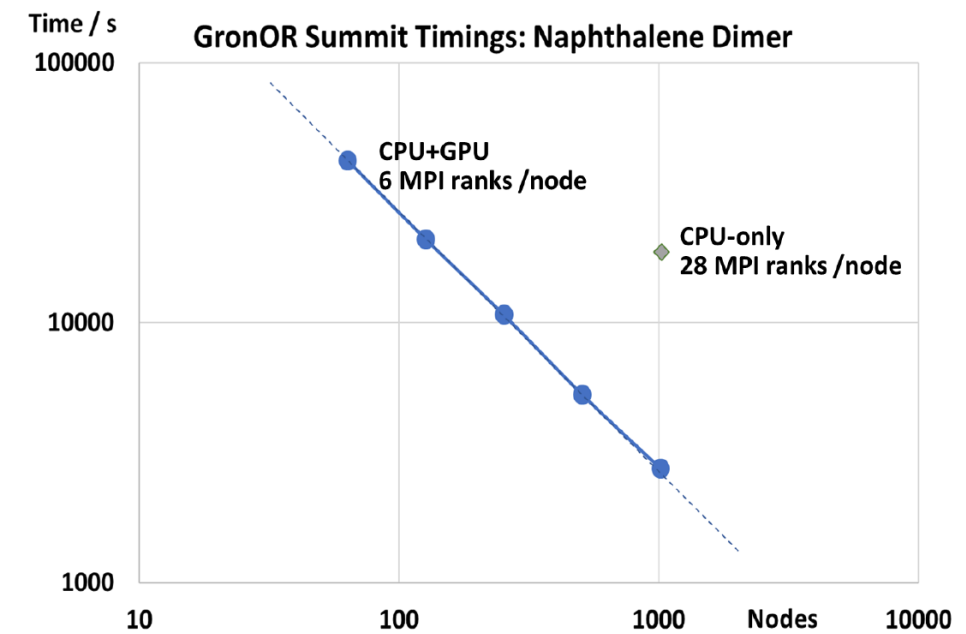
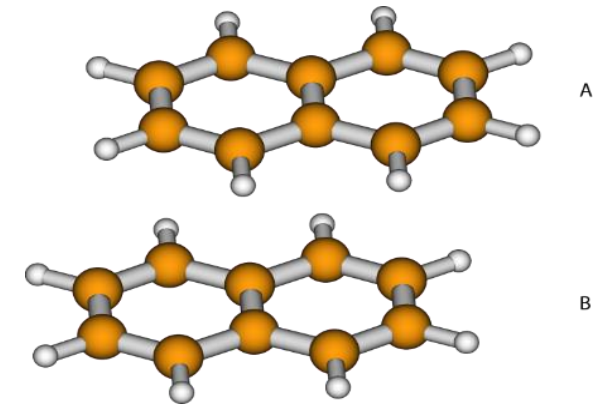
GronOR Application Readiness Status

Development work on scalability and load balancing

- OpenACC implementation for GPU off-loading
- Master-slave model with task based load balancing
- MPI parallelization with point-to-point non-blocking communication
- Avoid global synchronization and global reduction operations
- Fault resilient implementation
- GAMESS-UK and SYMOL for integrals and CASSCF vectors

Early Results on Summit

- Naphthalene molecules with asymmetric CASSCF configurations 44,88 and 6-311G basis set, leading to 112,867,800 Hamiltonian matrix elements
- GPU+CPU (6 MPI ranks per node) vs. CPU-only (28 MPI ranks per node) performance of 1024 node run on Summit for naphthalene dimer: **6.8x**
- Scalability is near linear on Summit up to 1080 nodes, which is close to the full Phase I system



Summit Early Results: Conclusions

- Interest in optimizing codes and starting early science on Summit has been enormous
- Applications running well on Titan tend to run well on Summit
- Porting and optimizing a non-accelerated code can be a multi-year effort
- Working with a new, developing software stack is challenging: multiple compilers, MPI library performance, job scheduler
- Programming is more complex: multiple GPUs per node, CPU SMT threading, NUMA domains, MPS, coordination of host code threading and device selection, NVRAM
- Partnership between code developers, OLCF staff and Center of Excellence is crucial

Acknowledgment

- CAAR Principal Investigators and their teams: David Bader (ACME), Lucas Visscher (DIRAC), Zhihong Lin (GTC), Salman Habib (HACC), Poul Jørgensen (LSDALTON), Jim Phillips (NAMD), Gaute Hagen (NUCCOR), Karol Kowalski (NWCHEM), Paul Kent (QMCPACK), Joe Oefelein (RAPTOR), Jeroen Tromp (SPECFEM), and CS Chang (XGC)
- Early Science Lol Principal Investigators and their teams: Remco Havenith (GRONOR), Dan Jacobson (COMET)
- OLCF Staff: Matt Norman, Dmitry Liakh, Bronson Messer, Wayne Joubert, Dmytro Bykov, Gustav Jansen, Ying Wai Li, Ramanan Sankaran, Judy Hill, and Ed D'Azevedo, and others
- CSEEN postdocs: Anikesh Pal, Amelia Fitzsimmons, Thom Papatheodore, Austin Harris, Micah Schuster, Andreas Tillack, Kalyana Gottiparthi, Yangkang Chen, and Stephen Abbott
- Center of Excellence Staff: Jaime Moreno, Leopold Grinberg, Cyrill Zeller, Eric Luo, David Appelhans, Matt Niemerg, Jeff Larkin, Stephen Abbott and many others

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.



**Questions?
Tjerk Straatsma
str@ornl.gov**