IXPUG FALL MEETING

TACC – SEPTEMBER 26, 2017

ARGONNE LEADERSHIP COMPUTING FACILITY UPDATE

DAVID E. MARTIN
Manager, Industry Partnerships and Outreach
Argonne Leadership Computing Facility

TIM WILLIAMS
Deputy Director of Science
Argonne Leadership Computing Facility

Argonne NATIONAL LABORATORY

ALCF THETA SYSTEM & EARLY SCIENCE PROGRAM (ESP)

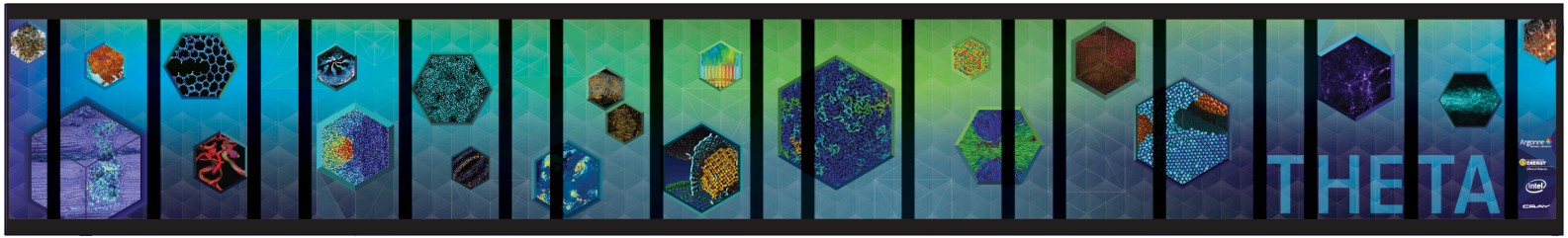
Argonne Leadership Computing Facility

THETA

- 3624 nodes
  - Xeon Phi 7230 (2nd gen.)
  - 16 GB MCDRAM
  - 192 GB DDR4
  - 128 GB SSD
- Peak 9.65 petaFLOPS
- Cray Aries interconnect
- 10 PB Lustre parallel file system

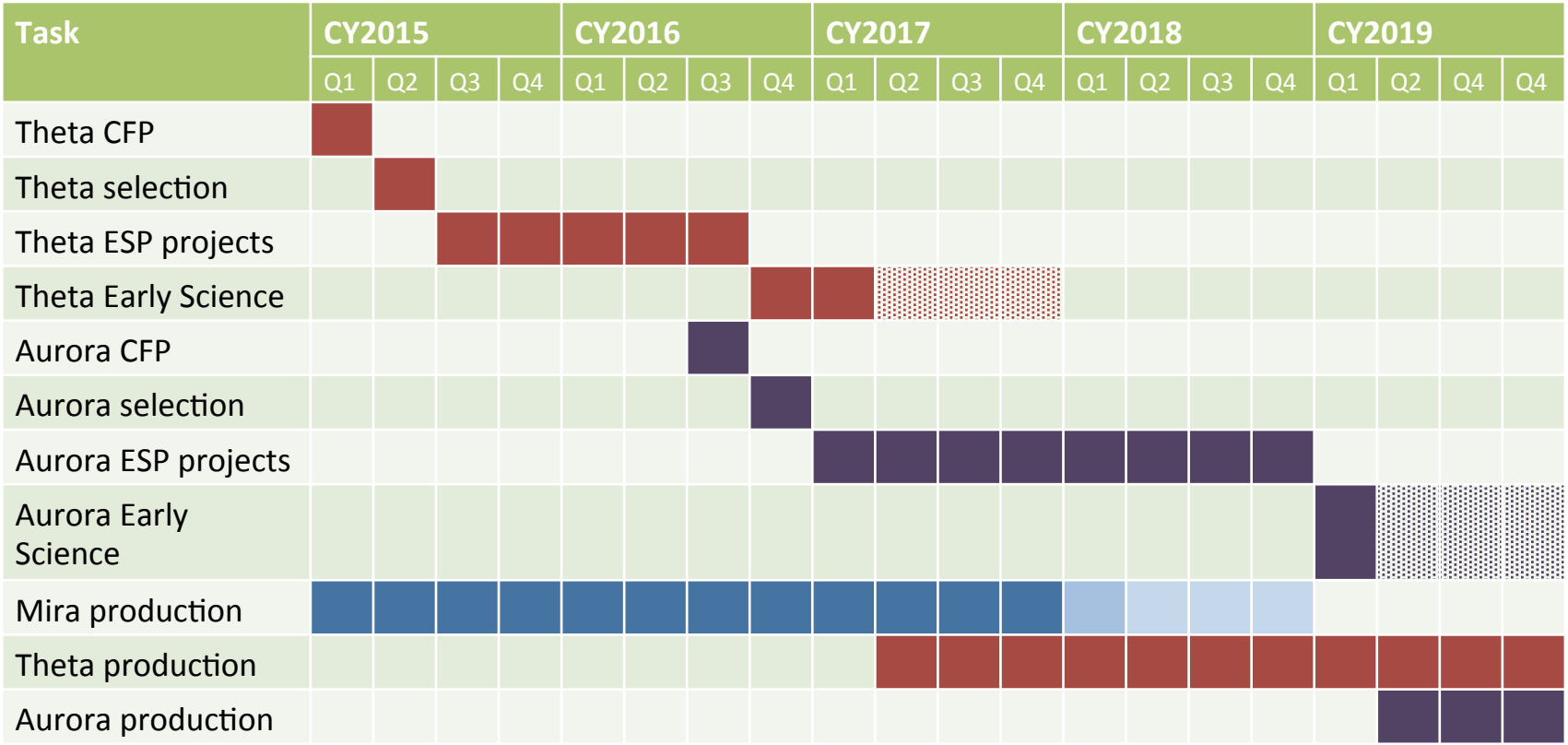
EARLY SCIENCE PROGRAM

- Theta dedicated for science runs: *just ended*
- 6 Tier 1 + 6 Tier 2 projects
- Optimize applications
- Solidify libraries & infrastructure
- Prep Theta for *science on day one*



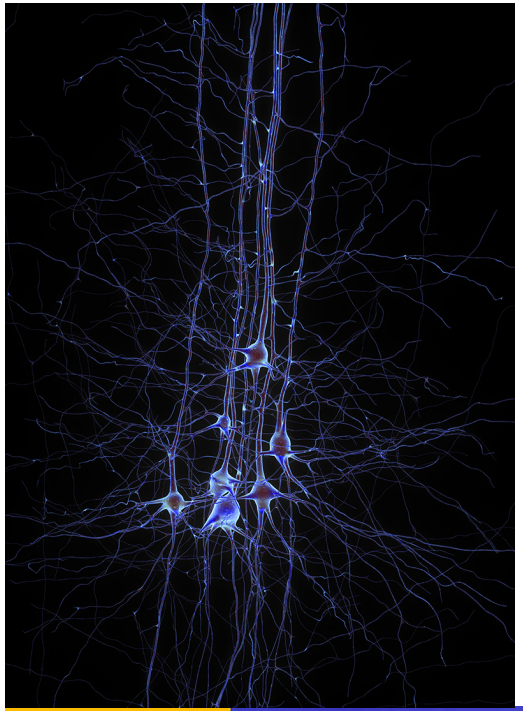
Argonne NATIONAL LABORATORY

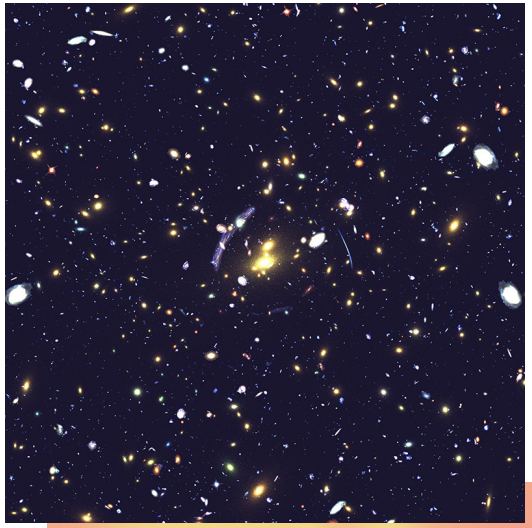
ESP Timeline

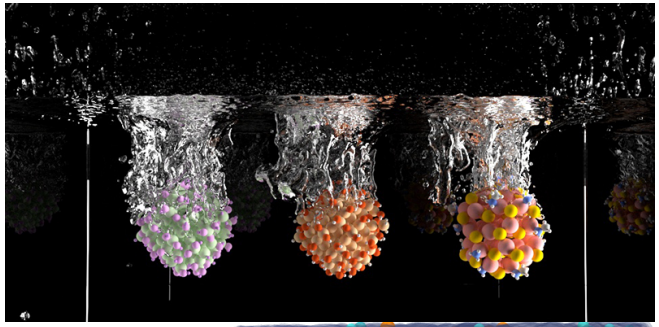


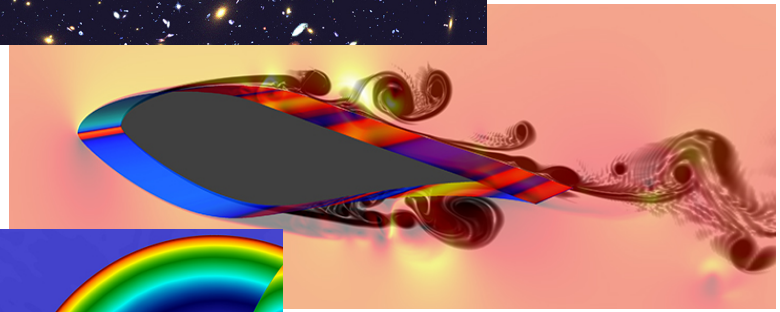
THETA ESP PROJECTS

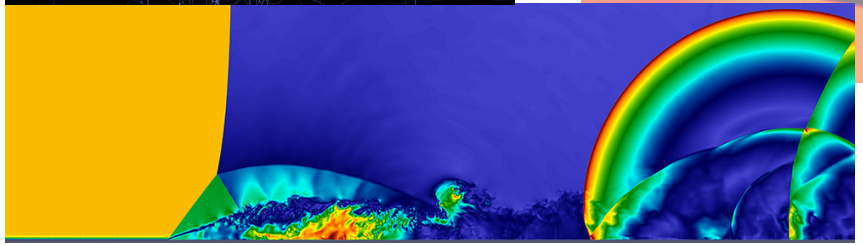
Tier 1

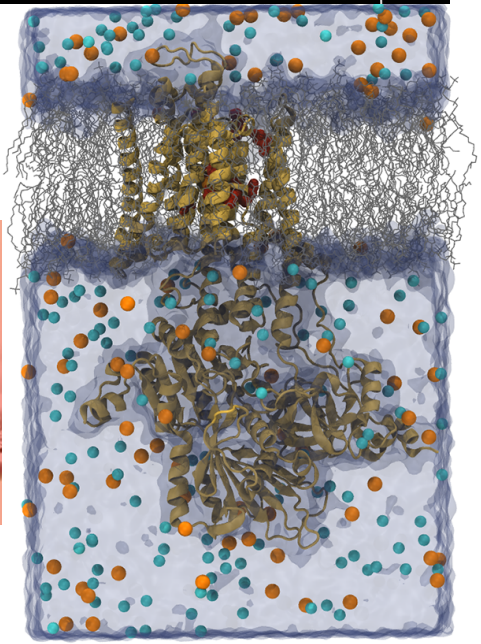






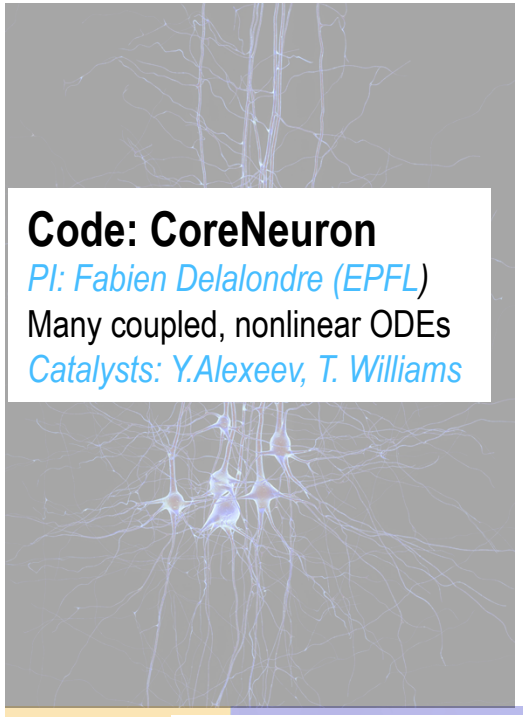




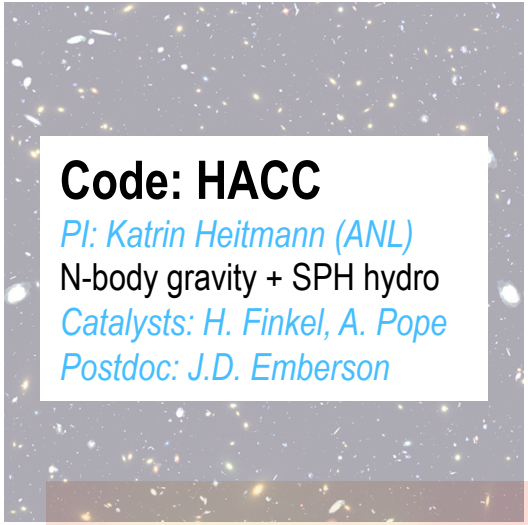


THETA ESP PROJECTS

Tier 1

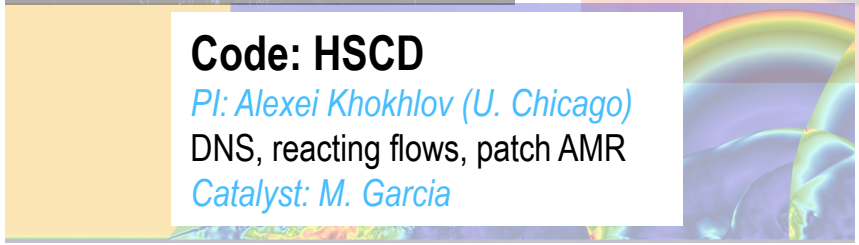


Code: CoreNeuron
PI: Fabien Delalondre (EPFL)
Many coupled, nonlinear ODEs
Catalysts: Y.Alexeev, T. Williams

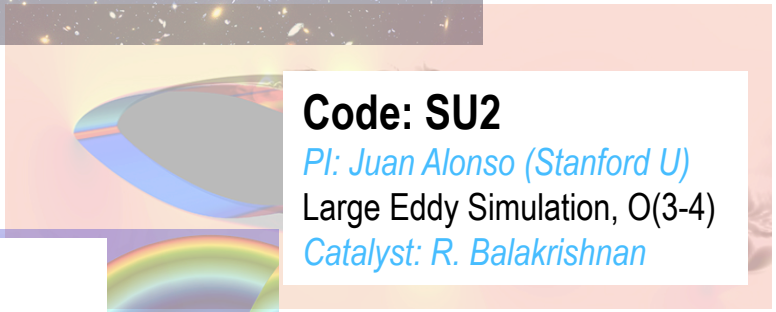


Code: HACC
PI: Katrin Heitmann (ANL)
N-body gravity + SPH hydro
Catalysts: H. Finkel, A. Pope
Postdoc: J.D. Emberson

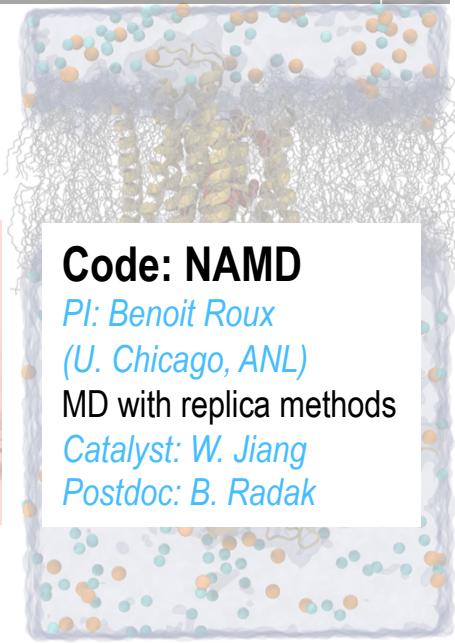
Codes: WEST & Qbox
PI: Giulia Galli (U. Chicago)
MBPT & ab initio MD
Catalyst: C. Knight
Postdoc: H. Zheng



Code: HSCD
PI: Alexei Khokhlov (U. Chicago)
DNS, reacting flows, patch AMR
Catalyst: M. Garcia



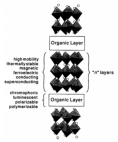
Code: SU2
PI: Juan Alonso (Stanford U)
Large Eddy Simulation, O(3-4)
Catalyst: R. Balakrishnan



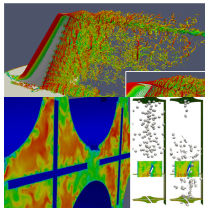
Code: NAMD
PI: Benoit Roux
(U. Chicago, ANL)
MD with replica methods
Catalyst: W. Jiang
Postdoc: B. Radak

THETA ESP PROJECTS

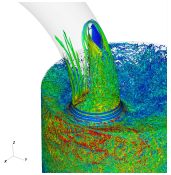
Tier 2



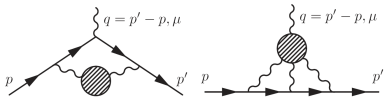
Codes: FHI-Aims & GAtor
PI: Volker Blum (Duke U.)
MBPT (DFT) & genetic algorithm
Catalyst: Álvaro Vázquez-Mayagoitia



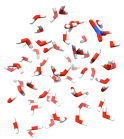
Code: PHASTA
PI: Kenneth Jansen (U. Colorado)
CFD, unstructured mesh
Catalyst: Hal Finkel



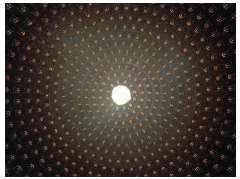
Code: Nek5000
PI: Christos Frouzakis (ETHZ)
Spectral element CFD with combustion
Catalyst: Scott Parker



Codes: MILC & CPS
PI: Paul Mackenzie (FNAL)
Lattice QCD
Catalyst: James Osborn



Code: GAMESS
PI: Mark Gordon (Iowa State U.)
FMO - quantum chemistry
Catalysts: Yuri Alexeev, Graham Fletcher



Code: GFMC
PI: Steven Pieper (ANL)
Greens Function Monte Carlo – nuclear
Catalyst: James Osborn

FREE ENERGY LANDSCAPES OF MEMBRANE TRANSPORT PROTEINS

Code: NAMD

PI: Benoit Roux (U. Chicago, ANL)

MBPT & ab initio MD

Catalyst: Wei Jiang

Postdoc: Brian Radak

Science Impact

- An atomistic picture of membrane transport proteins is a critical component of our understanding of a broad range of biological functions. This work will utilize computational models to provide both detailed visualizations of large protein motions as well as quantitative predictions into the energetics of these processes.

Numerical Methods/Algorithms

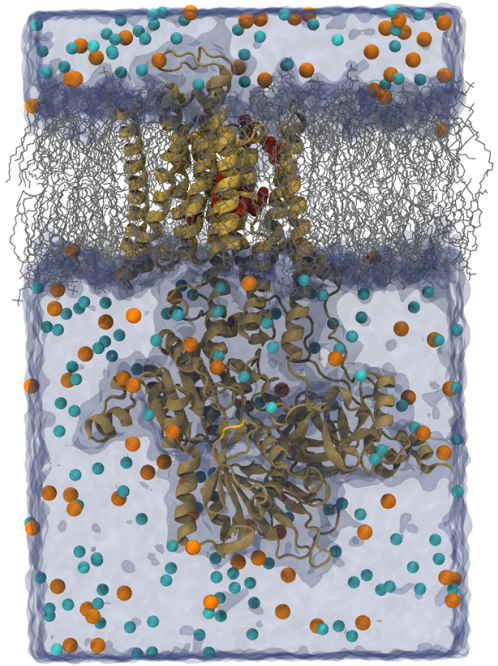
- Classical molecular dynamics simulations, including replica exchange and string methods with swarms of trajectories

Parallelism

- Charm++, an overdecomposition-based message-driven parallel programming model.

Application Development

- Implemented new hybrid algorithm combining Monte Carlo and MD
- Simulations should fit entirely within MCDRAM



The Na/K pump (yellow) is a P-type ATPase that spans the plasma membrane of animal cells. It acts to maintain the ionic gradient (orange and blue spheres) that gives rise to the cell potential, a critical component of cell machinery and signal transduction. This project will develop new simulation models and methodologies to study the sensitivity of key, pH-sensitive amino acid residues in the transmembrane region (red spheres) of the Na/K pump and other P-type ATPases, as well as related F/V-type ATPases. Brian Radak, University of Chicago; data courtesy of Huan Rui, University of Chicago

NAMD Code optimization and algorithm design on Theta

Generic charm++ machine layer on Aries interconnect

Efficient compiler generated vectorization for nonbond kernel by code restructuring

3X speedup per core or 12X per node relative BG/Q

Better strong scaling than Mira: 100M atom system strongly scale to whole Theta (16 racks on Mira)

Developed constant PH MD algorithm with built-in Python interface that enables on-the-fly topology conversion of molecular system

Free energy calculation with ensemble of constant PH MD trajectories scales to whole Theta

MPI/OPENMP PARALLELIZATION OF GAMESS FOR THE SECOND GENERATION OF INTEL® XEON PHI PROCESSOR

PI MARK GORDON (A COLLABORATION BETWEEN ANL, ISU, MSU, RSC, INTEL)

IMPACT AND APPROACH

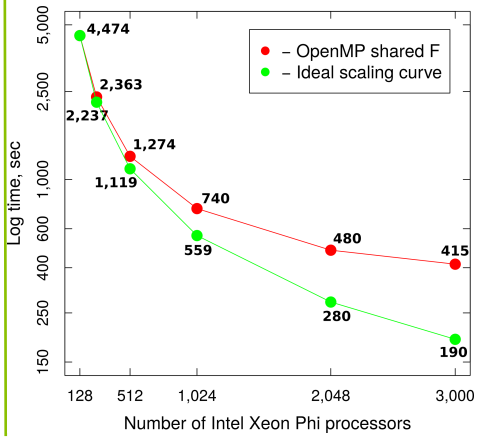
- GAMESS code - a general purpose quantum chemistry package. All major methods and a large number of properties are implemented
- One of the most popular quantum chemistry packages with an estimated user base of 100,000+ users worldwide
- Supported by INCITE, ALCC, ESP, ECP, and two IPCCs
- A parallel Fortran 77/95 code, which scales up to 750,000 cores using FMO method
- There is a major effort underway to rewrite code in Fortran 95, OpenMP threading of major methods, and vectorization of the integral packages

ACCOMPLISHMENTS

- Sped up HF method up to 6 times
- Reduced memory footprint by up to 200 times
- Scaled up the code to 3,000 KNL nodes on Theta
- Vectorized the Rys integral package resulting in up 2x speed up
- Implemented OpenMP parallelization in SCF driver
- Threaded energy and gradient code for RHF, UHF, ROHF, and HF-exchange part in hybrid DFT
- Results were published in SC17 technical paper, IPDPSW paper, the International Journal of High Performance Computing Applications
- Best SC17 poster nomination

ALCF CONTRIBUTIONS

- Yuri Alexeev (an ALCF computational scientist) led efforts for OpenMP threading of GAMESS
- Yuri Alexeev helped design and implement threaded HF code in a collaboration with Vladimir Mironov funded by IPCC
- All benchmarks were ran by Yuri Alexeev on Theta and JLSE

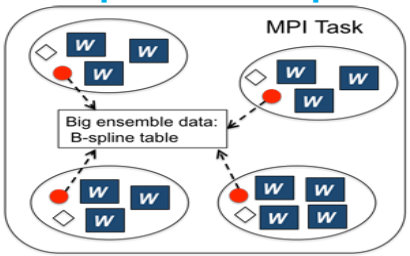


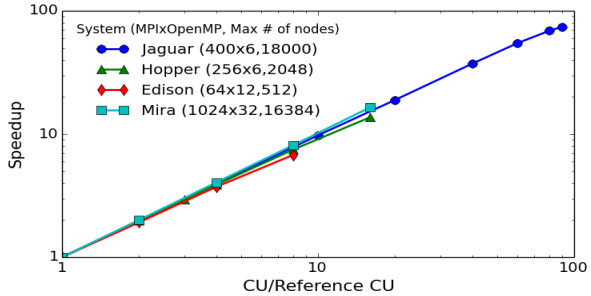
Publication: V. Mironov, Y. Alexeev, A. Moskovsky, K. Keipert, M. S. Gordon, M. Dmello, *Proceedings of the 2017 International Conference for High Performance Computing and Communications (SC17), Denver, CO, USA*, (2017)

ARCHITECTURE OF QMCPACK ON CLUSTERS OF SMP

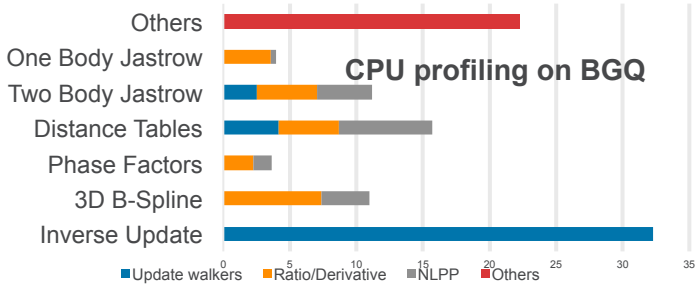
QMCPACK utilizes OpenMP/CUDA to optimize memory usage and to take advantage of the growing number of cores per SMP node or stream multiprocessors per GPU.

- Walkers within a MPI task is distributed among the cores in CPU or the SMs in GPU.
- Big common data is shared by all the walkers like wave function coefficients.





1: **for** MC generation $= 1 \cdots M$ **do**
2: **for** walker $= 1 \cdots N_w$ **do**
3: let $\mathbf{R} = \{\mathbf{r}_1 \ldots \mathbf{r}_N\}$
4: **for** particle $i = 1 \cdots N$ **do**
5: set $\mathbf{r}_i^{'} = \mathbf{r}_i + \delta$
6: let $\mathbf{R}^{'} = \{\mathbf{r}_1 \ldots \mathbf{r}_i^{'} \ldots \mathbf{r}_N\}$
7: **ratio** $\rho = \Psi_T(\mathbf{R}^{'})/\Psi_T(\mathbf{R})$
(One Body, Two Body, 3D B-Spline)
8: **derivatives** $\nabla_i \Psi_T, \nabla_i^2 \Psi_T$
(One Body, Two Body, 3D B-Spline)
9: **if** $\mathbf{r} \rightarrow \mathbf{r}'$ is accepted **then**
10: update state of a walker
(Inverse Update)
11: **end if**
12: **end for**{particle}
13: **local energy** $E_L =$ $\hat{H}\Psi_T(\mathbf{R})/\Psi_T(\mathbf{R})$
(One Body, Two Body, 3D B-Spline)
14: reweight and branch walkers
15: **end for**{walker}
16: update E_T and load balance
17: **end for**{MC generation}



Argonne NATIONAL LABORATORY

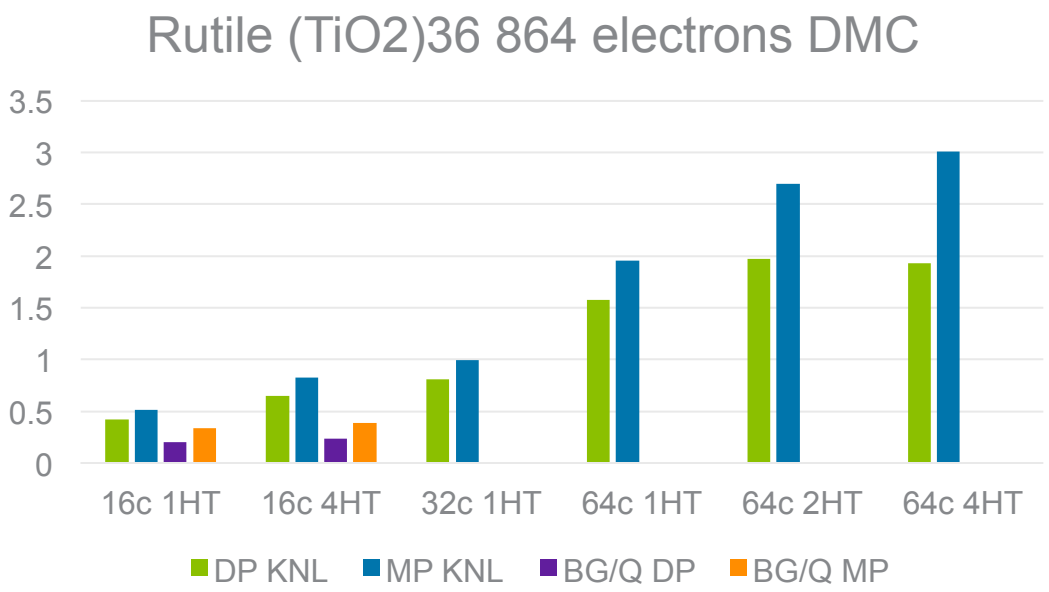
QMCPACK OPTIMIZATION STRATEGY

- Intel IPCC Support
- What we learned (From Optimization on IBM BGQ):
  - Straight forward application of QPX, SSE or Assembly brings 1.5~2x speedup
  - Better algorithm + Structures of Array (SoA) doubles the performance due to better memory bandwidth
  - Portable algorithm are possible but portable implementation is the challenge.
- What we do (Intel PCC / Exascale Computing Program)
  - Move from Double Precision to Mixed Precision
  - Move from Array of Structure (AoS) to SoA to AoSoA (Tiling)
  - Using nested threading at the walker level to reduce the time to solution
  - Rewrite completely Distance table function
  - Generate a Micro-QMC (Miniapp) to experiment with algorithms within the constraints of QMCPACK

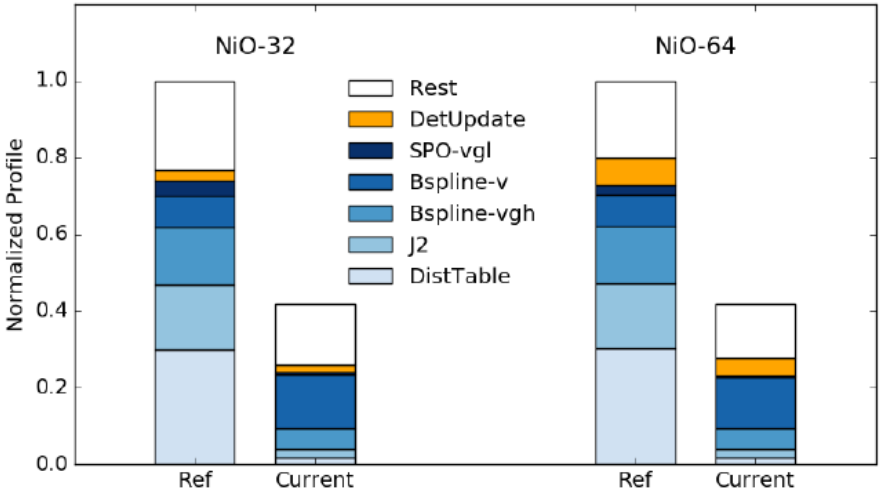
QMCPACK - DOUBLE TO SINGLE PRECISION

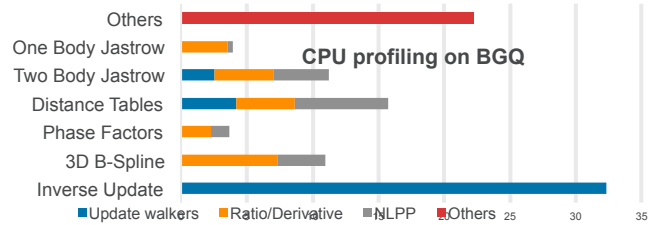
- Gain performance not only on KNL but also on BG/Q.

- Small core counts: 20% faster on KNL, gained from computing.
- Full node: 55% faster, gained from both computing and memory BW.
- Always: about 70% faster on BG/Q, gained from memory BW.



PERFORMANCE SUMMARY





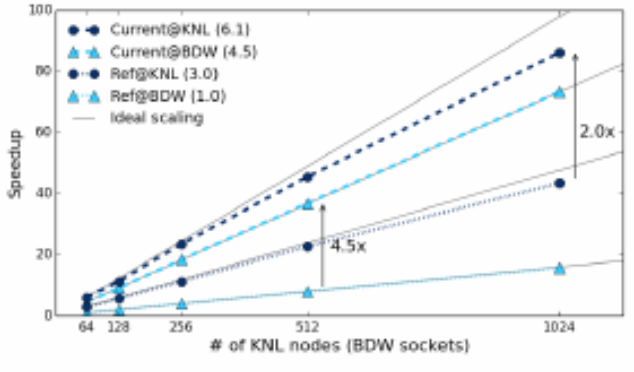


Figure 1: Strong scaling of NiO-64 benchmark on Trinity at LANL (KNL) and Serrano at SNL (BDW) systems. The performance is normalized by a reference throughput using 64 BDW sockets. Slopes of the ideal-scaling lines are provided in parentheses.

	NiO-32	NiO-64
N	384	768
N_{ion}	32	64
N_{ion}/uint cell	4	4
# of uint cells	8	16
Ion types (Z^*)	Ni(18), O(6)	
# of unique SPOs	144	240
FFT grid	80x80x80	
B-spline (GB)	1.3	2.1

(Embracing a new era of highly efficient and productive quantum Monte Carlo simulations) A. Mathuriya, Y. Luo, A. Benali, L. Shulenburger, R. Clay, J. Kim, accepted: Super Computing 17

(Optimization and parallelization of B-spline based orbital evaluations in QMC on multi/many-core shared memory processors), A. Mathuriya, Y. Luo, A. Benali, L. Shulenburger, J. Kim, International Parallel & Distributed Processing Symposium (Proceedings) 2017

OVERALL THETA ESP LESSONS LEARNED

- Structure of Arrays
- Strong scale to fit in MCDRAM
  - Successes with many MPI ranks per node (up to 64)
- Transition from BGQ (MPI + OpenMP) → KNL not generally painful
  - Adjust ranks/threads sweet spot
- Memory access looks like streaming?
  - `#pragma vector nontemporal`
- Use MKL FFT (multiple electronic structure codes)

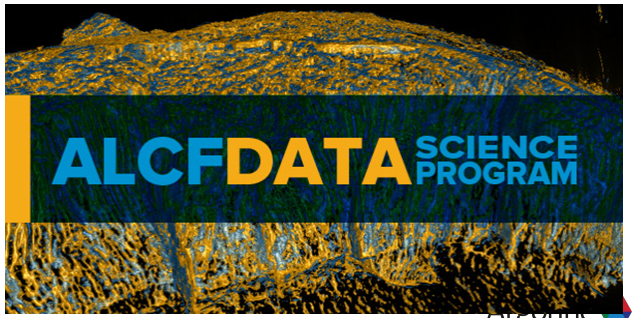
THETA ESP (AND OTHER) LESSONS LEARNED

- Running within MCDRAM? Cache mode as good as flat mode
  - Flat mode: `numactl -m 1` (allocate in HBM; error if spills out)

Code	Method	Runtime flat vs. cache
HACC	Tree N-body, particle-mesh	-0.1%
WEST	Many body perturbation theory	+8.98%
Qbox	Ab initio molecular dynamics	-6.6%
USQCD	Several Lattice QCD methods	Virtually no difference
NAMD	Classical molecular dynamics	No significant difference
QMCPACK	Quantum Monte Carlo	-4.8%
VSVB	Electronic structure	+4.2%, +0.59%

ALCF DATA SCIENCE PROGRAM (ADSP) OVERVIEW

- “Big Data” science problems that require the scale and performance of leadership computing resource
- The new initiative, targeted at big data problems that require the scale and performance of leadership-class supercomputers, will enable new science and novel usage modalities on these systems.
- Projects will cover a wide variety of application domains that span computational, experimental and observational sciences
- Focus on data science techniques including but not limited to statistics, machine learning, deep learning, UQ, image processing, graph analytics, complex and interactive workflows
- Two-year proposal period and will be renewed annually. Proposals will target **science** and **software technology** scaling for data science
- Projects receive ALCF staff support in Data and Computational Science. Tier-1 projects will be supported in part with postdoctoral scholars.
- Yearly call for proposal.
  Next deadline – Summer 2018 (Expected yearly call for proposals)
  https://www.alcf.anl.gov/alcf-data-science-program



ADSP SYSTEM RESOURCES

Mira – IBM BG/Q

- 49,152 nodes
- 786,432 cores
- 786 TB RAM
- 10 PF

Cooley - Cray/NVIDIA

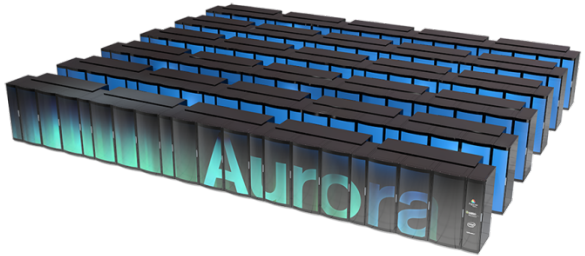
- 126 nodes (Haswell)
- 1512 cores
- 126 Tesla K80
- 48 TB RAM (3 TB GPU)

Theta- Cray/Intel

- 3240 nodes (KNL)
- 829,440 cores
- 50.6 TB MCDRAM
- 607.5 TB DDR4 RAM
- 414.7 TB SSD

Sage- Cray Urika-GX

- 32 nodes (Haswell)
- 8 TB DRAM
- 25.6 TB NVMe SSD
- BigData Analytics Stack



Over 180 PF peak performance
> 50,000 nodes with 3rd Generation Intel® Xeon Phi™ processor codename Knights Hill, > 60 cores
Over 7 PB total system memory

Data--Driven Molecular Engineering of Solar- powered Windows

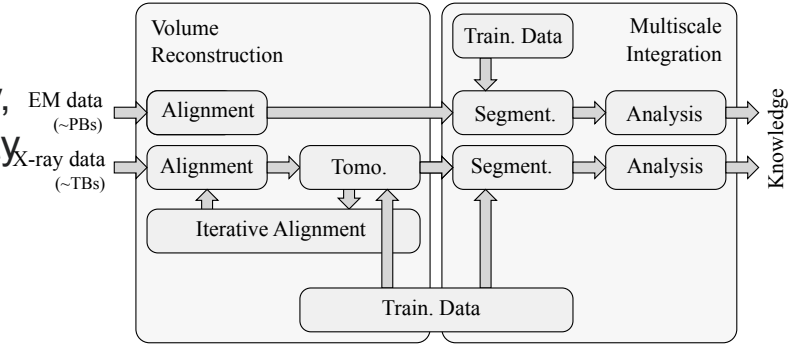
- PI: Prof. Jacqueline Cole, University of Cambridge, UK
- Co-PI – Alvaro Vasquez, Argonne National Laboratory, USA
- **Objectives:** Design of Dyne-sensitized cells (DSC), light- absorbing dye molecules needed to realize next- generation technology of solar- powered windows using large--scale data mining with machine learning.
- **Impact:** DSC are prospected to power ‘smart windows’ - windows that generate electricity from sunlight. These are expect to be an key component of buildings in future cities, in an entirely energy-sustainable fashion.

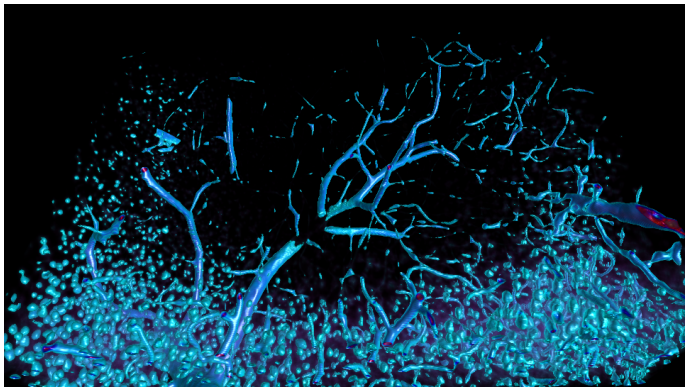
- **Approach:** A synergistic computational and experimental science approach wherein machine learning and data mining are used in conjunction with large-scale simulations and experiments to facilitate a material-by-design approach for DSC dye discovery.

ADSP CATEGORY: TIER-1

Large-scale computing and visualization on the connectomes of the brain

- **PI**: Doga Gursoy, Argonne National Laboratory
- **Other Participating Institutions:** Harvard University, Univ. of Chicago, Johns Hopkins University, University of Notre Dame, APL – Hopkins, Northwestern University
- **Objectives:** The development of a large-scale data and computational pipeline for brain science at extreme scale
- **Impact:** The scalable workflows will help facilitate gleaning invaluable knowledge about disease models such as Alzheimer's, autism spectrum disorder, etc., and enable advances in neuromorphic computing.
- **Approach:** an entirely new set of tools for understanding brain function and pathology

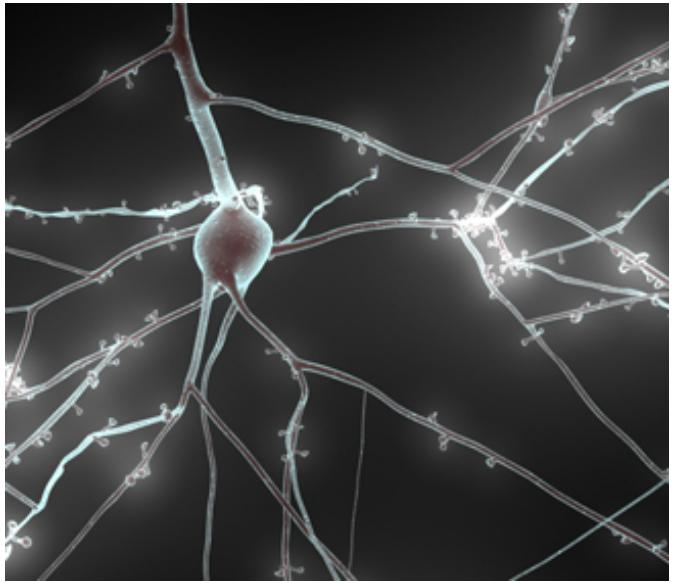




ADSP Category: Tier-1

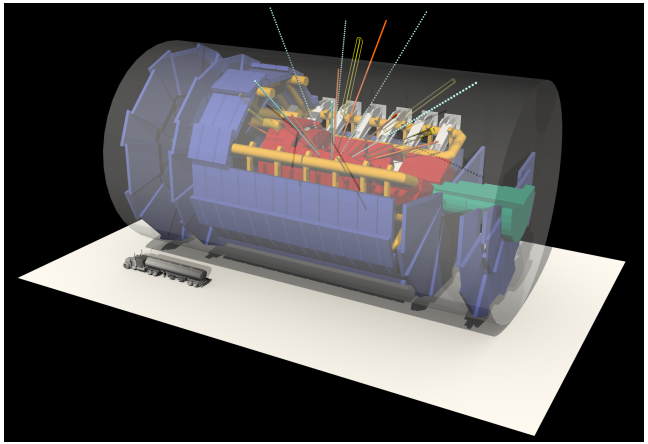
Leveraging Non-Volatile Memory, Big Data and Distributed Workflow Technology to Leap Forward Brain Modeling

- PI: Fabien Delalondre, EPFL, Switzerland
- **Objectives:** To facilitate and support complex computational neuroscience workflows by integration of new data storage paradigm, run times and big data technology
- **Impact:** Simulation and analysis at unprecedented scale of brain tissue models on ADSP systems, paving the way for future brain research and neuroscience breakthroughs
- **Approach:** Scale three components of the pipeline consisting of data management, data analysis and workflow management to fully use the ADSP resources.



ADSP Category: Tier-2

Accelerating LHC simulation workflows through adaptation for leadership systems



- **PI:** Taylor Childers, Argonne National Laboratory
- **Other Participating Institutions:** LBNL, Duke Univ., Univ. of Wisconsin
- **Objectives:** An end-to-end workflow to manage the data motion and job management to facilitate the ATLAS detector simulation on ADSP resources to process proton collision events.
- **Impact:** Leadership resources increases the analysis reach of LHC scientists enabling the discovery of new particle physics
- **Approach:** Develop an optimized workflow on ADSP resources to significantly accelerate the event generation and simulation on next-generation leadership systems.

ADSP Category: Tier-2