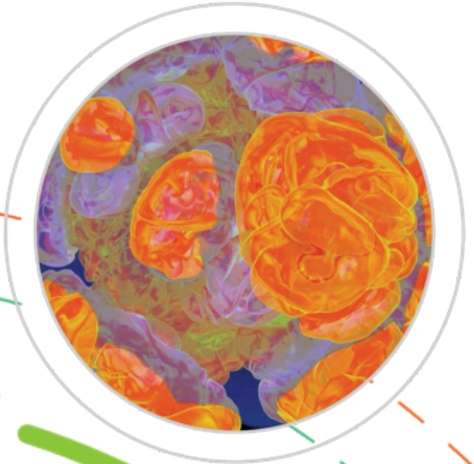
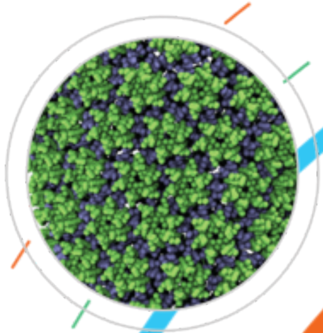
ALCF Early Science Program for Xeon Phi Supercomputers

Tim Williams
ALCF Deputy Director of Science

September 20, 2016 - IXPUG 2016 - ANL



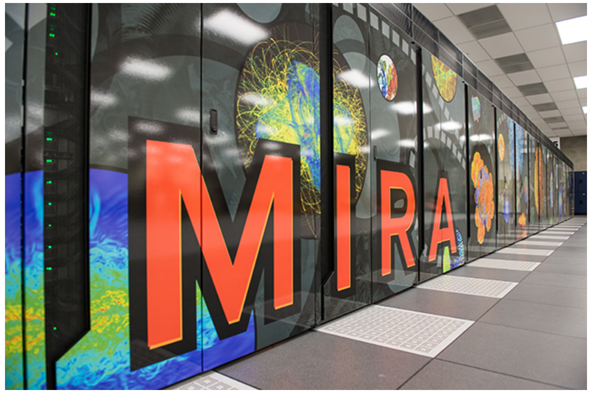


IXPUG
THE INTEL® XEON PHI™ USERS GROUP

Argonne Leadership Computing Facility

Argonne NATIONAL LABORATORY

Argonne Leadership Computing Facility



Mira

2012

IBM Blue Gene/Q

49,152 Power A2 nodes

10 petaFLOPS

768 TB RAM

Aurora

2018

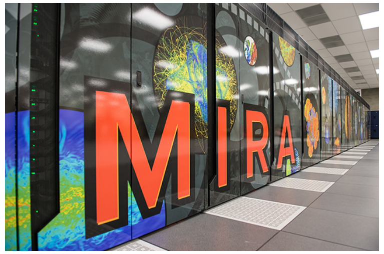
Intel-Cray system

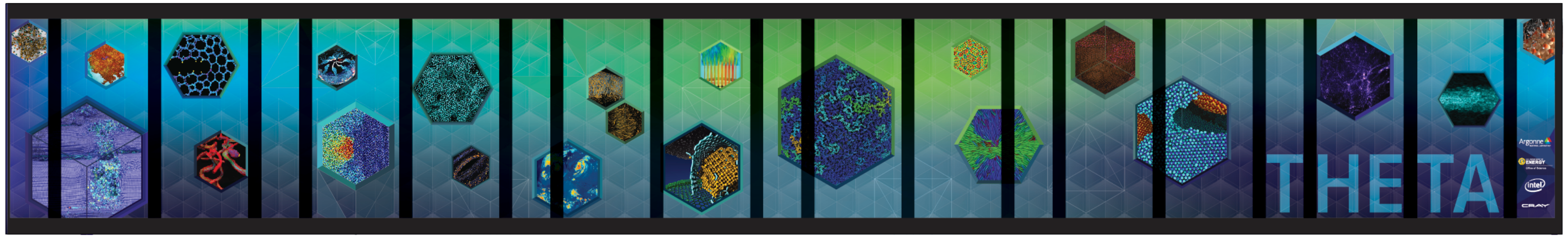
>50,000 KNH nodes

180 petaFLOPS

>7 PB {HBM + DRAM + NVM}

Argonne Leadership Computing Facility





Theta

2016

Intel-Cray system

3240 KNL-7230 nodes

8.62 petaFLOPS

673 TB {HBM + DRAM}

ALCF Early Science Program

Applications Readiness

- Prepare applications for next-gen system:
  - Architecture
  - Scale
- ~Two year lead time

Proposals

- Ambitious targeted science calculation
- Parallel performance
- Development needed
- Team

Support

PEOPLE

- Funded ALCF postdoc
- Catalyst staff member support
- Vendor experts

TRAINING

- Training on HW and programming
- Community workshop to share lessons learned

COMPUTE RESOURCES

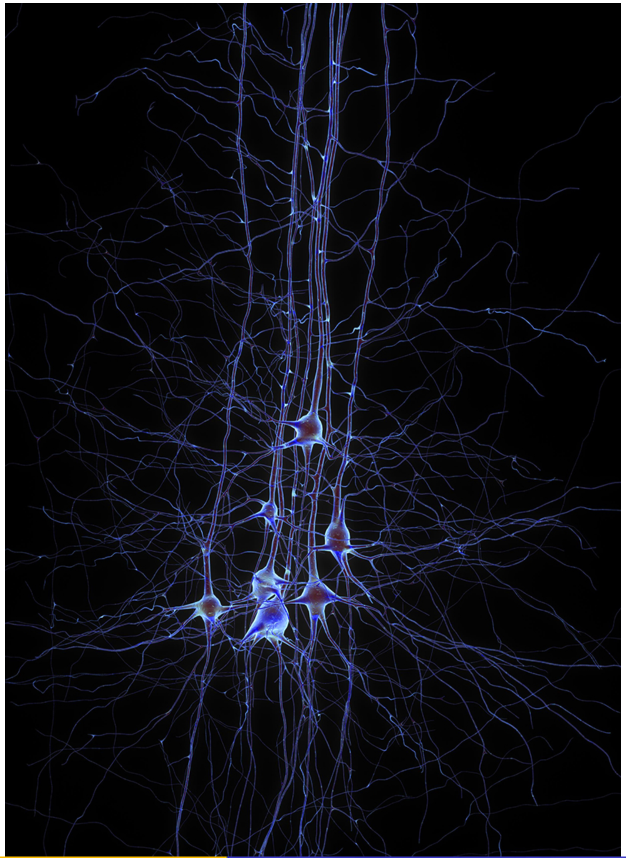
- Current ALCF systems
- Early next-gen hardware & simulators
- 3 months dedicated Early Science access
  - Pre-production (post-acceptance)
  - Large time allocation
  - Continued access for rest of year

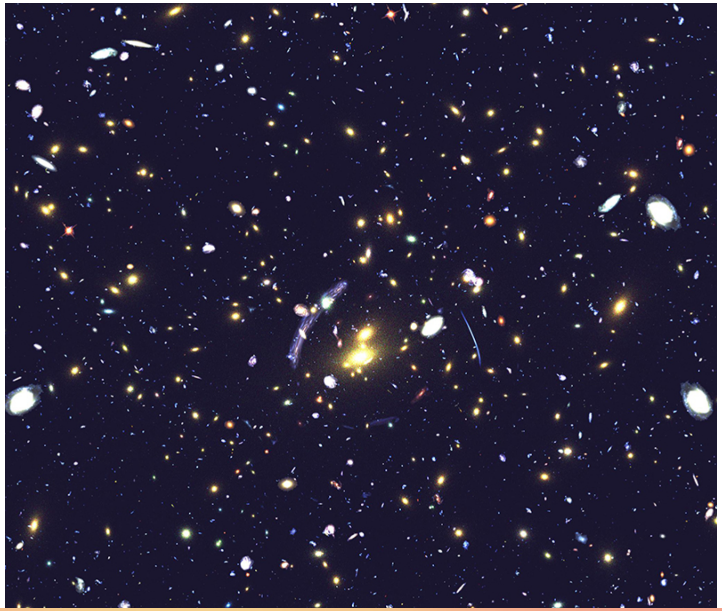
http://esp.alcf.anl.gov

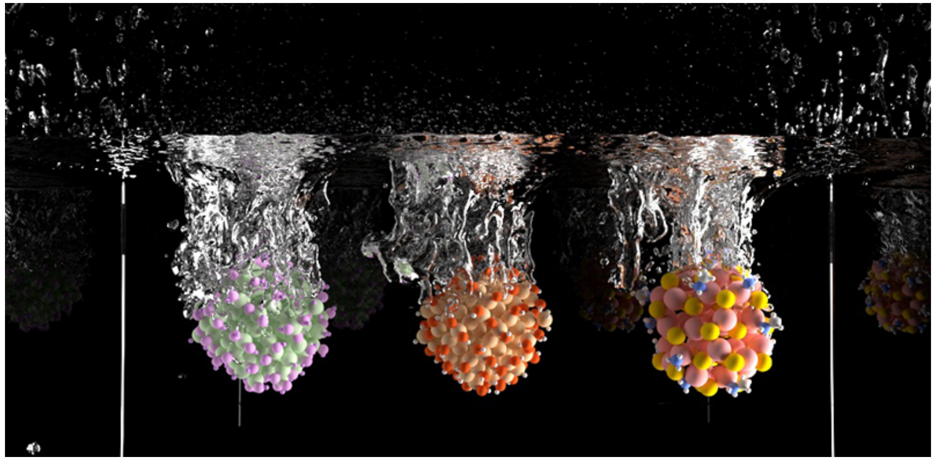
ESP Timeline

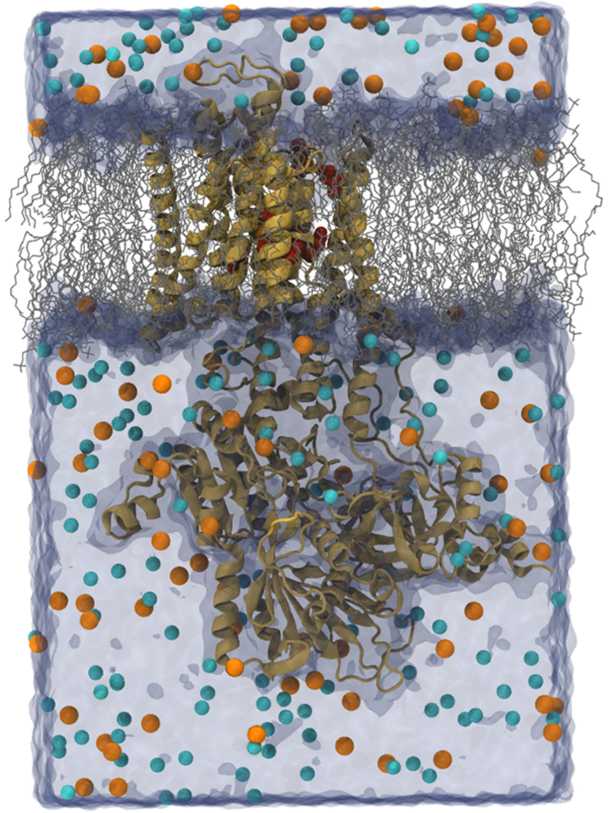
Task	CY2015				CY2016				CY2017				CY2018				CY2019			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q4	Q4
Theta CFP																				
Theta selection																				
Theta ESP projects																				
Theta Early Science																				
Aurora CFP																				
Aurora selection																				
Aurora ESP projects																				
Aurora Early Science																				
Mira production																				
Theta production																				
Aurora production																				

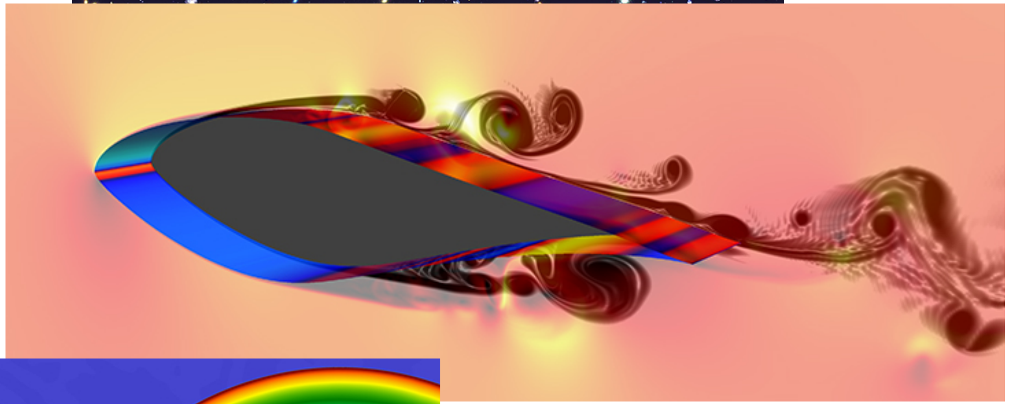
Theta ESP Projects

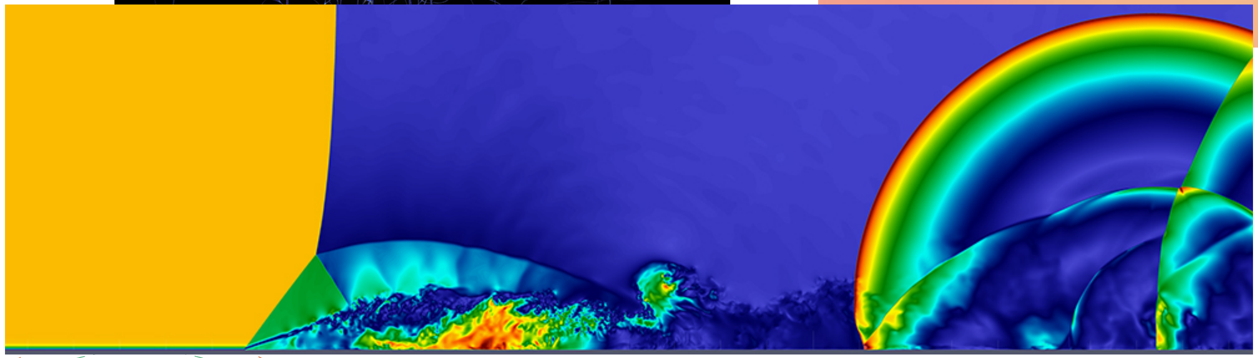




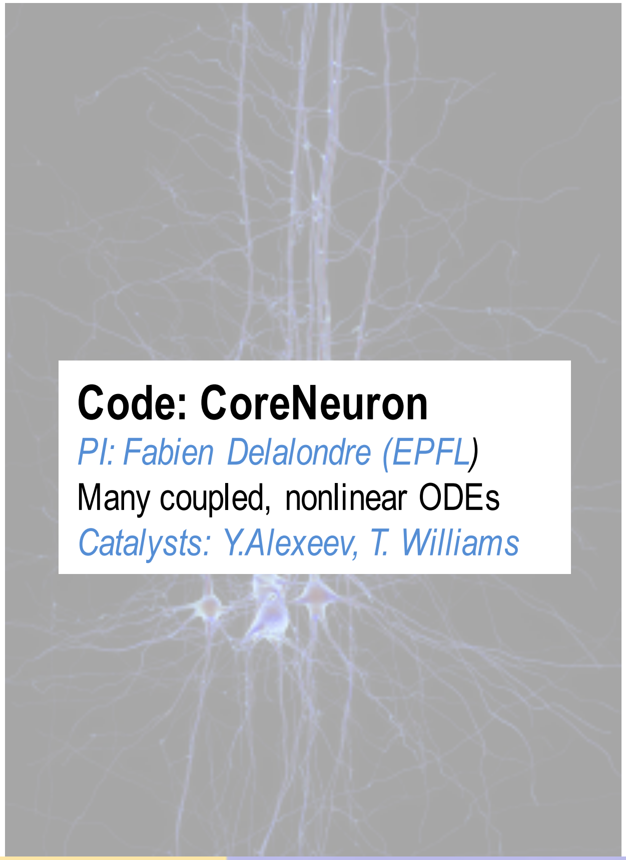




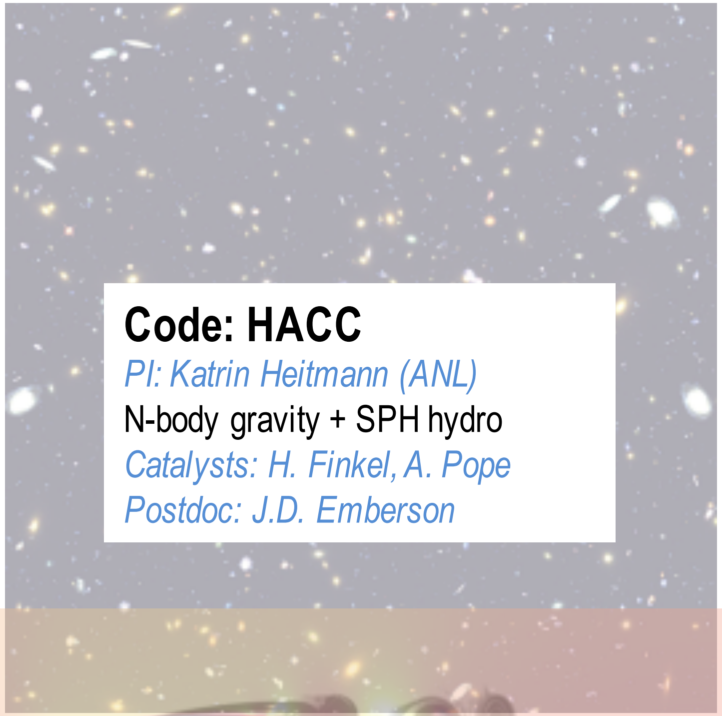




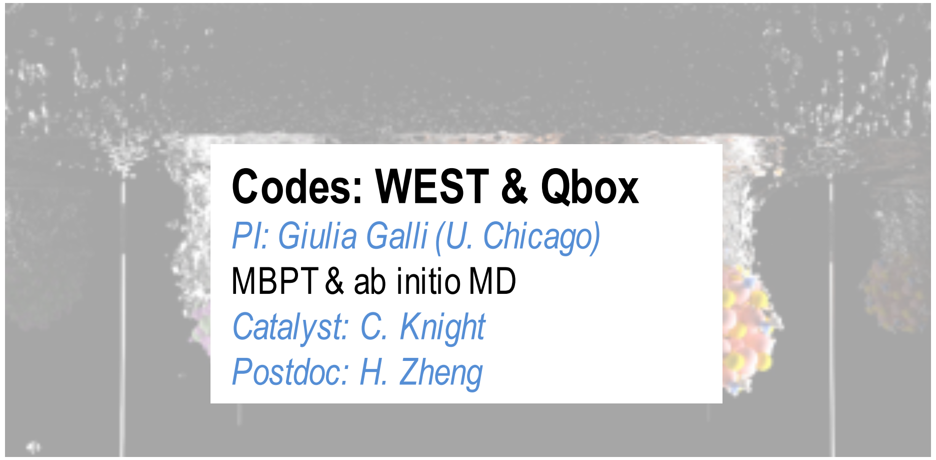
Theta ESP Projects



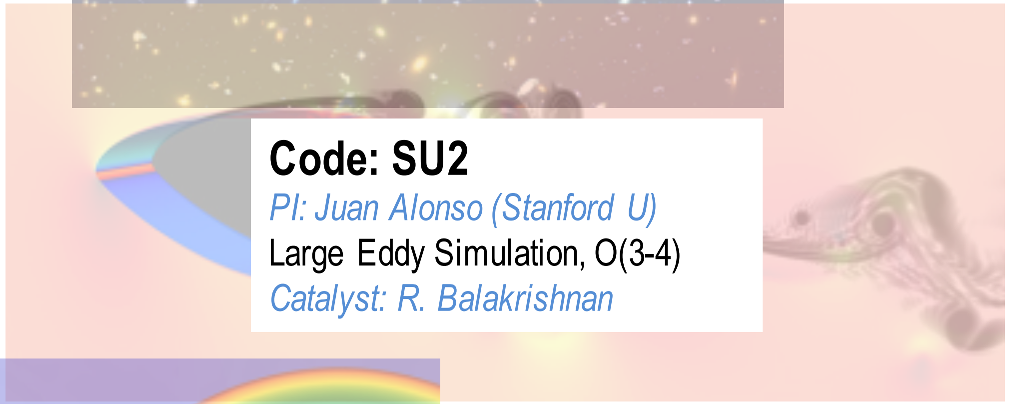
Code: CoreNeuron
PI: Fabien Delalondre (EPFL)
Many coupled, nonlinear ODEs
Catalysts: Y.Alexeev, T. Williams



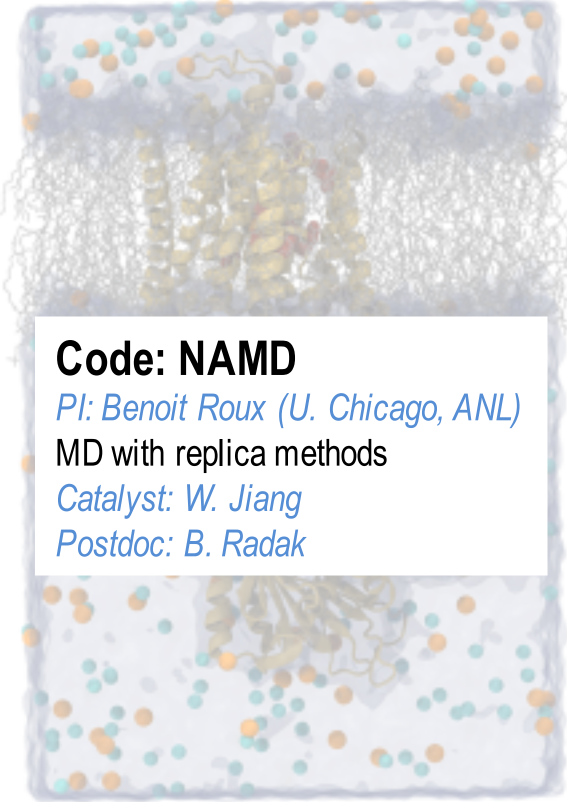
Code: HACC
PI: Katrin Heitmann (ANL)
N-body gravity + SPH hydro
Catalysts: H. Finkel, A. Pope
Postdoc: J.D. Emberson



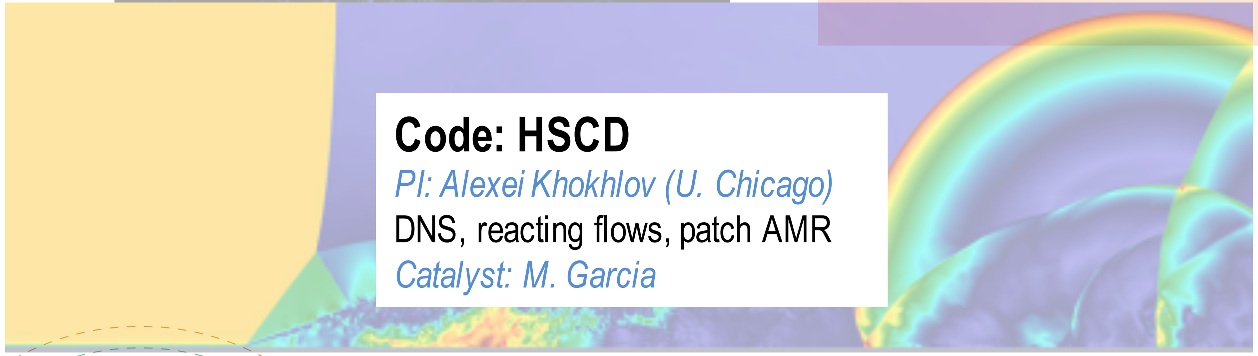
Codes: WEST & Qbox
PI: Giulia Galli (U. Chicago)
MBPT & ab initio MD
Catalyst: C. Knight
Postdoc: H. Zheng



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

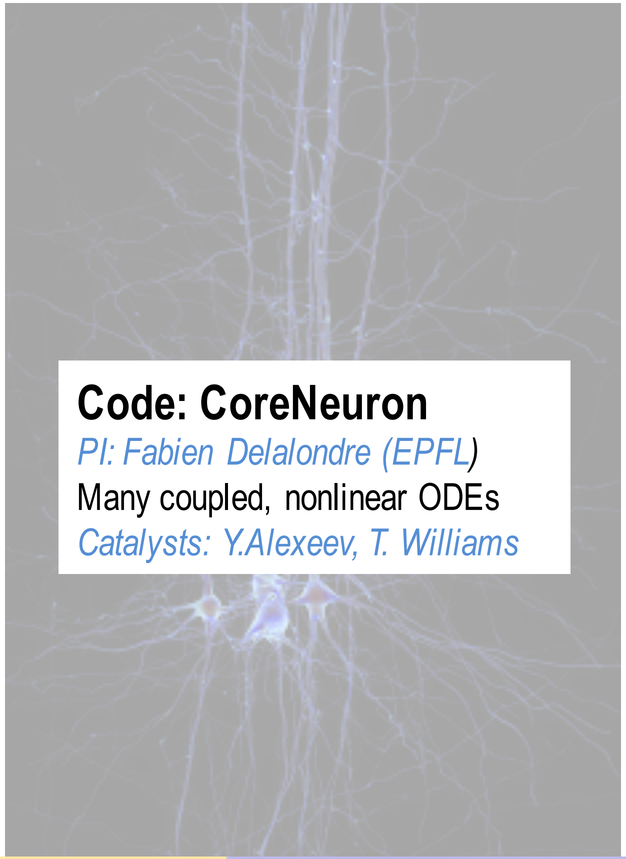


Code: NAMD
PI: Benoit Roux (U. Chicago, ANL)
MD with replica methods
Catalyst: W. Jiang
Postdoc: B. Radak



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

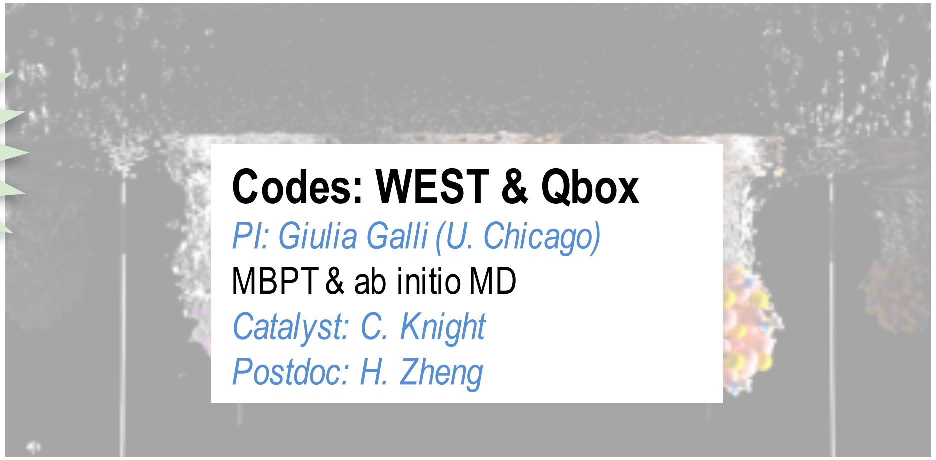
Theta ESP Projects



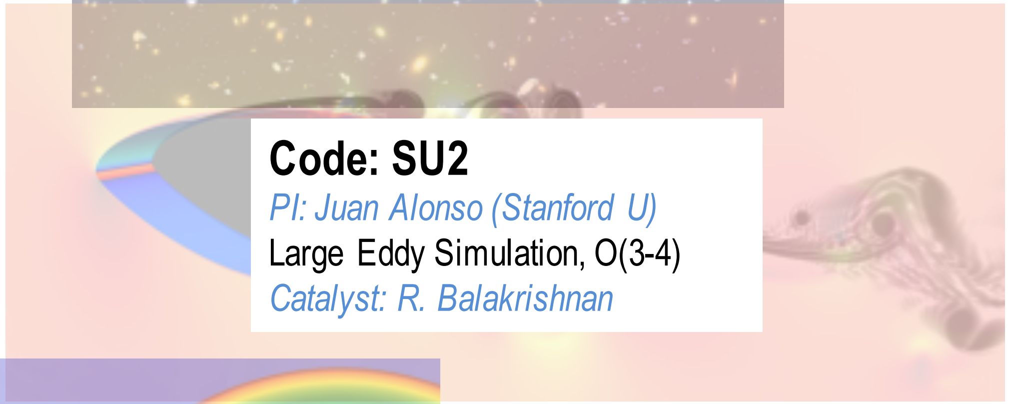
Code: CoreNeuron
PI: Fabien Delalondre (EPFL)
Many coupled, nonlinear ODEs
Catalysts: Y.Alexeev, T. Williams

Finkel+
Wed. 3:50

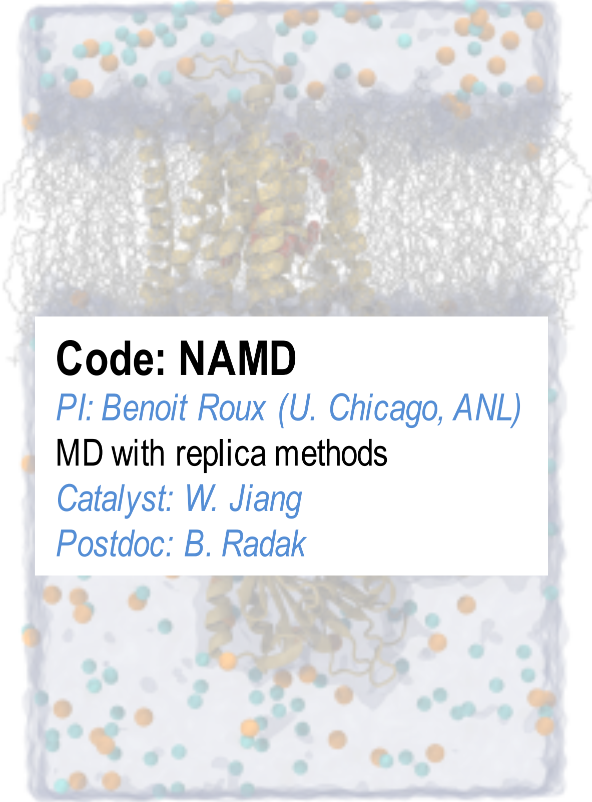
Code: HACC
PI: Katrin Heitmann (ANL)
N-body gravity + SPH hydro
Catalysts: H. Finkel, A. Pope
Postdoc: J.D. Emberson



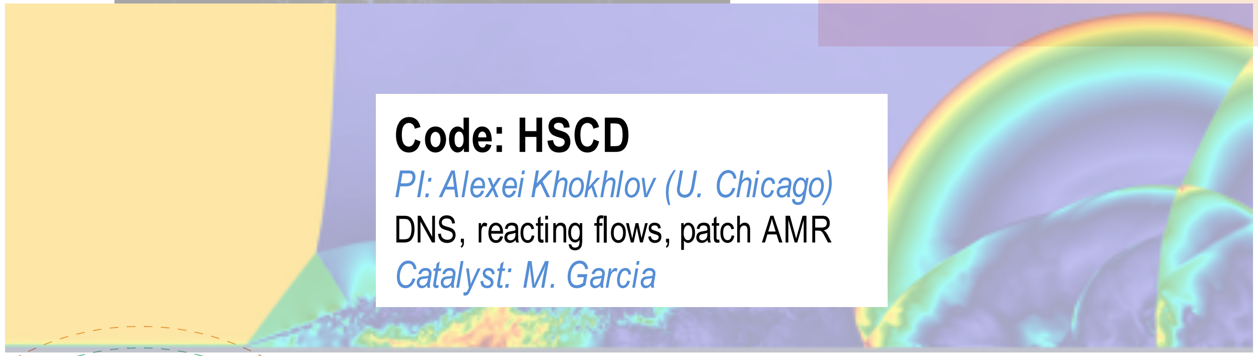
Codes: WEST & Qbox
PI: Giulia Galli (U. Chicago)
MBPT & ab initio MD
Catalyst: C. Knight
Postdoc: H. Zheng



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

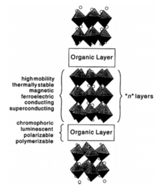


Code: NAMD
PI: Benoit Roux (U. Chicago, ANL)
MD with replica methods
Catalyst: W. Jiang
Postdoc: B. Radak

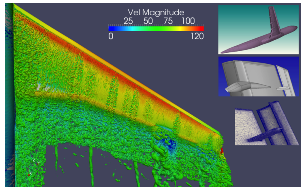


Code: HSCD
PI: Alexei Khokhlov (U. Chicago)
DNS, reacting flows, patch AMR
Catalyst: M. Garcia

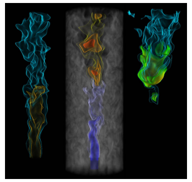
Tier 2 Theta ESP Projects



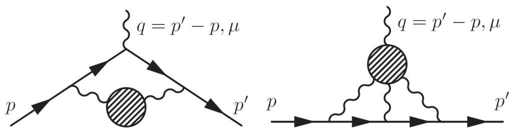
Codes: FHI-Aims & GAtor
PI: Volker Blum (Duke U.)
MBPT (DFT) & genetic algorithm
Catalyst: Álvaro Vázquez-Mayagoitia



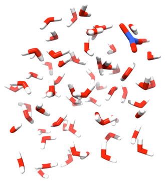
Code: PHASTA
PI: Kenneth Jansen (U. Colorado)
CFD, unstructured mesh
Catalyst: Hal Finkel



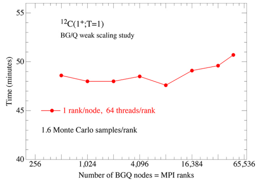
Code: Nek5000
PI: Christos Frouzakis (ETHZ)
Spectral element CFD with combustion
Catalyst: Scott Parker



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

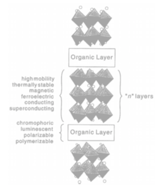


Code: GAMESS
PI: Mark Gordon (Iowa State U.)
FMO - quantum chemistry
Catalysts: Yuri Alexeev, Graham Fletcher

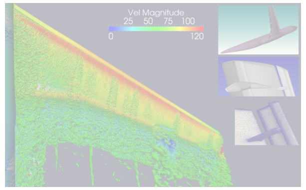


Code: GFMC
PI: Steven Pieper (ANL)
Greens Function Monte Carlo – nuclear
Catalyst: James Osborn

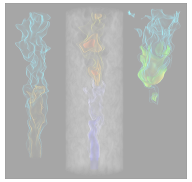
Tier 2 Theta ESP Projects



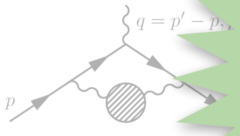
Codes: FHI-Aims & GAtor
PI: Volker Blum (Duke U.)
MBPT (DFT) & genetic algorithm
Catalyst: Álvaro Vázquez-Mayagoitia



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



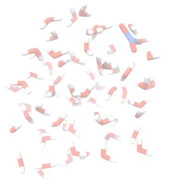
Code: Nek5000
PI: Christos Frouzakis (ETHZ)
Spectral element CFD with combustion
Catalyst: Scott Parker



Codes: MILC & CPS
I: Paul Mackenzie
attice QCD
atalyst: James Osbo

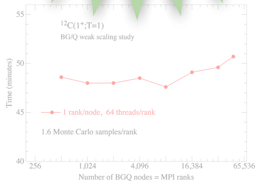
Osborn
Wed. 11:15

Li+
Wed. 3:50



Code: GAMESS
PI: Mark Gordon (Io
FMO - quantum chem
Catalysts: Yuri Alexee

Alexeev+
Wed. 11:15



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

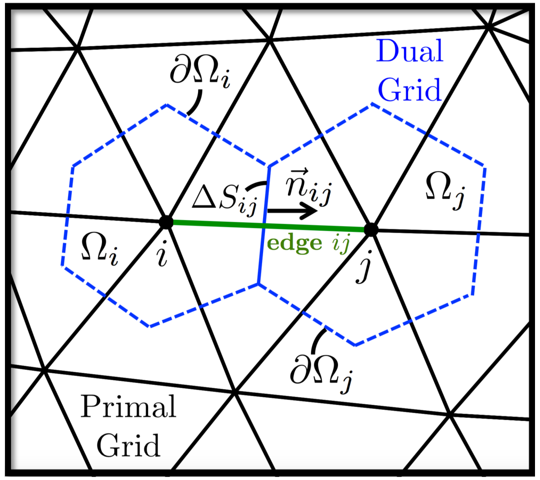
ALCF Theta ESP Hands-on Workshop

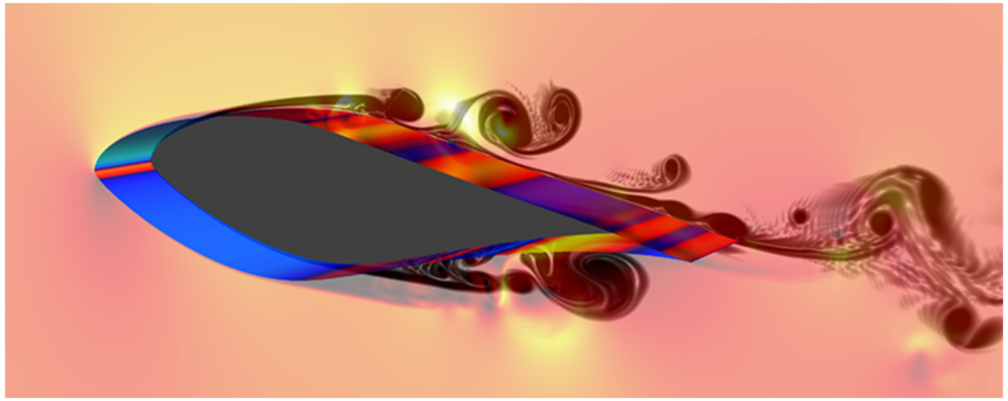
- 16-19 August 2016
- Developers from all 12 ESP projects
- Intel and Cray applications experts
- Two mornings of Intel, Cray developer environment presentations
- Hands-on
  - Profile, analyze, tune
  - Scaling studies

Scale-Resolving Simulations of Wind Turbines with SU2

Renewable Energy, Engineering, CFD
Codes: SU2

- PI: Juan J. Alonso (Stanford)
- Large Eddy Simulation (LES) of a few turbines plus tower
  - Third order finite volume
  - High order discontinuous Galerkin
- LES results feed reduced-order Kinematic Simulation for wind farm design
- SU2 evolving into high-end *open source* CFD package (community code)
  - Finite volume methods
  - Unstructured mesh





Gaurav Bansal, Anand Deshpande, Dheevatsa Mudigereb, others (Intel); Tom Economon, Francisco Palacios (Stanford)

Using Xeon Phi Features

Threads

1. Loop-level
2. **Single OMP parallel region at high level in program**

- 4 hot spots (edge loops)
  - Thread "owner" of edges
  - Edges touching shared node replicated (halo)
  - Many vert updates: static
  - Few vert updates: dynamic

Vectorization

- Outer loop (edges)
- Loop tiling

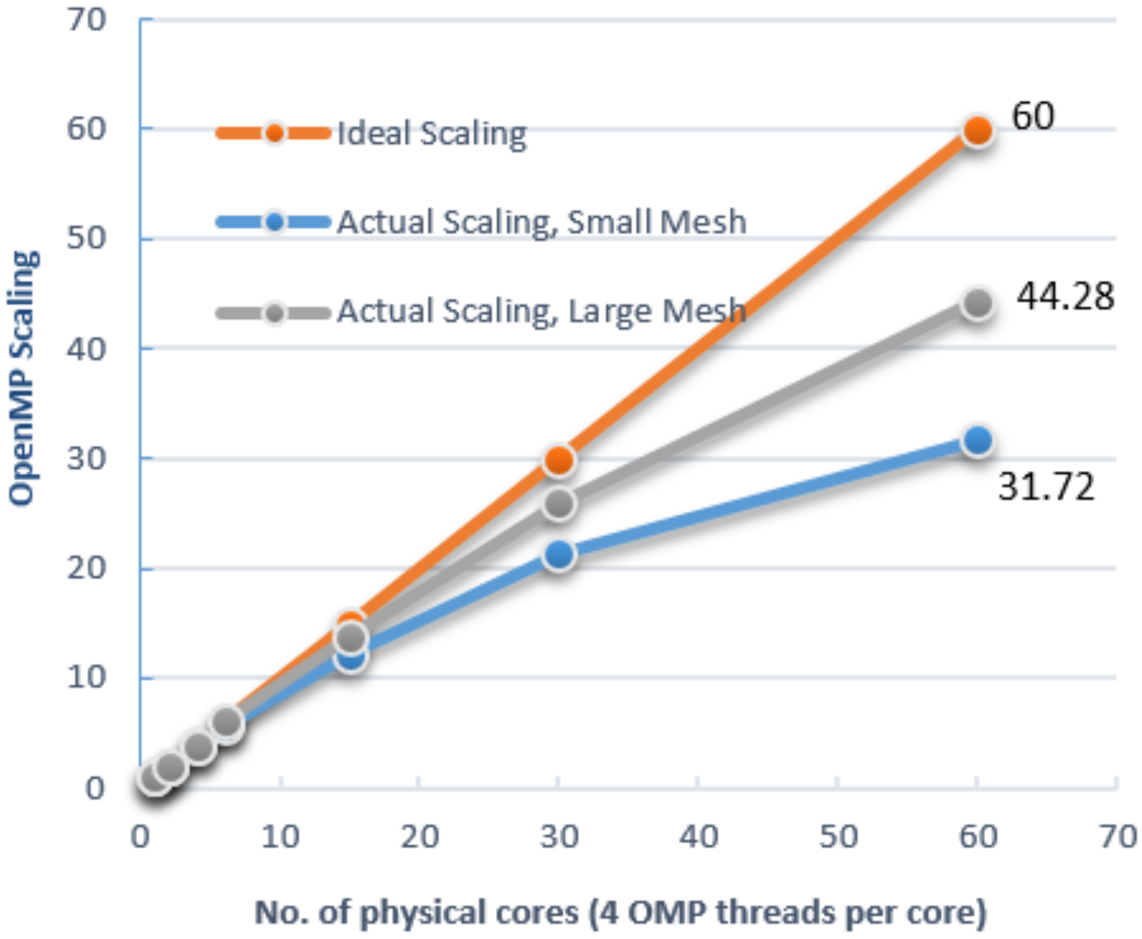
```
for (iEdge = 0; iEdge < nEdges; iEdge += VECSIZE) {
    for (ivec = 0; ivec < VECSIZE, ++ivec) {
```

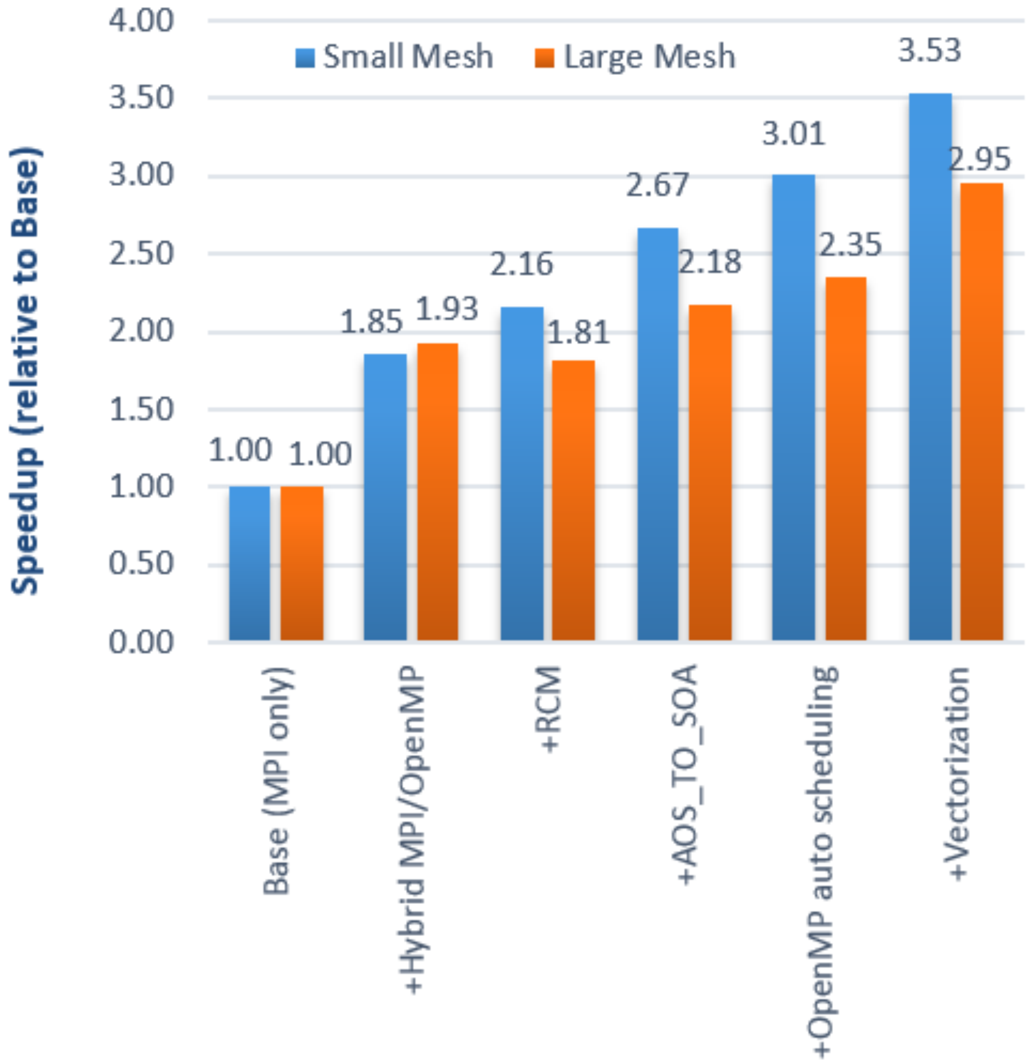
Memory Hierarchy

- Edge/vertex reordering
- Smart allocation
- Change AoS to SoA

http://dx.doi.org/10.1016/j.compfluid.2016.02.003 (Computers & Fluids, 2016)
http://dx.doi.org/10.2514/6.2015-1949 (AIAA, 2015)
https://software.intel.com/en-us/articles/high-performance-modern-code-optimizations-for-computational-fluid-dynamics (Intel Developer Zone, 2015)

KNC Performance





https://software.intel.com/en-us/articles/high-performance-modern-code-optimizations-for-computational-fluid-dynamics (Intel Developer Zone, 2015)

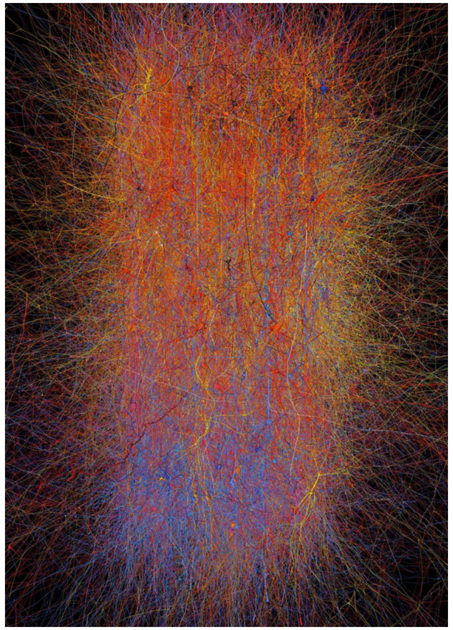
KNL Findings

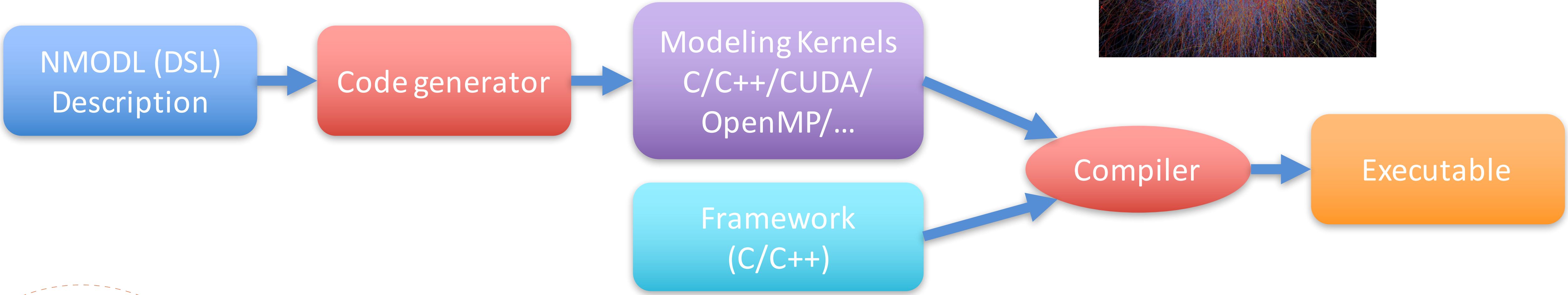
- All optimizations from KNC effort carried over to good effect on KNL
  - (thread scheduling still under study)
- Speedups on Broadwell also
- Additional KNL boost: run entirely in MCDRAM
- Altogether, optimizations led to >9X speedup on KNL

Large Scale Simulation of Brain Tissue: Blue Brain Project

Neuroscience
Codes: CoreNeuron

- PI: Fabien Delalondre (EPFL)
- Large set of nonlinear ODEs
  - Backward Euler time discretization
  - Spatial discretization: cylindrical *compartment* elements
  - Small linear system per neuron—quasi-tridiagonal





Four Science Cases

1. Microcircuit plasticity
   - Experience-dependent changes in synaptic connectivity
   - May be substrate for learning and memory
- A few neocortical columns:
   - 31,000 neurons
   - 37 million synapses

2. Neurorobotics
   - Electrical activity of rodent somatosensory cortex
     - Morphologically detailed neurons
     - 20 million neurons, 20 billion synapses
   - Embed model in simulated body
     - Study activity & plasticity in closed action/perception loop

3. Compare analysis of simulation results w/experiment
   - Electrical activity of rodent somatosensory cortex
     - Morphologically detailed neurons
     - 20 million neurons, 20 billion synapses
   - Increase accuracy & validity of model

4. Largest possible brain model on *Theta*
- Several seconds of biological time

Using KNL Features

Memory hierarchy

- Case 1 – mostly fit in L2 cache
- Case 2 – mostly fit in MCDRAM
- Case 3 – use DRAM
  - Cache mode
  - Flat: memory bound kernels in MCDRAM
    compute bound kernels in DRAM
- Case 4 – use node-local SSD

Vectorization

- Mostly memory bound, limits importance
- DSL generates
  1. Compiler directives
     - #pragma vector nontemporal
     - #pragma ivdep
  2. Vector intrinsic function calls
     not yet needed.

Threads

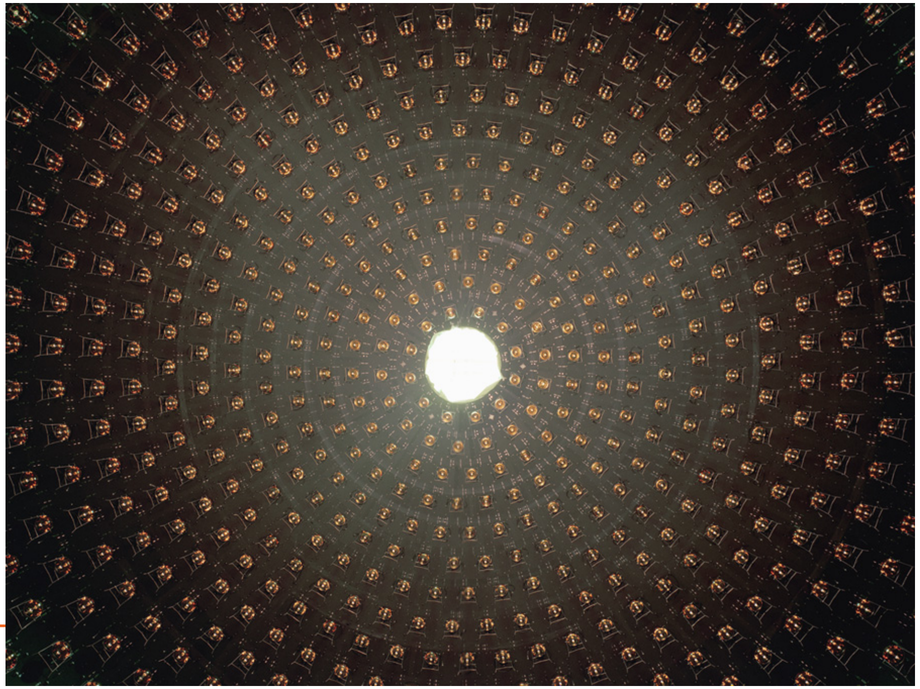
- Blue Gene/Q node: 1 MPI rank, 64 OpenMP threads
- Knights Corner Xeon Phi coprocessor: 1 rank, 120 threads
- More parallelism for case 1: across compartments

Quantum Monte Carlo Calculations in Nuclear Theory

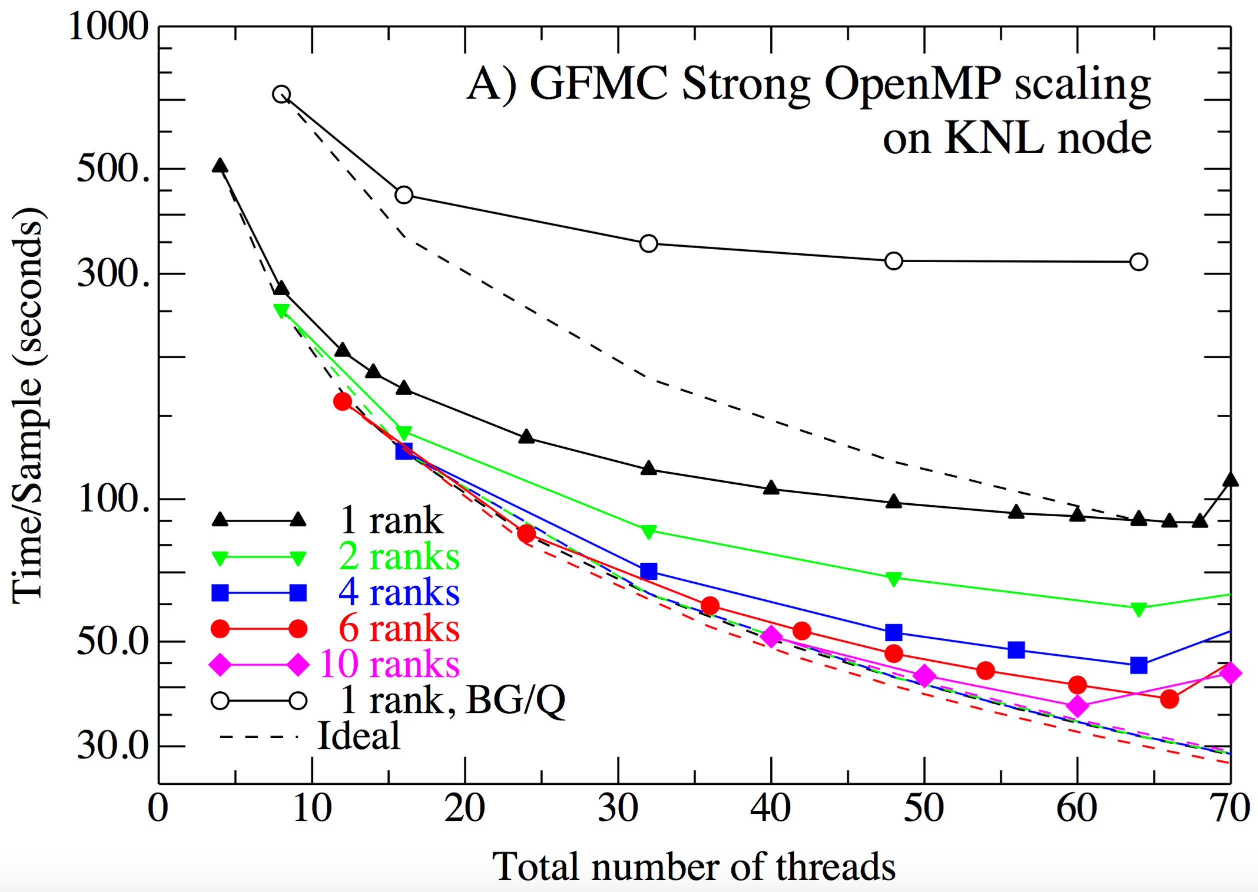
Nuclear Physics
Codes: GFMC

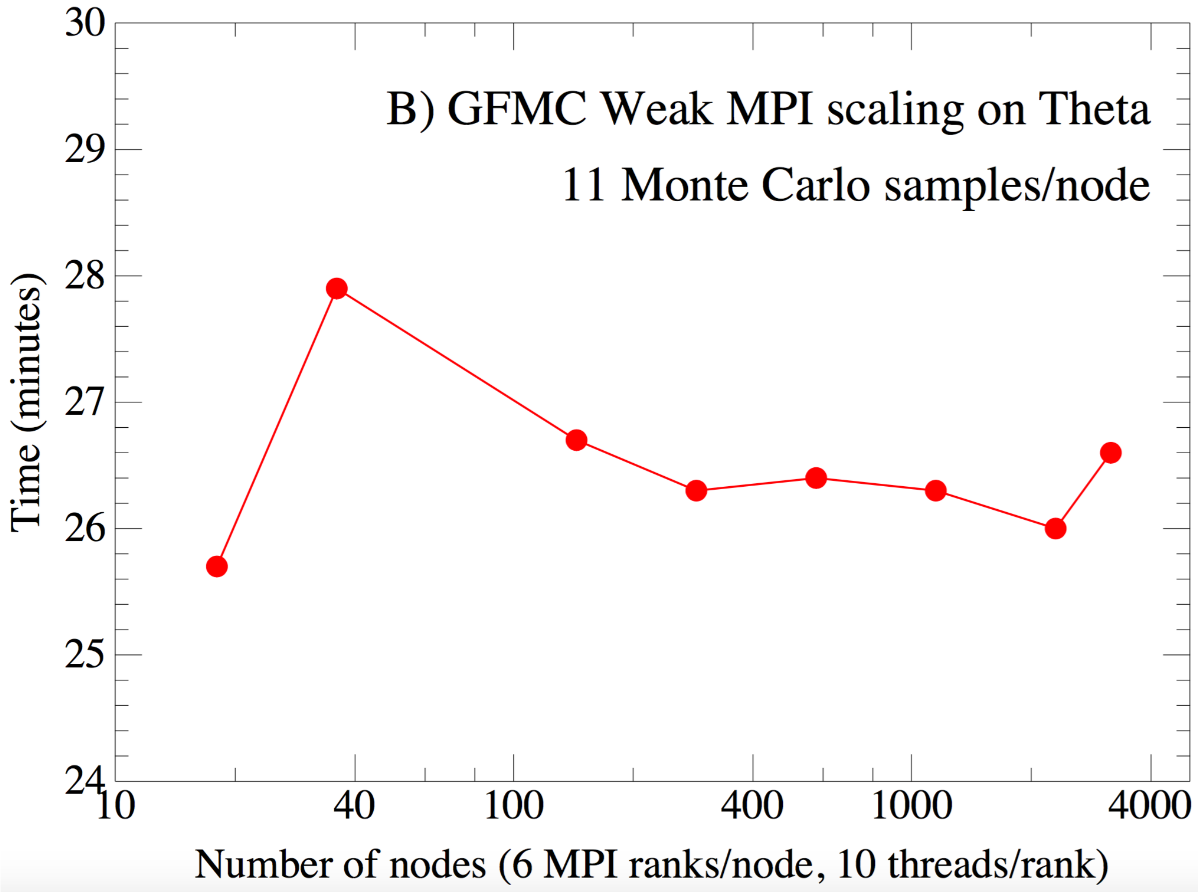
- PI: Steven Pieper (ANL)
- Parallelism:
  - ADLB library (Lusk, ANL)
  - DMEM distributed memory mgt system (MPI)
- Further plans:
  - Repetitive matrix-vector operations in HBM
  - Use MPI-3 shared memory
- Issues
  - Several OpenMP bugs (1 submitted)
  - Cray environment learning curve
    - Requires env var to allow MPI_THREAD_MULTIPLE

- Science
  - Greens Function Monte Carlo: solve many-body Schrödinger equation for light nuclei
  - ^{12}C charged-current response (MiniBooNE exp)
  - β decay in ^{14}O and ^{14}C



GFMC Performance Snapshot



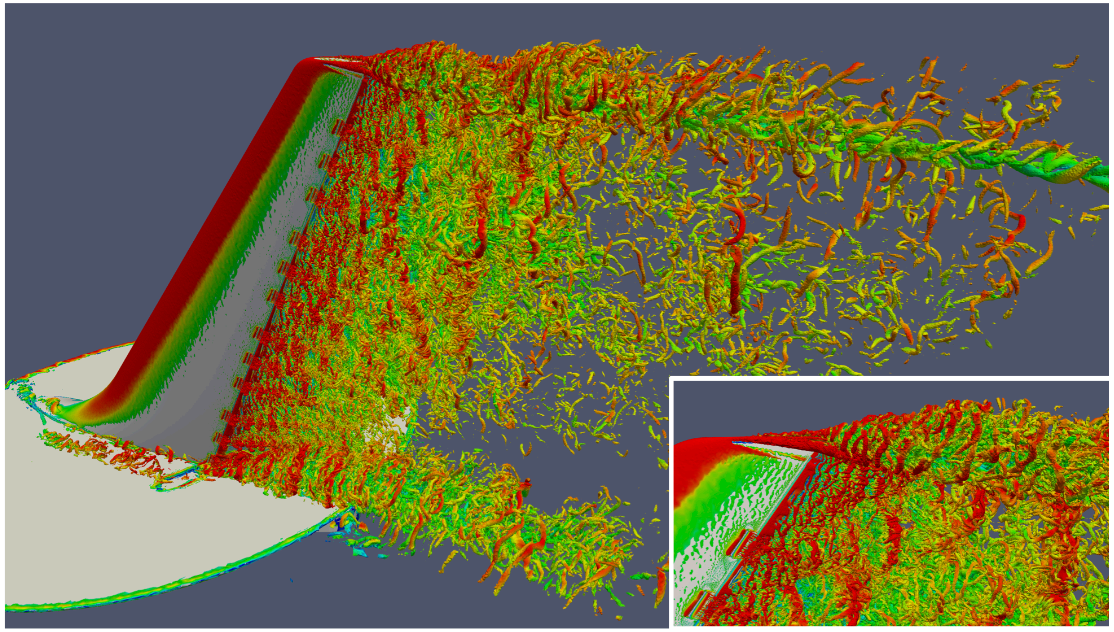


Extreme Scale Unstructured Adaptive CFD: From Multiphase Flow to Aerodynamic Flow Control

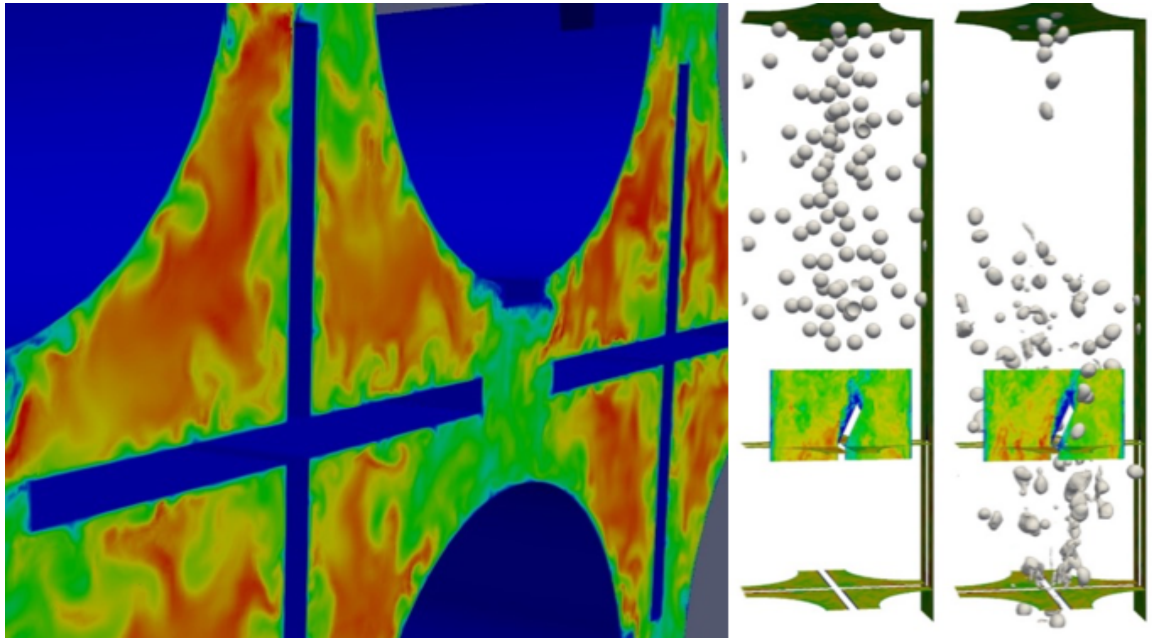
CFD, Aerodynamics, Nuclear Energy

Codes: PHASTA

- PI: Ken Jansen (U. Colorado)
- Navier-Stokes: compressible/incompressible, turbulent, unsteady
- 3D finite element, unstructured adaptive mesh
- Fully implicit in time



Active Flow control on vertical tail



Turbulent multiphase flow in reactor

Developers: Colorado, RPI, NCSU, Stanford, Michigan, UT-Austin, Berkeley, UCSD, USF, INRIA, and multiple industries.

PHASTA Parallel Strategy

- 2 key work components to implicit solver:
  - Equation formation (element loops sub-blocked for optimal cache-vectorization balance),
  - Equation solution (dominated by Sparse Ap, BLAS{1,2}).
- Base: Pure MPI (scaled to 3M processes).
- OpenMP: explore hybrid parallelism (Theading element block loops, Ap loops) => 80% efficiency.
- Advanced Dev: implementation in MPI Endpoints, MPI3.0+shared memory windows, and XSI shmem for on-node parallelism .
- Vectorization: Aggressive tuning under VTUNE and Advisor confirms high degree of vectorization: already 5x Mira core performance -> clear path to 2x more.
- 2 efficient solvers continuously improved: PETSc and native+mkl.

PHASTA Scaling on Theta

- 10B and 80B element mesh: August workshop Pure MPI scale out
  - 10B case scaled >95 % to 128Ki cores > 82% to 192Ki cores (relative to 16Ki cores). 51k elements per core
  - 80B case perfect to 192Ki cores; fits in MCDRAM with 1.2M elm. per core
- Proof that Pure MPI will scale to full Theta machine on the first Detached Eddy Simulation of active flow control at a ½ Flight Scale, vertical tail/rudder.