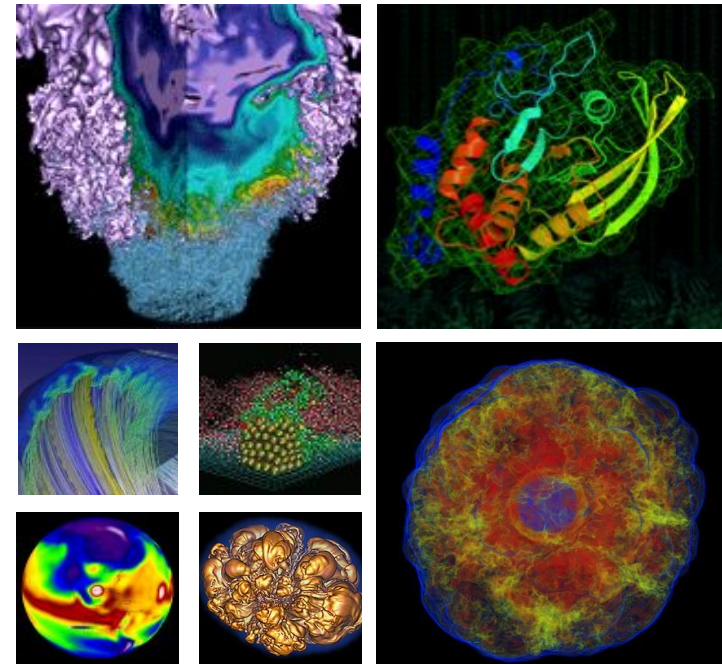


Many-Cores for the Masses: Lessons from 2 Years With the Cori System at NERSC



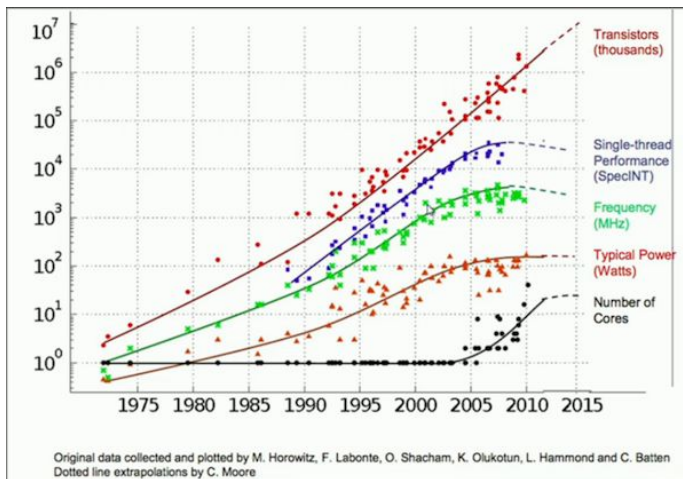
Jack Deslippe

Feb 2019

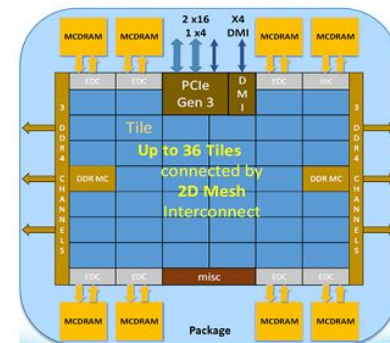
Change Has Arrived



Driven by power consumption and heat dissipation toward lightweight cores



Knights Landing Overview

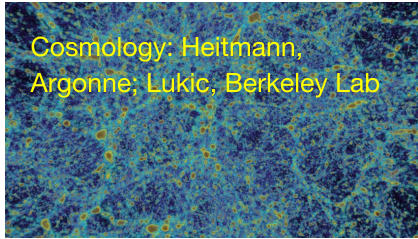


KNL: 215-230 W

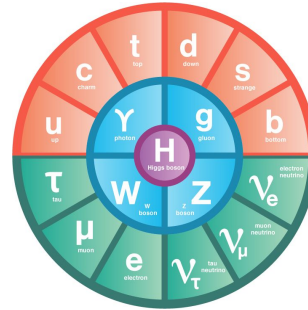
2-socket Haswell: 270 W

Cori, a 30 PFlop system, is an important resource to science in the U.S. because of new capabilities, but the Intel Xeon Phi many-core architecture will require a code modernization effort to use efficiently.

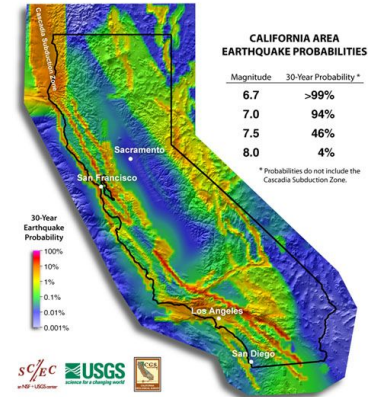
High Impact Science at Scale on Cori



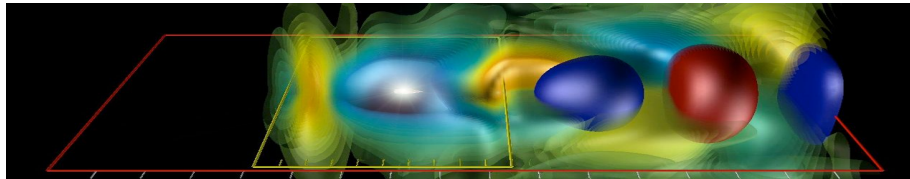
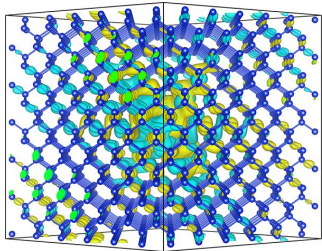
Strangeness
and Electric
Charge
Fluctuations in
Strongly
Interacting
Matter, Karsch,
Brookhaven



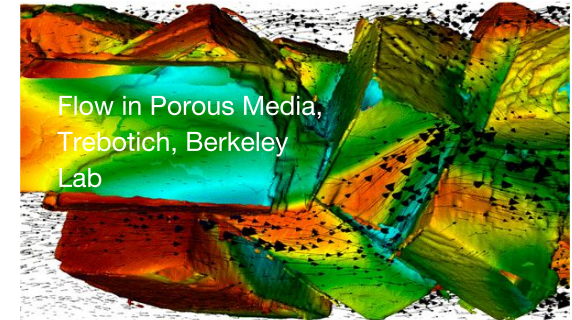
M8
Earthquake
on the San
Andreas
Fault, Goulet,
USC
Earthquake
Center



Optical Properties of Materials,
Louie, UC Berkeley



Asymmetric Effects in
Plasma Accelerators,
Vay, Berkeley Lab



Deep Learning on Cori KNL

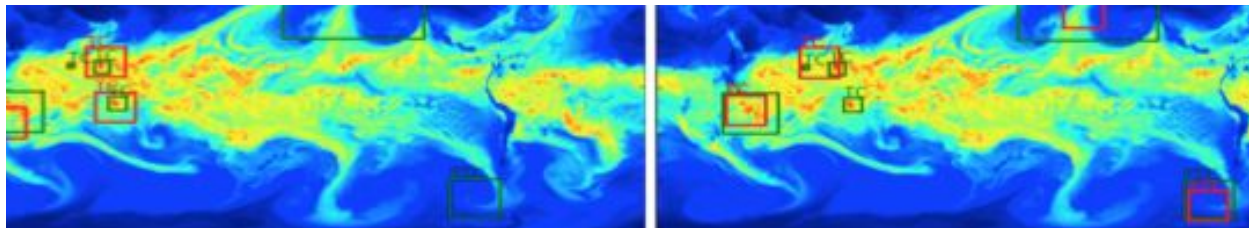


NERSC is actively exploring Deep Learning for Science

- Collaborating with leading vendors to optimize and deploy stack
- Collaborating with leading research institutions to develop methods
- Drive real science use cases

Deep Learning at 15 PF on NERSC Cori (Cray + Intel KNL)

- Trained in 10s of minutes on 10 terabyte datasets, millions of Images
- 9600 nodes, optimized on KNL with IntelCaffe and MKL (NERSC / Intel collaboration)
- Synch + Asynch parameter update strategy for multi-node scaling (NERSC / Stanford)



Identified extreme climate events using supervised (left) and semisupervised (right) deep learning. Green = ground truth, Red = predictions (confidence > 0.8). [NIPS 2017]

What is different about Cori?



Edison (“Ivy Bridge”):

- 5576 nodes
- 24 physical cores per node
- 48 virtual cores per node
- 2.4 - 3.2 GHz
- 8 double precision ops/cycle
- 64 GB of DDR3 memory (2.5 GB per physical core)
- ~100 GB/s Memory Bandwidth

Cori (“Knights Landing”):

- 9304 nodes
- 68 physical cores per node
- 272 virtual cores per node
- 1.4 - 1.6 GHz
- 32 double precision ops/cycle
- 16 GB of fast memory
96GB of DDR4 memory
- Fast memory has 400 - 500 GB/s
- No L3 Cache

Optimization Challenge and Strategy



Energy-Efficient Processors Have Multiple Hardware Features to Optimize Against:

- Many Cores
- Bigger Vectors
- New ISA
- Multiple Memory Tiers

It is easy for users to get bogged down in the weeds:

- How do you know what KNL hardware feature to target?
- How do you know how your code performs in an absolute sense and when to stop?

Optimization Challenge and Strategy



Energy-Efficient Processors Have Multiple Hardware Features to Optimize Against:

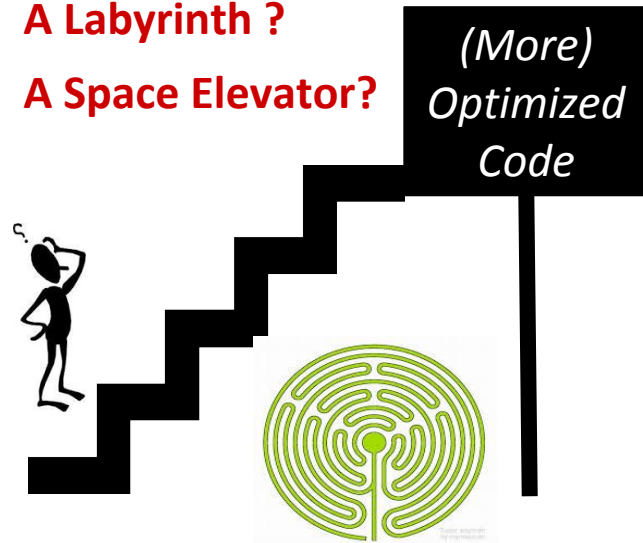
- Many Cores
- Bigger Vectors
- New ISA
- Multiple Memory Tiers

It is easy for users to get bogged down in the weeds:

- How do you know what KNL hardware feature to target?
- How do you know how your code performs in an absolute sense and when to stop?

Optimizing Code For Cori is Like?

- A. A Staircase ?
- B. A Labyrinth ?
- C. A Space Elevator?



Optimization Challenge and Strategy

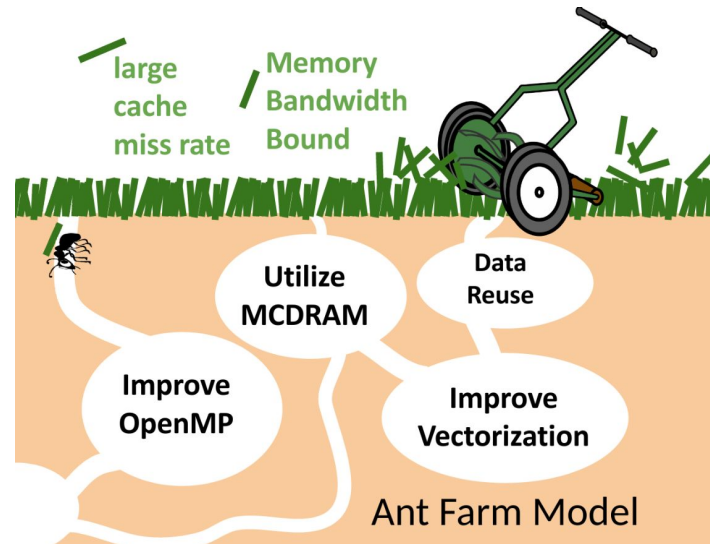


Energy-Efficient Processors Have Multiple Hardware Features to Optimize Against:

- Many Cores
- Bigger Vectors
- New ISA
- Multiple Memory Tiers

It is easy for users to get bogged down in the weeds:

- How do you know what KNL hardware feature to target?
- How do you know how your code performs in an absolute sense and when to stop?



Optimization Challenge and Strategy



Energy-Efficient Processors Have Multiple Hardware Features to Optimize Against:

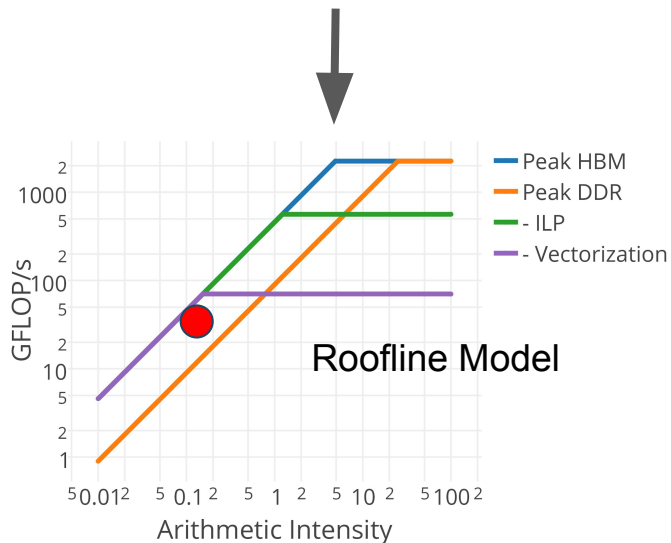
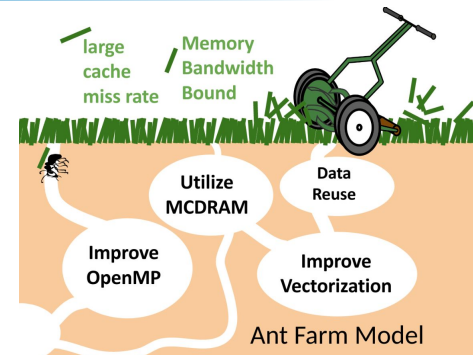
- Many Cores
- Bigger Vectors
- New ISA
- Multiple Memory Tiers

It is easy for users to get bogged down in the weeds:

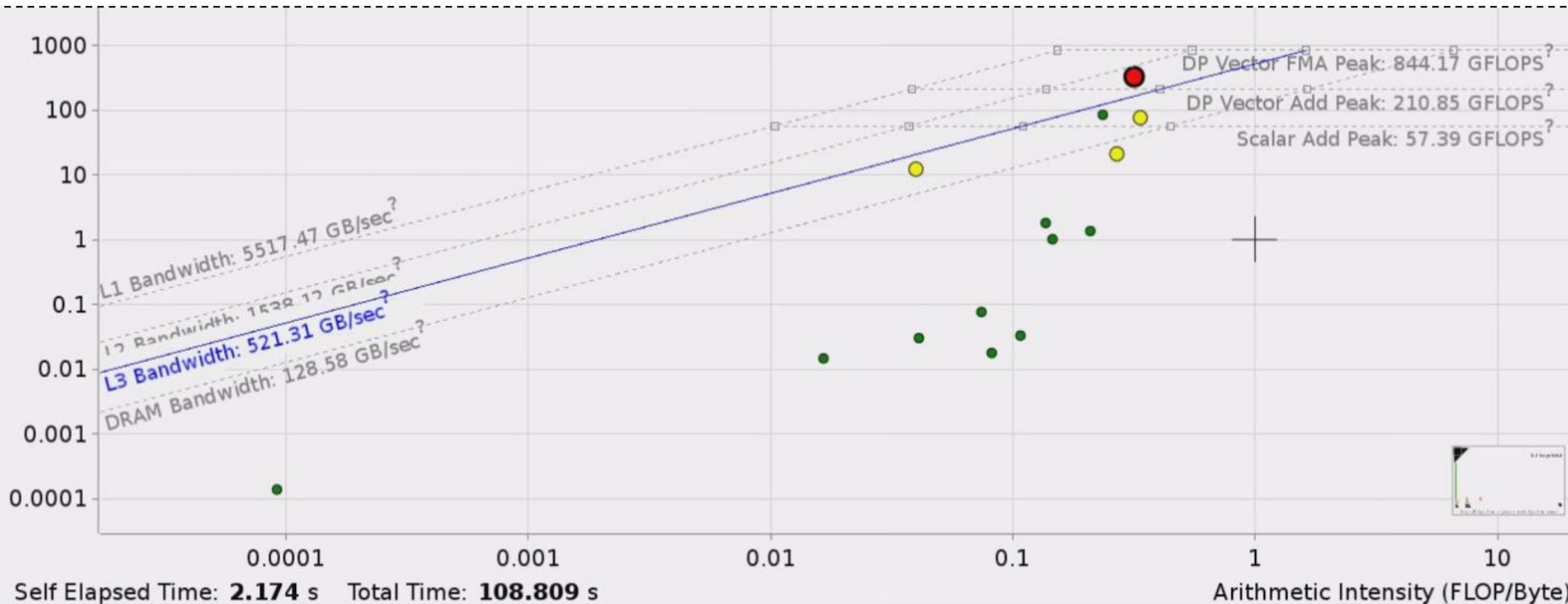
- How do you know what KNL hardware feature to target?
- How do you know how your code performs in an absolute sense and when to stop?

NERSC has developed tools and strategy for users to answer these questions:

- Designed simple tests that demonstrate code limits
- Use roofline as an optimization guide
- Training and documentation hub targeting all users



Tools CoDesign

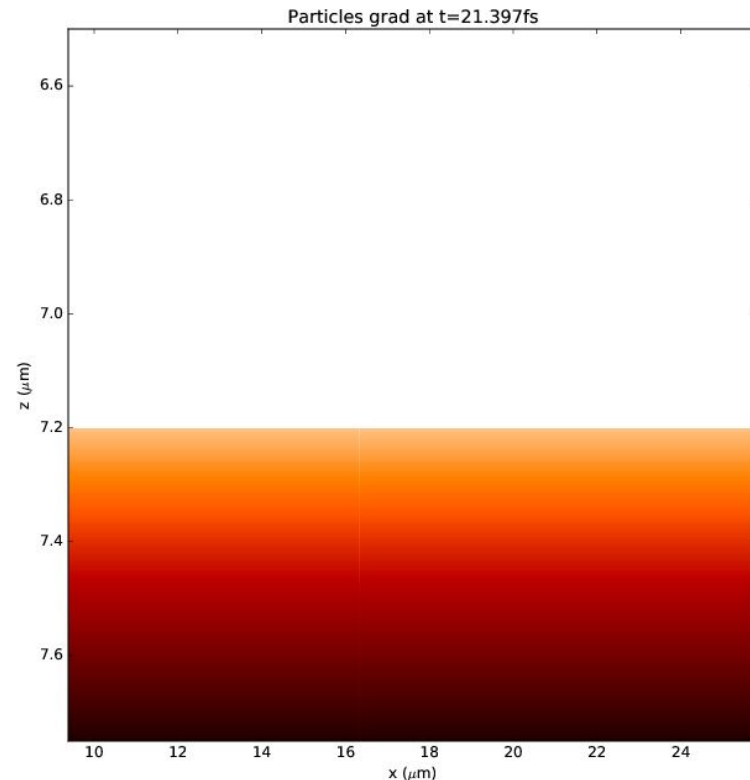
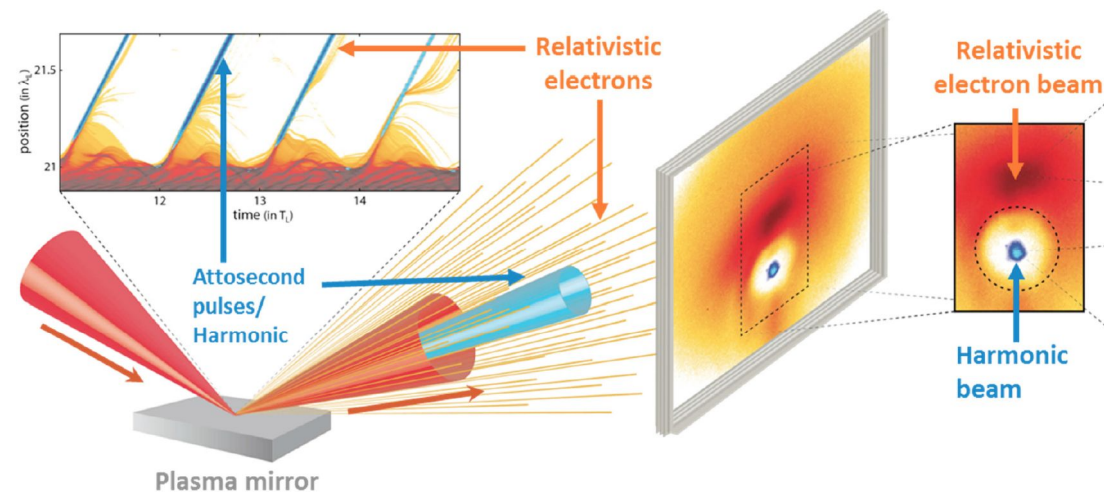


Intel Vector-Advisor Co-Design - Collaboration between NERSC, LBNL Computational Research, Intel

Example: WARP (Accelerator Modeling)



- Particle in Cell (PIC) Application for doing accelerator modeling and related applications.
- **Example Science:** Generation of high-frequency attosecond pulses is considered as one of the best candidates for the next generation of attosecond light sources for ultrafast science.



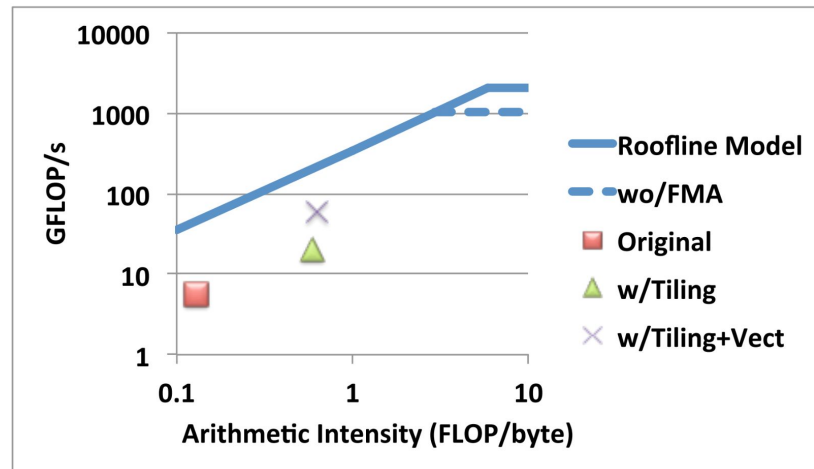
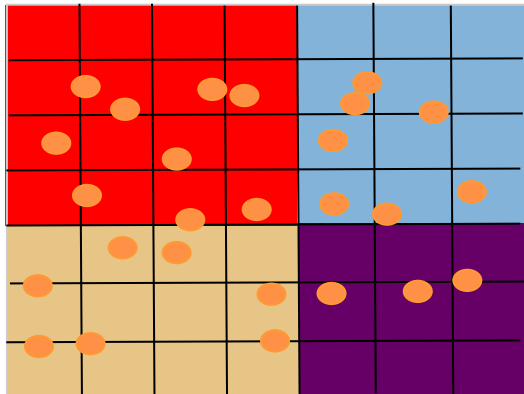
Animation from Plasma Mirror Simulations

Example: WARP (Accelerator Modeling)

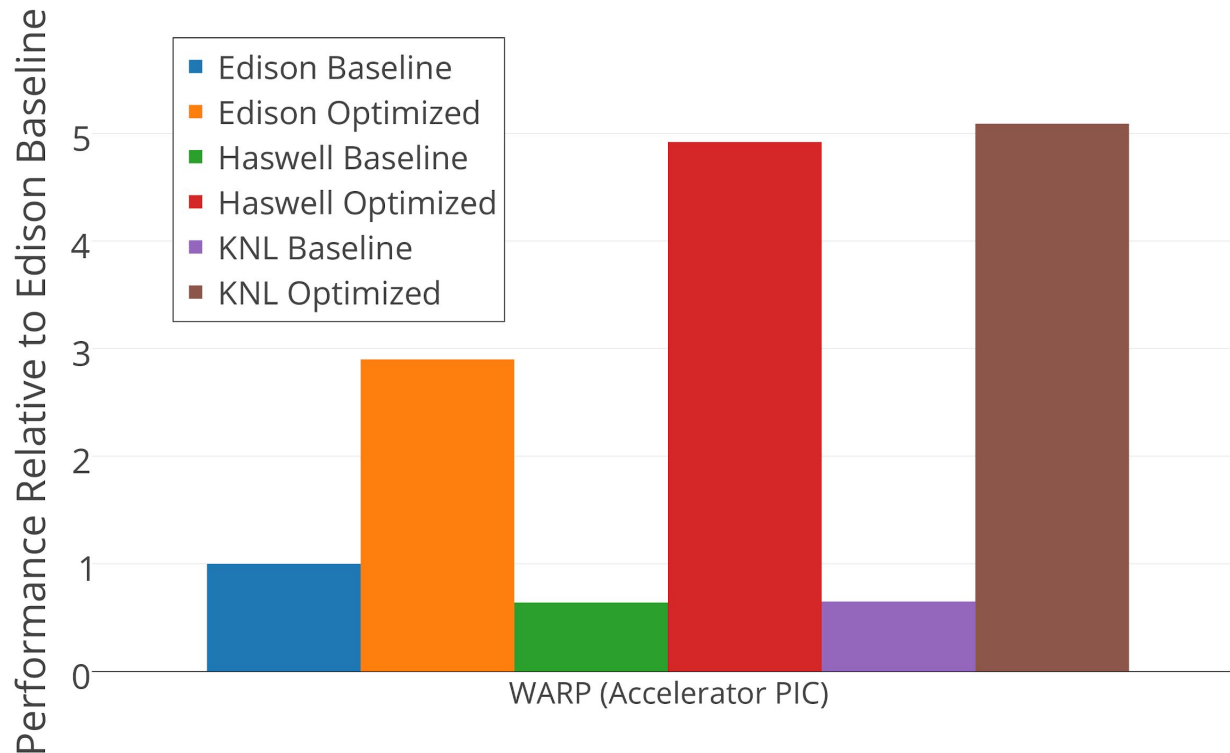


Optimizations:

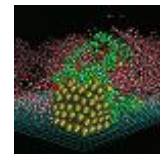
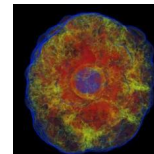
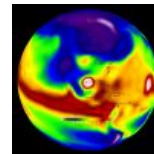
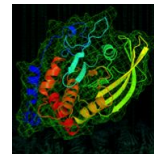
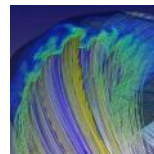
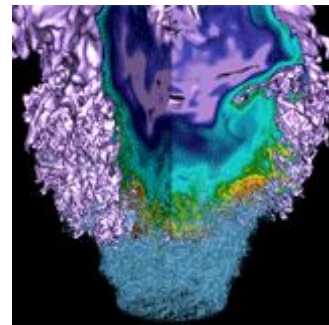
1. Add tiling over grid targeting L2 cache on both Xeon + Xeon-Phi Systems
2. Apply particle sorting + vectorization over particles (requires a number of datastructure changes)



Example: WARP (Accelerator Modeling)



KNL Performance



U.S. DEPARTMENT OF
ENERGY

Office of
Science

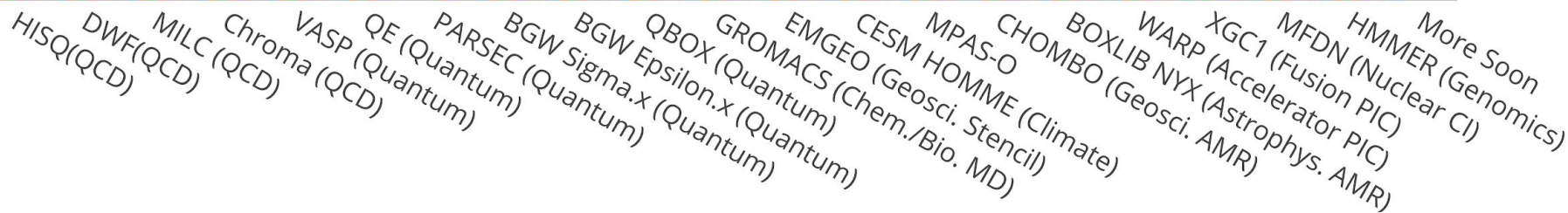


Preliminary NESAP Code Performance on KNL



Performance Relative to Edison Baseline

*Speedups from direct/indirect NESAP efforts as well as coordinated activity in NESAP timeframe



Preliminary NESAP Code Performance on KNL



PRELIMINARY

Code Speedups Via NESAP:

Haswell 2.3 x Faster W/ Optimization
KNL 3.5 x Faster W/ Optimization

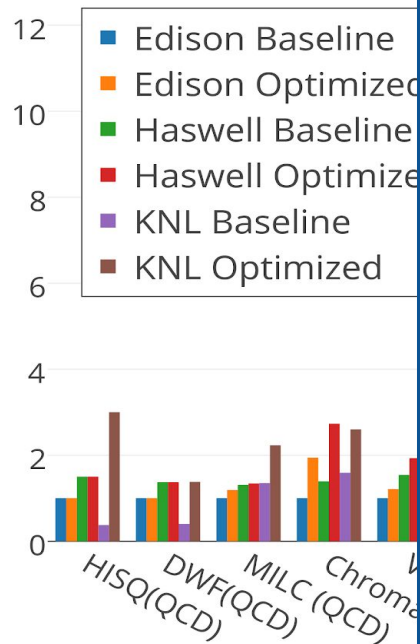
KNL / Haswell Performance Ratio

Baseline Codes 0.7 (KNL is slower)
Optimized Codes 1.1 (KNL is faster)
KNL Optimized /
Haswell Baseline **2.5**

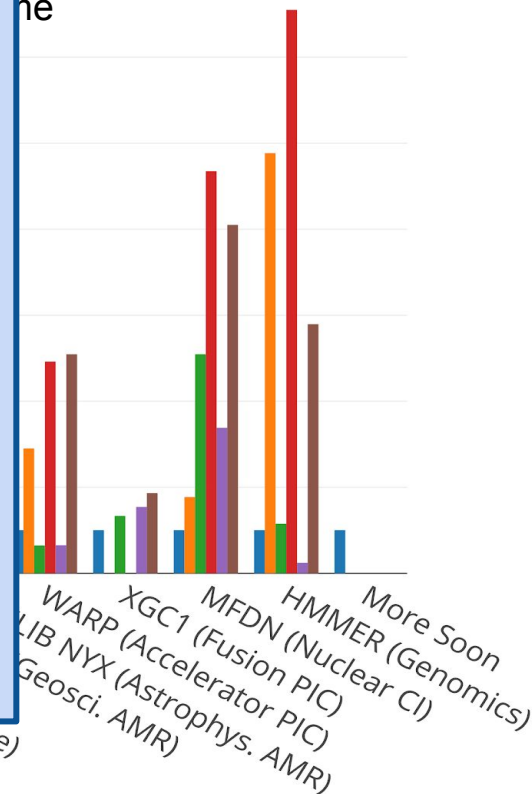
KNL / Ivy-Bridge (Edison) Performance Ratio

Baseline Codes 1.1 (KNL is faster)
Optimized Codes 1.8 (KNL is faster)
KNL Optimized /
Edison Baseline **3.4**

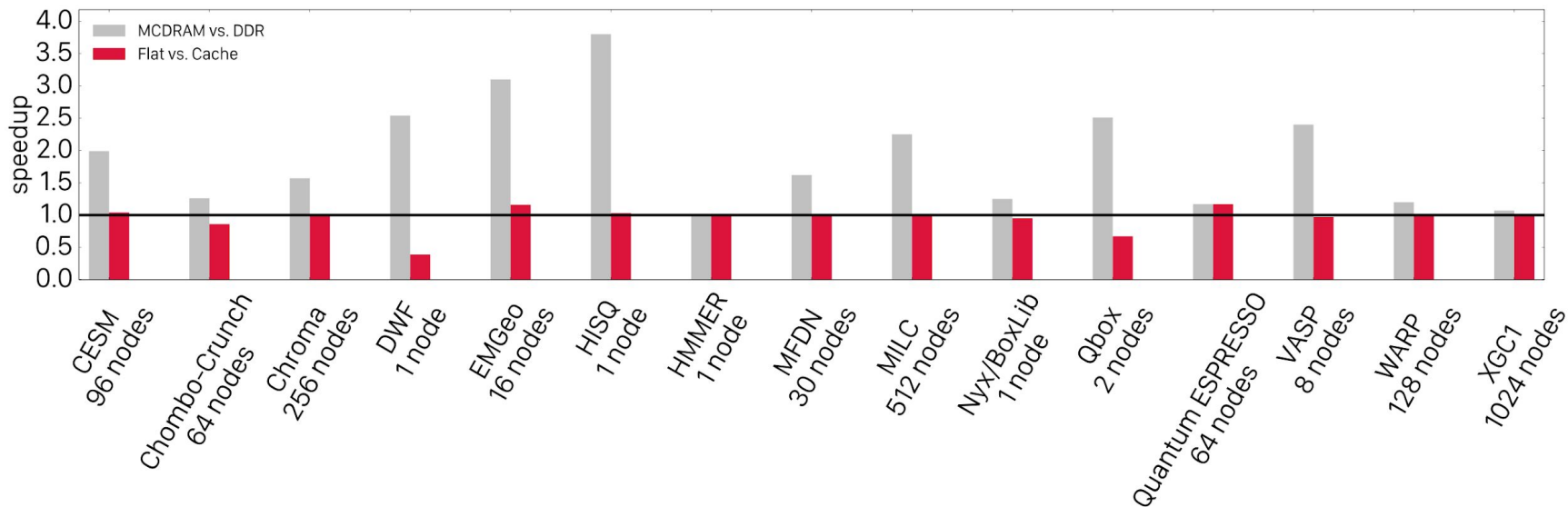
Performance Relative to Edison Baseline



as
ne



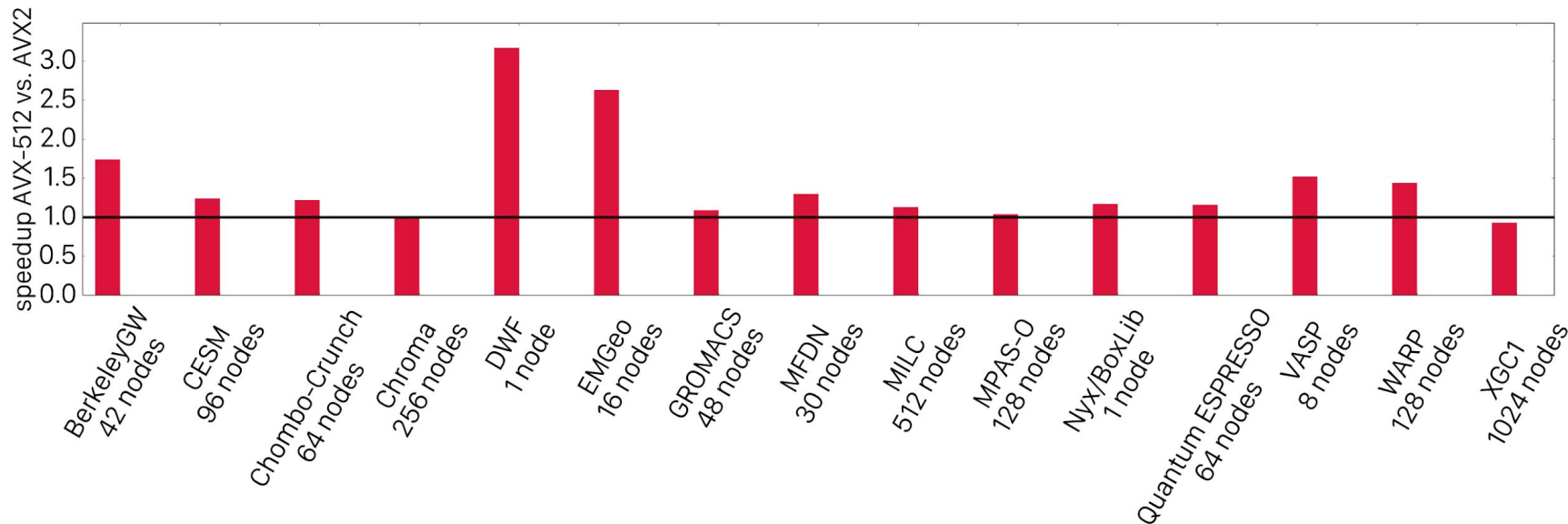
NESAP MCDRAM Effects



NESAP VPU Effects



AVX512 vs AVX2



What did we learn?



- It is crucial to understand what limits performance for your code/kernels. Tools like Advisor are necessary.
- To get good performance on KNL. One typically needs good MPI task or OpenMP thread scaling and depending on algorithm:
 - a) efficient vectorization (Codes with high AI)
 - b) efficient use of the MCDRAM (Codes with low AI)
 - c) both (Codes with AI near 1)
- The lack of an L3 cache on KNL can make cache blocking for L1/L2 more important. Particularly in latency-sensitive apps (e.g. indirect indexing)
- Cache mode provides nearly the same performance as flat mode (with directives) for most applications. However, cache-conflicts can be an issue with some apps.
- MPI apps tend to stop scaling at the same number of ranks on Xeon and Xeon-Phi (often characterized by the algorithm). This translates to lower node counts on Xeon-Phi. Additional, parallelism needs to be exploited - usually expressed as OpenMP.

The Payoff: Large Scale Science on Cori

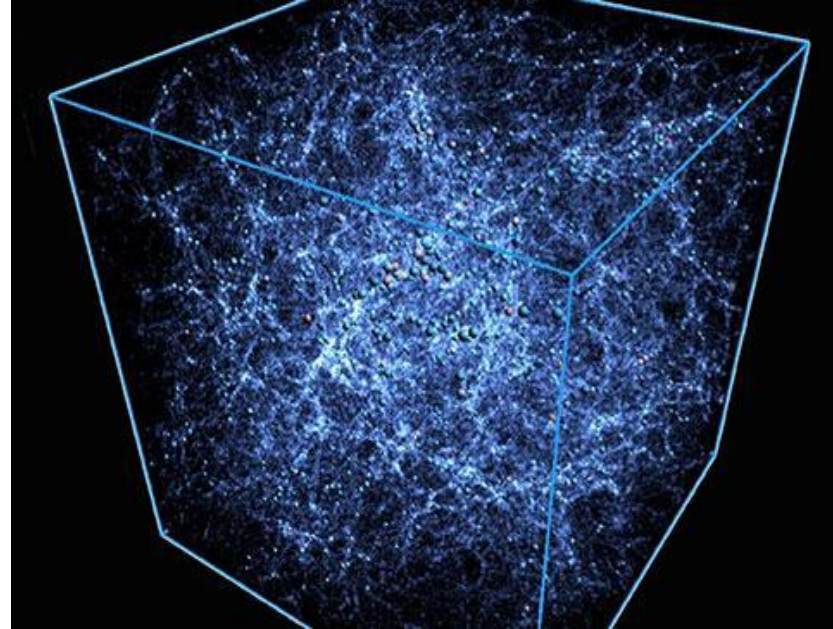


3-Pt Correlation On 2B Galaxies Recently Completed on Cori

- NESAP For Data Prototype (Galactos)
- First anisotropic, 3-pt correlation computation on 2B Galaxies from Outer Rim Simulation
- Solves an open problem in cosmology for the next decade (LSST will observe 10B galaxies)
- Can address questions about the nature of dark-energy and gravity
- Novel $O(N^2)$ algorithm based on spherical harmonics for 3-pt correlation

Scale:

- 9600+ KNL Nodes (Significant Fraction of Peak)



Large Scale Science Being Done on Cori



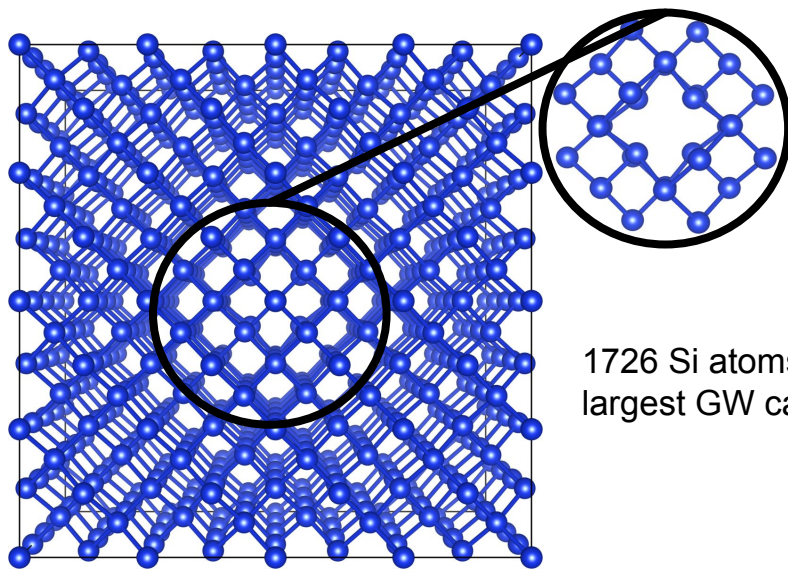
Defect States in Materials:

Important material properties are often determined by the effects of defects. Require large calculations to isolate defect states and require beyond DFT in LDA/GGA.

(Quantum ESPRESSO and BerkeleyGW)

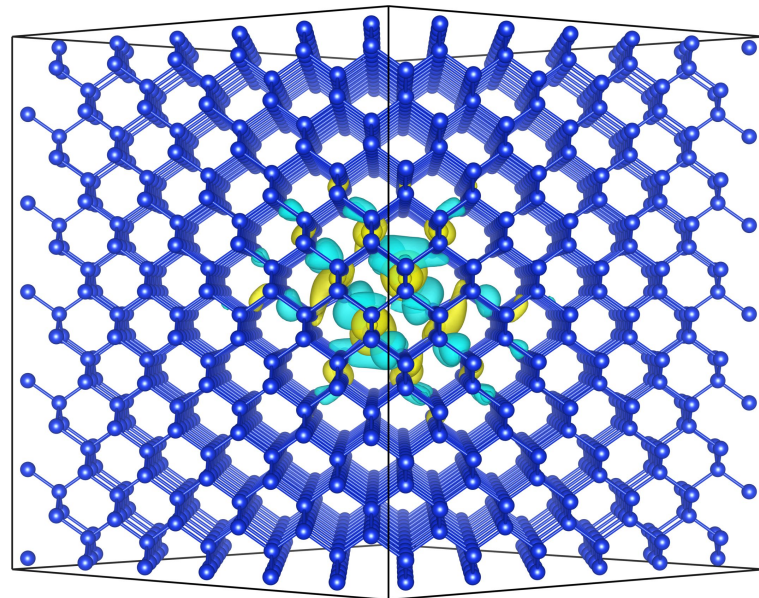
Scale:

Simulated on Cori with up to 9600 KNL Nodes -
Large percentage of peak performance obtained
> 10 PFLOPS.



Di-vacancy defect
and localized
defect orbital in
crystalline Silicon.

1726 Si atoms (~7K electrons) is
largest GW calculation published



Understanding multi-scale, non-thermal edge physics in fusion reactors.

Edge physics crucially affects both the core energy production and the erosion rate of edge materials. Both are keys to creating clean energy from fusion reactions.

PI: C.S. Chang, Princeton Plasma Physics Lab

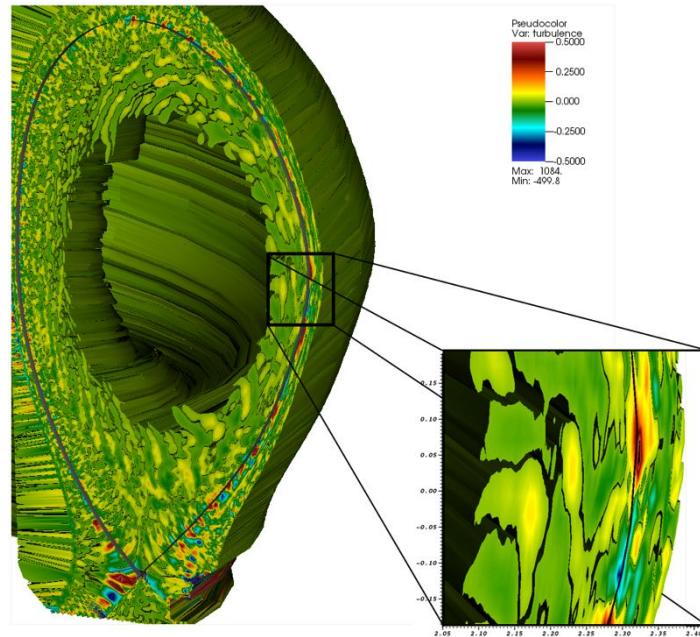
NERSC Hours on KNL: 71 million

Max Concurrency: 4,096 nodes (279K cores)

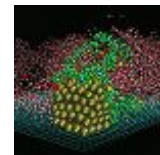
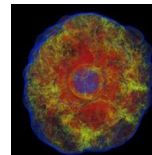
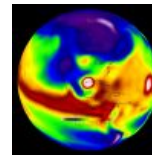
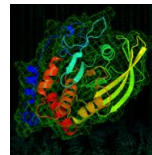
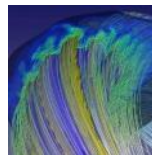
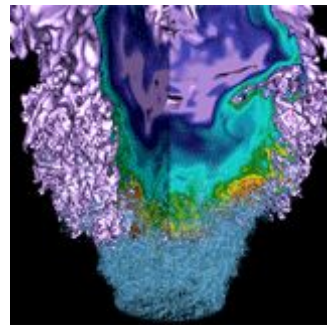
Code: XGC1

NERSC contact: Tuomas Koskela

NESAP



END, Thank you!



U.S. DEPARTMENT OF
ENERGY

Office of
Science



Large Scale Science Being Done on Cori



Deep Learning on Cori:

- NESAP for Data Prototype
- Supervised Classification for LHC datasets
- Pattern discovery for climate datasets
- Production DL stack (IntelCaffe, MLSL)
- Convolutional architectures optimized on KNL with IntelCaffe and MKL
- Synch + Asynch parameter update strategy for multi-node scaling

Scale:

- 9600 KNL nodes on Cori
- 10 Terabyte datasets
- Millions of Images
- 10's of Minutes to Train



Machine learning techniques can automatically detect patterns in simulation data. Applied to climate and particle-physics data from collider experiments.

Parting observations



- Cori is successfully enabling large scale science at NERSC
- Individual core performance is lower, as expected, but aggregate performance is on par or better than Edison and Cori/Haswell
- The vast majority of users are running with KNLs in Quadrant+Cache mode on Cori

Largest Ever Quantum Simulation

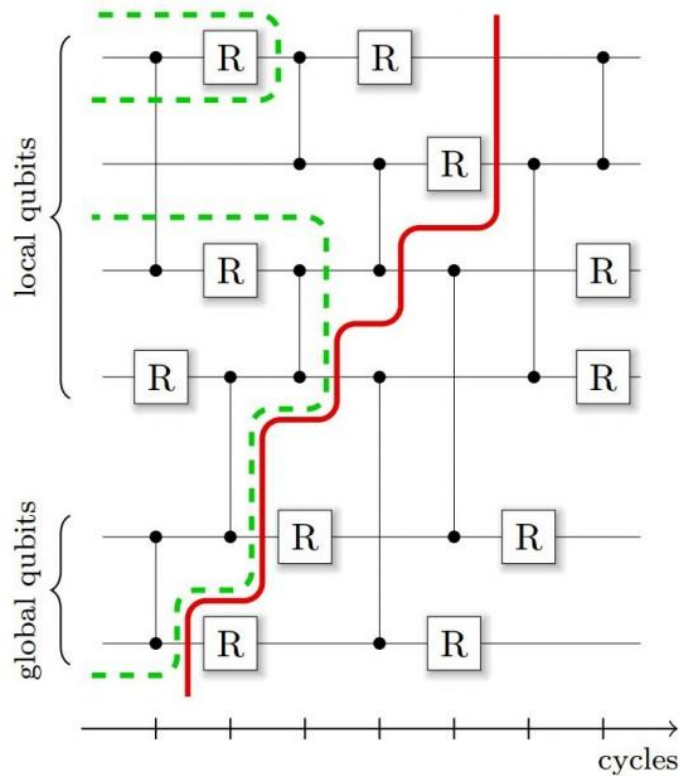


Largest Ever Quantum Computing Simulation

- 45 Qubit simulation is largest ever quantum computing simulation ever
- Simulations are important for validating prototype quantum computers devices
- Team lead by ETH scientists collaborators at Google, LBNL's Computational Research Division

Scale:

- >8000 KNL nodes
- 0.5 Petabytes of memory used ($\sim 2^{45}$)
- 0.43 PetaFLOPS (Bandwidth bound)



The Photosystem I (PS1) is a membrane protein complex that captures solar energy and stores it in the complex.

It can couple with H2ase to produce Hydrogen from sunlight and water. Using MD simulations to study PS1/H2ase interactions in detail and learn how to produce a clean fuel.

PI: Jeremy Smith, Oak Ridge National Laboratory

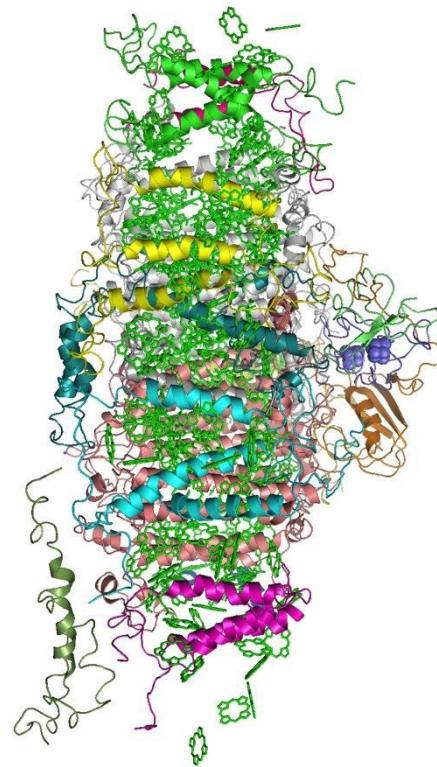
NERSC Hours on KNL: 54 million

Max Concurrency: 1,080 nodes (73K cores)

Code: GROMACS

NERSC contact: Zhengji Zhao

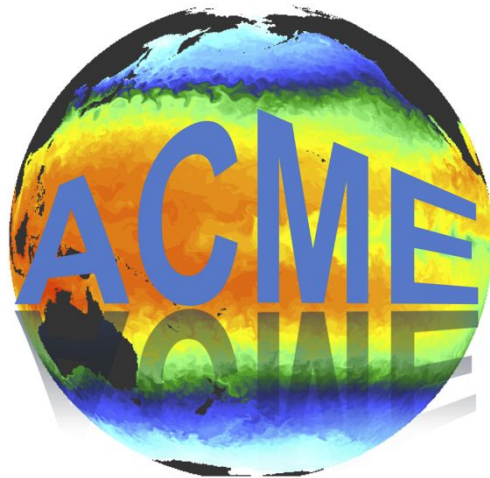
NESAP



Water Cycle: How will the water cycle evolve in a warmer climate?

Biogeochemistry: How will terrestrial and coastal ecosystems drive natural sources & sinks of CO₂ and CH₄ in a warmer world?

Cryosphere system: How will more extreme storms enhance the coastal impacts of sea level rise?



PI: Lai-Yung Ruby Leung, Pacific NW National Lab

NERSC Hours on KNL: 109 million

Max Concurrency: 8,192 nodes (557K cores)

Code: ACME

NERSC contact: Helen He

NESAP

Designing new, cheaper, better, and environmentally benign catalysts for production and chemical utilization of hydrogen, for production of hydrocarbon fuels, and for low temperature fuel cells.

PI: Manos Mavrikakis, U. of Wisconsin, Madison

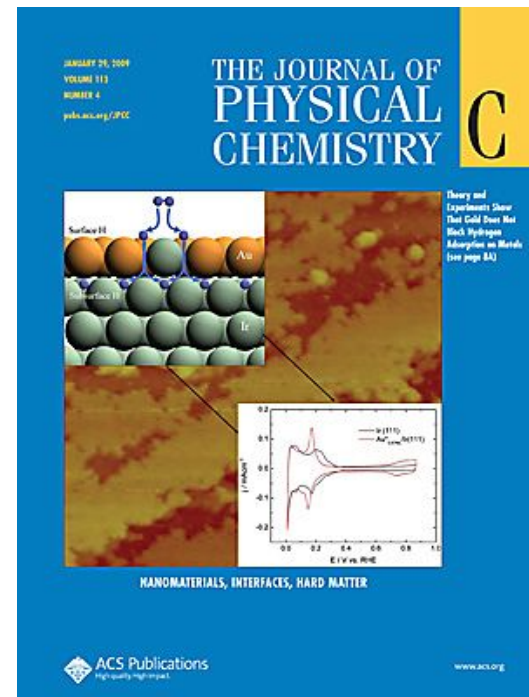
NERSC Hours on KNL: 201 million

Max Concurrency: 64 nodes (4.3K cores)

Code: VASP

NERSC contact: Zhengji Zhao

NESAP



Imaging subsurface geophysical properties in 3D and relating these properties to critical geological processes relevant to energy exploration and carbon sequestration.

PI: Jeff Newman, Berkeley Lab

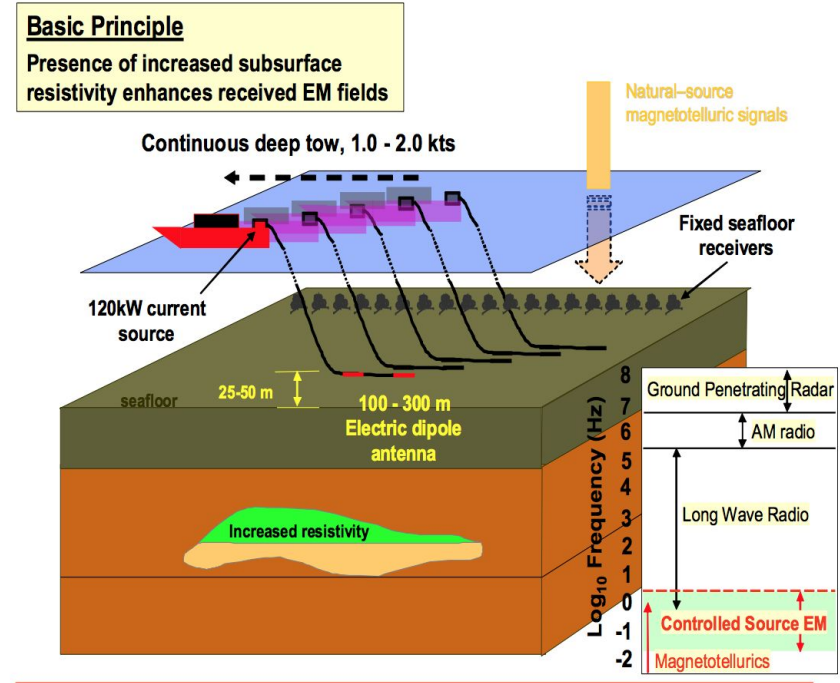
NERSC Hours on KNL: 38 million

Max Concurrency: 520 nodes (35K cores)

Code: EMGeo

NERSC contact: Thorsten Kurth

NESAP



“In the lab” materials synthesis and discovery can take 20 years to get to market.

The materials project is accelerating the way materials discovery is done by creating a high-throughput computing environment together with a searchable, interactive database of computed materials properties.



PI: Kristin Ceder-Persson, Berkeley Lab

NERSC Hours on KNL: 109 million

Max Concurrency: 8 nodes (544 cores, 150K runs)

Code: VASP

NERSC contact: Zhengji Zhao

NESAP

Studying three of the most exciting areas of current astrophysics: supernovae, extrasolar planets, and active galactic nuclei.

Modeling radiative transfer in atmospheres of very low mass stars and giant planets to understand observations.

PI: Eddie Baron, U. Oklahoma

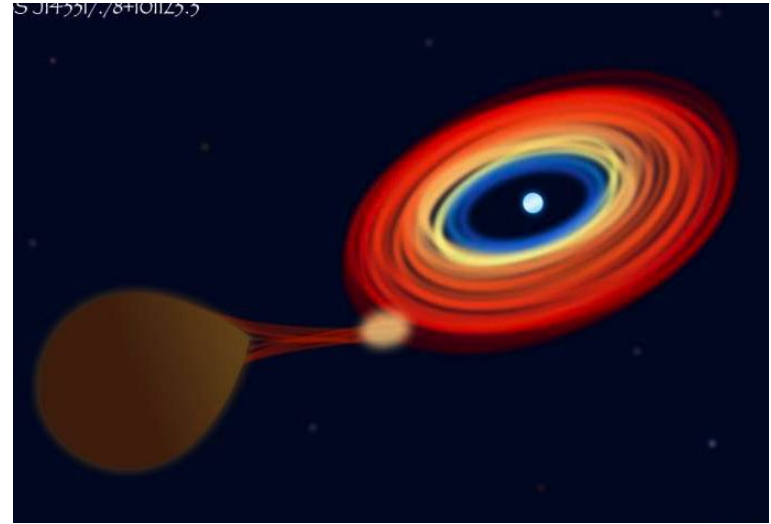
NERSC Hours on KNL: 55 million

Max Concurrency: 8,064 nodes (548K cores)

Code: phoenix

NERSC contact: Brian Friesen

NESAP



Nature paper on discovery of an irradiated brown-dwarf companion to an accreting white dwarf.

Compute predictions of the standard model for kaon decays and mixings, quantities that offer highly-visible, exciting prospects for the discovery of physics beyond the standard model of particle physics.

PI: Norman Christ, Columbia University

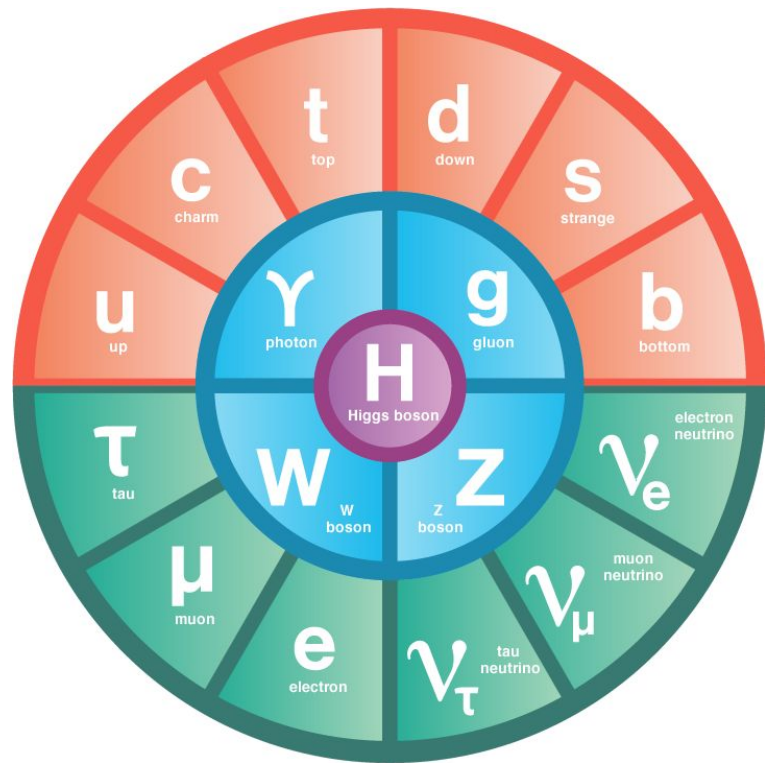
NERSC Hours on KNL: 430 million

Max Concurrency: 2,449 nodes (166K cores)

Code: DWF Inverter

NERSC contact: Woo-Sun Yang

NESAP



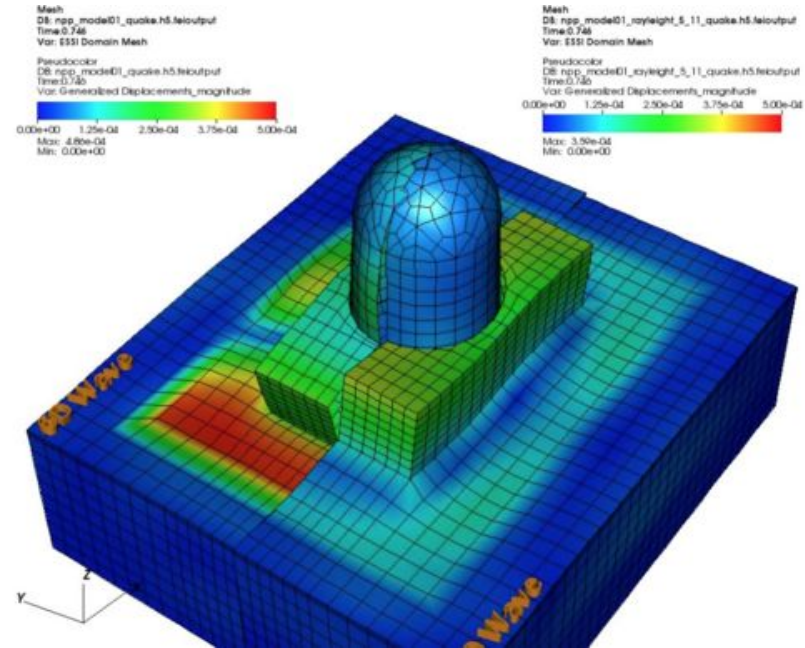
This project is focused on the development of advanced nonlinear modeling and simulation for seismic analysis of nuclear facilities. The ESSI nonlinear finite element program is being extended to include nonlinear structural elements necessary for a fully coupled nonlinear analysis of soil-structure systems.

PI: David McCallen, Berkeley Lab

NERSC Hours on KNL: 8 million


Max Concurrency: 4,096mnodes (278K cores)

Code: SW4




NERSC as a Documentation Hub





Powering **Scientific Discovery** Since 1974

Site Map | My NERSC |  Share

search...

HOME ABOUT SCIENCE AT NERSC SYSTEMS **FOR USERS** NEWS & PUBLICATIONS R & D EVENTS LIVE STATUS TIMELINE

FOR USERS

- » Live Status
- » User Announcements
- » My NERSC
- » Getting Started
- » Connecting to NERSC
- » Accounts & Allocations
- » Computational Systems
 - Cori
 - Updates and Status
 - Cori Timeline
 - Configuration
 - Getting Started
 - Programming
 - Running Jobs
 - Burst Buffer
 - Cori Intel Xeon Phi Nodes
 - Application Porting and Performance**
 - Getting Started and Optimization Strategy
 - Application Case Studies
 - Profiling Your Application
 - Improving OpenMP Scaling
 - Measuring and Understanding Memory Bandwidth
 - Vectorization
 - Using on-package memory
 - Using High Performance Libraries and Tools
 - Dungeon Session Worksheet
 - KNL White Boxes
- NESAP
- NERSC-8 Procurement
- Programming models
- File Storage and I/O
- Edison
- PDSF
- Genepool

Home » For Users » Computational Systems » Cori » Application Porting and Performance

APPLICATION PORTING AND PERFORMANCE

Many applications will need code modifications in order to run efficiently on Cori's Intel Xeon Phi "Knights Landing" manycore processors. Applications need to have good thread scalability to take advantage of the 68-core Xeon Phi processor, a data structure layout that can effectively use the 16 GB of onboard MCDRAM can help memory, and loop structures that exploit the 512-bit vector units. In the web pages that follow we document strategies that can help you improve your application's performance. While achieving good performance on Cori may take some work, the good news is that optimizations made for Cori will very likely improve your code's performance on other architectures.

Getting Started and Optimization Strategy »

The purpose of this page is to get you started thinking about how to optimize your application for the Knights Landing (KNL) Architecture that will be on Cori. This page will walk you through the high level steps and give an example using a real application that runs at NERSC. How Cori Differs From Edison There are several important differences between the Cori (Knight's Landing) node architecture and the Edison (Ivy Bridge) node architecture that require special attention from application... [Read More »](#)

Application Case Studies »

NERSC staff along with engineers have worked with NESAP applications to prepare for the Cori-Phase 2 system based on the Xeon Phi "Knights Landing" processor. We document the several optimization case studies below. Our presentations at ISC 16 IXPUG Workshop can all be found: <https://www.ixpug.org/events/ixpug-isc-2016> Other pages of interest for those wishing to learn optimization strategies of Cori Phase 2 (Knights Landing): Getting Started Measuring Arithmetic Intensity Measuring and... [Read More »](#)

Profiling Your Application »


There are a number of tools which can help users profile applications to determine the best strategy for improving code performance on Cori. Some popular tools are Vtune, CrayPat, and MAP. [Read More »](#)

Improving OpenMP Scaling »


Each processor on Cori will have over 60 processor cores with 4 hardware threads each. Efficient thread scalability will be important to achieving good performance. [Read More »](#)

Measuring and Understanding Memory Bandwidth »

It is important to understand if your application is memory bandwidth bound, memory latency bound or compute bound. Understanding the characteristics of your application will determine what tactics you use to optimize your application. [Read More »](#)



Powering **Scientific Discovery** Since 1974

Site Map | My NERSC |  Share

search...

HOME ABOUT SCIENCE AT NERSC SYSTEMS **FOR USERS** NEWS & PUBLICATIONS R & D EVENTS LIVE STATUS TIMELINE

FOR USERS

- » Live Status
- » User Announcements
- » My NERSC
- » Getting Started
- » Connecting to NERSC
- » Accounts & Allocations
- » Computational Systems
 - Cori
 - Updates and Status
 - Cori Timeline
 - Configuration
 - Getting Started
 - Programming
 - Running Jobs
 - Burst Buffer
 - Cori Intel Xeon Phi Nodes
 - Application Porting and Performance
 - Getting Started and Optimization Strategy
 - Application Case Studies**
 - EMGEO Case Study
 - BerkeleyGW Case Study
 - QPhIX Case Study
 - WARP Case Study
 - MFDn Case Study
 - BoxLib Case Study
 - VASP Case Study
 - CESM Case Study
 - Chombo-Crunch Case Study
 - HMMER3 Case Study
 - Early application case studies
 - ISC16 IXPUG Performance Workshop
 - Quantum ESPRESSO Exact Exchange Case Study
 - XGCI Case Study
 - Profiling Your Application

Home » For Users » Computational Systems » Cori » Application Porting and Performance » Application Case Studies

APPLICATION CASE STUDIES

NERSC staff along with engineers have worked with NESAP applications to prepare for the Cori-Phase 2 system based on the Xeon Phi "Knights Landing" processor. We document the several optimization case studies below.

Our presentations at ISC 16 IXPUG Workshop can all be found: <https://www.ixpug.org/events/ixpug-isc-2016>

Other pages of interest for those wishing to learn optimization strategies of Cori Phase 2 (Knights Landing):

- [Getting Started](#)
- [Measuring Arithmetic Intensity](#)
- [Measuring and Understanding Memory Bandwidth](#)
- [Vectorization](#)

EMGEO Case Study »

June 20, 2016
Early experiences working with the EMGeo geophysical imaging applications. [Read More »](#)

BerkeleyGW Case Study »

Code Description and Science Problem BerkeleyGW is a Materials Science application for calculating the excited state properties of materials such as band gaps, band structures, absorption spectroscopy, photoemission spectroscopy and more. It requires as input the Kohn-Sham orbitals and energies from a DFT code like Quantum ESPRESSO, PARATEC, PARSEC etc. Like such DFT codes, it is heavily dependent on FFTs, Dense Linear algebra and tensor contraction type operations similar in nature to those... [Read More »](#)

QPhIX Case Study »

June 20, 2016
Background QPhIX [1,2,3] is a library optimized for Intel(R) manycore architectures and provides sparse solvers and slash kernels for Lattice QCD calculations. It supports the Wilson dslash operator with and without clover term as well as Conjugate Gradient [4] and BiCGStab [5] solvers. The main task for QPhIX is to solve the sparse linear system where the Dslash kernel is defined by Here, U are complex, special unitary, 3x3 matrices (the so-called gauge links) which depend on lattice site x... [Read More »](#)

WARP Case Study »

Update A more complete summary is now available at <https://picsar.net/> Background WARP is an accelerator code that is used