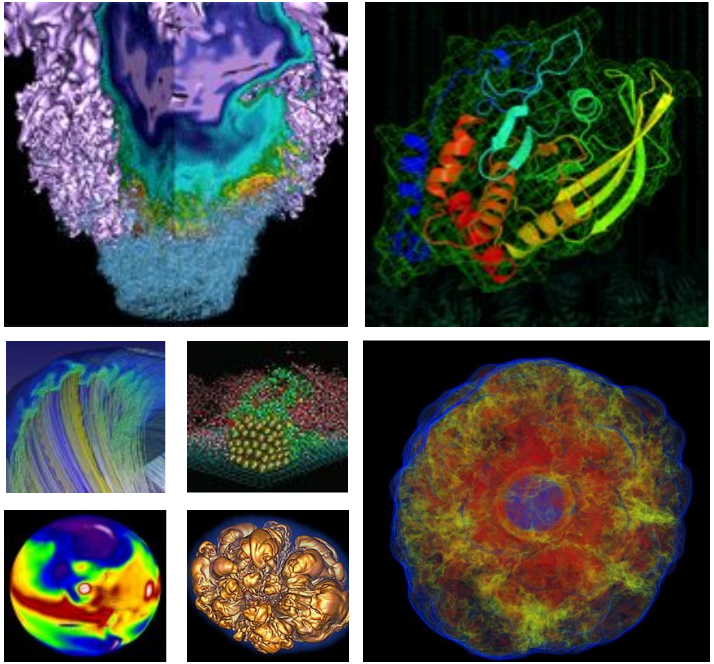
Many-Cores for the Masses: Lessons from 2 Years With the Cori System at NERSC



Jack Deslippe

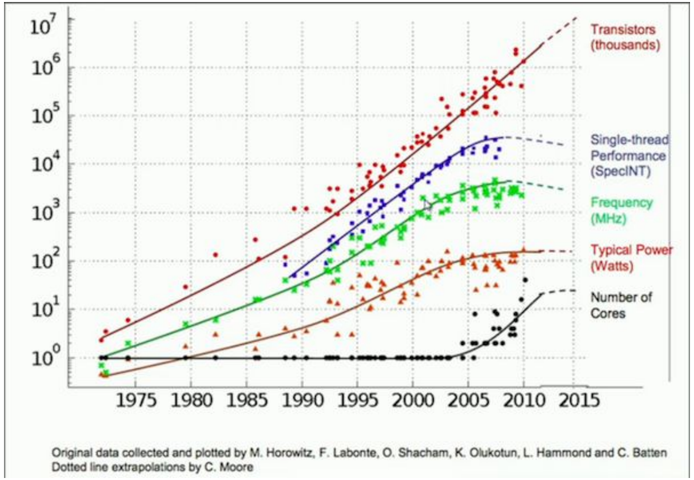
April 2019

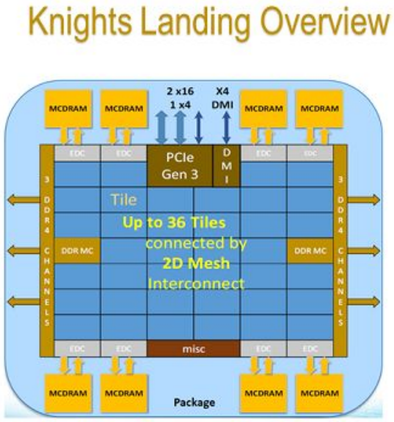
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Change Has Arrived

Driven by power consumption and heat dissipation toward lightweight cores



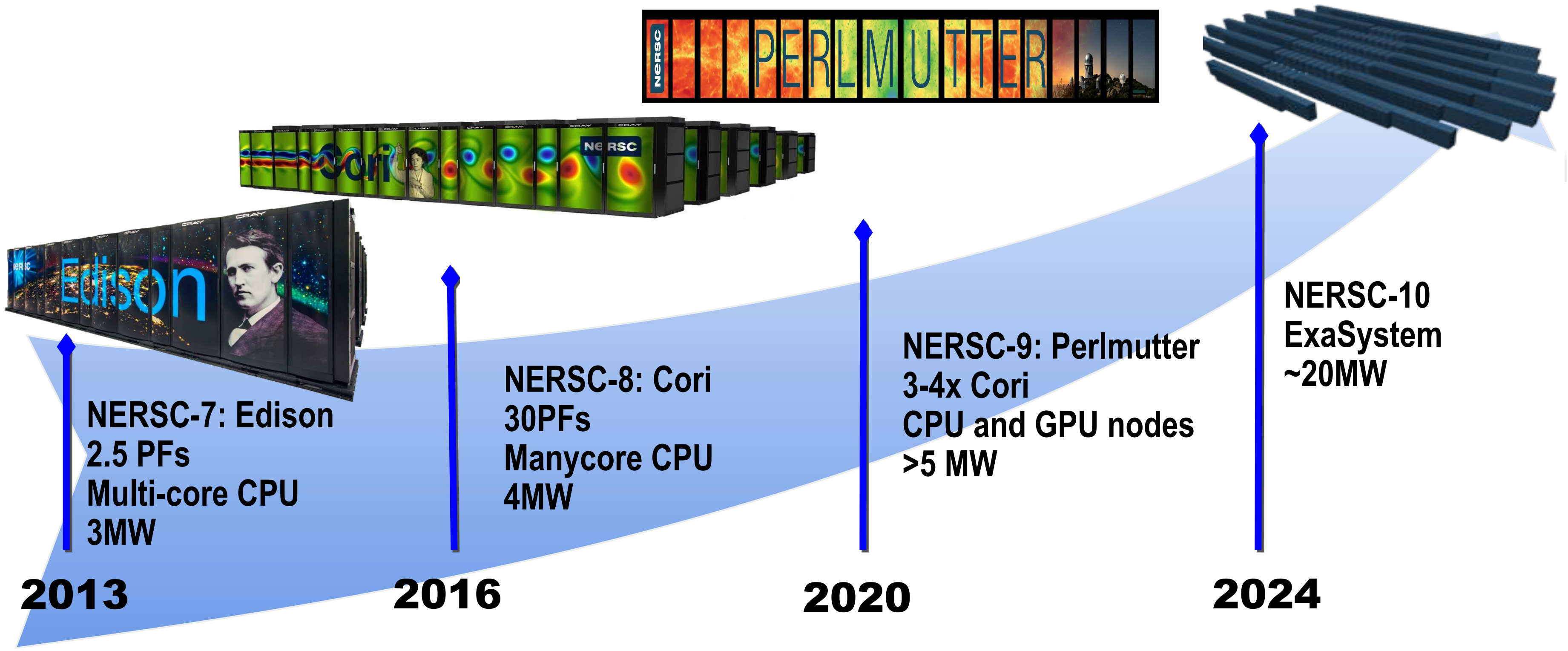


KNL: 215-230 W

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.

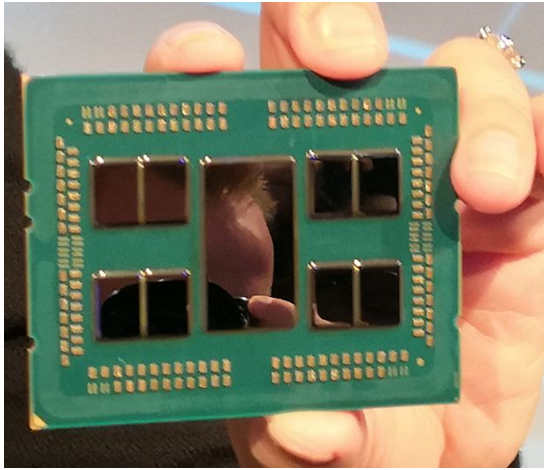
NERSC Systems Roadmap



AMD CPU nodes

"Rome" specs

- ~64 cores
- AVX2 SIMD (256 bit)
  (Perlmutter will have Milan)

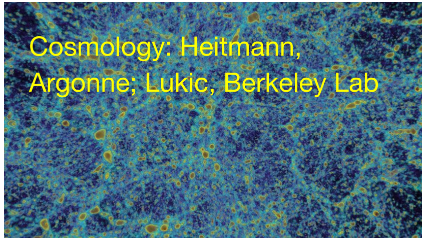


1 Slingshot connection

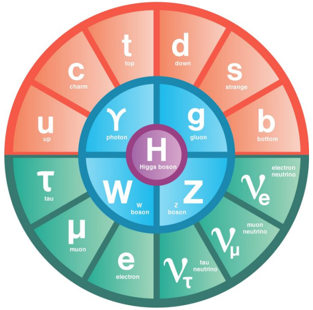
- 1x25 GB/s

~1 Cori (if your problem runs on Cori today it will work on Perlmutter)

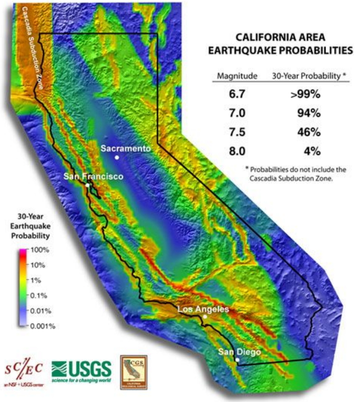
High Impact Science at Scale on Cori

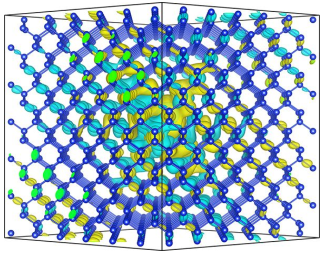

Cosmology: Heitmann, Argonne; Lukic, Berkeley Lab

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



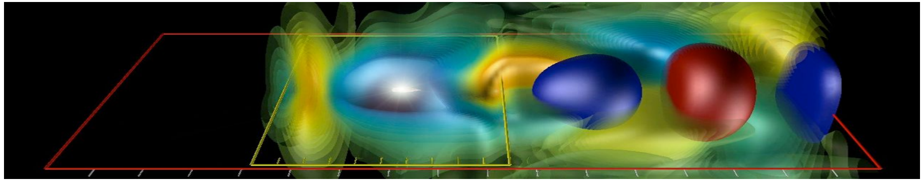
M8 Earthquake on the San Andreas Fault, Goulet, USC Earthquake Center




Optical Properties of Materials, Louie, UC Berkeley


Magnetic Reconnection, Stanier, Los Alamos


Flow in Porous Media, Trebotich, Berkeley Lab


Asymmetric Effects in Plasma Accelerators, Vay, Berkeley Lab

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:

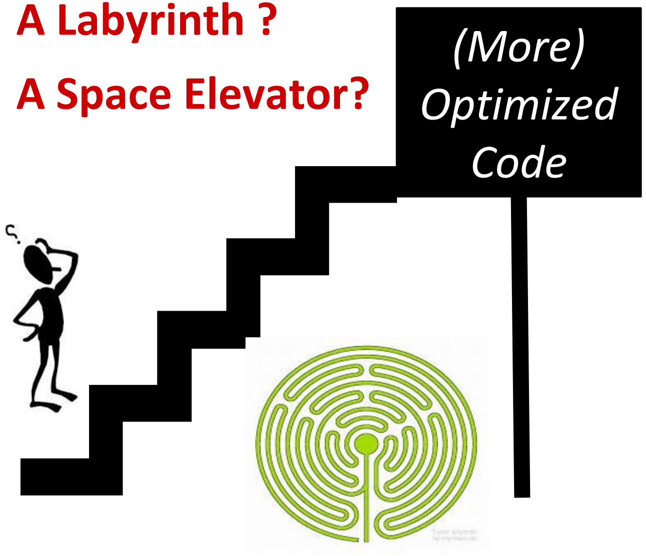
- How do you know what KNL hardware feature to target?
- How do you know how your code performs in an absolute sense and when to stop?

Optimizing Code For Cori is Like?

A. **A Staircase ?**

B. **A Labyrinth ?**

C. **A Space Elevator?**



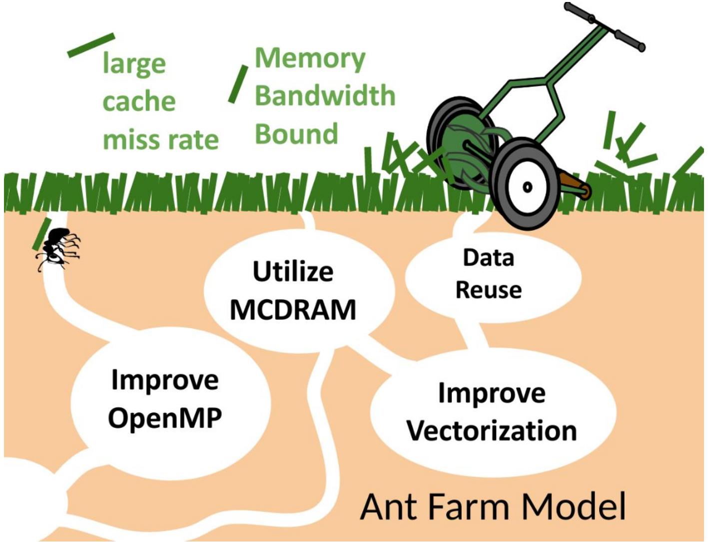
Optimization Challenge and Strategy

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:

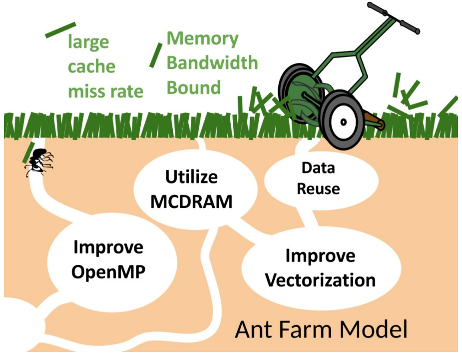
- Many Cores
- Bigger Vectors
- New ISA
- Multiple Memory Tiers

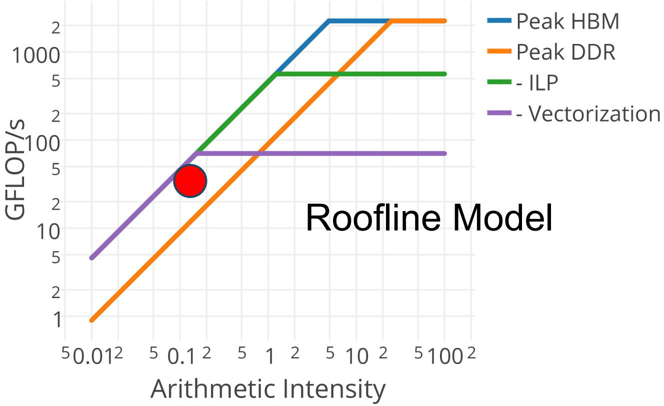
It is easy for users to get bogged down in the weeds:

- 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?

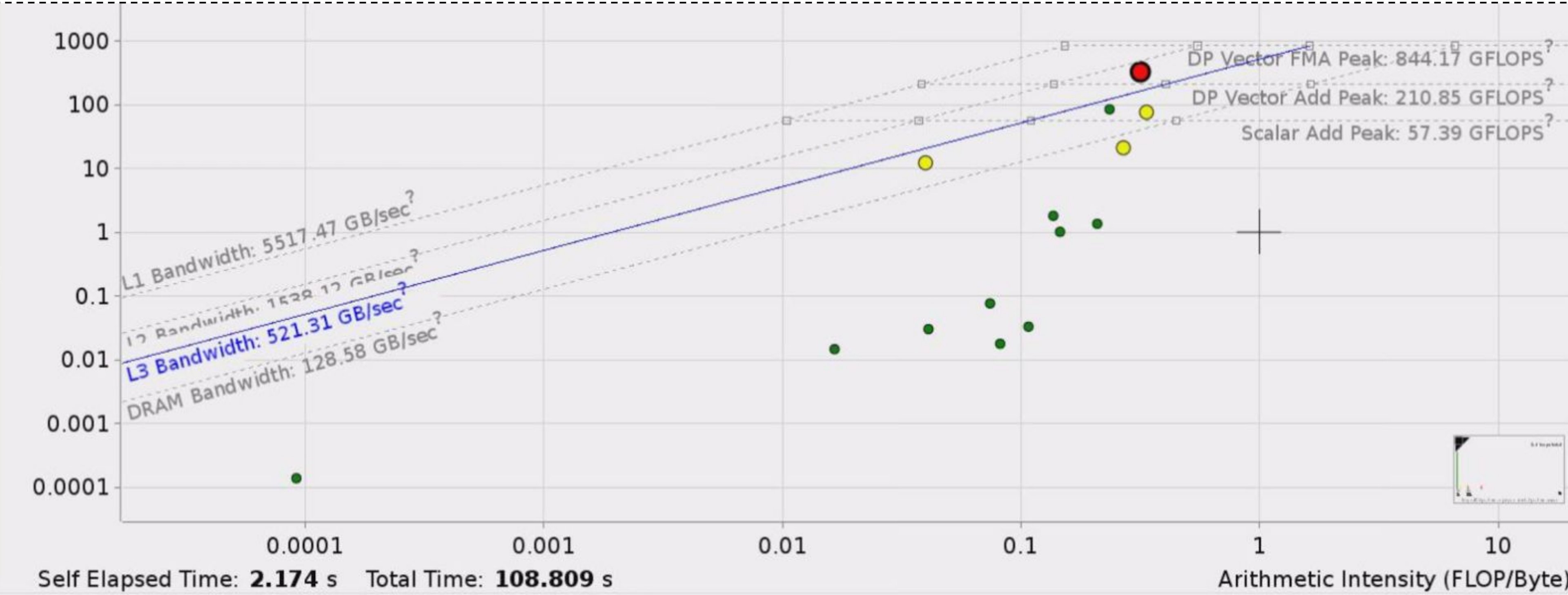
NERSC has developed tools and strategy for users to answer these questions:

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





Tools CoDesign



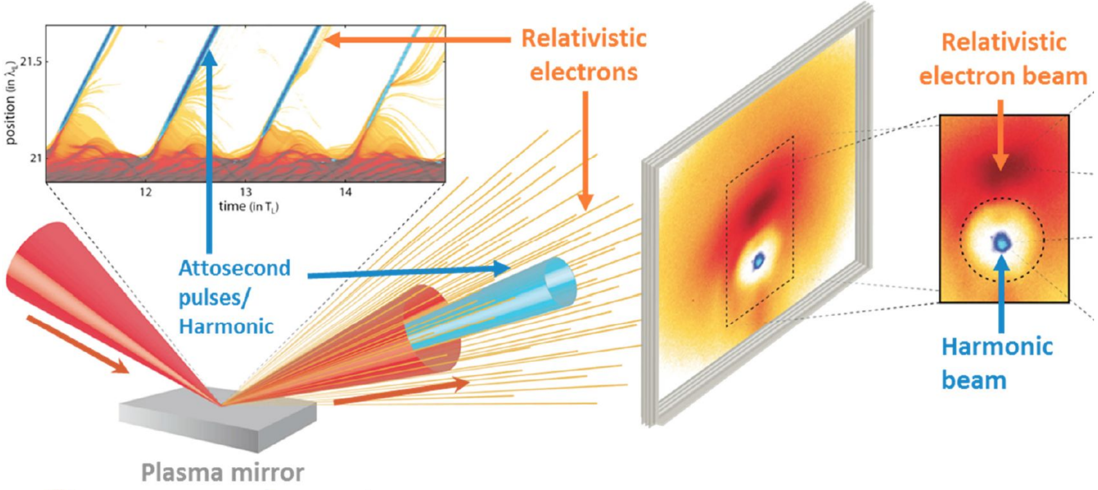
Intel Vector-Advisor Co-Design - Collaboration between NERSC, LBNL Computational Research, Intel

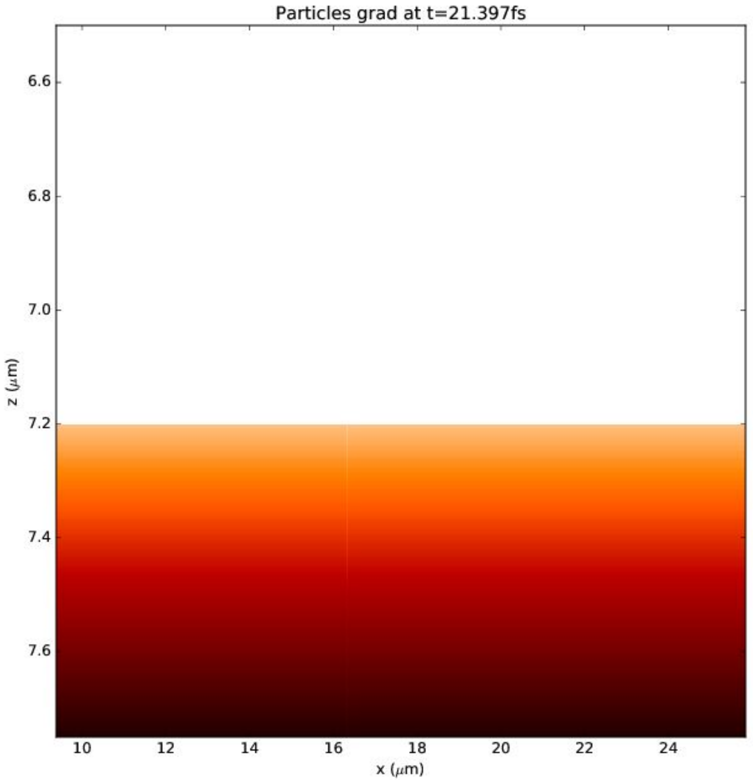
https://www.nersc.gov/users/software/performance-and-debugging-tools/advisor/#toc-anchor-6

Example: WARP (Accelerator Modeling)

- Particle in Cell (PIC) Applicaion 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.



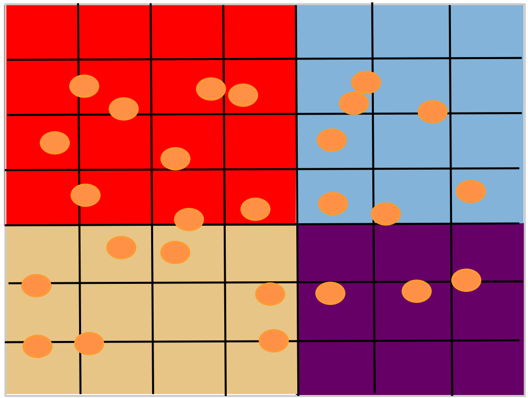


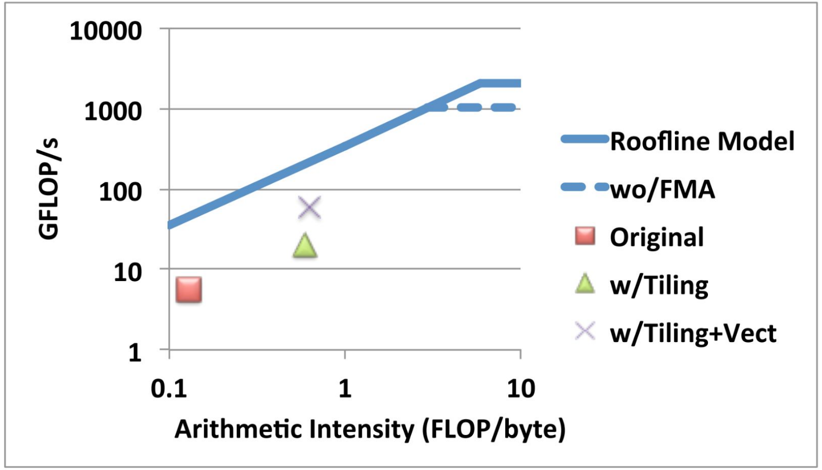
Animation from Plasma Mirror Simulations

Example: WARP (Accelerator Modeling)

Optimizations:

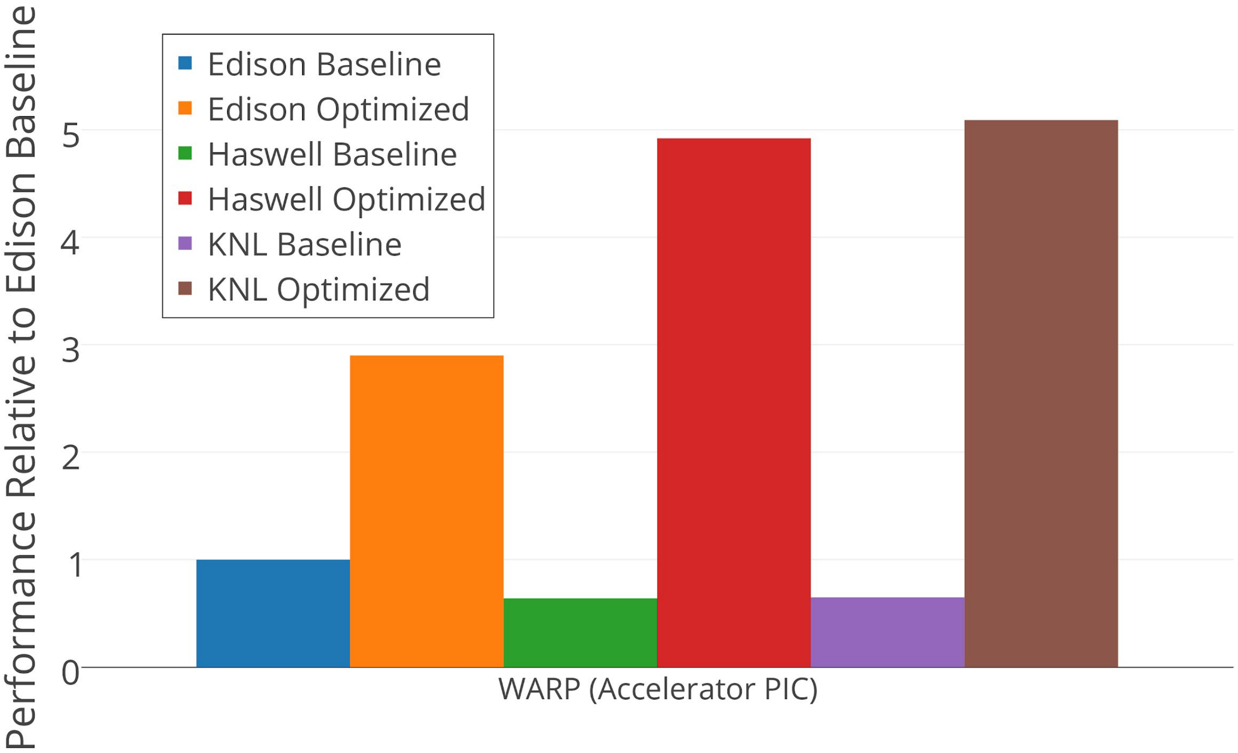
1. Add tiling over grid targeting L2 cache on both Xeon + Xeon-Phi Systems
2. Apply particle sorting + vectorization over particles (requires a number of datastructure changes)



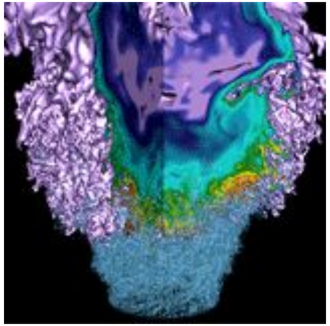


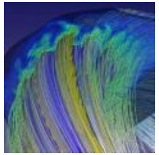
KNL Roofline

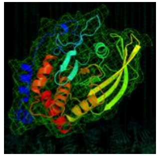
Example: WARP (Accelerator Modeling)

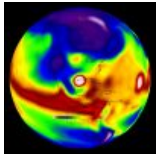


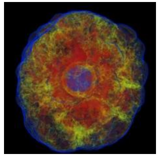
KNL Performance

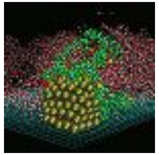








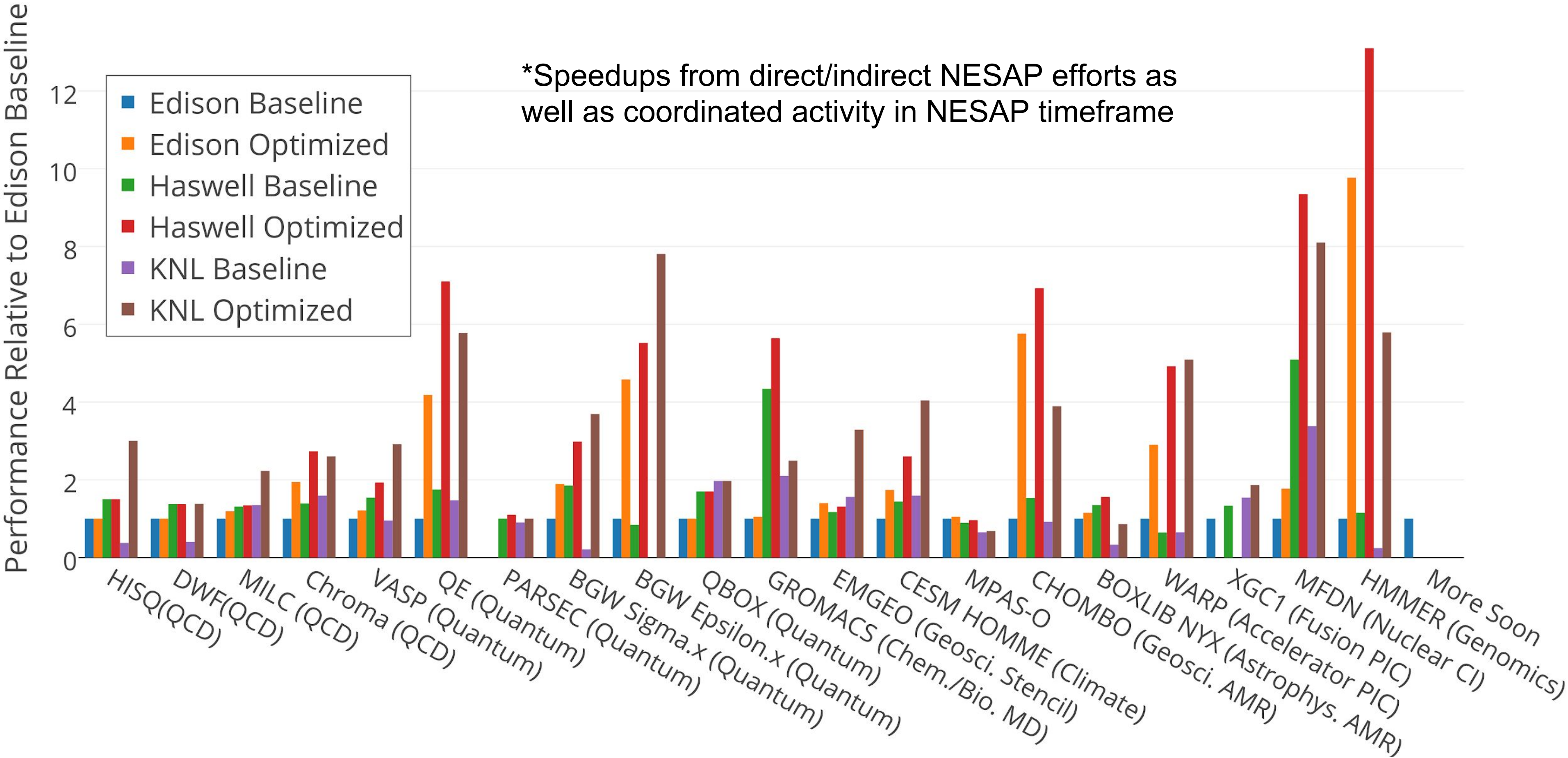




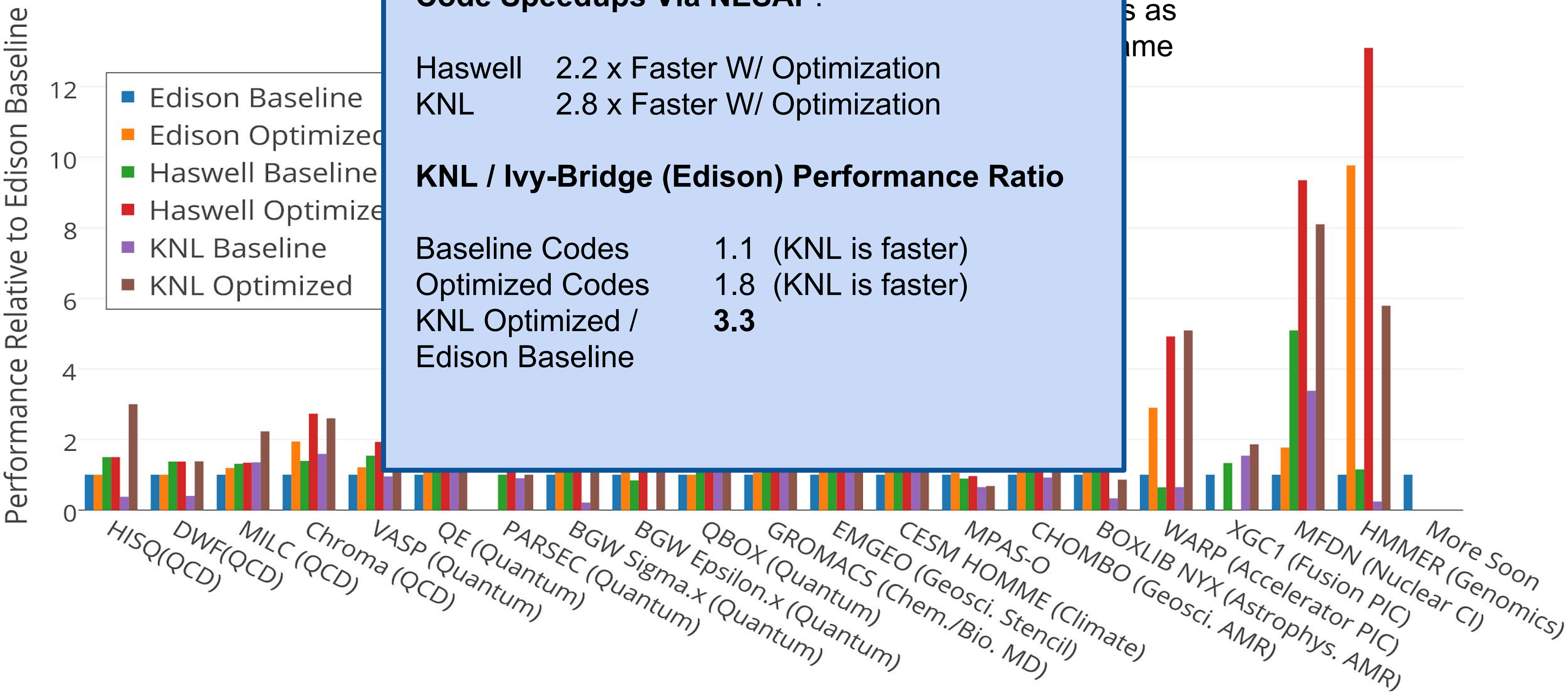
NERSC

Preliminary NESAP Code Performance on KNL

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



Preliminary NESAP Code Performance on KNL



PRELMINARY

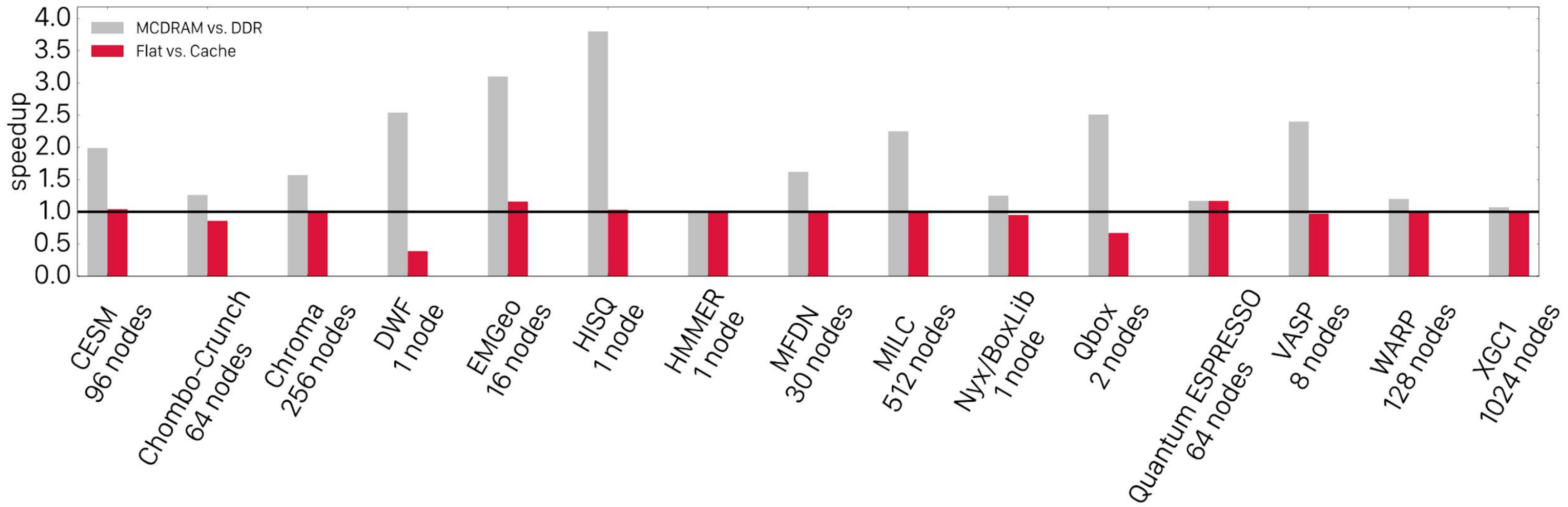
Code Speedups Via NESAP:

Haswell 2.2 x Faster W/ Optimization
KNL 2.8 x Faster W/ Optimization

KNL / Ivy-Bridge (Edison) Performance Ratio

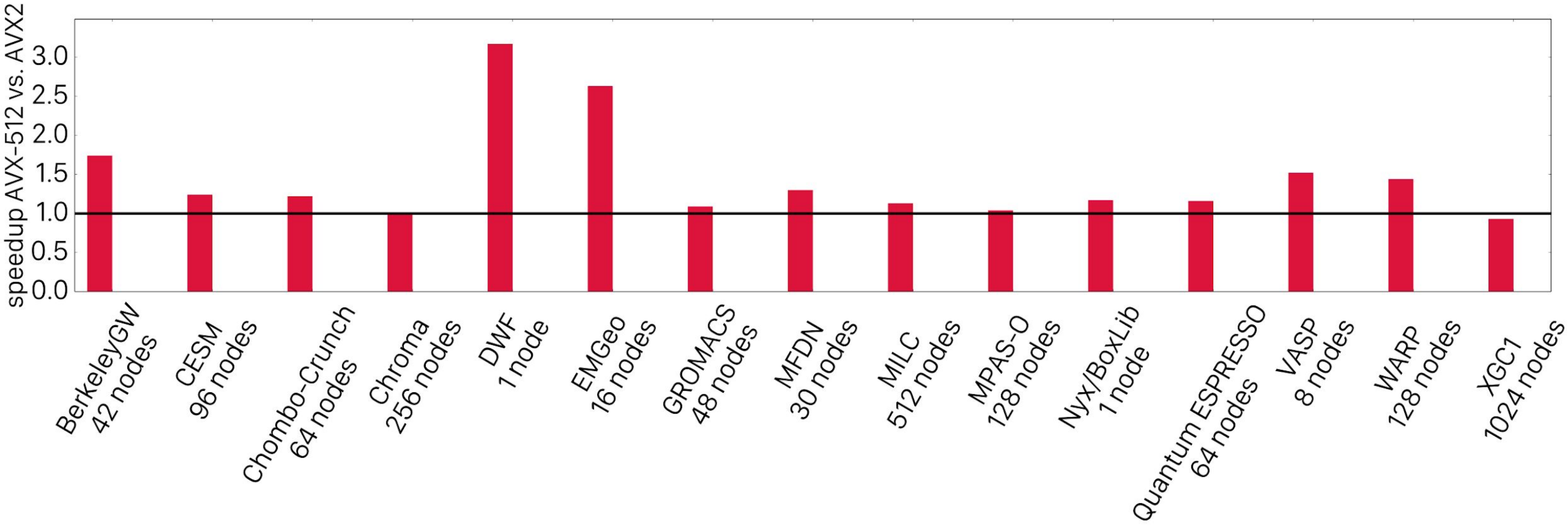
Baseline Codes	1.1	(KNL is faster)
Optimized Codes	1.8	(KNL is faster)
KNL Optimized / Edison Baseline	**3.3**	

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 Craypat, 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)

- 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

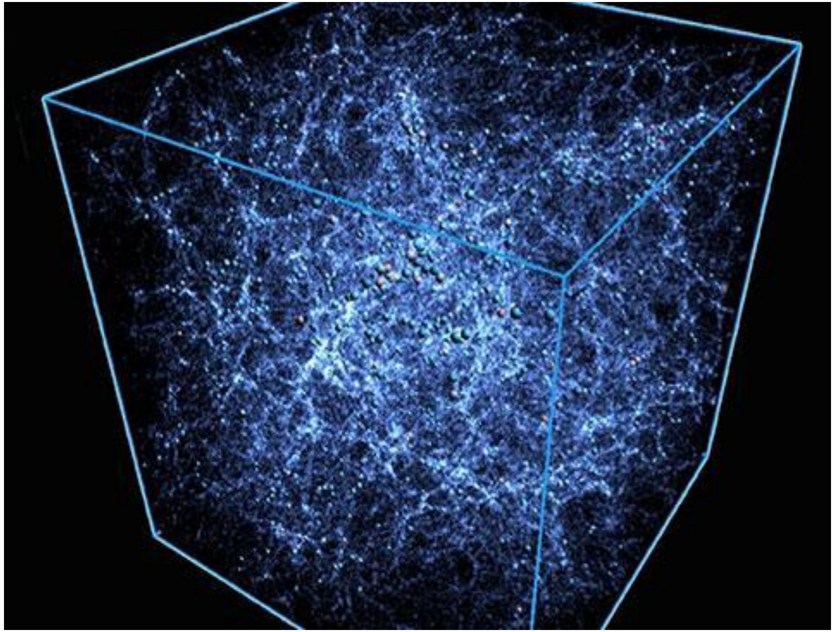
NERSC

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)

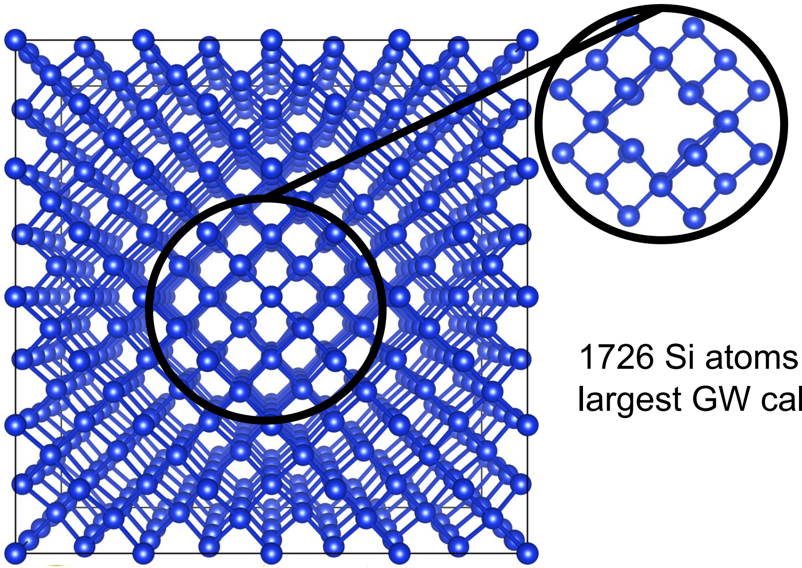


Large Scale Science Being Done on Cori

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.

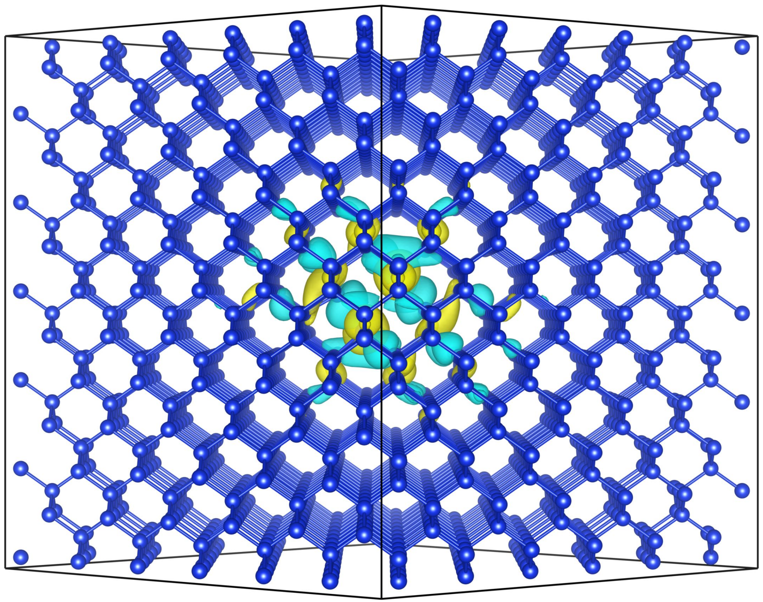
(Quantum ESPRESSO and BerkeleyGW)

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



END, Thank you!