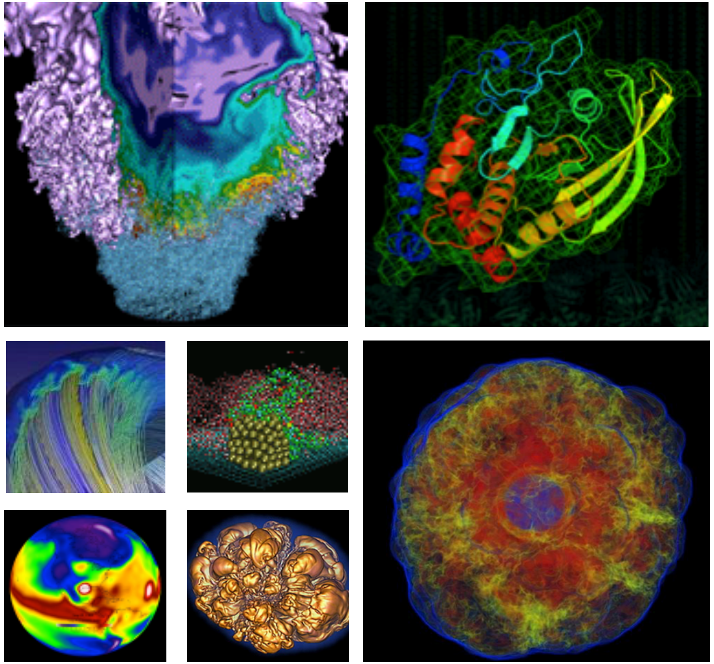
Many-Cores for the Masses: A Year With the Cori System at NERSC



NERSC

Richard Gerber
Jack Deslippe

Intel HPC Developer Conference
November, 2017

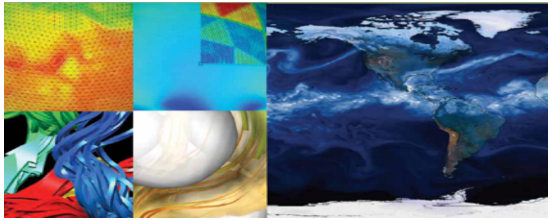
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

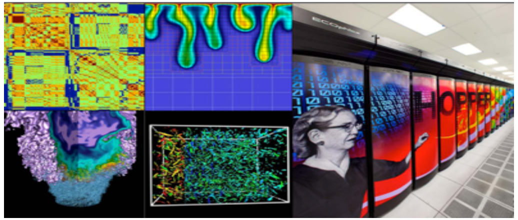
NERSC: Mission HPC for DOE Office of Science

U.S. DEPARTMENT OF ENERGY | Office of Science

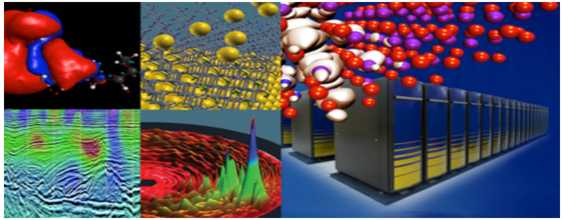
Largest funder of physical science research in U.S.



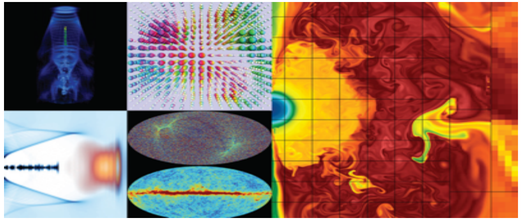
Bio Energy, Environment



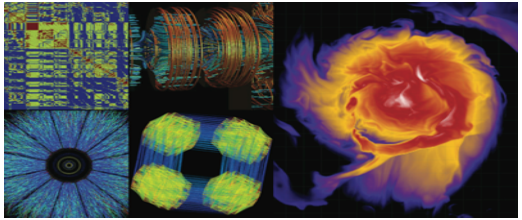
Computing



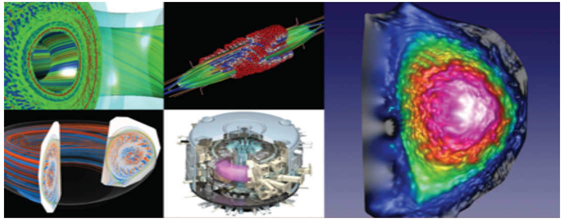
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics



Nuclear Physics



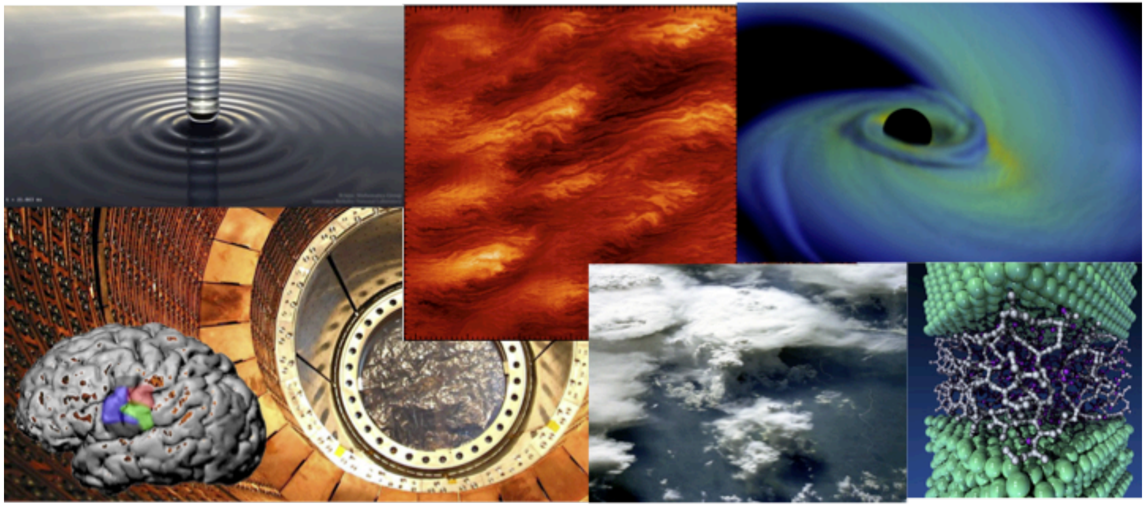
Fusion Energy, Plasma Physics

6,000 users, 700 projects, 700 codes, 48 states, 40 countries, universities & national labs

Focus on Science

NERSC users produce publish more than any other center in the world*; ~2K/year

1,036 citations via Web of Science in 2017 so far (underestimate!)



2017 to date

nature
International weekly journal of science

5 in Nature
30 in Nature Comm.
70 in 12 journals

Science
AAAS

4 in Science

11 in PNAS

PNAS
Proceedings of the National Academy of Sciences of the United States of America www.pnas.org

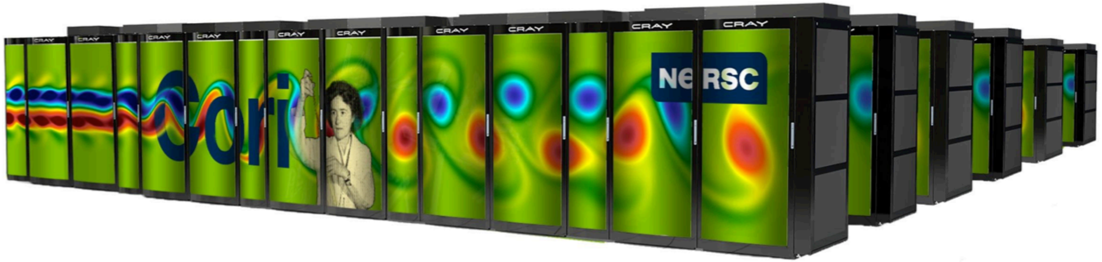
6 Nobel-prize winning users

* as far as I can tell

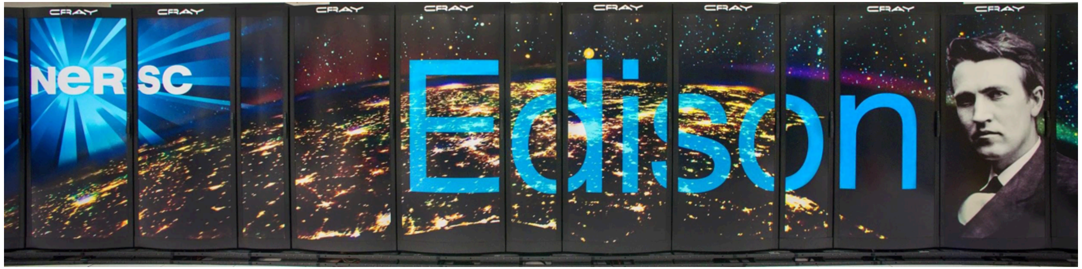
High Performance Computing Systems

Cori

9,300 Intel Xeon Phi “KNL” manycore nodes
2,000 Intel Xeon “Haswell” nodes
700,000 processor cores, 1.2 PB memory
Cray XC40 / Aries Dragonfly interconnect
30 PB Lustre Cray Sonexion scratch FS
1.5 PB Burst Buffer
Haswell: ~1 B NHrs/yr; KNL: ~6 B NHrs/yr



#6 on June 2017 Top 500 list

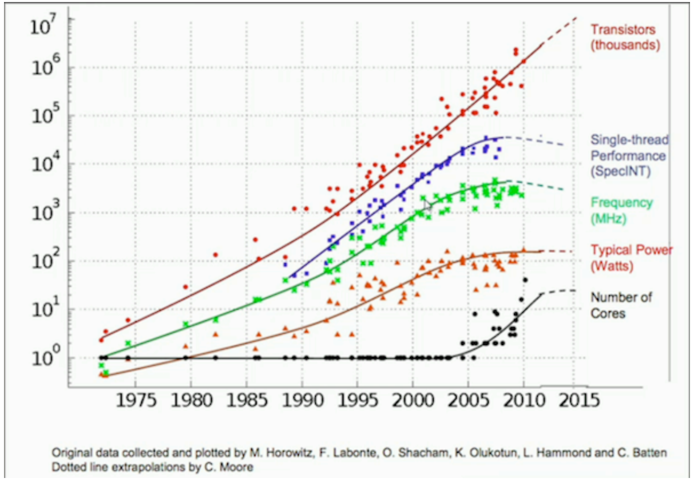


Edison

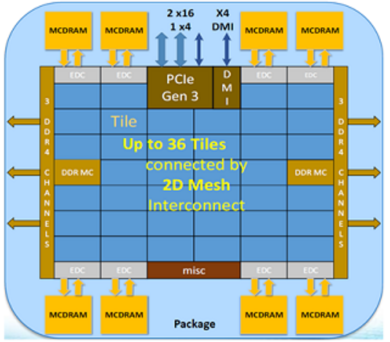
5,560 Ivy Bridge Nodes / 24 cores/node
133 K cores, 64 GB memory/node
Cray XC30 / Aries Dragonfly interconnect
6 PB Lustre Cray Sonexion scratch FS
Edison: ~2 B NHrs/yr

Change Has Arrived

Driven by power consumption and heat dissipation toward lightweight cores



Knights Landing Overview



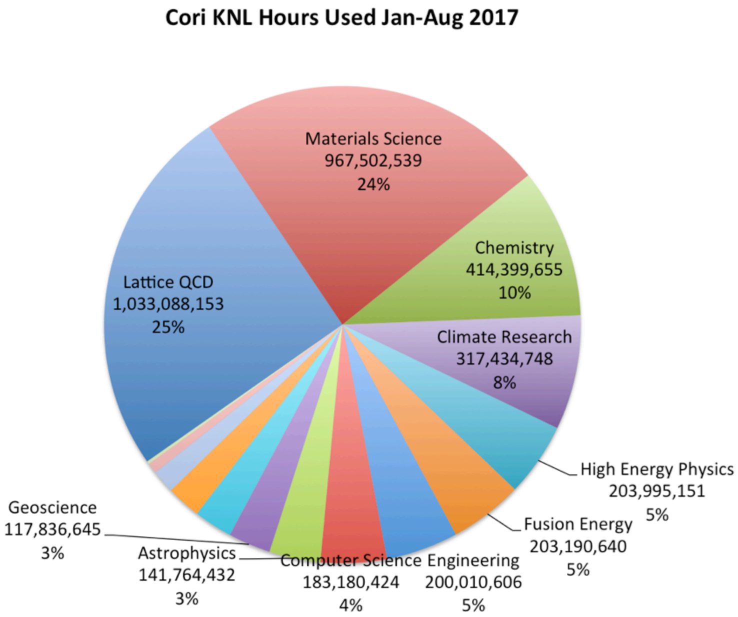
KNL: 215-230 W

2-socket Haswell: 270 W

Cori, a 30 PFlop system, will be a boon 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.

The Good News Is ...

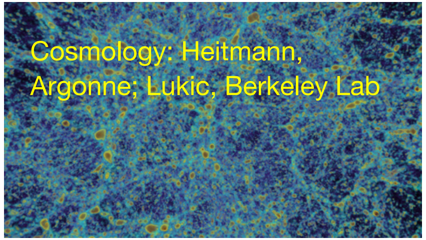
Adoption of KNL has been good and Cori KNL nodes are fully used by DOE Office of Science Researchers



- 150 projects have used > 1 M NERSC Hours
- 233 projects have used > 100 K NERSC Hrs
- Still leaves ~500 projects to move over
- 32% of hours use > 1,024 nodes (69K cores)
- ~~6~~ Gordon Bell submissions using Cori KNL
  7
  Massively Parallel 3D Image Reconstruction, Wang et al.

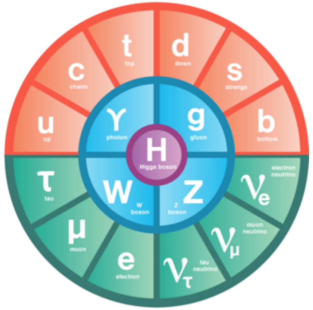
High Impact Science at Scale Projects

NERSC

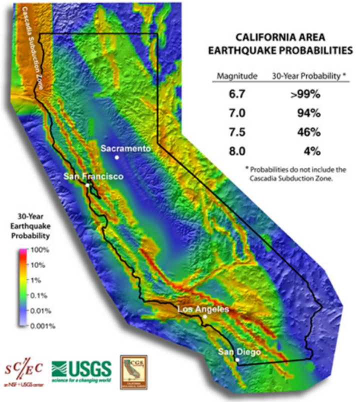


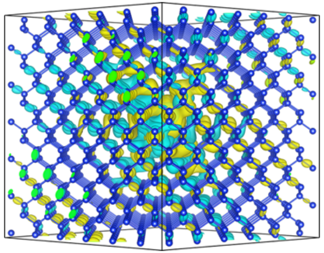
Cosmology: Heitmann, Argonne; Lukic, Berkeley Lab

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



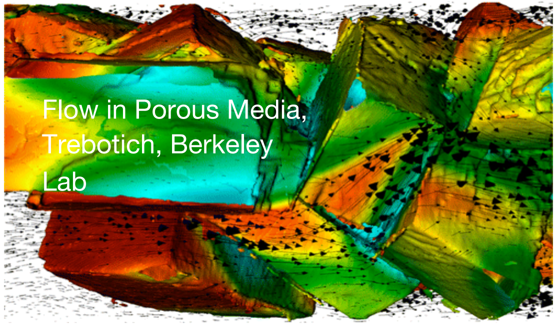
M8 Earthquake on the San Andreas Fault, Goulet, USC Earthquake Center



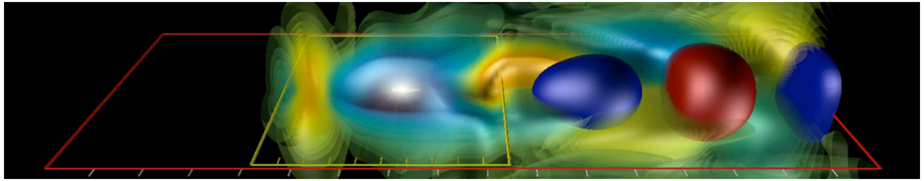


Optical Properties of Materials, Louie, UC Berkeley

Magnetic Reconnection, Stanier, Los Alamos



Flow in Porous Media, Trebotich, Berkeley Lab



Asymmetric Effects in Plasma Accelerators, Vay, Berkeley Lab

Deep Learning on Cori KNL

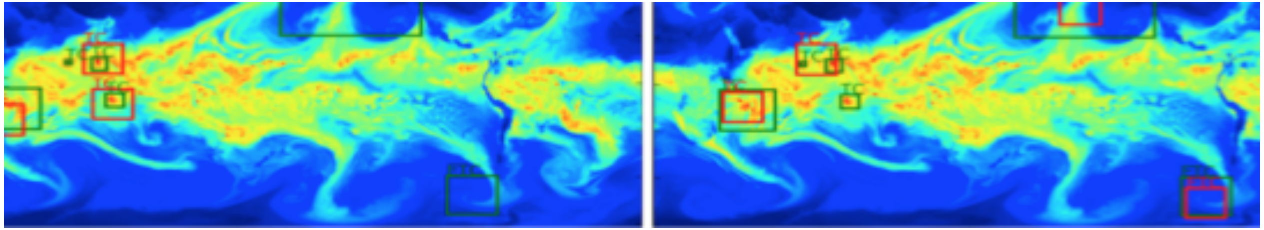
NERSC is actively exploring Deep Learning for Science

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

See Prabhat's plenary talk Sunday morning

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

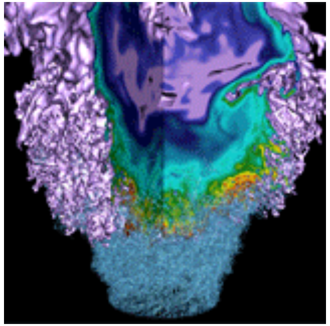
- Trained in 10s of minutes on 10 terabyte datasets, millions of Images
- 9600 nodes, optimized on KNL with IntelCaffe and MKL (NERSC / Intel collaboration)
- Synch + Asynch parameter update strategy for multi-node scaling (NERSC / Stanford)

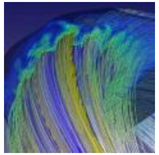


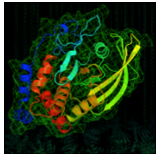
Identified extreme climate events using supervised (left) and semisupervised (right) deep learning. Green = ground truth, Red = predictions (confidence > 0.8). [NIPS 2017]

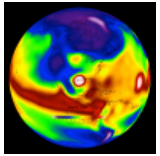
How Did We Get Here?

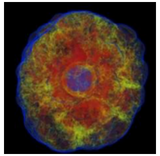
NESAP - NERSC Exascale Science Apps Program

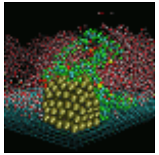












NERSC

U.S. DEPARTMENT OF ENERGY | Office of Science

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

NERSC Exascale Scientific Application Program (NESAP)

Goal: Prepare DOE Office of Science users for many core

Partner closely with ~20 application teams and apply lessons learned to broad NERSC user community.

20 applications cover (or serve as proxies for) > 50% of NERSC hours used in 2016

Activities:

- Close interactions with vendors
- Developer Workshops
- Early engagement with code teams
- Postdoc Program
- Leverage community efforts IXPUG
- Dungeon Sessions
- Early Access To KNL

Optimization Challenge and Strategy

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

- Many (Heterogeneous) 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 (Heterogeneous) Cores
- Bigger Vectors
- New ISA
- Multiple Memory Tiers

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

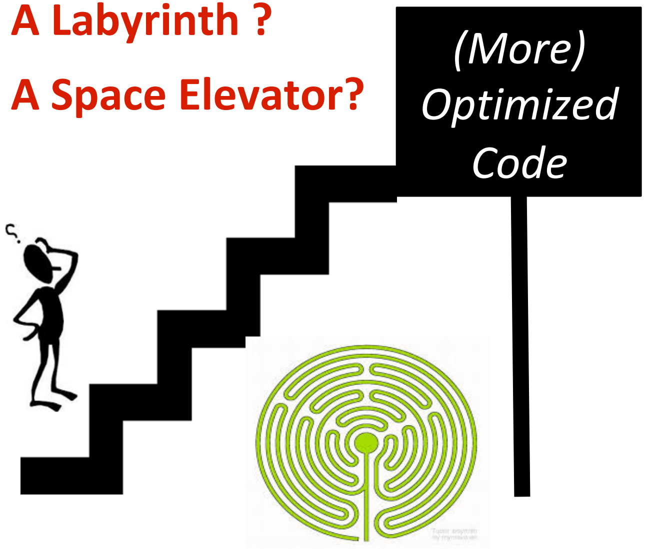
- How do you know what KNL hardware feature to target?
- How do you know how your code performs in an absolute sense and when to stop?

Optimizing Code For Cori is Like?

A. A Staircase ?

B. A Labyrinth ?

C. A Space Elevator?



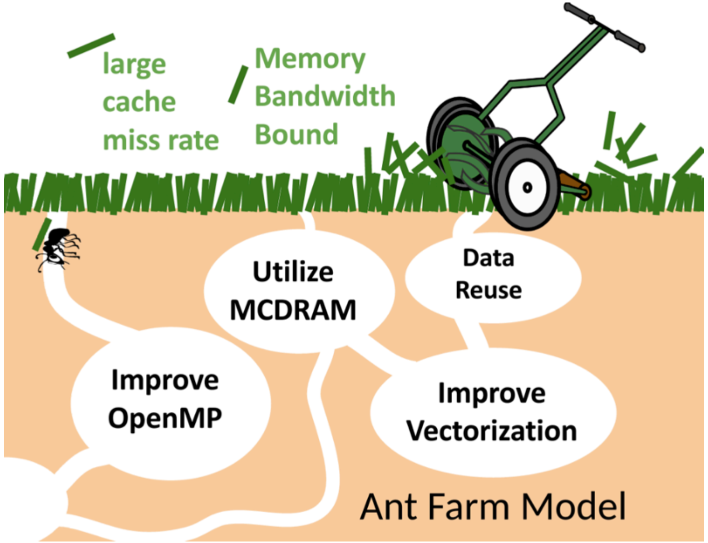
Optimization Challenge and Strategy

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

- Many (Heterogeneous) 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:

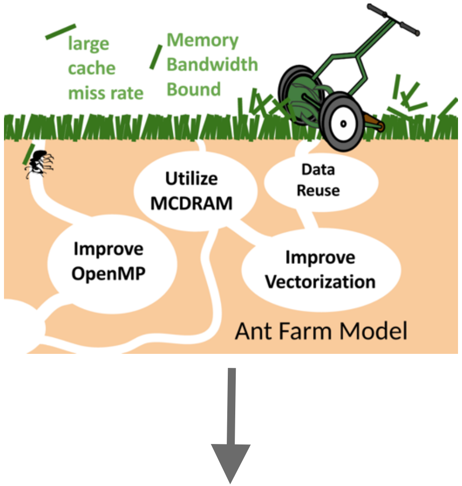
- Many (Heterogeneous) Cores
- Bigger Vectors
- New ISA
- Multiple Memory Tiers

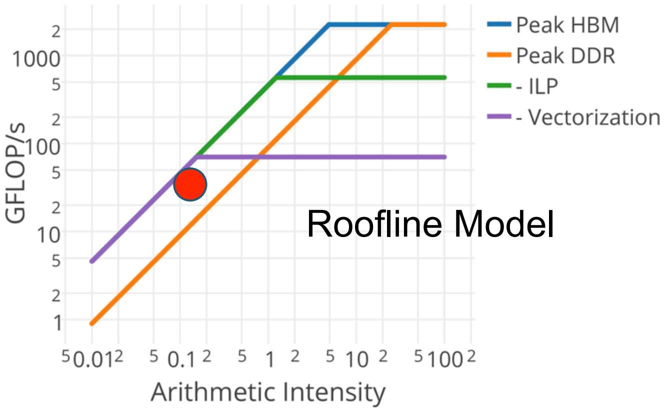
It is easy for users to get bogged down in the weeds:

- 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?

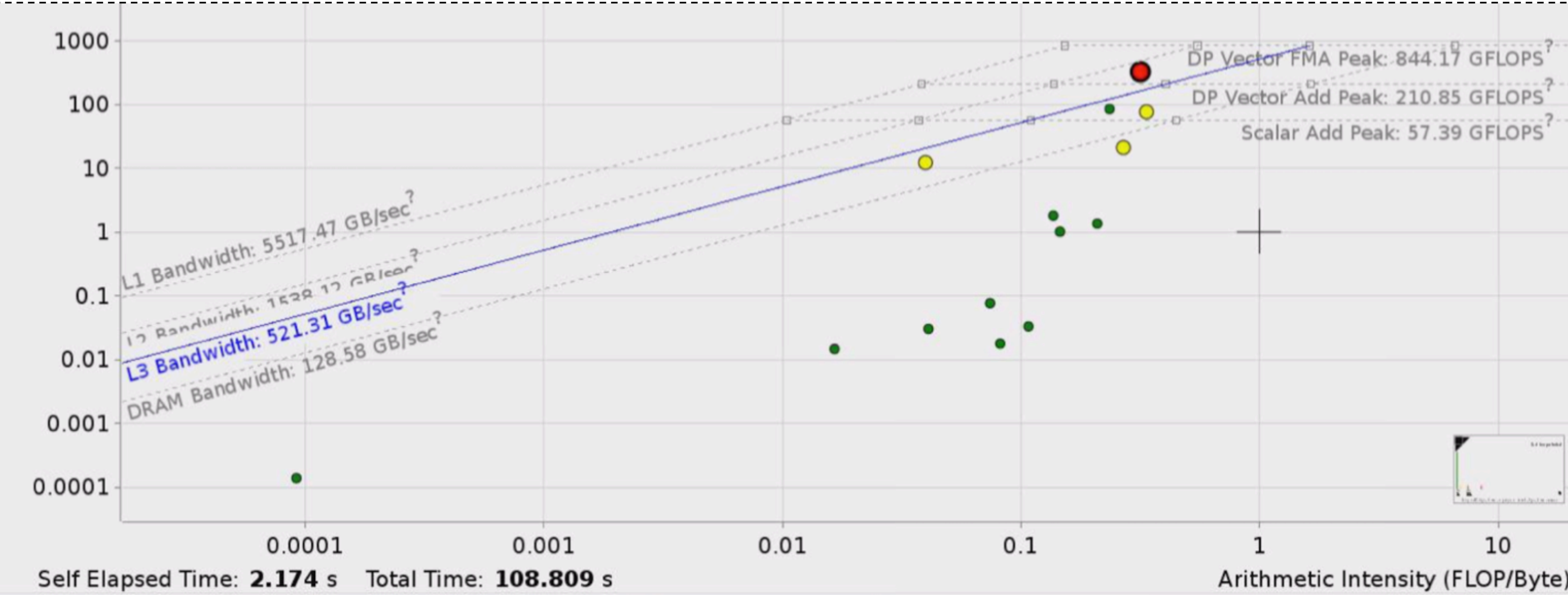
NERSC has developed tools and strategy for users to answer these questions:

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





Tools CoDesign

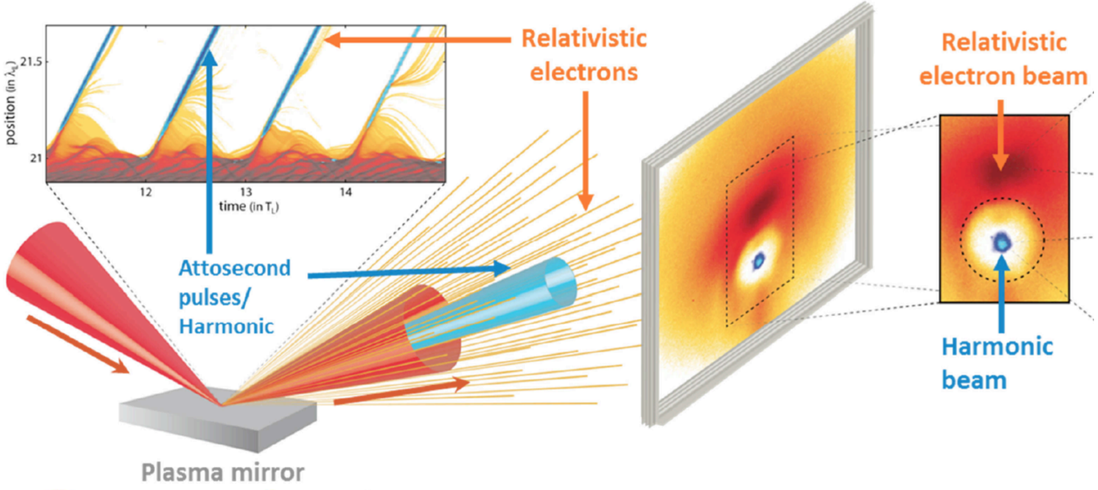


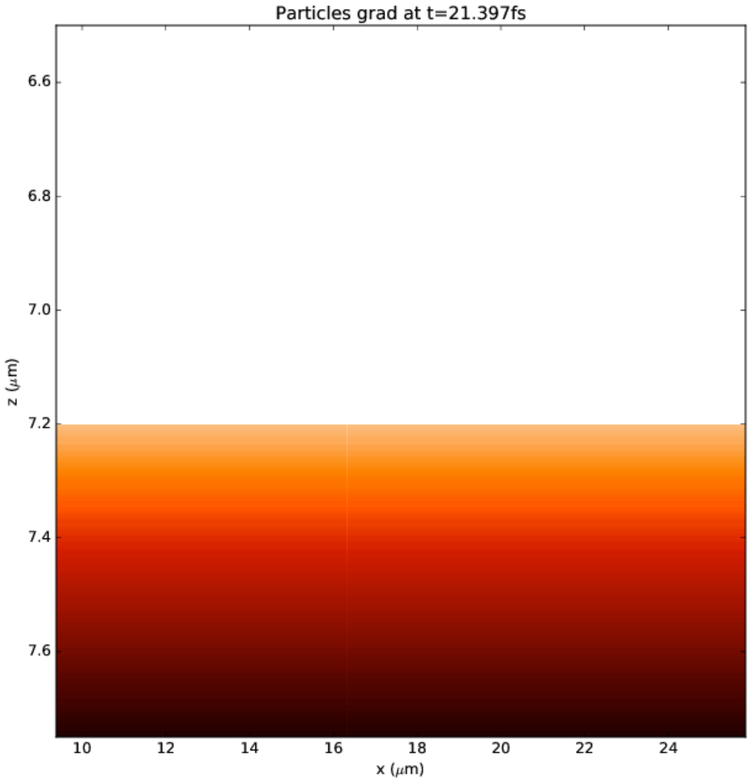
Intel Vector-Advisor Co-Design - Collaboration between NERSC, LBNL Computational Research, Intel

Example: WARP (Accelerator Modeling)

Particle in Cell (PIC) Applicaion for doing accelerator modeling and related applications. Developed library PICSAR for

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.



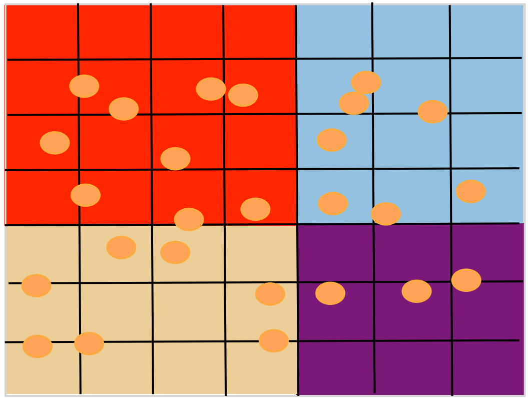


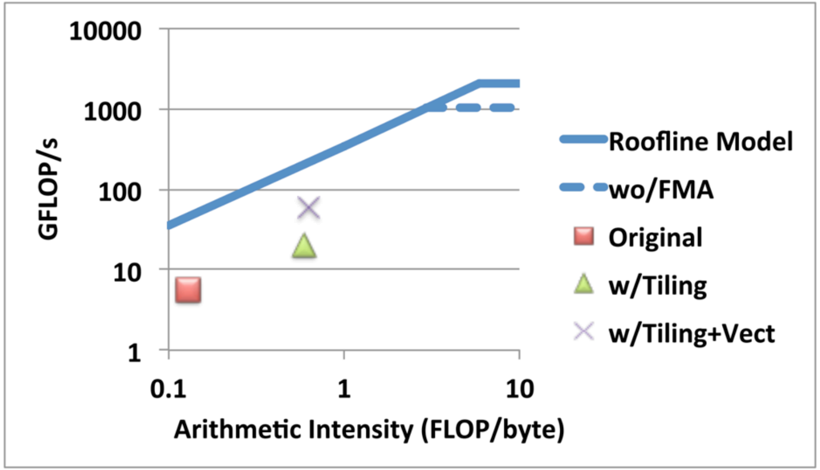
Animation from Plasma Mirror Simulations

Example: WARP (Accelerator Modeling)

Optimizations:

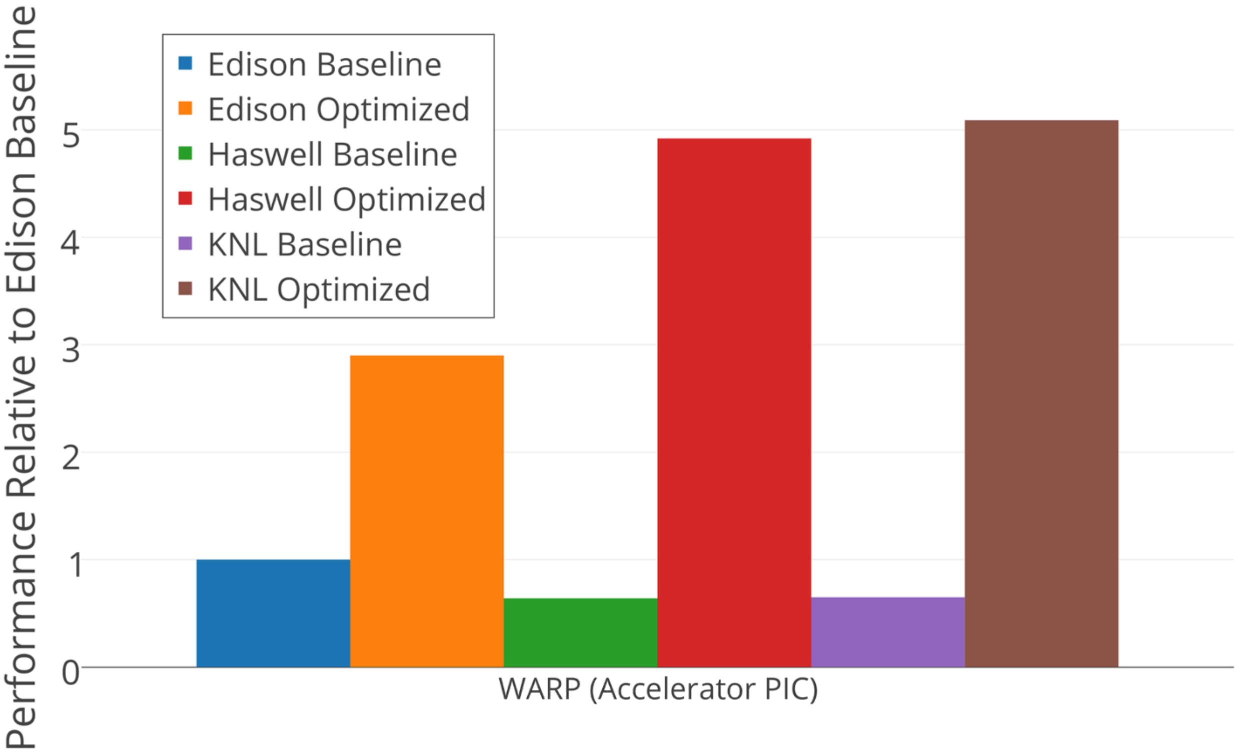
1. Add tiling over grid targetting L2 cache on both Xeon + Xeon-Phi Systems
2. Apply particle sorting + vectorization over particles (requires a number of datastructure changes)



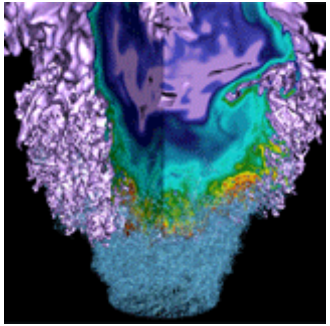


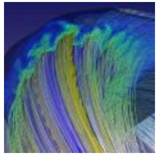
KNL Roofline

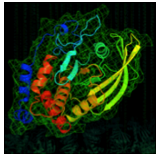
Example: WARP (Accelerator Modeling)

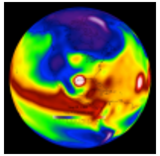


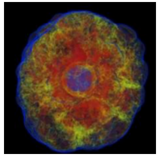
KNL Performance

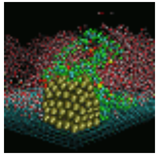




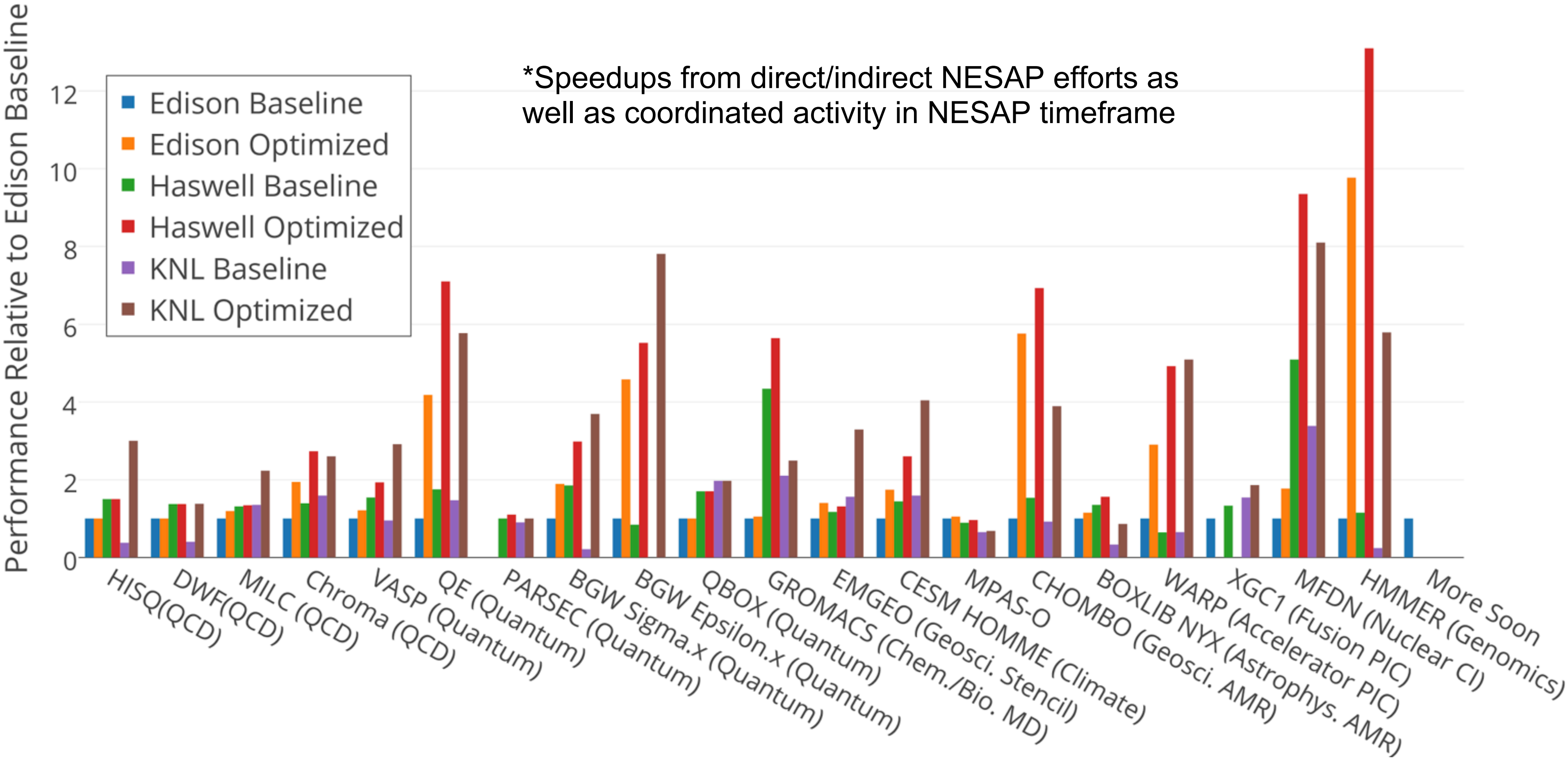




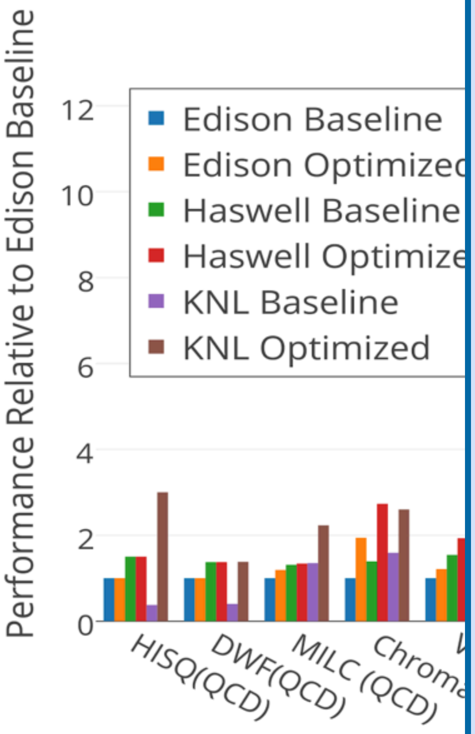




Preliminary NESAP Code Performance on KNL



Preliminary N … IL



PRELMINARY

Code Speedups Via NESAP:

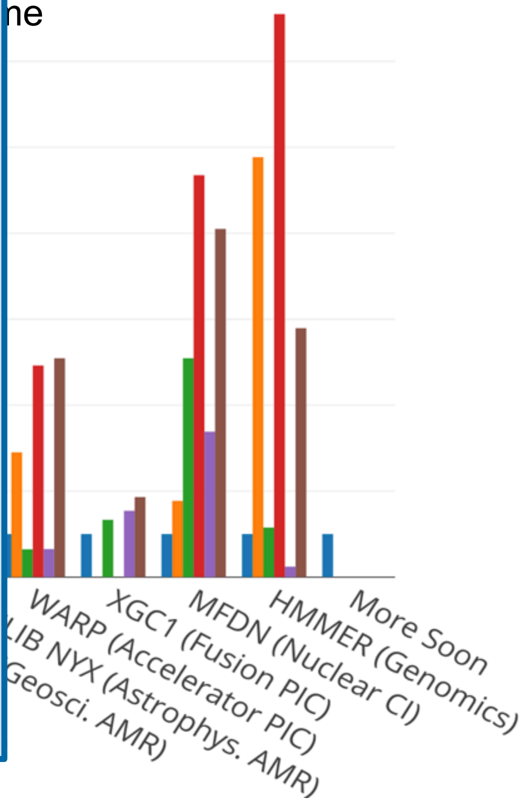
Haswell 2.3 x Faster W/ Optimization
KNL 3.5 x Faster W/ Optimization

KNL / Haswell Performance Ratio

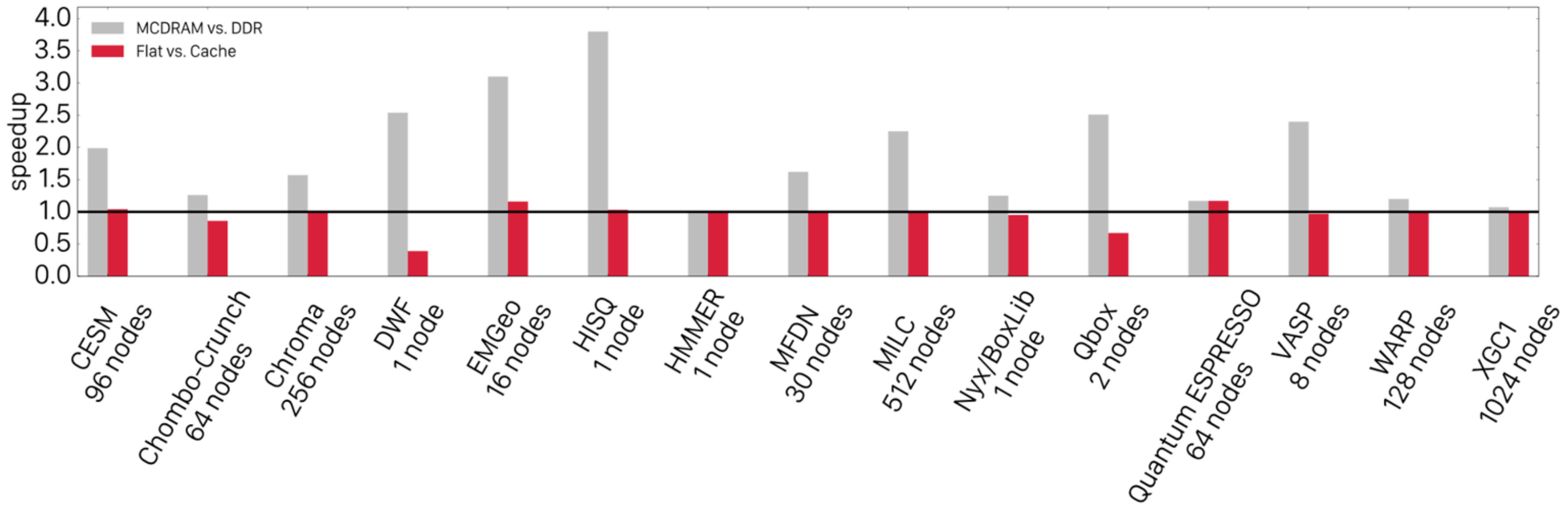
Baseline Codes	0.7	(KNL is slower)
Optimized Codes	1.1	(KNL is faster)
KNL Optimized / Haswell Baseline	**2.5**	

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**	

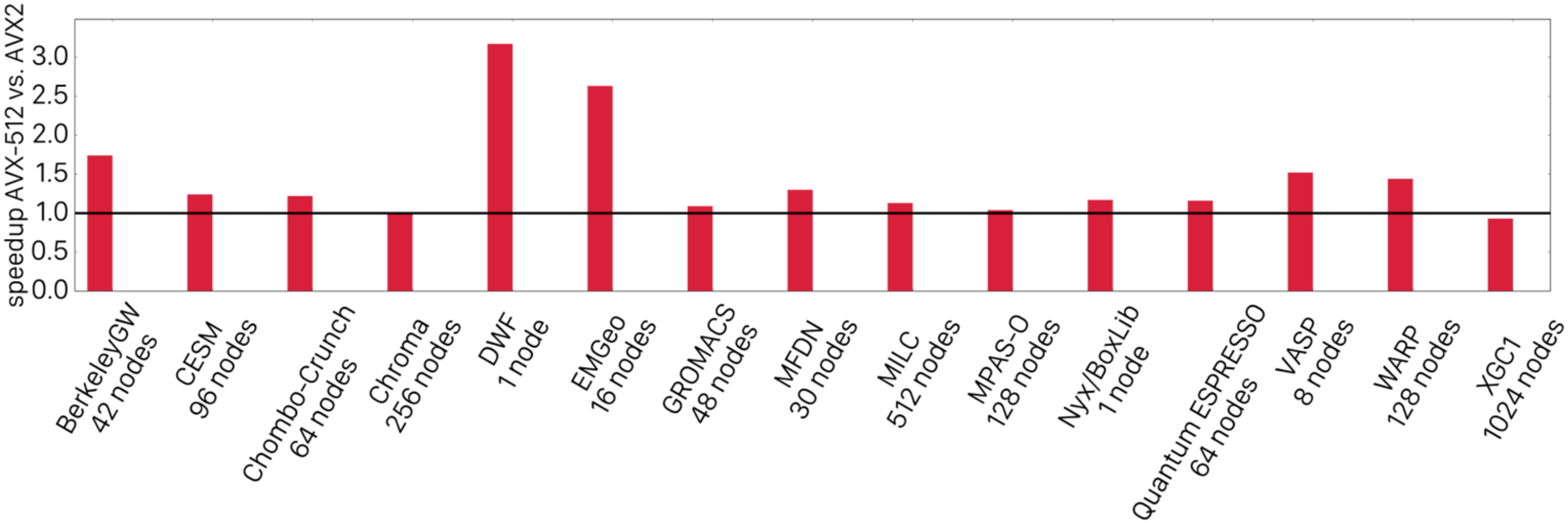


NESAP MCDRAM Effects



NESAP VPU Effects

AVX512 vs AVX2



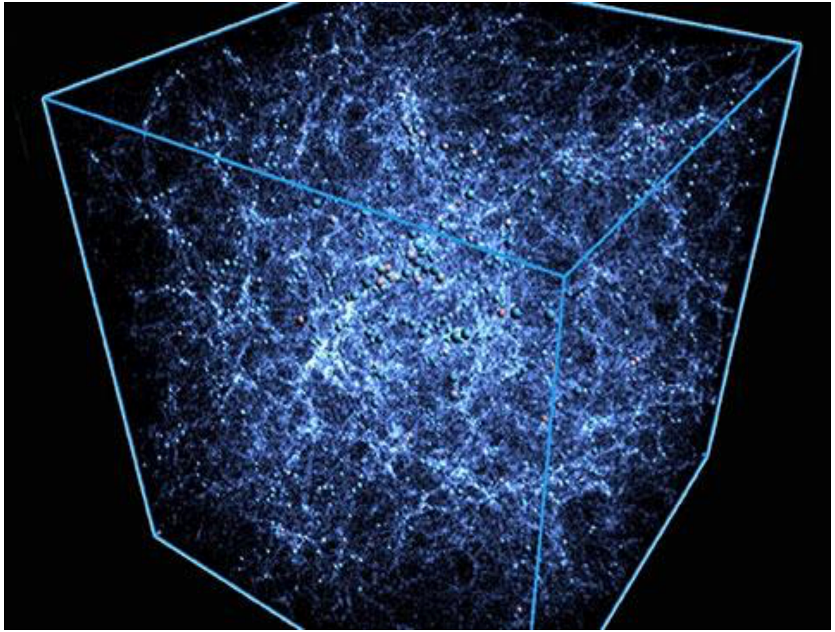
What did we learn?

- 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 task/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

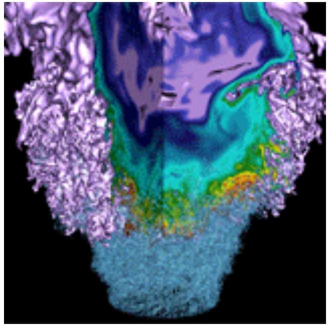
- NESAP For Data Prototype (Galactos)
- First anisotropic, 3-pt correlation computation on 2B Galaxies from Outer Rim Simulation
- Solves an open problem in cosmology for the next decade (LSST will observe 10B galaxies)
- Can address questions about the nature of dark-energy and gravity
- Novel $O(N^2)$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

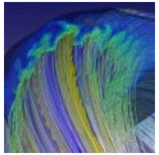


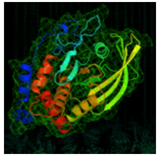
Scale:

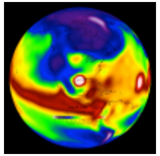
- 9600+ KNL Nodes (Significant Fraction of Peak)

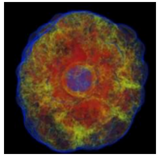
END, Thank you!

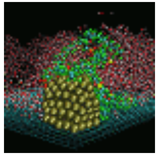












NERSC