Science on Cori - Hot off the Presses

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)
- Novel $O(N^2)$ algorithm based on spherical harmonics for 3-pt correlation

Scale:
- 9600+ KNL Nodes (Significant Fraction of Peak)
Defect States in Materials:
Important to material properties of transistors etc.

Defect properties require extremely large calculations to isolate defect states and high accuracy method.

Scale:
Simulated on Cori with up to 9600 KNL Nodes - Near Perfect Strong and Weak Scaling. Large percentage of peak performance obtained > 10 PFLOPS.

1726 Si atoms (~7K electrons) is largest GW calculation published
NERSC: the Mission HPC Facility for DOE Office of Science Research

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 & nat. labs
DOE SC Users are Coming From Traditional CPU Systems

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
+1.5 PB Burst Buffer 1.5 TB/s

Cori Haswell Nodes
1,900 Haswell Nodes / 32 cores/node
52 K cores, 128 GB memory/node
Cray XC40 / Aries Dragonfly interconnect
24 PB Lustre Cray Sonexion scratch FS
1.5 PB Burst Buffer
Cori KNL Nodes

Cray XC40 system with 9,600+ Intel Knights Landing compute nodes

68 cores / 96 GB DRAM / 16 GB HBM

Support the entire Office of Science research community

Begin to transition workload to energy efficient architectures
## What is different about Cori for NERSC Users?

<table>
<thead>
<tr>
<th>Edison (Ivy-Bridge):</th>
<th>Cori (KNL):</th>
</tr>
</thead>
<tbody>
<tr>
<td>● 5000+ Nodes</td>
<td>● 9600+ Nodes</td>
</tr>
<tr>
<td>● 12 Cores Per CPU / 24 Per Node</td>
<td>● 68 Physical Cores Per CPU</td>
</tr>
<tr>
<td>● 24 HW Threads Per CPU</td>
<td>● 272 HW Threads Per CPU</td>
</tr>
<tr>
<td>● 2.4-3.2 GHz</td>
<td>● 1.2-1.6 GHz</td>
</tr>
<tr>
<td>● Can do 8 Double Precision Operations per Cycle</td>
<td>● Can do 32 Double Precision Operations per Cycle</td>
</tr>
<tr>
<td>● 64 GB Memory Per Node</td>
<td>● 16 GB of Fast Memory</td>
</tr>
<tr>
<td>● ~100 GB/s Memory Bandwidth</td>
<td>96GB of DDR Memory</td>
</tr>
<tr>
<td></td>
<td>● MCDRAM ~450 GB/s Memory Bandwidth</td>
</tr>
<tr>
<td></td>
<td>○ Configurable as Memory or Cache</td>
</tr>
</tbody>
</table>
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.

Learned from ALCF ESP and OLCF CAAR

Activities:

- Close interactions with vendors
- Developer Workshops
- Early engagement with code teams
- Postdoc Program
- Leverage community efforts IXPUG
- Dungeon Sessions
- Early Access To KNL
Breakdown of Application Hours
Edison at Start of NESAP

NESAP Tier-1, 2 Code
NESAP Proxy Code or Tier-3 Code
NESAP Staffing at NERSC

Our Post-Docs are Going On to Benefit the HPC Community:

Mathieu Lobet - La Maison de la Simulation (CEA), France

Brian Friesen - NERSC (Applications Performance Group)
    LBNL CRD (US ECP AMR Development)

Tareq Malas - Intel (Applications Engineer)

More post-docs graduating soon ...
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
Collaboration on Tools

Intel Vector-Advisor Co-Design - Collaboration between NERSC, LBNL Computer Research Division and Intel
Facilities are Key to App-Readiness Success

- We Provide Venue for Successful Interaction Between Apps and Vendors
- Engagement point with IXPUG, User Communities (Exascale Workshops at CRT)
- Advocates for Standards and Portability
- Host for a number of Facility and Vendor Training (Vectorization, OpenMP, Tools/Compilers)
- Documentation Point for a Massive Amount of Lessons Learned about Tools and Architecture (VTune, SDE, HBM etc.) and Case Studies from Dungeon Sessions
Example: WARP (Accelerator Modeling)

Particle in Cell (PIC) Application for doing accelerator modeling and related applications.

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)

- Edison Baseline
- Edison Optimized
- Haswell Baseline
- Haswell Optimized
- KNL Baseline
- KNL Optimized

Performance Relative to Edison Baseline

Higher is Better
Preliminary NESAP Code Performance on KNL

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

Higher is Better
**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**

<table>
<thead>
<tr>
<th></th>
<th>Baseline Codes</th>
<th>Optimized Codes</th>
<th>KNL Optimized / Haswell Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Codes</td>
<td>0.7 (KNL is slower)</td>
<td>1.1 (KNL is faster)</td>
<td>2.5</td>
</tr>
<tr>
<td>Optimized Codes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KNL Optimized /</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haswell Baseline</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**KNL / Ivy-Bridge (Edison) Performance Ratio**

<table>
<thead>
<tr>
<th></th>
<th>Baseline Codes</th>
<th>Optimized Codes</th>
<th>KNL Optimized / Edison Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Codes</td>
<td>1.1 (KNL is faster)</td>
<td>1.8 (KNL is faster)</td>
<td>3.4</td>
</tr>
<tr>
<td>Optimized Codes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KNL Optimized /</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edison Baseline</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
NESAP VPU Effects

KNL AVX and FMA Speedups

- AVX512
- FMA
Wider Vectors > Complex ISA
More Cores > Wider Vectors > Complex ISA
More Bandwidth > More Cores > Wider Vectors > Complex ISA
Faster Clockspeed

> 

More Bandwidth > More Cores > Wider Vectors > Complex ISA
Faster Clockspeed

> 

More Bandwidth > More Cores > Wider Vectors > Complex ISA
Demand for KNL Nodes is High

Daily usage on Cori KNL nodes

Queue backlog on KNL nodes
## All DOE Offices Represented in Top 10 Projects on KNL

<table>
<thead>
<tr>
<th>DOE Office</th>
<th>Project Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASCR</td>
<td>Prabhat - NERSC Data and Analytics</td>
</tr>
<tr>
<td>BER</td>
<td>Lai-Yung Ruby Leung - Accelerated Climate Modeling for Energy</td>
</tr>
<tr>
<td>BES</td>
<td>Paul Kent - Extending the capabilities of Quantum Espresso for Cori</td>
</tr>
<tr>
<td>FES</td>
<td>CS Chang - Center for Edge Physics Simulation: SciDAC-3 Center</td>
</tr>
<tr>
<td>HEP</td>
<td>Doug Touissant - Quantum Chromodynamics with four flavors of dynamical quarks</td>
</tr>
<tr>
<td>NP</td>
<td>Norman Christ - Domain Wall Fermions and Highly Improved Staggered Quarks for Lattice QCD</td>
</tr>
</tbody>
</table>
Deep-Learning on Full Cori System

- 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

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

EXTRA SLIDES
Quantum Supremacy Simulations

• 45 Qubit simulation is largest ever quantum computing simulation ever
• Previously largest calculation was 42 (complexity is exponential)
• Simulationas are important for validating prototype quantum computers devices
• Team lead by ETH scientists collaborators at Google, LBNL’s Computational Research Divisic

Scale:
- >8000 KNL nodes
- 0.5 Petabytes of memory used (~2^45)
- 0.43 PetaFLOPS (Bandwidth bound)
Creating a Catalog of All Objects in the Sky

- The world’s largest scientific generative model has been developed
- Core statistical procedure for scalable inference has been implemented in Julia
- Joint inference across multiple images and instruments is conducted to produce uncertainties in parameter estimates of celestial bodies
- DESI instrument will use these estimates for target selection
- Code written in Julia, optimized for execution on KNL

Scale:
- 9000+ KNL Nodes
- 55 TByte SDSS dataset
- High-productivity language able to reach petascale