Preparing NERSC Applications for Perlmutter as an Exascale Waypoint

Charlene Yang
Application Performance Specialist
Pawsey Supercomputing Centre
Mar 26 2019
NERSC is the mission High Performance Computing facility for the DOE SC

- 7,000 Users
- 800 Projects
- 700 Codes
- 2000 NERSC citations per year

Simulations at scale

Data analysis support for DOE’s experimental and observational facilities

Photo Credit: CAMERA
NERSC has a dual mission to advance science and the state-of-the-art in supercomputing

- We collaborate with computer companies years before a system’s delivery to deploy advanced systems with new capabilities at large scale
- We provide a highly customized software and programming environment for science applications
- We are tightly coupled with the workflows of DOE’s experimental and observational facilities – ingesting tens of terabytes of data each day
- Our staff provide advanced application and system performance expertise to users
Perlmutter was announced 30 Oct 2018

“Continued leadership in high performance computing is vital to America’s competitiveness, prosperity, and national security,” said U.S. Secretary of Energy Rick Perry. “This advanced new system, created in close partnership with U.S. industry, will give American scientists a powerful new tool of discovery and innovation and will be an important milestone on the road to the coming era of exascale computing.”

"We are very excited about the Perlmutter system," said NERSC Director Sudip Dosanjh. “It will provide a significant increase in capability for our users and a platform to continue transitioning our very broad workload to energy efficient architectures. The system is optimized for science, and we will collaborate with Cray, NVIDIA and AMD to ensure that Perlmutter meets the computational and data needs of our users. We are also launching a major power and cooling upgrade in Berkeley Lab’s Shyh Wang Hall, home to NERSC, to prepare the facility for Perlmutter.”
NERSC Systems Roadmap

NERSC-7: Edison Multicore CPU
2013

NERSC-8: Cori
Manycore CPU
NESAP Launched: transition applications to advanced architectures
2016

NERSC-9: CPU and GPU nodes
Continued transition of applications and support for complex workflows
2020

NERSC-10: Exa system
2024

NERSC-11: Beyond Moore
2028

Increasing need for energy-efficient architectures
Cori: A pre-exascale supercomputer for the Office of Science workload

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

1,600 Haswell processor nodes
NVRAM Burst Buffer 1.5 PB, 1.5 TB/sec
30 PB of disk, >700 GB/sec I/O bandwidth
Integrated with Cori Haswell nodes on Aries network for data / simulation / analysis on one system
## Perlmutter is a Pre-Exascale System

<table>
<thead>
<tr>
<th>Pre-Exascale Systems</th>
<th>Exascale Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2013</strong></td>
<td><strong>2021-2023</strong></td>
</tr>
<tr>
<td><strong>2016</strong></td>
<td></td>
</tr>
<tr>
<td><strong>2018</strong></td>
<td></td>
</tr>
<tr>
<td><strong>2020</strong></td>
<td></td>
</tr>
</tbody>
</table>

### Pre-Exascale Systems
- **Mira**
  - Argonne
  - IBM BG/Q
- **Theta**
  - Argonne
  - Intel/Cray KNL
- **Titan**
  - ORNL
  - Cray/Nvidia K20
- **Sequoia**
  - LLNL
  - IBM BG/Q
- **Trinity**
  - LANL/SNL
  - Cray/Intel Xeon/KNL
- **Summit**
  - ORNL
  - IBM/NVIDIA P9/Volta

### Exascale Systems
- **LBNL**
  - Cray/NVIDIA/AMD
- **A21**
  - Argonne
  - Intel/Cray
- **Frontier**
  - ORNL
  - TBD
- **Crossroads**
  - LANL/SNL
  - TBD
- **TBD**
  - LLNL
  - TBD

### Timeline
- **2013**
- **2016**
- **2018**
- **2020**
- **2021-2023**
NERSC-9 will be named after Saul Perlmutter

• Winner of 2011 Nobel Prize in Physics for discovery of the accelerating expansion of the universe.
• Supernova Cosmology Project, lead by Perlmutter, was a pioneer in using NERSC supercomputers combine large scale simulations with experimental data analysis
• Login “saul.nersc.gov”
Perlmutter: A System Optimized for Science

- GPU-accelerated and CPU-only nodes meet the needs of large scale simulation and data analysis from experimental facilities
- Cray “Slingshot” - High-performance, scalable, low-latency Ethernet-compatible network
- Single-tier All-Flash Lustre based HPC file system, 6x Cori’s bandwidth
- Dedicated login and high memory nodes to support complex workflows
From the start NERSC-9 had requirements of simulation and data users in mind

- All Flash file system for workflow acceleration
- Optimized network for data ingest from experimental facilities
- Dedicated workflow management and interactive nodes
- Real-time scheduling capabilities
- Supported analytics stack including latest ML/DL software
- System software supporting rolling upgrades for improved resilience
AMD “Milan” CPU
- ~64 cores
- “ZEN 3” cores - 7nm+
- AVX2 SIMD (256 bit)

8 channels DDR memory
- >= 256 GiB total per node

1 Slingshot connection
- 1x25 GB/s

~ 1x Cori
4x NVIDIA “Volta-next” GPU
- > 7 TF
- > 32 GiB, HBM-2
- NVLINK

1x AMD CPU
4 Slingshot connections
- 4x25 GB/s

GPU direct, Unified Virtual Memory (UVM)
2-3x Cori
Slingshot Network

- **High Performance scalable interconnect**
  - Low latency, high-bandwidth, MPI performance enhancements
  - 3 hops between any pair of nodes
  - Sophisticated congestion control and adaptive routing to minimize tail latency

- **Ethernet compatible**
  - Blurs the line between the inside and the outside of the machine
  - Allow for seamless external communication
  - Direct interface to storage
Perlmutter has an All-Flash Filesystem

- **Fast across many dimensions**
  - 4 TB/s sustained bandwidth
  - 7,000,000 IOPS
  - 3,200,000 file creates/sec
- **Usable for NERSC users**
  - 30 PB usable capacity
  - Familiar Lustre interfaces
  - New data movement capabilities
- **Optimized for NERSC data workloads**
  - NEW small-file I/O improvements
  - NEW features for high IOPS, non-sequential I/O
Analytics and Workflow Integration

- **Software**
  - Optimized analytics libraries, includes Cray Analytics stack
  - Collaboration with NVIDIA for Python-based data analytics support
  - Support for containers

- **NERSC-9 will aid complex end-to-end workflows**
  - Slurm co-scheduling of multiple resources and real-time/deadline scheduling
  - Workflow nodes: container-based services
    - Connections to scalable, user workflow pool (via Spin) with network/scheduler access
  - High-availability workflow architecture and system resiliency for real-time use-cases
SSI-based Design
Perlmutter: A System Optimized for Science

- GPU-accelerated and CPU-only nodes meet the needs of large scale simulation and data analysis from experimental facilities
- Cray “Slingshot” - High-performance, scalable, low-latency Ethernet-compatible network
- Single-tier All-Flash Lustre based HPC file system, 6x Cori’s bandwidth
- Dedicated login and high memory nodes to support complex workflows

How do we optimize the size of each partition?
NERSC System Utilization (Aug’17 - Jul’18)

- 3 codes > 25% of the workload
- 10 codes > 50% of the workload
- 30 codes > 75% of the workload
- Over 600 codes comprise the remaining 25% of the workload.
GPU Readiness Among NERSC Codes (Aug’17 - Jul’18)

**Breakdown of Hours at NERSC**

- **Enabled:** Most features are ported and performant (32%)
- **Kernels:** Ports of some kernels have been documented (10%)
- **Proxy:** Kernels in related codes have been ported (19%)
- **Unlikely:** A GPU port would require major effort (14%)
- **Unknown:** GPU readiness cannot be assessed at this time (25%)

A number of applications in NERSC workload are GPU enabled already.

We will leverage existing GPU codes from CAAR + Community.
Select codes to represent the anticipated workload

• Include key applications from the current workload.
• Add apps that are expected to contribute significantly to the future workload.

**Scalable System Improvement**
Measures aggregate performance of HPC machine

• How many more copies of the benchmark can be run relative to the reference machine
• Performance relative to reference machine

**SSI - Sustained System Improvement**
Hetero system design & price sensitivity: Budget for GPUs increases as GPU price drops

Chart explores an isocost design space

- Vary the budget allocated to GPUs
- Assume GPU enabled applications have performance advantage = 10x per node, 3 of 8 apps are still CPU only.
- Examine GPU/CPU node cost ratio

<table>
<thead>
<tr>
<th>GPU / CPU $ per node</th>
<th>SSI increase vs. CPU-Only (@ budget %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:1</td>
<td>None</td>
</tr>
<tr>
<td>6:1</td>
<td>1.05x @ 25%</td>
</tr>
<tr>
<td>4:1</td>
<td>1.23x @ 50%</td>
</tr>
</tbody>
</table>

Circles: 50% CPU nodes + 50% GPU nodes
Stars: Optimal system configuration.

Explore an isocost design space

- Assume 8:1 GPU/CPU node cost ratio.
- Vary the budget allocated to GPUs
- Examine GPU / CPU performance gains such as those obtained by software optimization & tuning. 5 of 8 codes have 10x, 20x, 30x speedup.

<table>
<thead>
<tr>
<th>GPU / CPU perf. per node</th>
<th>SSI increase vs. CPU-Only (@ budget %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10x</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>No justification for GPUs</td>
</tr>
<tr>
<td>20x</td>
<td>1.15x @ 45%</td>
</tr>
<tr>
<td></td>
<td>Compare to 1.23x for 10x at 4:1 GPU/CPU cost ratio</td>
</tr>
<tr>
<td>30x</td>
<td>1.40x @ 60%</td>
</tr>
<tr>
<td></td>
<td>Compare to 3x from NESAP for KNL</td>
</tr>
</tbody>
</table>

NESAP Overview
Application Readiness Strategy for Perlmutter

NERSC’s Challenge

How to enable NERSC’s diverse community of 7,000 users, 800 projects, and 700 codes to run on advanced architectures like Perlmutter and beyond?
Application Readiness Strategy for Perlmutter

NERSC Exascale Science Application Program (NESAP)

- http://nersc.gov/users/application-performance/nesap/
- Engage ~25 Applications
- up to 17 postdoctoral fellows
- Deep partnerships with every SC Office area
- Leverage vendor expertise and hack-a-thons
- Knowledge transfer through documentation and training for all users
- Optimize codes with improvements relevant to multiple architectures
NESAP for Perlmutter will extend activities from NESAP for Cori

1. Identifying and exploiting on-node parallelism
2. Understanding and improving data-locality within the memory hierarchy

Knowledge and skills of multi/many-core optimization on HSW/KNL transferrable to AMD CPUs

What’s New for NERSC Users?

1. Heterogeneous compute elements - NVIDIA GPUs
2. Identification and exploitation of even more parallelism
3. Data locality again, host/device

Emphasis on performance-portable programming approach: Continuity from Cori through future NERSC systems
• 6 NESAP for Data apps will be continued. Additional apps focused on experimental facilities.
• 5 ECP Apps Jointly Selected (Participation Funded by ECP)
• Open call for proposals. Reviewed by a committee of NERSC staff, external reviewers and input from DOE PMs.
  – App selection will contain multiple applications from each SC Office and algorithm area
  – Additional applications (beyond 25) will be selected for second tier NESAP with access to vendor/training resources and early access

Simulation
~12 Apps

Data Analysis
~8 Apps

Learning
~5 Apps
NESAP for Perlmutter Timeline

- **NESAP-2 Call for Proposals**: (Oct. 2018) HPC, Data/Analytics, Learning
- **Code Team Selection**: (Dec. 2018)
- **Dungeon Sessions Begin**
- **Finalize Edison Reference Numbers**
- **System Delivery**
- **Early Access**

2018:
- **NESAP for Data (6 Existing Apps)**
- **Begin engagement with ECP applications**
- **NESAP-2**

2019:
- Dungeon Sessions Begin
- Finalize Edison Reference Numbers

2020:
- System Delivery

2021:
Resources Available to NESAP Awardees

- 1 hackathon session per quarter (Center of Excellence)
  - NERSC, Cray, NVIDIA engineer attendance
  - 3-4 code teams per hackathon
- Cray/NVIDIA engineer time before and after sessions
  - 6-week ‘ramp-up’ period with code teams and Cray/NVIDIA to ensure everyone is fully prepared to work on hackathon day 1
  - Tutorials/deep dives into GPU programming models, profiling tools, etc
- NESAP postdocs (NERSC will hire up to 17)
- NERSC application performance specialist attention
- General programming, performance and tools training
- Early access (Perlmutter and GPU testbed)

Tier 1 apps have higher priority than Tier 2!
## Support for NESAP Teams

<table>
<thead>
<tr>
<th>Benefit</th>
<th>Tier 1</th>
<th>Tier 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early Access to Perlmutter</td>
<td>yes</td>
<td>eligible</td>
</tr>
<tr>
<td>Hack-a-thon with vendors</td>
<td>yes</td>
<td>eligible</td>
</tr>
<tr>
<td>Training resources</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Additional NERSC hours from Director’s Reserve</td>
<td>yes</td>
<td>eligible</td>
</tr>
<tr>
<td>NERSC funded postdoctoral fellow</td>
<td>eligible</td>
<td>no</td>
</tr>
<tr>
<td>Commitment of NERSC staff assistance</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>
Training, Case Studies and Documentation

- For those teams not in NESAP, there will be a robust training program

- Lessons learned from deep dives from NESAP teams will be shared through case studies and documentation
Programming & Performance Portability
Performance Portability Strategies

- Conditional compilation
- Directives: OpenMP, OpenACC
- Libraries: Use a library when possible
- Abstractions: Kokkos, Raja
- General-purpose high-level programming languages: UPC, Coarray Fortran
- DSLs: NMODL for neuroscience

Good coding practices
- Modularity, some high-level abstractions
- Data structures flexibly allocatable to different memory spaces
- Task level flexibility so work can be allocated to different compute elements (GPU & CPU)
OpenMP is the most popular non-MPI parallel programming technique

- Results from ERCAP 2017 user survey
  - Question answered by 328 of 658 survey respondents
- Total exceeds 100% because some applications use multiple techniques
NERSC, NVIDIA to Partner on Compiler Development for Perlmutter System

MARCH 21, 2019

The National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory (Berkeley Lab) has signed a contract with NVIDIA to enhance GPU compiler capabilities for Berkeley Lab’s next-generation Perlmutter supercomputer.

In October 2018, the U.S. Department of Energy (DOE) announced that NERSC had signed a contract with Cray for a pre-exascale supercomputer named “Perlmutter,” in honor of Berkeley Lab’s Nobel Prize-winning astrophysicist Saul Perlmutter. The Cray Shasta machine, slated to be delivered in 2020, will be a heterogeneous system...
Ensuring OpenMP is ready for Perlmutter CPU+GPU nodes

- NERSC will collaborate with NVIDIA to enable OpenMP GPU acceleration with PGI compilers
  - NERSC application requirements will help prioritize OpenMP and base language features on the GPU
  - Co-design of NESAP-2 applications to enable effective use of OpenMP on GPUs and guide PGI optimization effort

- We want to hear from the larger community
  - Tell us your experience, including what OpenMP techniques worked / failed on the GPU
  - Share your OpenMP applications targeting GPUs
Engaging around Performance Portability

NERSC is working with PGI/NVIDIA to enable OpenMP GPU acceleration

NERSC Hosted Past C++ Summit and ISO C++ meeting on HPC.

NERSC is a Member


Work with NVIDIA and SLATE to better support SCALAPACK

NERSC is leading development of performanceportability.org
Performance Portability

There is no consensus on the definition or metric for performance portability, but…

**DOE SC Facility Definition (performanceportability.org)**

An application is performance portable if it achieves a consistent ratio of the actual time to solution to either the best-known or the theoretical best time to solution on each platform with minimal platform specific code required.

**Performance portability metric and preliminary work [1-2]**

\[
\Phi(a, p, H) = \begin{cases} 
\frac{|H|}{\sum_{i \in H} e_i(a, p)} & \text{if } i \text{ is supported, } \forall i \in H \\
0 & \text{otherwise}
\end{cases}
\]

Actual Application Performance

\[
e_i(a, p) = \frac{P_i(a, p)}{\min(F_i, B_i \times I_i(a, p))}
\]

Architectural Efficiency

Max Attainable Performance defined by Roofline [3]

Methodology to Measure Perf. Port.

1. Measure empirical compute and bandwidth ceilings: ERT [3]
2. Measure application performance: SDE and LIKWID on KNL; NVPROF on V100

Performance = \( \frac{\text{SDE or } \text{nvprof FLOPs}}{\text{Runtime}} \), Arithmetric Intensity = \( \frac{\text{SDE or } \text{nvprof FLOPs}}{\text{LIKWID or } \text{nvprof Data Movement}} \)

An example: GPP kernel from BerkeleyGW (Roofline)

Methodology to Measure Perf. Port.

An example: GPP kernel from BerkeleyGW (Perf. Port. Scores)

<table>
<thead>
<tr>
<th>Architectural Efficiency</th>
<th>nw=1</th>
<th>nw=2</th>
<th>nw=3</th>
<th>nw=4</th>
<th>nw=5</th>
<th>nw=6</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNL</td>
<td>84.98%</td>
<td>77.50%</td>
<td>66.77%</td>
<td>55.28%</td>
<td>46.56%</td>
<td>39.65%</td>
</tr>
<tr>
<td>V100</td>
<td>97.36%</td>
<td>91.50%</td>
<td>76.70%</td>
<td>65.44%</td>
<td>65.07%</td>
<td>66.38%</td>
</tr>
<tr>
<td>Performance Portability</td>
<td>90.76%</td>
<td>83.92%</td>
<td>71.39%</td>
<td>59.93%</td>
<td>54.28%</td>
<td>49.65%</td>
</tr>
<tr>
<td>KNL</td>
<td>82.06%</td>
<td>72.95%</td>
<td>73.74%</td>
<td>78.72%</td>
<td>81.28%</td>
<td>82.81%</td>
</tr>
<tr>
<td>V100</td>
<td>92.88%</td>
<td>92.88%</td>
<td>97.43%</td>
<td>98.91%</td>
<td>1</td>
<td>99.73%</td>
</tr>
<tr>
<td>Performance Portability</td>
<td>87.14%</td>
<td>81.72%</td>
<td>83.95%</td>
<td>87.67%</td>
<td>89.93%</td>
<td>90.49%</td>
</tr>
</tbody>
</table>

- Roofline captures application bottlenecks, machine balance, problem size, etc.
- Perf. Port. metric captures performance changes across architectures.

Summay
Perlmutter: A System Optimized for Science

- Cray Shasta System providing 3-4x capability of Cori system
- First NERSC system designed to meet needs of both large scale simulation and data analysis from experimental facilities
  - Includes both NVIDIA GPU-accelerated and AMD CPU-only nodes
  - Cray Slingshot high-performance network will support Terabit rate connections to system
  - Optimized data software stack enabling analytics and ML at scale
  - All-Flash filesystem for I/O acceleration
- Robust readiness program for simulation, data and learning applications and complex workflows
- Delivery in late 2020
Thank You!