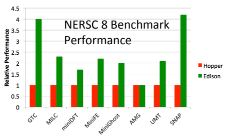
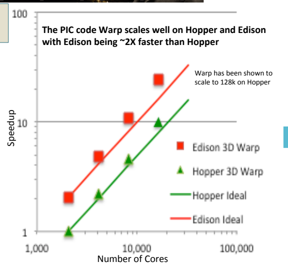
Plasma Physics Simulations on Next Generation Platforms

A. Koniges, R. Gerber, D. Skinner, Y. Yao, Y. He, D. Grote, J-L Vay, H. Kaiser, and T. Sterling

APS Division of Plasma Physics Annual Meeting, Denver, CO, November 2013

Cray XC30 with Intel Ivy Bridge 12-core processors
Aries interconnect with Dragonfly topology





Edison is very similar to Hopper, but with 2-5 times the performance per core on most codes

NERSC's plasma physics users started running production codes immediately on Edison

	Hopper	Edison	Mira	Titan
Peak Flops (PF)	1.29	2.4	10.0	5.26 (CPU) 21.8 (GPU)
CPU cores	152,408	124,800	786,432	299,008 (GPU) 18,688 (CPU)
Frequency (GHz)	2.1	2.4	1.6	2.2 (CPU) 0.7 (GPU)
Memory (TB) Total / Per-node	217 / 32	333 / 64	786 / 15	598 / 32 (CPU) 112 / 6 (GPU)
Memory BW* (TB/s)	331	530.4	1406	614 (CPU) 3,270 (GPU)
Memory BW/ node* (GB/s)	52	102	29	33 (CPU) 175 (GPU)
Filesystem	2 PB 70 GB/s	6.4 PB 140 GB/s	35 PB 240 GB/s	10 PB 240 GB/s
Peak Bisection BW (TB/s)	5.1	11.0	24.6	11.2
Sq ft	1956	1200	~1500	4352
Power (MW Linpack)	2.91	1.9	3.95	8.21

NERSC's Dirac GPU Cluster (sub-cluster of Carver)

Dirac has 50 GPU nodes containing 2 Intel 5530 2.4 GHz, 8MB cache, 5.86GT/sec QPI Quad core Nehalem processors (8 cores per node) and 24GB DDR3-1066 Reg ECC memory.

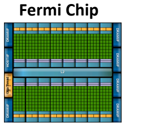
44 nodes: 1 NVIDIA Tesla C2050 (code named Fermi) GPU with 3GB of memory and 448 parallel CUDA processor cores.

4 nodes: 1 C1060 NVIDIA Tesla GPU with 4GB of memory and 240 parallel CUDA processor cores.

1 node: 4 NVIDIA Tesla C2050 (Fermi) GPU's, each with 3GB of memory and 448 parallel CUDA processor cores.

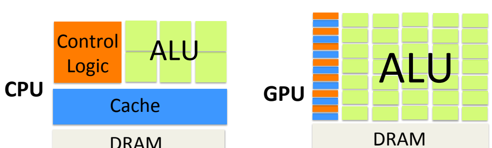
1 node: 4 C1060 Nvidia Tesla GPU's, with 4GB of memory and 240 parallel CUDA processor cores.

Fermi Chip



Dirac has been converted to Scientific Linux 6.3 (SL6), which enables new capabilities such as OpenCL

For users that want to continue using cuda you need to load the latest cuda and SL6 flavor of gcc:
module unload cuda
module load cuda/5.5
module unload pgi
module load gcc-sl6



- Less than 20 cores
- Designed for general programming

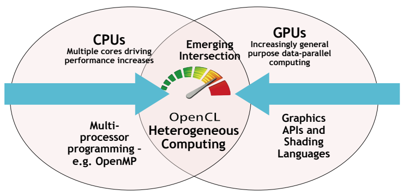
- More than 500 cores
- Optimized for SIMD (same-instruction-multiple-data) problems

OpenCL is also good for systems using Xeon Phi Coprocessors

Babbage is a NERSC internal cluster containing the Intel Xeon Phi coprocessor, which is sometimes called a Many Integrated Core (MIC) architecture. (Babbage is not available to general NERSC users.) Babbage has one login node and 45 compute nodes with two MIC cards and two Intel Xeon "host" processors within each compute node.

Stampede at TACC was deployed in January with a base cluster comprised of 6,400 nodes with Intel Xeon E5 processors, providing 2.2 petaflops and another cluster comprised of 6,880 Intel Xeon Phi coprocessors that add 7 petaflops of performance.

Programming Heterogeneous Platforms



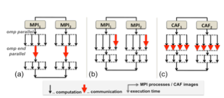
OpenCL – Open Computing Language

Open, royalty-free standard for portable, parallel programming of heterogeneous CPUs, GPUs, and other processors

SC13 Tutorial next week in Denver
OpenCL: A Hands-on Introduction
Tim Mattson, Simon McIntosh-Smith, Alice Koniges

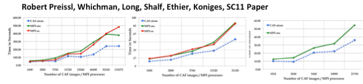
Extending MPI + OpenMP and new approaches

A new "shifter algorithm" using a combination of MPI, OpenMP, and CAF gives significant performance improvement on 130K cores



a) Classical hybrid MPI/OpenMP
b) Extension – MPI thread teams for work distribution and collective MPI function calls
c) Hybrid PGAS (CAF) / OpenMP allows ALL OpenMP threads per team to make communication calls to the thread-safe PGAS communication layer

Robert Preissl, Whichman, Long, Shalf, Ethier, Koniges, SC11 Paper



Requires interoperability of MPI, OpenMP, and PGAS; takes advantage of new network design

Particle-Grid algorithms combined with developing programming languages are appropriate for next generation platforms

- A broad family of computations using discrete-particle methods already perform at extremely high scalability
- Exascale will be constrained by lock-step nature
- Consider new and rethought algorithms that break away from traditional lock-step programming
- Compute-send;compute-send=>limited overlap
- HPX runtime system implementation exposes intrinsic parallelism and latency hiding
- Use a message-driven work-queue based approach to finer grain parallelism based on lightweight constraint-based synchronization

A combination of new OS+runtime+languages with proven event-driven models can surpass performance of traditional time-step models

Lightweight multi-threading
- Divides work into smaller tasks
- Increases concurrency

Message-driven computation
- Move work to data
- Keeps work local, stops blocking

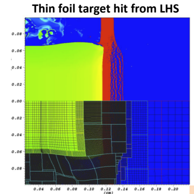
Constraint-based synchronization
- Declarative criteria for work
- Event driven
- Eliminates global barriers

Data-directed execution
- Merger of flow control and data structure

Shared name space
- Global address space
- Simplifies random gathers

Edison supports complex multi-physics plasma simulations with advanced tool sets

Thin foil target hit from LHS

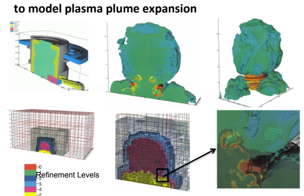


- 3D ALE hydrodynamics
- AMR (use 3X refinement)
- With 6 levels, vol ratio 10^7 to 1
- Material interface reconstruction
- Anisotropic stress tensor
- Material failure with history
- Ion/laser deposition
- Thermal conduction
- Radiation diffusion
- Surface tension

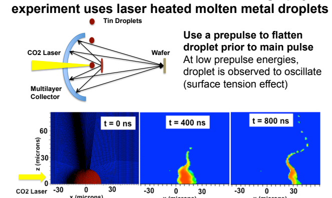
ALE - Arbitrary Lagrangian Eulerian
AMR - Adaptive Mesh Refinement

ALE-AMR is an open science code and has no export control restrictions

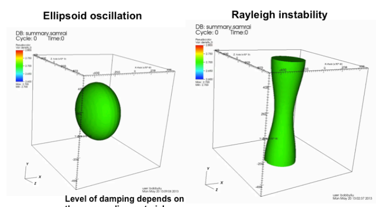
The use of AMR with six levels of refinement is critical to model plasma plume expansion



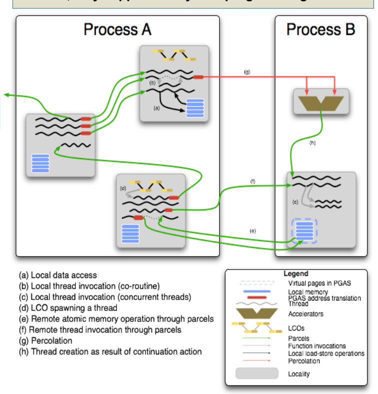
ALE-AMR used to model extreme UV lithography experiment uses laser heated molten metal droplets



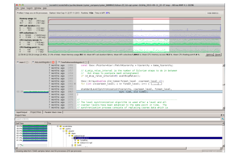
Surface tension models based on volume fractions are being validated with analytic test cases



Exascale operating systems, e.g., ParalleX Execution Model, may support totally new programming models



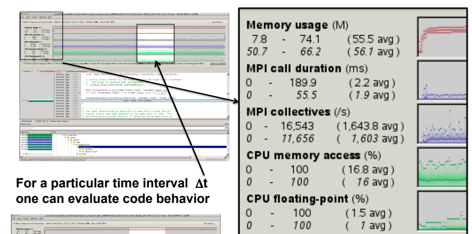
MAP developed by Allinea is available on Edison



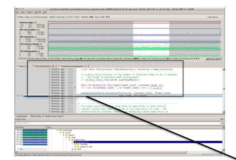
Screen capture of MAP window with memory usage, MPI calls, etc. as a function of time shown along the top

- Large multiphysics codes like ALE-AMR have complex make/build scripts to load a significant number of supporting libraries
- It is important that performance analysis tools can work in this environment and can be accessed in a relatively painless manner

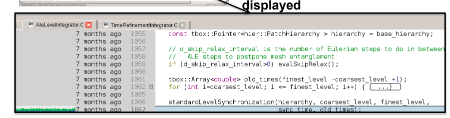
MAP developed by Allinea is available at NERSC



For a particular time interval Δt one can evaluate code behavior



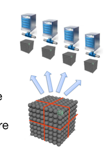
The source code associated with the majority of communication or computation also can be displayed



NERSC is exploring new ways to share and analyze terabytes of science data

SciDB Testbed at NERSC

★Large SciDB Cluster
★10+ Science Projects
★Multiple Science Fields: Climate, Bioinformatics, Astronomy, ...

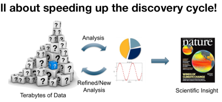


The Collaboration/Support Model

★We form partnerships with science projects
★Not only provide hardware/software
★But also hold their hands to
★Load the 1st batch of data
★Do the 1st round of analysis

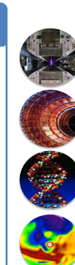
We Welcome New Science Projects!

It's all about speeding up the discovery cycle!



Contact: Yushu Yao (yyao@lbl.gov)

Science Gateways http://portal.nersc.gov



Web-accessible Data Depots
- High Performance filesystems, fast subselection and reductions
- OpenDAP, FastBIT, HDF available on portal.nersc.gov

Web-based Team Analytics
- Collaboratively build data-driven models. Team data science.
- MongoDB, SQL, SciDB, Rstudio, machine learning

HPC & HPC Execution Engines
- Marshal high-throughput workflows and ensembles
- Flexibly adapt concurrency between HPC and HTC needs

Data as a tool for Web Apps
- Modern APIs (NEWT, REST, etc.) to deliver data downstream
- Science-aware data policies and APIs w/ production reliability

- This work was supported by the Director, Office of Science, Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.