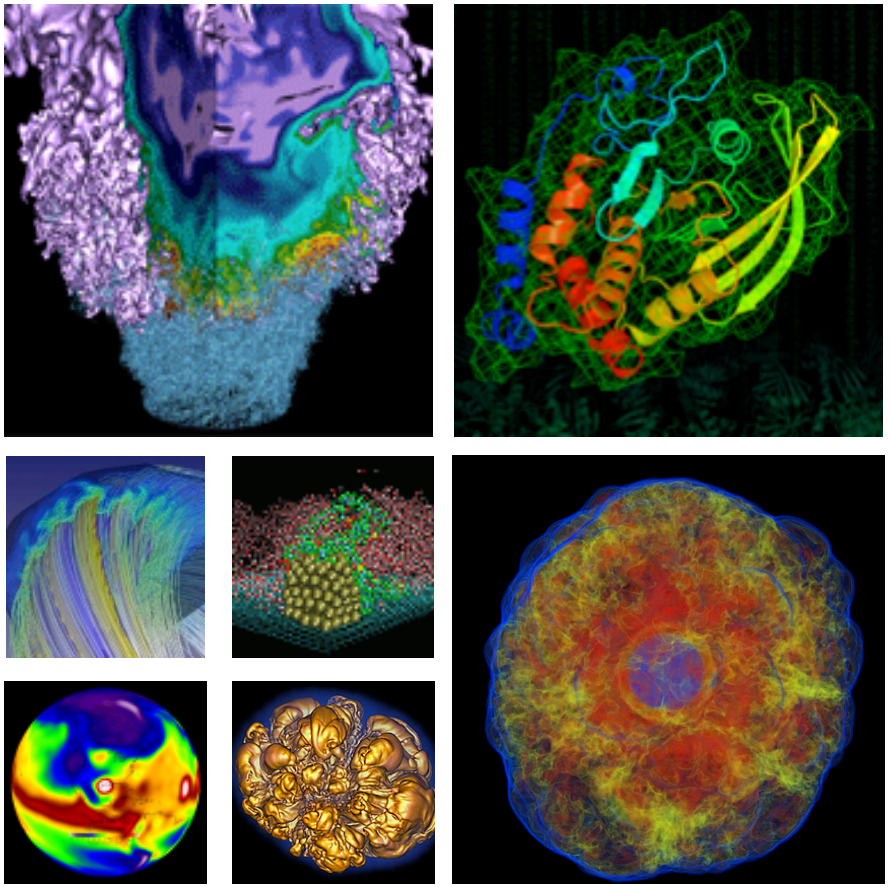
Shifter in the Wild



Lisa Gerhardt

Data and Analytics Group, NERSC

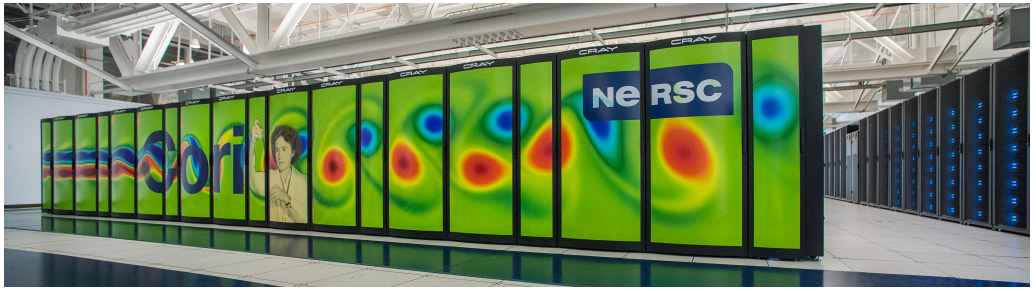
November XX, 2016

NERSC: DOE's Scientific Computing Center

- 6,000 active users, 750+ codes, 2000+ paper/year
- Biology, Energy, Environment
- Computing
- Materials, Chemistry, Geophysics
- Particle Physics, Cosmology
- Nuclear Physics
- Fusion Energy, Plasma Physics

HPC is Awesome

- **Cori Cray XC40**
  - Data-intensive (32-core Haswells, 128GB) partition
  - Compute-intensive (68-core KNLs, 90GB) partition
  - ~10k nodes, ~700k cores
- **Edison Cray XC30**
  - 2.5PF
  - 357TB RAM
  - ~5000 nodes, ~130k cores
- **High speed parallel file system**
  - >10 PB project file system (GPFS)
  - >38 PB scratch file system (Lustre)
  - >1.5 PB Burst Buffer (flash)
- **High Speed Aires interconnect**
  - Some speed numbers here

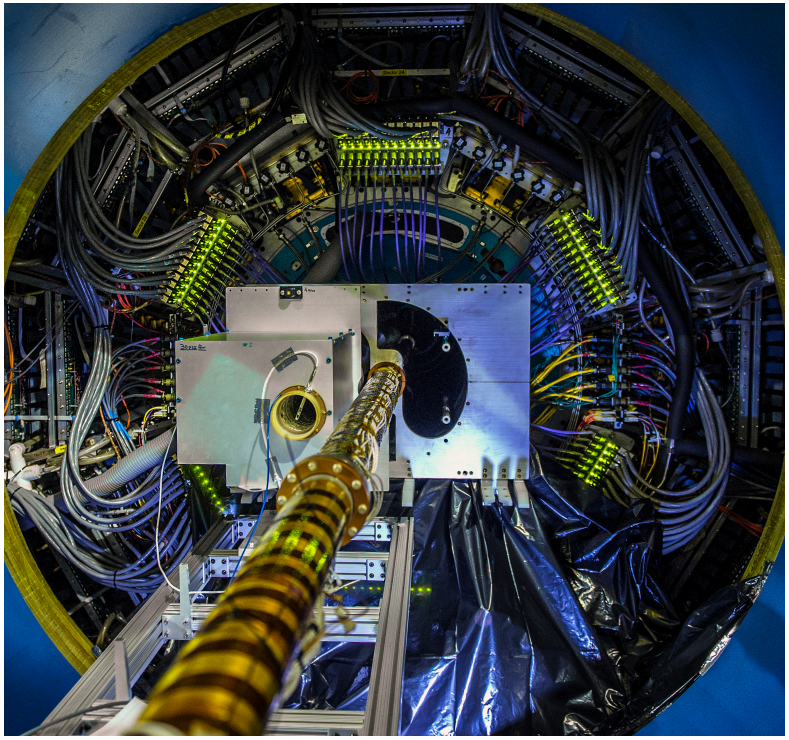


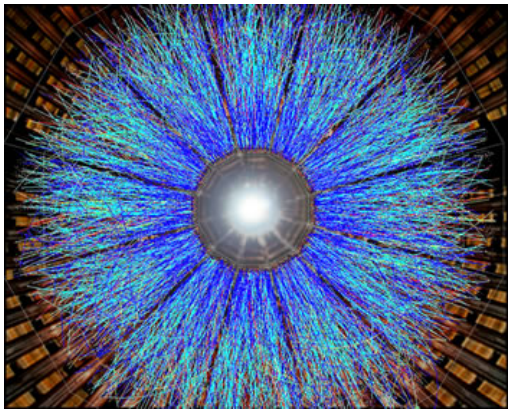
HPC is Awkward

- **No local disk**
  - Breaks a lot of standard Linux work flows
- **Minimal OS**
  - Designed to accelerate parallel software
  - Many "expected" Linux tools are absent
  - Runs SUSE, and doesn't upgrade often
- **Different file systems have different responses**
  - Sometimes unclear to users where is the best place to put their software
- **Many groups have turned to Shifter to over come these obstacles**

Probing The Nucleus

- **STAR at Brookhaven, NY**
  - smashing nuclei into each other to understand their component parts
- **Data analysis and simulation**
- **Why Shifter?**
  - Difficult software dependencies (32-bit libraries)





How'd they do it?

- **Started with a publically available Scientific Linux 6.4 image**
- **Installed dependencies via rpm**
  - MySQL, MPI
- **Installed custom STAR software**
  - ROOT, etc.
  - Bundled them up from another SL6 system where they were already installed, untarred them into the image

Leveraging Shifter for Easy Scalability

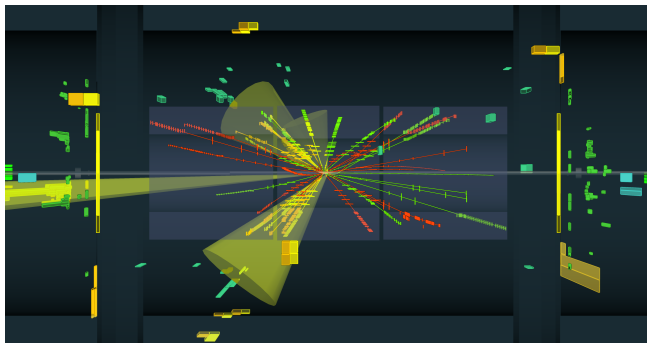
- **Shifter has capability to loop mount an xfs file**
  - Backed by Lustre, but all metadata actions are limited to a single node, so access is very fast
- **STAR needs to read conditions from a ~100 MB MySQL database**
  - Running 32 individual jobs / node
- **Initially, copied DB to Lustre, but first query timed out after 30 minutes**
- **Copied DB to Shifter's xfs**
  - Initial copy ~5 minutes (multi hour job)
  - Query was instantaneous
- **Used this functionality to quickly scale up without re-engineering their workflow**

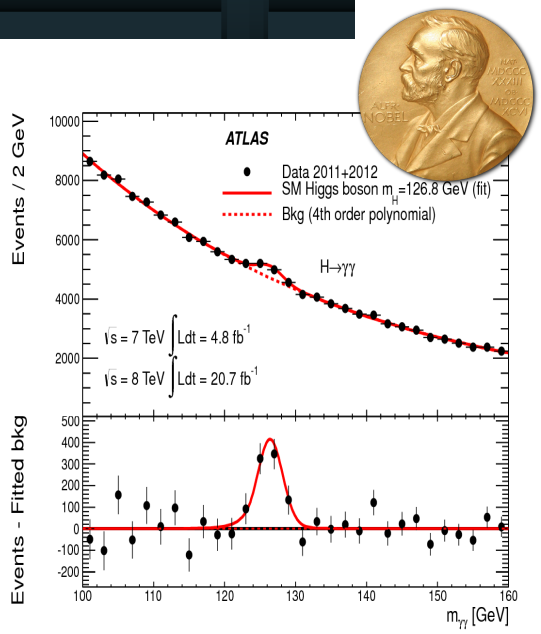
Exploring the Universe

- **Dark Energy Survey: Visualizing the universe**
  - Measuring the expansion history of the universe to understand the nature of Dark Energy.
- **Data analysis code: identify objects (stars, galaxies, quasars, asteroids etc) in images, calibrate, measure their properties.**
- **Why Containers?**
  - Complicated software stack – runs on laptops to supercomputers
  - Python-based code; lots of imports

Probing the Fundamentals of Matter

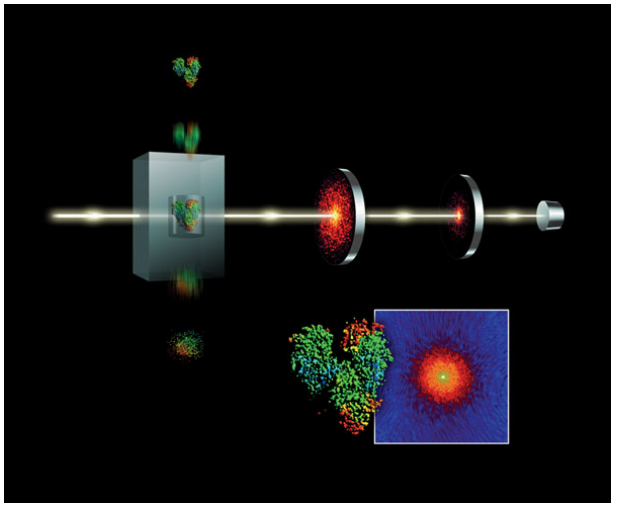
- **Large Hadron Collider (LHC)**
  - A billion proton-proton collisions per second and multi-GB of data per second.
- **Data analysis, simulation, multi-site data and computing pool**
- **Why Shifter?**
  - Complicated software stack: Needs FUSE and elevated permissions to run
  - Monster images: ~300 – 500 GB and 10s of millions of inodes
  - Integrated framework for running with images at all computing sites





Imaging the Heart of Things

- **LCLS: Linac Coherent Light Source at SLAC**
  - Using X-rays to image nanoscale particles and understand chemistry on the natural timescale of reactions
- **Realtime image analysis based on python stack (tomo.py)**
- **Why Shifter?**
  - Many library imports, complicated software stack



Shifter Enables Science

- **Shifter is making scientific analysis easier at NERSC**
  - Shifter framework can be extended to other Cray systems and shifter images can be run at any "Docker-friendly" computing center
  - Successful use across many scientific disciplines

National Energy Research Scientific Computing Center

Loop Mounted FS for Super Fast I/O

- **Shifter can mount an xfs file system on each node**
  - Created when job starts and destroyed when job ends
  - Cray "local disk"
  - Excellent I/O rates:
    - Backed by the Lustre file system, metadata operations are all confined to a single node
    - Also good for "bad IO"