Shifter: Containers for HPC

Wahid Bhimji, Shane Canon, Doug Jacobsen, Lisa Gerhardt, Mustafa Mustafa, Jeff Porter, Vakho Tsulaia (LBNL and NERSC)
CHEP 2016
NERSC: DOE’s Scientific Computing Center

• 7,000 active users, 750+ codes, 2000+ paper/year
• Biology, Energy, Environment
• 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, 1600 nodes
  – 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
  – 5 PB project file system (GPFS)
  – 28 PB scratch file system (Lustre)
  – 1.5 PB Burst Buffer (flash)*

• High Speed Aires interconnect

*“Extreme I/O on HPC for HEP using the Burst Buffer at NERSC” Wahid Bhimji et al. Tuesday, Track 6
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 a new software package called Shifter to over come these obstacles**
NERSC has implemented Docker-like container technology through Shifter
- Supports Docker and other images (vmware, ext4, squashfs, etc.)
- Intended for complex code stacks and codes with shared libraries
- Users can create custom images in desired OS
- Upload image to docker hub and pull down on HPC system
- Hooks to tie into the batch system

Work of Doug Jacobsen and Shane Canon at NERSC
Running CVMFS Shifter Image

- As proof of concept created “Mega” CVMFS shifter image
  - Full CVMFS stack pulled down and deduped with uncmvms software stack. 1 – 3 TB ext4 file uncompresssed, 300 GB compressed w/squashfs

- Use Shifter to load job
  - Add a single flag to batch script “--image=<image name>”
  - ATLAS cvmfis repository is found at /cvmfs/atlas.cern.ch like normal

- Tested with ATLAS G4 simulations and Analysis Software (QuickAna)
  - Simulation load times scale well out to 500 nodes (16,000 cores)

"Production Experience with the ATLAS Event Service", Vakho Tsulaia et al., Tuesday Track 4
ALICE: CVMFS and parrot in Shifter

- Using parrot to load from alien cache. All tests in a Shifter image with:
  - Full CVMFS stack filled with `uncvmfs`
  - Parrot reading from full CVMFS stack preinstalled with `cvmfs_preload`
  - Parrot reading from CVMFS stack installed on Lustre parallel file system
  - Parrot reading from remote squid server (prefilled cache on Lustre)

Start up times are negligible for all but remote squid server. Shifter based solution best at higher concurrencies.

Work of Jeff Porter and Markus Fasel at LBNL
**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”

![Graph showing write bandwidth and millions of IOPS vs processes]

- **57 GiB/s!**
Loop FS: STAR Case Study*

- Installed STAR software into Shifter image
- Ran 16 – 32 STAR jobs on a single node
- Each job must access a calibration DB
  - $O(100\ \text{MB})$
  - Roughly 30,000 queries during processing
- First DB query did not finish in 30 minutes when DB was placed on Lustre file system
- Copied DB into Shifter loop FS
  - Copy time: ~3 minutes (out of 10 – 15 hour job)
  - Cache time was nearly instantaneous
- Using Shifter Loop FS for a local DB allows trivial scalability of job concurrency

*“STAR Data Production at NERSC/Cori, an adaptable Docker container approach for HPC”, Mustafa Mustafa, Thurs 16:30 poster session
Shifter is Not Just for Cray

• Successfully installed at
  – ANL: IBM Blue Gene/Q, Cobalt scheduler
  – CSCS: IBM Blue Gene/Q, SLURM scheduler
  – VLSCI: IBM Blue Gene/Q, SLURM scheduler
  – Berkeley Institutional Cluster: Dell Linux farm (SL6), Warewulf scheduler

• Shifter is open source through a BSD license
  – Shifter-hpc google group for those interested in installing the framework on their systems
Shifter Enables Science

- A number of HEP and astronomy groups are already investigating Shifter at NERSC
  - ATLAS, ALICE, STAR, CMS, LSST, LCLS, Dayabay
- New framework and innovation are making scientific analysis easier at NERSC
  - Custom environment for reproducible science
  - Super fast IO layer
  - Successful use across many scientific disciplines
  - Cray will make Shifter a mainstream capability for their systems
  - Framework can be extended to other HPC systems

- Related talks and posters
  - “Production Experience with the ATLAS Event Service”, Vakho Tsulaia et al., Tuesday Track 4
  - “STAR Data Production at NERSC/Cori, an adaptable Docker container approach for HPC”, Mustafa Mustafa, Thurs 16:30 poster session
National Energy Research Scientific Computing Center