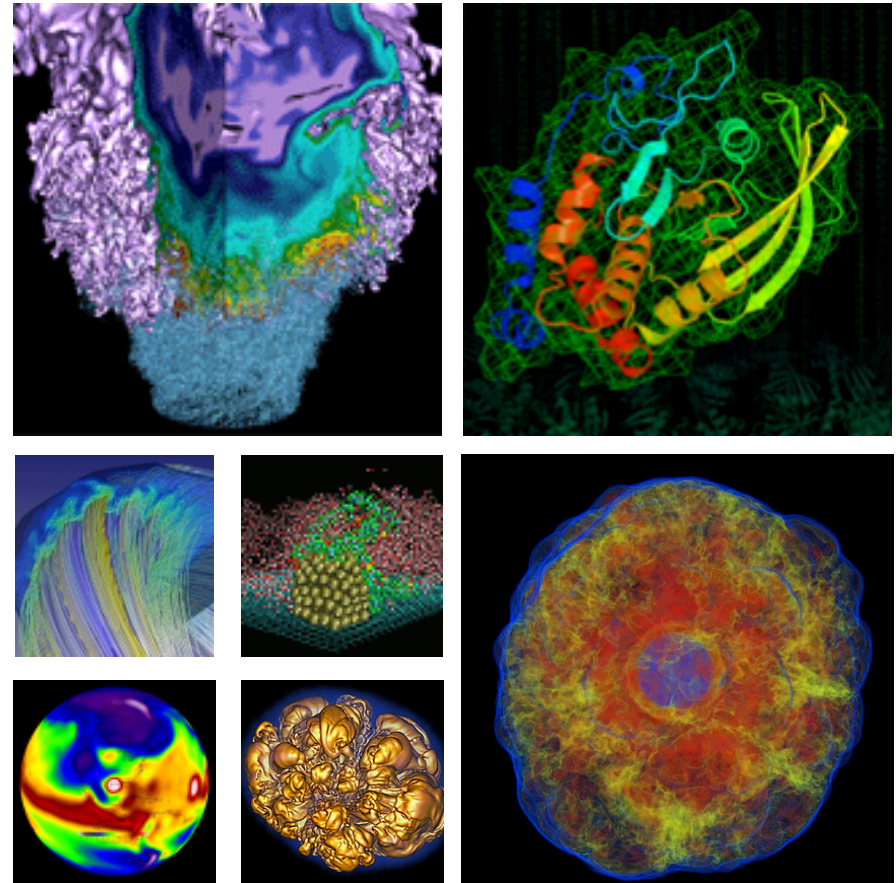


Shifter at NERSC



Shane Canon
Lisa Gerhardt
NERSC Data and Analytics Services

December 19, 2016

Shifter: Bringing Containers to HPC



- **Docker: open source, automated container deployment service**



- Docker containers wrap up a piece of software in a complete filesystem that contains everything it needs to run (code, runtime, system tools and libraries)
- Guaranteed to operate the same, regardless of the environment in which it is running
- **NERSC has partnered with Cray to deliver Docker-like container technology through a new software package known as **Shifter****

Shifter at NERSC



- **Secure and scalable way to deliver containers to HPC**
- **Implemented on Cori and Edison**
- **Supports Docker images and other images (vmware, ext4, squashfs, etc.)**
- **Basic Idea**
 - Users create custom images in desired OS
 - Upload image to docker hub and pull down on HPC system
 - Hooked into the batch system



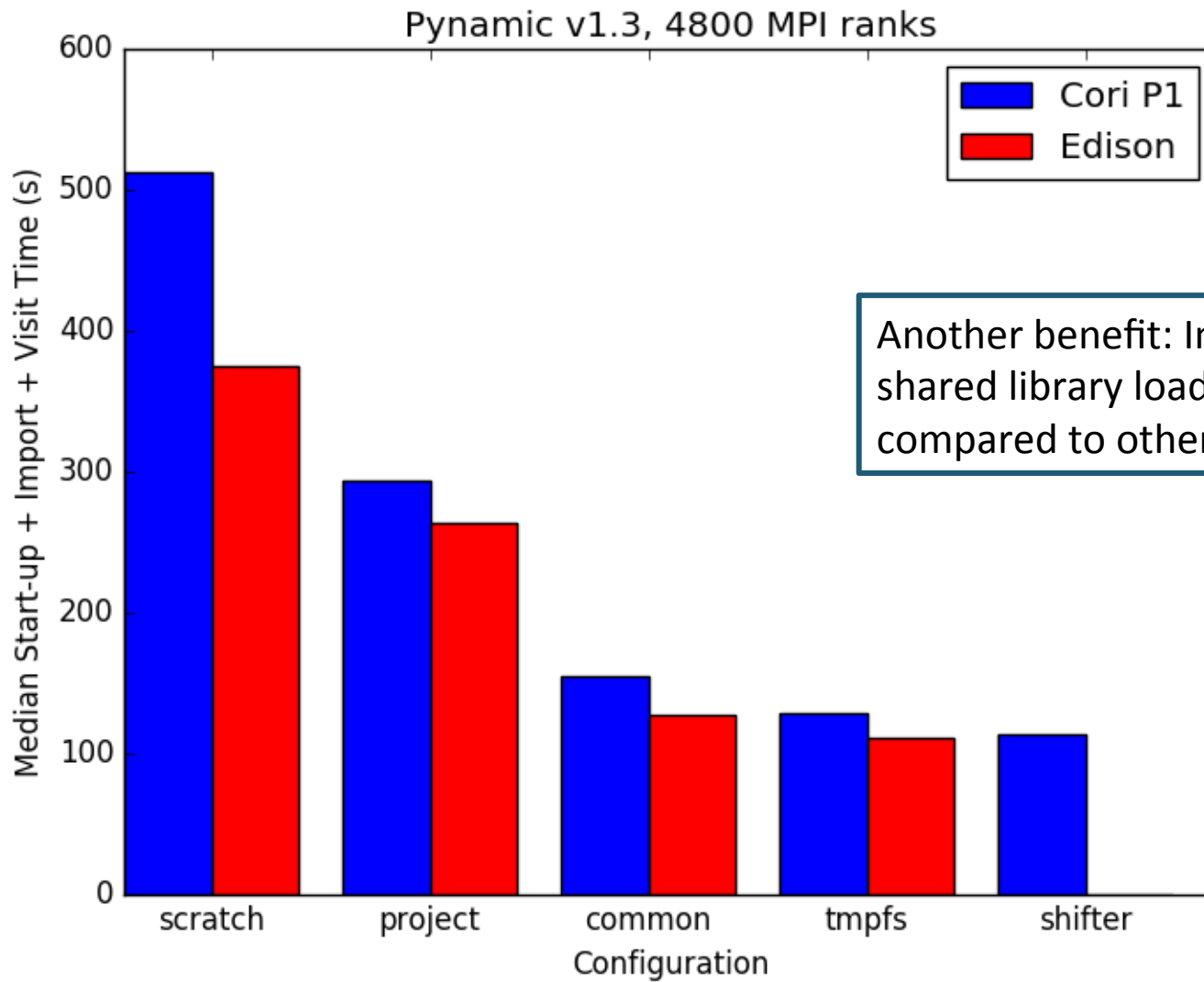
<http://www.nersc.gov/users/software/using-shifter-and-docker/>

Why Use Shifter?



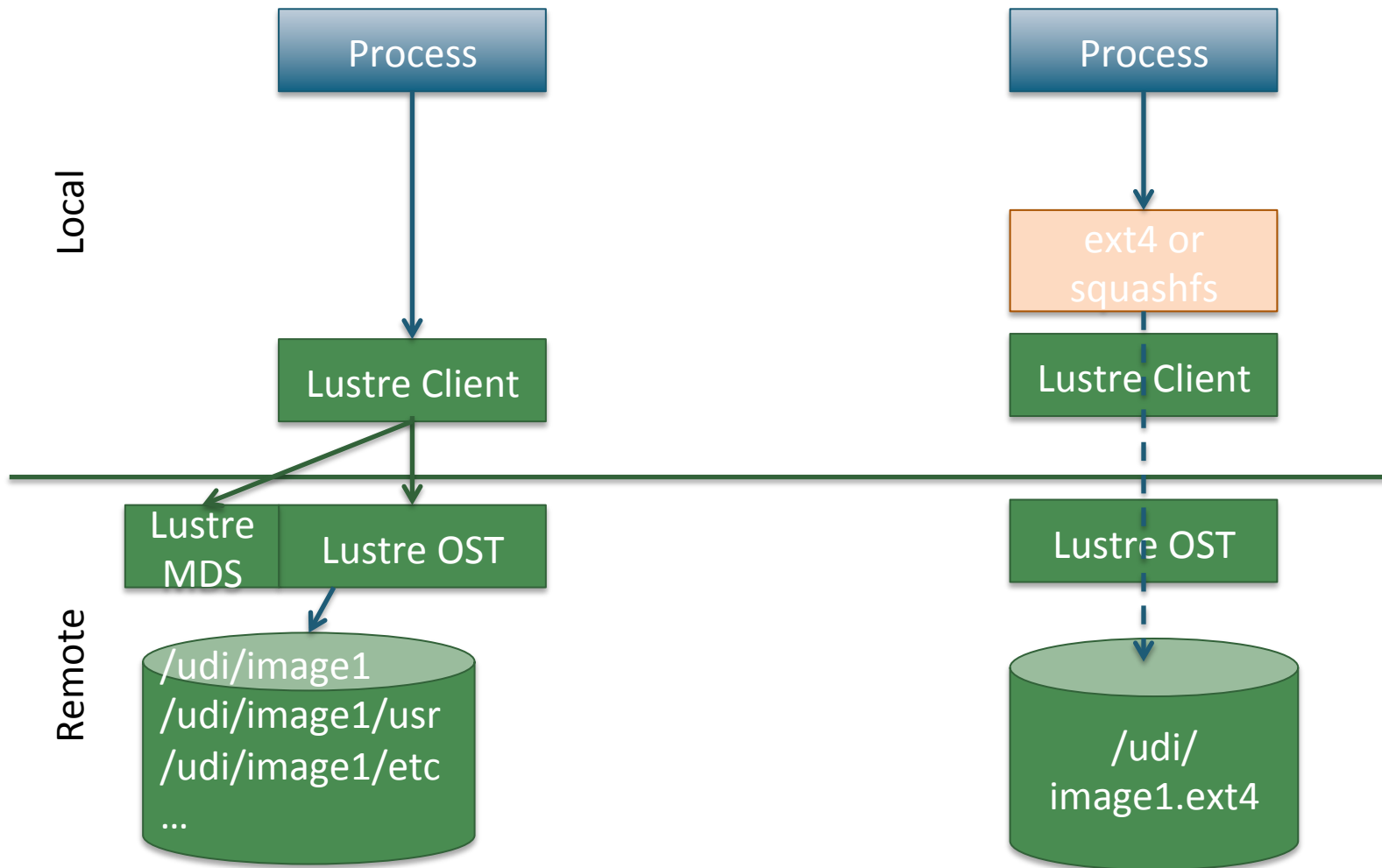
- **Shifter allows you to fully customize your operating environment**
 - Want SL 6.X with 32 bit libraries? Use Shifter
 - Have a very complicated software stack with lots of dependencies? Use Shifter
- **Portability**
 - Can volume mount directories into shifter images
 - Have an /input and /output that are linked to directories in your scratch directory
 - Images are NERSC-independent, can be run anywhere

Shifter is Fast



Another benefit: Improved shared library loading times compared to other file system

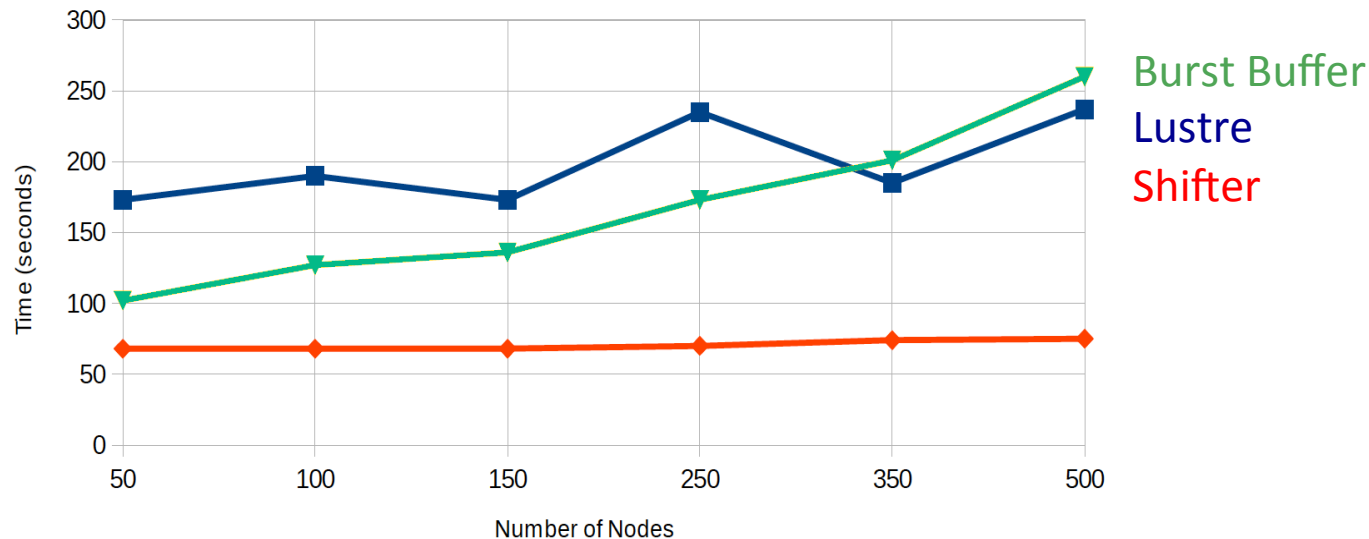
File System flow – Traditional vs Shifter



Even Big Images Load Quickly



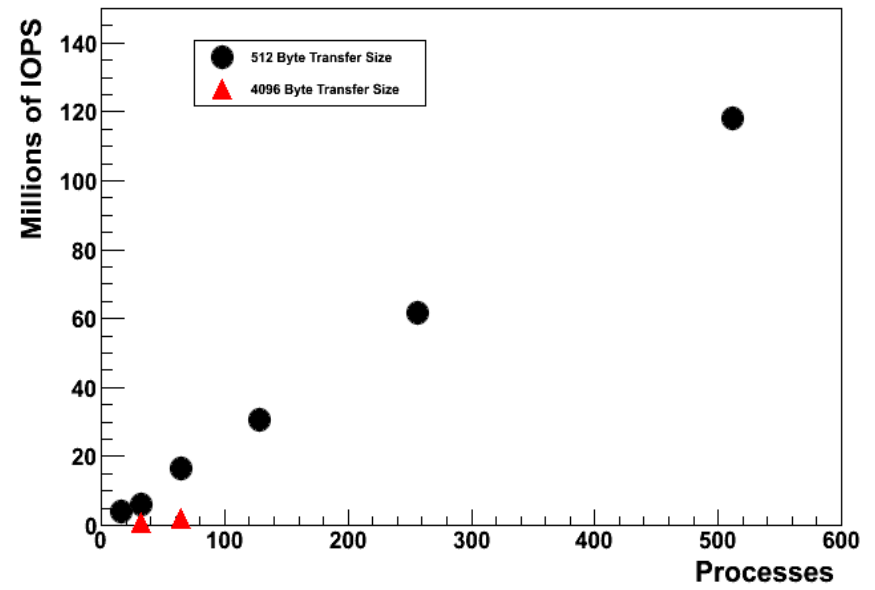
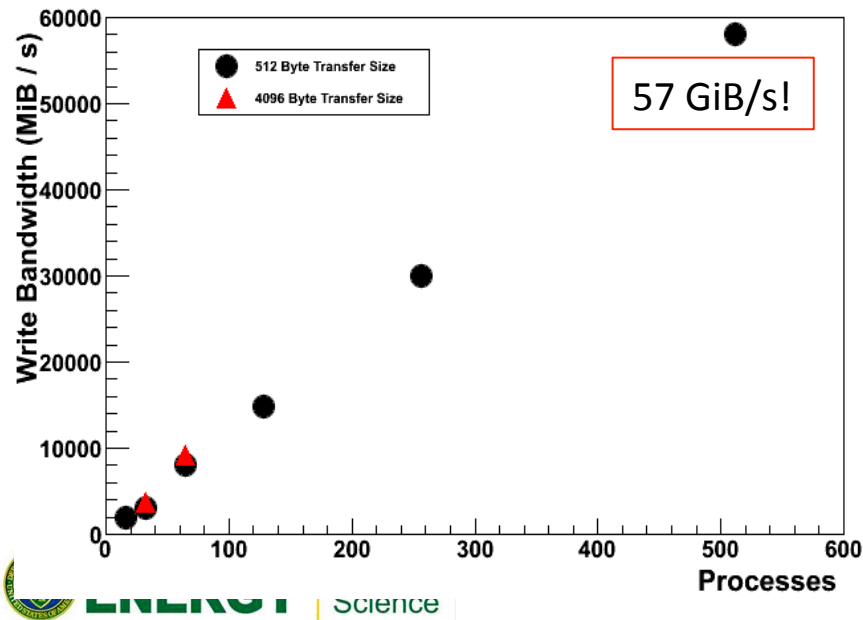
- **As proof of concept created “Mega” CVMFS shifter image**
 - Full CVMFS stack pulled down and deduped with uncvmfs software stack. 1 – 3 TB ext4 file uncompressed, 300 GB compressed w/ squashfs
- **Use Shifter to load job**
 - Add a single flag to batch script “--image=<image name>”
 - ATLAS cvmfs repository is found at /cvmfs/atlas.cern.ch like normal



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
 - Compute node “local disk”
 - Excellent I/O rates:
 - Small databases
 - Also good for “bad IO”



Shifter and MPI



- **MPI communication over Aries network is available by default for Shifter on Cori**
- **In image, build and link against standard MPICH libraries**
- **Cray libraries swapped in at run time by front loading LD_LIBRARY_PATH**

Shifter: How to Use It



- **Create a Docker image and upload it to Dockerhub**

```
FROM ubuntu:15.10
RUN apt-get update && apt-get install -y autoconf automake gcc g++ make
gfortran && apt-get clean all
RUN mkdir /build/
COPY ./input_files/Python-2.7.11.tgz /build
COPY ./input_files/mpich-3.2.tar.gz /build
COPY ./input_files/mpi4py-1.3.1.tar.gz /build
RUN cd /build && tar xvzf Python-2.7.11.tgz
RUN cd /build/Python-2.7.11 && ./configure && make -j4 && make install
RUN cd /build && tar xvzf mpich-3.2.tar.gz
RUN cd /build/mpich-3.2 && ./configure;make -j4 && make install
RUN cd /build && tar xvzf mpi4py-1.3.1.tar.gz
RUN cd /build/mpi4py-1.3.1 && python setup.py build && python setup.py
install
RUN cd /build && rm Python-2.7.11.tgz && rm mpich-3.2.tar.gz && rm
mpi4py-1.3.1.tar.gz
```

Shifter at NERSC: Demo



- **Shifter is being successfully used by many users including users from HEP, NP, and BES**
- **Future Shifter plans**
 - Ability to overlay multiple shifter images
 - Private shifter images for groups with access limitations
- **Shifter is an easy way to improve performance and get portability for your science environment**



National Energy Research Scientific Computing Center

Pulling Down an Image from Dockerhub



On Cori or Edison

```
shifterimg pull docker:lgerhardt/mpi-test:v5
```

Format is source:image_name: tag

PUBLIC REPOSITORY

[lgerhardt/mpi-test](#) ☆

Last pushed: 4 months ago

Repo Info **Tags** Collaborators Webhooks Settings

Tag Name	Compressed Size	Last Updated
v5	202 MB	4 months ago
v4	172 MB	4 months ago
v3	656 MB	4 months ago
v1	201 MB	8 months ago

Existing Shifter Images



shifterimg images

```
cori11> shifterimg images
cori  docker  READY  ce20c473cd  2015-12-11T09:53:47  centos:7
cori  docker  READY  17583c7dd0  2015-12-04T08:32:58  busybox:latest
cori  docker  READY  92b9f2dbf7  2015-12-04T10:48:12  scanon/shanetest:latest
cori  docker  READY  d120eb68e8  2015-12-04T15:36:13  jcorrea/spot2:v1
cori  docker  READY  621a496b0d  2015-12-08T06:26:58  registry.services.nersc.gov/
cori  docker  READY  d55e68e6cc  2015-12-22T09:16:46  ubuntu:14.04
cori  docker  READY  86314f27ce  2015-12-09T00:33:46  registry.services.nersc.gov/p
cori  docker  READY  d6bf89bd0b  2016-02-08T09:32:43  registry.services.nersc.gov:r
cori  docker  READY  b345ea6cf6  2015-12-09T07:57:41  registry.services.nersc.gov/n
cori  docker  READY  33200b4db5  2015-12-12T00:00:06  kbase/kbase_base:develop
cori  docker  READY  e5ff6dcec0  2015-12-14T11:03:53  registry.services.nersc.gov/p:
cori  docker  READY  109b72e23c  2015-12-15T05:50:38  fedora:23
cori  docker  READY  2bc6cdd62f  2015-12-15T06:06:25  miguelgila/wlwg_wn:201512
cori  custom  READY  6378f30b51  2015-12-16T17:13:54  atlas_cvmfs:latest
cori  docker  READY  9ff944a24c  2015-12-18T03:08:09  marius311/run_sim:latest
cori  docker  READY  56e5a8a6e0  2016-01-07T15:24:40  marius311/mpitest:latest
cori  docker  READY  df9675993b  2015-12-21T14:47:19  paterno/centos67-art_v1_1:
cori  docker  READY  c0f009e667  2015-12-22T09:37:10  fenicsproject:dev:latest
cori  docker  READY  0f73fcfb8a  2015-12-22T13:08:00  fenicsproject/stable:latest
cori  docker  READY  f1c24227b7  2015-12-22T10:54:20  jbkowalkowski/art_test:late
cori  custom  READY  3f2d105bb2  2015-12-23T17:06:47  cms_cvmfs:latest
cori  docker  READY  e51ffe812d  2016-01-07T11:46:19  paterno/centos-uboone_v04
cori  custom  READY  9e6c507dae  2016-01-11T22:40:36  cvmfs_test:latest
cori  custom  READY  8e24498895  2016-01-12T16:45:12  cvmfs_mpitest:latest
cori  custom  READY  e5da39eb02  2016-01-13T17:37:11  alice_cvmfs:late
```

Running A Job in A Shifter Image



```
#!/bin/bash
#SBATCH --image=docker:image_name:latest
#SBATCH --volume="/global/cscratch1/sd/lgerhard:/output"
#SBATCH --volume="/global/cscratch1/sd/lgerhard/shifter_tmp:/tmp
:perNodeCache=size=200G"

#SBATCH --nodes=1
#SBATCH --partition=regular
#SBATCH -C haswell

srun -n 32 shifter python myPythonScript.py args
```

Many more commands at
<http://www.nersc.gov/users/software/using-shifter-and-docker/using-shifter-at-nersc/>