# A scalable tool for deploying Linux containers in high-performance computing

September 1 2015, by Kathy Kincade

The explosive growth in data coming out of experiments in cosmology, particle physics, bioinformatics and nuclear physics is pushing computational scientists to design novel software tools that will help users better access, manage and analyze that data on current and next-generation high-performance computing (HPC) architectures.

One increasingly popular approach is container-based computing, designed to support flexible, scalable computing. Linux containers, which are just now beginning to find their way into the HPC environment, allow an application to be packaged with its entire software stack, including portions of the base operating system files, user environment variables and application "entry points."

With the growing adoption of container-based computing, new tools to create and manage containers have emerged—notably Docker, an open-source, automated container deployment service. Docker containers wrap up a piece of software in a complete filesystem that houses everything it needs to run, including code, runtime, system tools and system libraries. This guarantees that it will always operate the same, regardless of the environment in which it is running.

While Docker has taken the general computing world by storm in recent years, like container-based computing it has yet to be fully recognized in HPC. To facilitate the use of both of these tools in HPC, NERSC is enabling Docker-like container technology on its systems through a new, customized software package known as Shifter. Shifter, designed as a

scalable method for deploying containers and user-defined images in an HPC environment, was developed at NERSC to improve flexibility and usability of its systems for increasingly data-intensive workloads. It is initially being tested on NERSC's Edison system—a Cray XC30 supercomputer—by users in particle physics and nuclear physics and will eventually be made available as an open source tool for the general HPC community.

Here, NERSC's Doug Jacobsen, Shane Canon, Lisa Gerhardt and Deborah Bard—who have been instrumental in developing, deploying and testing Shifter—discuss how Shifter works and how it will help the scientific community better utilize resources at NERSC and other supercomputing facilities and increase their scientific productivity in the process.

## Q. How is container-based computing changing the way applications are developed and deployed in HPC?

Canon: That remains to be seen, but my expectation is that, similar to the way it is having an impact in the enterprise space, it will carry over into scientific computing and HPC as well. One reason I think it will be powerful is because it is a productivity enhancer. It makes it easier for users to develop something locally on their laptop and push to another place. Another key factor is scientific reproducibility; being able to take that image and know that you can reliably instantiate it over and over or share it with others is really powerful. I think the enterprise world wants the same things but for different reasons – for them it is about compliance and stability, while for scientists it's about reproducibility and verifiability. But it wouldn't surprise me if, five years from now, the way people build and deliver software is through something like containers, that it becomes the predominant way people share their software, both in general and in HPC.

## Q. With the growing adoption of container-based computing, Docker is gaining popularity in the HPC community as well. What motivated you to develop Shifter rather than modify Docker for use on HPC systems?

Jacobsen: Shifter is strongly focused on the needs and requirements of the HPC workload, which means it can deliver the functionality we are looking for while still meeting the overall performance requirements and constraints that a large-scale HPC platform imposes. Shifter allows the user to supply us with a Docker image that we can then convert to a form that is easier and more efficient for us to distribute to the compute nodes. We are leveraging the user interface that Docker makes available to people to create their software images and leveraging that ecosystem, but not directly using their software internally. Shifter it is not a replacement for Docker functionality; it is specifically focused on the HPC use cases.

Canon: What Docker has done is develop a framework that makes it easy for people to create images and then publish those to something like Dockerhub, which makes it easy for them to share. So we are leveraging that and trying to preserve the things we think are most useful for scientists. We felt it was important for the implementation to make it easy for users to create a software environment and then instantiate that on our systems and also leverage what is happening in the Docker ecosystem so they can use some of the existing images out there or publish to Docker and share with other scientists but also easily run on NERSC systems. So we are preserving the best parts of Docker.

## Q. How does Shifter work?

Jacobsen: Shifter works by converting user- or staff-generated images in Docker, virtual machines or CHOS (another method for delivering flexible environments) to a common format that provides a tunable point to allow images to be distributed on the Cray supercomputers at NERSC. Through the user interface, a user can select an image from their Dockerhub account or private Docker registry. An image manager at NERSC then automatically converts the image to the common format based on an automated analysis of the image. The image is then copied to the Lustre scratch filesystem and the user can begin submitting jobs—all of which run entirely within the container—specifying which image to use.

## Q. In addition to enabling user-defined images and automating the image conversion process, what other advantages does Shifter bring to NERSC users?

Jacobsen: What makes this software a big deal is that it is enabling science on our systems that has been inaccessible in the past. For example, for data-intensive users such as researchers with experimental apparatus who want to analyze data versus just-run simulations on our systems—their codes typically tend to be very different from the way most calculations on Edison run. They tend to have very large, complex software stacks with many different dependencies.

With Shifter, users can prepare a Docker container on their own system, bring it onto the Edison system through already constructed pipelines and it just works. Applications go from not working on Edison to immediately working. And as a critical side benefit, Shifter provides a lot of performance benefits to data-intensive codes that rely on many different dependencies because of the way the software shifts to the compute nodes. With Shifter, it is very performant to start them up and run them. Previously, we relied a lot on the centralized resources at

NERSC to make that happen.

Bard: I've been working with Shifter since I first came to NERSC this summer, evaluating it for running simulations for the LSST (Large Synoptic Survey Telescope). One of the things I have learned is that you have root access when you use Docker, which is not good because you can accidentally screw things up. This is a barrier to running Docker in a lot of places (in HPC) that Shifter fixes, which is huge. People don't want to deploy Docker because of security issues, but Shifter controls the external connections in a way that means it works at NERSC. And they are going to make it open source, which is brilliant. When you're thinking about software for a large collaboration, for example, you want to be able to develop a software environment that people can run anywhere, and with Shifter you can run it safely anywhere.

I've also been learning about how to incorporate Shifter into workflow tools, and it is very easy, which is nice. I am particularly interested in how we are going to be supporting it from the users' perspective, not just within large collaborations but for all users of Cori (NERSC's next-generation Cray X40). With Shifter, users will be able to get running very quickly on Cori.

## Q. Are there certain applications or science domains in which Shifter will have greater impact, or is it designed to improve data management across the board?

Canon: I think it can work for a large range of applications. Initially it's more important for some of the nontraditional and data-intensive areas because they are the ones that often come in with the most challenging software requirements, and trying to take those requirements and deploy their software on an HPC system can sometimes be very challenging. It's

the thing they stumble on right out of the gate. They don't even get to the point where they are effectively running their applications because they can't get all the prerequisite software requirements satisfied first.

What's happened is that as datasets have gotten larger these communities' computing demands have grown, so that they now have problems that are similar in scale to traditional HPC users. In the past maybe they could have gotten by just running something on a workstation or very small cluster, but now they have problems that have gotten big enough that they can take advantage of facilities like NERSC. But then when they tried to make that leap they would struggle. We may be able to work around the problems, but a lot of the time it is very time consuming and tedious. So now we believe we have a solution—Shifter—that will allow them to get past those problems more quickly, and the early results are very encouraging.

## Q. Can you share some of those results and/or successes?

Jacobsen: There are two ways in which Shifter has already been successful. First, we have two major groups beginning to use it right now: the LCLS experimental facility at SLAC, and the high energy physics community at CERN. The LCLS, for example, has its own software environment, and it is rather challenging to take it out of its context and put it into our environment. LCLS spent a lot of time trying to adapt to Edison, which they eventually did, but it took them months to make a Docker image. Using Shifter, however, we were able to create a Docker image in one day and demonstrated that the staff effort in migrating applications is greatly reduced, which was the original purpose of Shifter: to make our system more adaptable to external software.

The other benefit is that, because of the way we present images to the

Edison system in Shifter, it turns out that the software can load much faster than before. So in the case of the LCLS, before Shifter once an image had been ported to NERSC it could sometimes take up to half an hour just for the software to start. With Shifter, everything starts in a matter of seconds—somewhere between 5-20 seconds. So this results in much better utilization of our resources.

Gerhardt: With scientists from the Large Hadron Collider's (LHC) ALICE, ATLAS and CMS experiments at CERN, we are testing Shifter in conjunction with their CVMFS (CERN Virtual Machine File System) software package. I've been working to bring the CVMFS software onto Edison, but it's a huge software repository. For example, with ATLAS (one of two general-purpose detectors at the LHC), if we do a straight-up rsync over scratch, we're working with more than 3.5 TB of data and 20 million inodes. We found that Shifter is a good tool for handling this because you can build a filesystem image on the local node and deliver startup times around half a minute on a single node. So with Shifter the jobs run as efficiently, if not more so, than they do in their current configurations and without the user having to jump through any special hoops. It just works.

Canon: From a big-picture perspective, Shifter is really about trying to enable and simplify the process of science. Scientists really struggle with the fact that they create some sort of code or simulation and it can be really difficult for another user to replicate the computing environment that was used. It's just as challenging as it used to be for scientists to replicate experimental conditions. Shifter is potentially one way to address that challenge.

## Q: One last question: are you going to make this available to other centers? How can people learn more about it?

Canon: Doug has already gone through the steps to open source it and release it through a BSD (Berkeley Software Distribution) license. The intent is that others can download it and use it at their centers. While that might take away from it being a unique capability for NERSC, we think in the end it is important for scientists because the more available this capability is for users, the more they will adopt it and make it a standard for how they operate.

Also, as we've been developing Shifter we've been discussing it with Cray, and hopefully this is something that will become a mainstream capability for Cray systems. The fact that they are working with us on this means they recognize the potential for it.

We also plan to conduct demonstrations of Shifter in the Department of Energy booth at SC15 in November.

Provided by Lawrence Berkeley National Laboratory