Distributed Environment Modules for the OSG Virtual Cluster

Background

We start from the premise that to scale up to large numbers of campus users, we need to make OSG look familiar if not identical to existing campus environments. For two reasons:

- To minimize the "gap" between the campus cluster and the OSG, for the benefit of user ease of adoption
- To provide campus research computing center directors and consultants with a value-added service they in-turn provide to their users and support locally.

We have spoken about the OSG XSEDE service as providing a virtual cluster, for example at https://www.xsede.org/web/guest/OSG-User-Guide, which is our best usage guide for campus users. Yet, lets say we compare this virtual cluster with other XSEDE resources (https://www.xsede.org/high-performance-computing) and campus clusters out there. What we see is something quite different.

While we emphasize DHTC, overlays, the OSG consortium, what is and what is not a good fit, etc., we say little of the software environment, and what tools or applications users might find (or expect to be installed) on the virtual cluster, how to move data from and to the virtual cluster, application development, work queues, etc.

By default, most campuses provide and market "HPC clusters" to their users, irrespective of workloads (whether they are parallel or "serial" HTC applications). And, we find campus HPC centers always have a section describing the software environment, and far and away the most common tool uses is **Environment Modules** http://modules.sourceforge.net/, or tools which evolve from the original, e.g. https://www.tacc.utexas.edu/tacc-projects/lmod.

Indeed, consider its widespread use:

- University of Nebraska campus clusters, https://hcc-docs.unl.edu/display/HCCDOC/Module+Commands
- Illinois Campus Cluster: https://campuscluster.illinois.edu/user_info/doc/#modules
- UChicago Campus Cluster: http://rcc.uchicago.edu/resources/modules.html
- XSEDE Stampede: https://www.tacc.utexas.edu/user-services/user-guides/stampede-user-guide#compenv
- XSEDE SDSC Gordon: https://www.xsede.org/sdsc-gordon#modules
- UC San Diego, RCI http://rci.ucsd.edu/computing/jobs/modules.html, and http://rci.ucsd.edu/computing/system-info/software.html
- Indiana University clusters, http://kb.iu.edu/data/bcwy.html , http://kb.iu.edu/data/bcqt.html#software, https://cybergateway.uits.iu.edu/iugateway/modulesInfo?machine=bigred2
- XSEDE Future Grid: https://www.xsede.org/web/guest/futuregrid#hpc-compenv
- Purdue campus cluster: https://www.rcac.purdue.edu/userinfo/resources/hansen/userguide.cfm#app_module
- Clemson Palmetto cluster: http://citi.clemson.edu/palmetto/pages/userguide.html
- NYU: https://wikis.nyu.edu/display/NYUHPC/Union+Square
- UCLA: (interesting campus computing program link, https://idre.ucla.edu/hpc/shared-cluster-program, and why can't we get on this page, https://idre.ucla.edu/hpc/additional-computing-resources), their

software: http://hpc.ucla.edu/hoffman2/software/software.php (no user guides available) I also think they've got a relic campus grid there.

- Harvard's Odyssey cluster,https://rc.fas.harvard.edu/resources/odyssey-quickstart-guide/ and https://rc.fas.harvard.edu/resources/module-list/
- NERSC, https://www.nersc.gov/users/software/nersc-user-environment/modules/
- Titan, https://www.olcf.ornl.gov/kb_articles/using-modules/

Indicated Solution

So we are thinking of the virtual cluster equivalent, or ***Distributed Environment Modules*** where, after a survey and study of software commonly installed on HPC clusters (even with an eye towards "parallel" scripts which can take advantage of multi-core slots), and the XSEDE Campus Bridging Yum rpm repo: https://www.xsede.org/web/xup/knowledge-base/-/kb/document/bdwx (and this list of modules: https://software.xsede.org/packages/cb/centos6/x86_64/README.0.0.7).

Here we leverage OASIS, distributing self-resolved collections of software with extremely few dependencies on the host compute system (e.g. *6/x88_64) , and provide a simple command line environment with the same syntax, look and feel as:

module [*switches*] [*sub-command*] *[sub-command-args]*

from http://modules.sourceforge.net/man/module.html.

We should have a well-defined user software support page including:

- A description of currently 'installed' software, and the basics of how to load specific versions of software into your environment.
- Instructions for compiling and linking user applications
- procedures for requesting new software to be installed

An OSG Software Librarian is appointed to manage collections; we are thinking of three types of collections:

1) The Common Core Collection (i.e. our best collection for the virtual cluster)
2) The XSEDE campus bridging collection
3) Campus Series A, B, … where these can be collections requested specifically by a campus to more closely follow their own environment

Note any of these could be used. The only requirement is dependency consistency within a collection.

Yes we have to deal with licensed software and proprietary compilers We just need document what's possible, what's not.

Yes, the modules list can be in the thousands. More maintaining versions. But there are obvious starting points.

We have the issue of OASIS supporting versus non-OASIS supporting resource targets. We have Parrot for the latter, though we may find additional work is needed there to address potential shortcomings.

I believe this is the most important thing we can do to reach existing campus researchers, and it is all very doable. This would ben an opportunity to collaborate with TACC and Harvard.