2014-04-10 Virtualized platforms for research

Readings

Initial readings announcement:

Dav Clark of the D-Lab, which maintains Collaboratool / Berkeley Computational Environment -- compute environments for portable, reproducible data science -- will kick off discussion that will also include Ben Gross of IST-EIS (Endpoint Engineering and Infrastructure). We hope that Owen McGrath of ETS (Educational Technology Services) will also be able to join us.

Short readings are:

About Collaboratool ("meta-features")

About Docker -- Docker is software that permits applications to run in a container that can itself be reproduced and run on any of multiple machines & platforms

A use case describing how Docker has been applied in a higher-ed instructional context

About Packer -- Packer is a tool for creating machine images -- please read the four short "About" section pages starting with the linked page, http://www.packer.io/intro

Follow-up announcement:

Ben Gross of IST-EIS notes his team's use of Vagrant for provisioning reproducible and portable dev environments. The author of Vagrant is the same fellow who wrote Packer (Mitchell Hashimoto), who weighs in on a great Stack Overflow back-and-forth that looks at differences between and appropriate applications of Docker and Vagrant. Ben also points out a nice summary with examples on the topic of Advanced Provisioning With Packer For Docker And Vagrant by UCLA CS student Matthew McKeen, which will help to orient readers to the relationships between these tools/projects.

Notes

Participating: ~20 individuals, representing D-Lab, ETS, Econometrics/Statistics, EECS, IST-EIS, and RIT

Dav Clark (D-Lab, cf. BCE) led a white-board exercise to elicit major points in spectrum of what virtualization might mean:

* Containerize application
* Compute environment
* Full virtualized environments

Another way to parse:

* environments for development
* environments for (parallel) evaluation - a.k.a. grading
* environments for reproducible research
* environments for instruction
* environments for SaaS/PaaS offerings (example: Collection Space deployments at UCB)

EIS offers, per Ben Gross -- vis-a-vis infrastructure for managing and updating 9000 machines, pushing updates to 6000 machines/month:

* Citrix -- containerized browsers for ~20 enterprise apps
* Win 7 64 and (soon) Mac standard workstation images installable on bare metal and/or on VMs

Observation that this group -- diverse, experienced -- does not have a set clean schema for classifying use cases and solution types ...

Outreach an issue: current users have been in the habit for years; many faculty other than these don't know these resources exist

Another need: beefier provisioning than on a laptop, VM virtualized on more powerful resources, snapshots that can be taken and resumed later ... for computational research that requires more than a laptop's worth of CPU, memory, and/or storage. To this end, Docker as a potential solution -- not only app management, resource quotas -- but incremental image management. Branching. Perhaps specific to research use cases. Snapshotted VMs gives you big/heavy artifact to manage, to move around on a network; Docker lets you snapshot something smaller, record diffs and move those around -- easier to manage movement on network.

Issue on reconstitution of snapshots: managing updates after reconstituting an environment that is a couple of years old
Closing questions / follow-up:

- How to continue this discussion? Community of practice? Listserve on this topic for campus-interested parties?
- Possibility of BCE (D-Lab) on software central (IST-EIS)? Ben and Dav to follow-up
- If BCE addresses an environment that uses only open software, who is tilting at the windmill use cases that require licensure?