www. chameleoncloud.org

IMPROVING REPRODUCIBILITY WITH CLOUDS AND NOTEBOOKS

Kate Keahey

Mathematics and CS Division, Argonne National Laboratory

CASE, University of Chicago

keahey@anl.gov

November 8, 2019
GEFI Workshop, Coimbra, Portugal

REPRODUCIBILITY DILEMMA

Should I invest in making my experiments repeatable?

Should I invest in more new research instead?

- Challenges
  - Actionable digital artifacts: configurations, scientific practices and processes
  - Context in which they can be shared: instruments, resources, etc.
  - Publication, discovery, indexing, etc.
- Towards intentional up front shareable research

CHAMELEON AS A SCIENTIFIC INSTRUMENT

- We like to change: testbed that adapts itself to your experimental needs
  - Deep reconfigurability (bare metal) and isolation (CHI) – but also ease of use (KVM)
  - CHI: power on/off, reboot, custom kernel, serial console access, etc.
- We want to be all things to all people: balancing large-scale and diverse
  - Large-scale: ~large homogenous partition (~15,000 cores), 5 PB of storage distributed over 2 sites (now +1!) connected with 100G network…
  - …and diverse: ARMs, Atoms, FPGAs, GPUs, Corsa switches, etc.
- Cloud++: leveraging mainstream cloud technologies
  - Powered by OpenStack with bare metal reconfiguration (Ironic) + "special sauce"
  - Chameleon team contribution recognized as official OpenStack component
- We live to serve: open, production testbed for Computer Science Research
  - Started in 10/2014, testbed available since 07/2015, renewed in 10/2017
  - Currently 3,500+ users, 500+ projects, 100+ institutions

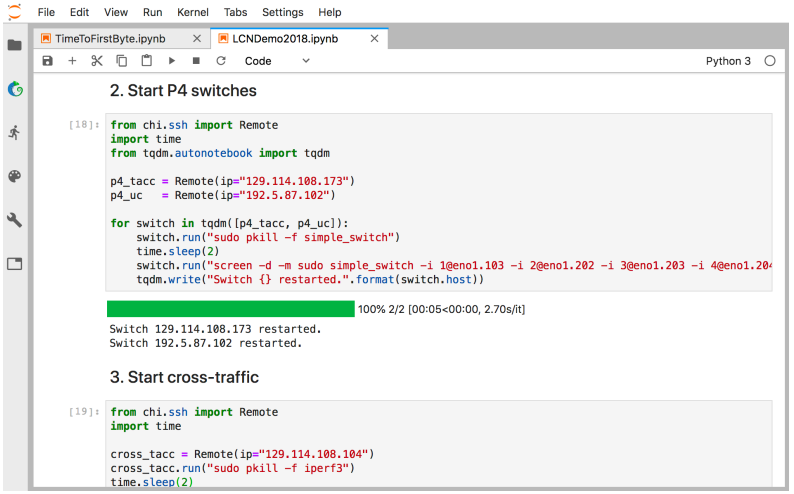
BEYOND THE INSTRUMENT: AN ECOSYSTEM FOR REPEATABILTY AND SHARING

- Clouds/testbeds generate a wealth of shareable artifacts
  - Images, orchestration templates, tools, etc.
  - Clouds/testbeds as “players” for common artifacts

- Repeatability/replicability features
  - Testbed versioning (>50 versions of the testbed)
  - Appliance/digital artifact versioning
  - Experiment Precis: a history command analogue: uses testbed logging data to reconstruct a set of actions
- Documenting a scientific process
  - Imperative, non-transactional, version controlled, etc.
  - Orchestration versus notebooks

CHAMELEON JUPYTER INTEGRATION

- Combining the ease of notebooks and the power of a shared platform
  - Storytelling with Jupyter: ideas/text, process/code, results
  - Chameleon: sophisticated experimental containers in need of "storytelling"
- JupyterLab server for our users
  - Go to jupyter.chameleoncloud.org and use with your Chameleon credentials
- Chameleon/Jupyter integration
  - Python/bash interfaces to the testbed, storing and sharing, Chameleon credentials
  - Named containers
- Templates of existing experiments



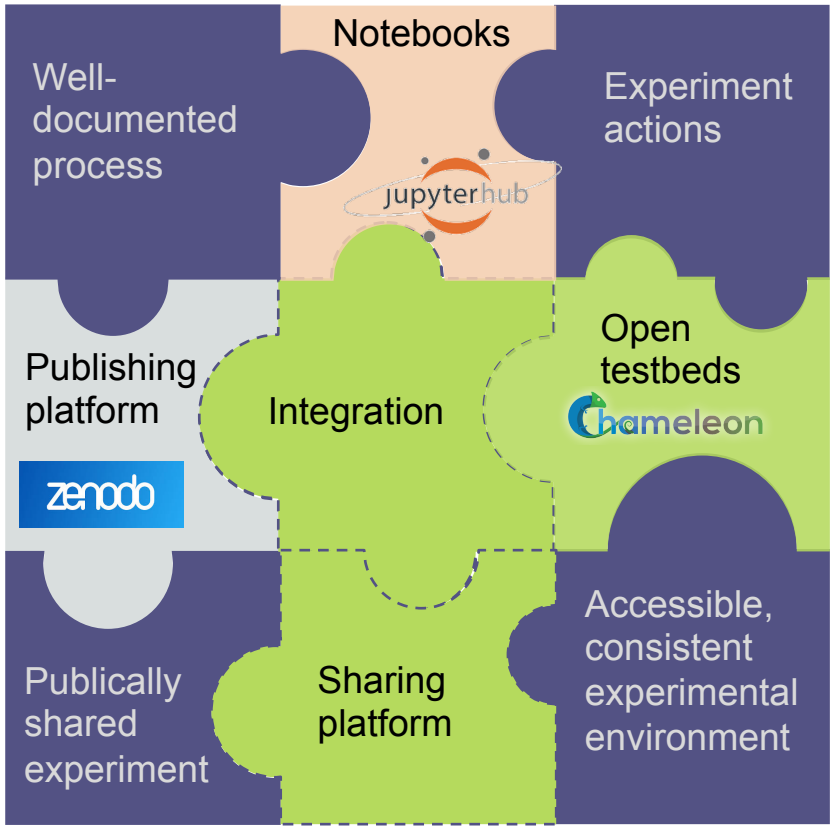
Screencast of a complex experiment: https://vimeo.com/297210055

"A Case for Integrating Experimental Containers with Notebooks", CloudCom 2019

SHARING, PUBLISHING, LEVERAGING

- We now have everything we need to share experiments
  - Ways to establish an experimental environment + player
  - Ways to document an experimental process
- But wait… how do I actually share them?
  - Send mail, Chameleon object store, github…
  - Publishing via Zenodo: store your experiments and make them citable via DOIs
- Creating bridges, integration
  - Import/Export from/to Zenodo
- Making research findable: the sharing platform

SC19 Poster: Sharing and Replicability of Notebook-Based Research on Open Testbeds


Notebooks
Well-documented process
Experiment actions
jupyterhub
Publishing platform
zenodo
Integration
Open testbeds
chameleon
Publically shared experiment
Sharing platform
Accessible, consistent experimental environment

PARTING THOUGHTS

- Logistical barriers stunt creativity and ambition
- Towards a Digital Research Ecosystem: a meeting place of users and providers sharing resources and research
  - Clouds/testbeds are more than just experimental platforms; they create a "common denominator" that can eliminate much complexity that goes into systematic experimentation, sharing, and reproducibility
  - Notebooks + testbeds provide both the sharing underpinnings (common artifacts) and the ability to document the process
  - Digital era publishing tools facilitate sharing
- Leveraging new digital artifact and sharing patterns towards up front shareable research