Open Source Configuration of Bioinformatics Infrastructure

John Chilton1, Pratik Jagtap1, Benjamin Lynch1, Brad Chapman2, Timothy Griffin3
1 University of Minnesota Supercomputing Institute
2 Harvard School of Public Health
3 University of Minnesota

Beyond CloudBioLinux
Implemented on top of Git repositories

This presentation: http://bit.ly/boSc2013

CloudBioLinux Extensions
Extended Git to allow use of Puppet modules and Chef cookbooks.
Puppet/CloudBioLinux and Chef modules packaged.
Remote server cloned from central repository.
Shorten up, update, apply to remote server.

Community?
http://cloudbio.org/community

High-Throughput
Single command for high-quality
Puppet/BioLinux modules for use with
CloudBioLinux. Galaxy integration or
Institutional repositories.

Initial Applications
- LWR
- BioCloudCentral

...we can do better
with Puppet and Chef

Background

Traditional

Testable

Composable

High-Throughput
Open Source Configuration of Bioinformatics Infrastructure

John Chilton1, Pratik Jagtap1, Benjamin Lynch1, Brad Chapman2, Timothy Griffin3
1 University of Minnesota Supercomputing Institute
2 Harvard School of Public Health
3 University of Minnesota

CloudBioLinux Extensions
- Extended DSL to allow use of Puppet modules and Chef cookbooks
- Puppet can be easily integrated into any workflow or infrastructure
- Chef can be easily integrated into any workflow or infrastructure

Community?
http://www.cloudbio.org/community

This presentation:
CloudBioLinux a Start but...

Fabric is library used by CBL to remotely run install commands.

Fabric is great at recreating identical deployments on multiple machines.

The Problem: Different institutions/teams want to build different environments with applications configured differently.

Fabric is NOT a configuration management tool.

Core Idea

Configuring complex applications is hard!

Building on open-source frameworks can simplify this task.

CloudBioLinux (aka CloudMan) is an example.

Background
Core Idea

Configuring complex applications is hard!

Building on open source frameworks can simplify this task.

CloudBioLinux (& CloudMan) is an example.
CloudBioLinux

Packages (YAML)
Can be OS packages, language libraries, or custom installs

bio.nextgen:
- bio-linux-fastqc
- fastx-toolkit
- maq
- plink

bio.proteomics:
- xsltproc
- libxml-sax-expat-perl
- libgd2-xpm-dev
- libbz2-dev

Fabric (Python) Methods

```
@if not installed('blast!
  def install_blast(env):
    """BLAST Blast-like Fast Accurate Search Tool.
    """
    default_version = "2.2.26"
    version = env.get('tool.version', default_version)
    major_version_regex = "\d+\.(\d+).*"
    major_version = re.search(major_version_regex, version).group(0)
    url = "http://downloads.sourceforge.net/project/blast/blast/15/blast-x.tar.gz"
    # major_version, version
    .get_install_url(env, configure_make)
```
Fabric (Python) Methods

```python
@if_not_installed("bfast")
def install_bfast(env):

    # BFAST: Blat-like Fast Accurate Search Tool.

    default_version = "0.7.0a"
    version = env.get("tool_version", default_version)
    major_version_regex = "\d+\\d+\d+"
    major_version = re.search(major_version_regex, version).group(0)
    url = "http://downloads.sourceforge.net/project/bfast/bfast/\$s/bfast-%s.tar.gz"
       % (major_version, version)
    _get_install(url, env, _configure_make)
```
CloudBioLinux a Start but...

Fabric is library used by CBL to remotely run install commands.

Fabric is great at recreating identical deployments on multiple machines.

The Problem: Different institutions/teams want to build different environments with applications configured differently.

Fabric is NOT a configuration management tool.
High Level

Fabric is a low-level procedural library. Chef & Puppet are DSLs with higher level constructs for services, dependencies, packages, etc...

Built-in easy templating (great for config files).

Testable

Great unit testing frameworks available.

...we can do better

with Puppet and Chef!

Composable

Applications broken down into packages that can be easily shared.

Huge wealth of existing best practice configurations exist. Apache, Firewalls, etc...
High Level

Fabric is a low-level procedural library. Chef & Puppet are DSLS with higher level constructs for services, dependencies, packages, etc...

Built-in easy templating (great for config files).
Composable

Applications broken down into packages that can be easily shared.

Huge wealth of existing best practice configurations exist.

Apache, Firewalls, etc...
Testable

Great unit testing frameworks available.
CloudBioLinux Extensions

Extended CBL to allow use of Puppet modules and Chef cookbooks.
  Puppet/Chef remotely installed as needed, packages are
  bundled up, shipped to remote server, and applied to server.

Integrates with existing CBL structure for ‘properties’ and ‘packages’.
  Can set Puppet and Chef properties via Fabric

Can define what modules/cookbooks configured
  via new YAML package types.
Initial Applications

- **LWR**
LWR is a tool to stage and run Galaxy jobs on remote servers.
https://bitbucket.org/readyforcloud/lwr

- **Globus**
The Globus Toolkit provides utilities for federated data transfer, identity management, etc.
https://bitbucket.org/biocompute/globus
http://www.globus.org

- **BioCloudCentral**
Django application allowing users to easily launch CloudBioLinux and Cloudman instances

https://github.com/biocompute/puppet-biocloudcentral

Powers https://biocloudecentral.msf.uran.edu allowing end users to easily launch Galaxy-P instances on Amazon.
LWR is a tool to stage and run Galaxy jobs on remote servers.

https://lwr.readthedocs.org/

Hope to get this tightly integrated into CloudMan instances by default, potentially a path forward for cloud bursting Galaxy instances.

Puppet module for configuring LWR has been integrated into CloudBioLinux.

https://github.com/bioconfig/puppet-lwr
Globus

The Globus Toolkit provides utilities for federated data transfer, identity management, etc...

https://github.com/bioconfig/chef-globus

Fork of the Globus Provision Chef recipes.


Instructions for using GridFTP to transfer data into CBL instance created with Gloubs.
BioCloudCentral

Django application allowing users to easily launch CloudBioLinux and CloudMan instances

https://github.com/bioconfig/puppet-biocloudcentral

Powers https://biocloudcentral.msi.umn.edu allowing end users to easily launch Galaxy-P instances on Amazon.
Beyond CloudBioLinux

Implemented on top of git submodules

```
github.com:chapmanb/cloudbiolinux.git
config/
puppet/
modules/
lwr
biocloudcentral
apache
.....
github.com:bioconfig/puppet-lwr.git
```

```
github.com:bioconfig/puppet-biocloudcentral.git
```

```
github.com:puppetlabs/puppetlabs-apache.git
```

**Upshot:**
They can be easily integrated the same way by institutions or teams with their own Chef or Puppet repositories or by tools such as Globus Provision.
Beyond CloudBioLinux

Implemented on top of git submodules

github.com:chapmanb/cloudbiolinux.git
  config/
  puppet/
  modules/
    lwr
    biocloudcentral
    apache
    ....

...
Beyond Cloud BioLinux

Implemented on top of git submodules

github.com:chapmanb/cloudbiolinux.git
  config/
  puppet/
    modules/
      lwr
      biocloudcentral
      apache
    ...

github.com:bioconfig/puppet-lwr.git

github.com:bioconfig/puppet-biocloudcentral.git

github.com:puppetlabs/puppetlabs-apache.git

Upshot:
They can be easily integrated the same way by institutions or teams with their own Chef or Puppet repositories or by tools such as Globus Provision.
Community?

biopython, bioperl, biojava...

bioconfig?

http://github.com/bioconfig/XXXXX

Clearing house for high quality interoperable modules for use with CloudBioLinux, Globus Provision, or institutional repositories.