

# Open Source Configuration of Bioinformatics Infrastructure

John Chilton<sup>1</sup>, Pratik Jagtap<sup>1</sup>, Benjamin Lynch<sup>1</sup>, Brad Chapman<sup>2</sup>, Timothy Griffin<sup>3</sup>

<sup>1</sup> University of Minnesota Supercomputing Institute

<sup>2</sup> Harvard School of Public Health

<sup>3</sup> University of Minnesota

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

**Initial Applications**

- LWR
- Globus
- StoCloudCentral

**Beyond CloudBioLinux**  
Implemented on top of git submodules

**Update:** 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

**Community?**

biopython, bioport, bioconda, bioconig?

<http://github.com/bioconig/xxxxx>

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

**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 "preparates" and "packages". Can set Puppet and Chef repositories via Fabric

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

**High Level**

Helps to easily troubleshoot (great for remote install)

**Testable**

and test before deployment

**...we can do better with Puppet and Chef!**

**Composable**

Application for those users who partition that can be easily shared

Helps with existing tool services (bioinformatics, genomics, proteomics, etc...)

**Background**

CloudBioLinux is a Linux distribution that is specifically designed for bioinformatics infrastructure. It is based on CentOS and includes a variety of bioinformatics tools and services.

CloudBioLinux is a Linux distribution that is specifically designed for bioinformatics infrastructure. It is based on CentOS and includes a variety of bioinformatics tools and services.

# Open Source Configuration of Bioinformatics Infrastructure

John Chilton<sup>1</sup>, Pratik Jagtap<sup>1</sup>, Benjamin Lynch<sup>1</sup>, Brad Chapman<sup>2</sup>, Timothy Griffin<sup>3</sup>

1 University of Minnesota Supercomputing Institute  
 2 Harvard School of Public Health  
 3 University of Minnesota

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

**Initial Applications**

- LWR
- Globus
- BioCloudCentral

**Beyond CloudBioLinux**  
 Implemented on top of git submodules

- github.com/opensciencegrid/opensciencegrid
- github.com/opensciencegrid/opensciencegrid
- github.com/opensciencegrid/opensciencegrid
- github.com/opensciencegrid/opensciencegrid
- github.com/opensciencegrid/opensciencegrid

**Update:**  
 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?**  
 bioinformatics, bioinformatics, bioinformatics

<http://github.com/opensciencegrid/opensciencegrid>

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

**CloudBioLinux Extensions**

Extended CSL 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 CSL structure for 'preparises' and 'packages'.  
 Can set Puppet and Chef properties via Fabric.  
 Can define what modules/cookbooks configured via new YAML package types.

**...we can do better with Puppet and Chef!**

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

**Testable**  
 Small test suites, resources available

**Composable**  
 Applications for data, don't write packages that use too many packages.  
 Many months of existing best practice configurations of puppet, chef, etc.

**Background**

CloudBioLinux is a DSL that...  
 ...  
 ...  
 ...

## Core Idea

[bit.ly/prodcloudman-slides](http://bit.ly/prodcloudman-slides)

Configuring complex applications is hard!

Building on open source frameworks can simplify this task.

CloudBioLinux (& CloudMan) is an example.



## 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.*

# Background

# Core Idea

[bit.ly/prodcloudman-slides](http://bit.ly/prodcloudman-slides)

Configuring complex applications is hard!

Building on open source frameworks can simplify this task.

CloudBioLinux (& CloudMan) is an example.

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

```
My-system
- bio-issue-headers
- matplotlib
- perl
- python
- python-scipy
- software
- libtool-cxx-external
- libtool-cxx-internal
- libtool-cxx
```

**Fabris (Python) Methods**

```
#!/usr/bin/perl
# This script is a wrapper
# for the fab command
# It will run the fab command
# with the arguments passed
# to it.
# Usage: ./fab.py [options]
# Options:
# -h, --help            show this help message and exit
# -v, --verbose          verbose output
# -q, --quiet            quiet output
# -s, --ssh-key=KEY     ssh key file
# -u, --user=USER        ssh user
# -p, --port=PORT        ssh port
# -O, --options=OPTIONS  ssh options
```

# 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('bfast')
def install_bfast(env):
    """BFAST: Blat-like Fast Accurate Search Tool.
    http://sourceforge.net/apps/mediawiki/bfast/index.php?title=Main\_Page
    """
    default_version = "0.7.0a"
    version = env.get("toolVersion", 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)
```

# Fabric (Python) Methods



```
@_if_not_installed("bfast")
def install_bfast(env):
    """BFAST: Blat-like Fast Accurate Search Tool.
    http://Sourceforge.net/apps/mediawiki/bfast/index.php?title=Main\_Page
    """
    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://downloadSourceforge.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.**

# ...we can do better with Puppet and Chef!

## 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.

## 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://hr.readthedocs.org/>  
*Here to get this tightly integrated into CloudMan instances by default, potentially a path forward for cloud hosting 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>  
Part of the Globus Provision Chef recipes.  
<http://bit.ly/cbl-gridftp>  
Instructions for using GridFTP to transfer data into CBL instances created with Globus.

## - 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.



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>



# Gloubs

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.

<http://bit.ly/cbl-gridftp>

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

.....

...

github.com:bioconfig/puppet-lwr.git

github.com:bioconfig/puppet-biocloudcentral.git

github.com:puppetlabs/puppetlabs-apache.git

ot:

can be easily integrated the same way by institutions or teams with their own Chef or puppet repositories or by tools such as Globus



# 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.



# 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.