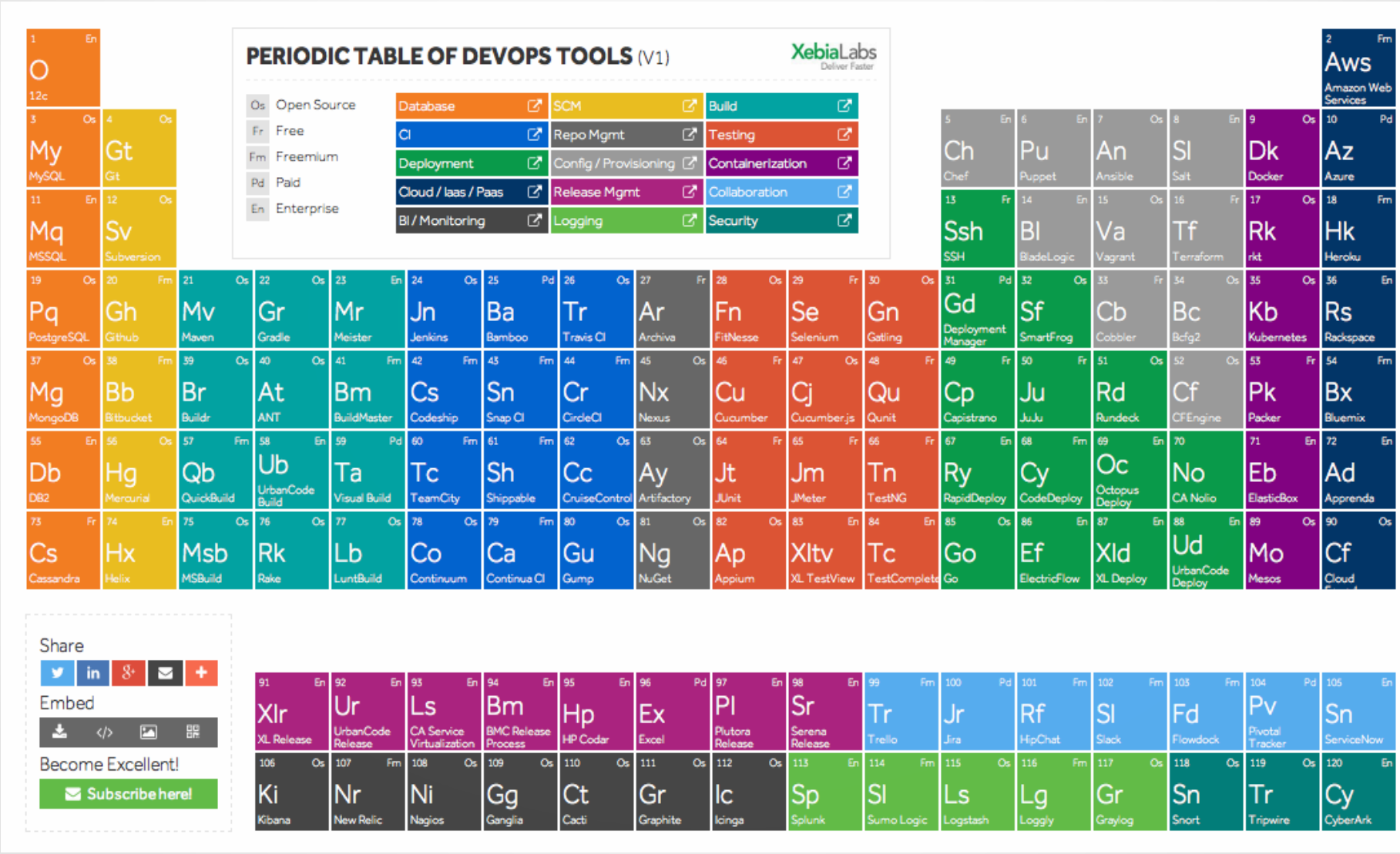
Consortium des Equipements
de Calcul Intensif
en Fédération Wallonie-Bruxelles

Introduction to Scientific Software Deployment and Development

damien.francois@uclouvain.be
October 2018

http://www.ceci-hpc.be/training.html

What is this?



https://xebialabs.com/periodic-table-of-devops-tools/

Goal of this session:

"Give you access to the same tools
the professionals are using
for **developing** and **deploying** programs."

Dev's toolkit :

1. Programming languages
2. Good practices/principles/style
3. Text editor
4. Source control management
5. Debuggers / Profilers
6. Databases
7. Packaging / Distributing tools
8. Comments and documentation
9. Tests
10. Licensing

1. Programming language

- Good reasons for choosing language X:
  - it offers useful paradigms for your problem
  - it offers high-level constructs/tools - e.g. for parsing arguments
  - it offers (directly or indirectly) useful libraries - e.g. for linear algebra
- Ok reasons for choosing language X:
  - standard in your community – easier to get accepted
- Bad reasons for choosing language X:
  - it runs fast – probably needs high skills to be fast
  - it is the language you already know

1. Programming language

Be aware of the 'other' paradigm...

Imperative – “Do this” BASIC, Assembly	Declarative – “I need this” SQL
Structured – Subroutines, scopes C, FORTRAN77 algorithms + data : good for explicit computing	Functional – Pure functions, lazy evaluation Haskell, Scala functions o functions : good for reasoning
Object-Oriented – Encapsulation, Inheritance, ... C++, Python objects + messages : good for modeling	Logic – Predicates and rules Prolog, Datalog facts + rules : good for searching

1. Programming language

C

```c
void f(int a[], int lo, int hi)
{
  int h, l, p, t;

  if (lo < hi) {
    l = lo;
    h = hi;
    p = a[hi];

    do {
      while ((l < h) && (a[l] <= p))
          l = l+1;
      while ((h > l) && (a[h] >= p))
          h = h-1;
      if (l < h) {
          t = a[l];
          a[l] = a[h];
          a[h] = t;
      }
    } while (l < h);

    a[hi] = a[l];
    a[l] = p;

    f( a, lo, l-1 );
    f( a, l+1, hi );
  }
}
```

Haskell

```haskell
qsort []     = []

qsort (p:xs) = (qsort lesser) ++ [p] ++ (qsort greater)
    where
        lesser  = filter (< p) xs
        greater = filter (>= p) xs
```

Purely functional
Static strong typing
Lazy evaluation

2. Good practices

- Write for humans, not for computers
- Use the appropriate language(s)
- Organize for change, and make incremental changes
- Plan for mistakes, automate testing
- Automate repetitive tasks
- Use modern source-code management system
- Document the design and purpose, not the implementation
- Optimize only when it works already

2. Good practices

Paul F. Dubois. 1999. Ten Good Practices in Scientific Programming. *Computing in Science and Engg.* 1, 1 (January 1999), 7-11. DOI=10.1109/MCISE.1999.743610 http://dx.doi.org/10.1109/MCISE.1999.743610

Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, et al. (2014) Best Practices for Scientific Computing. PLoS Biol 12(1): e1001745. doi:10.1371/journal.pbio.1001745

Dubois PF, Epperly T, Kumfert G (2003) Why Johnny can't build (portable scientific software). Comput Sci Eng 5: 83–88. doi: 10.1109/mcise.2003.1225867

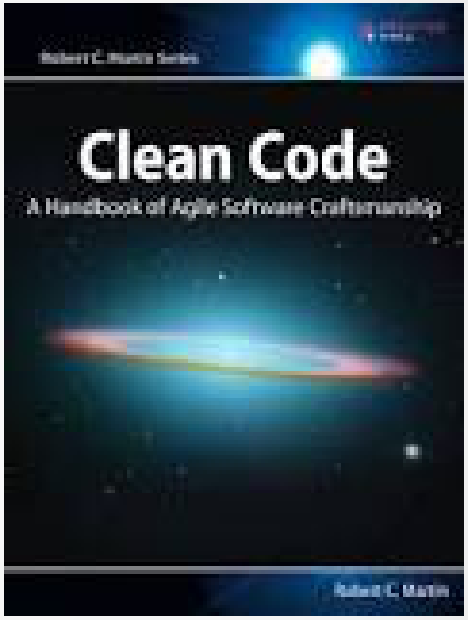
Prlić A, Procter JB (2012) Ten Simple Rules for the Open Development of Scientific Software. PLoS Comput Biol 8(12): e1002802. doi:10.1371/journal.pcbi.1002802

Victor R. Basili, Jeffrey C. Carver, Daniela Cruzes, Lorin M. Hochstein, Jeffrey K. Hollingsworth, Forrest Shull, Marvin V. Zelkowitz, "Understanding the High-Performance-Computing Community: A Software Engineer's Perspective," IEEE Software, vol. 25, no. 4, pp. 29-36, July/August, 2008

Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK (2017) Good enough practices in scientific computing. PLoS Comput Biol 13(6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510

2. Good coding principles

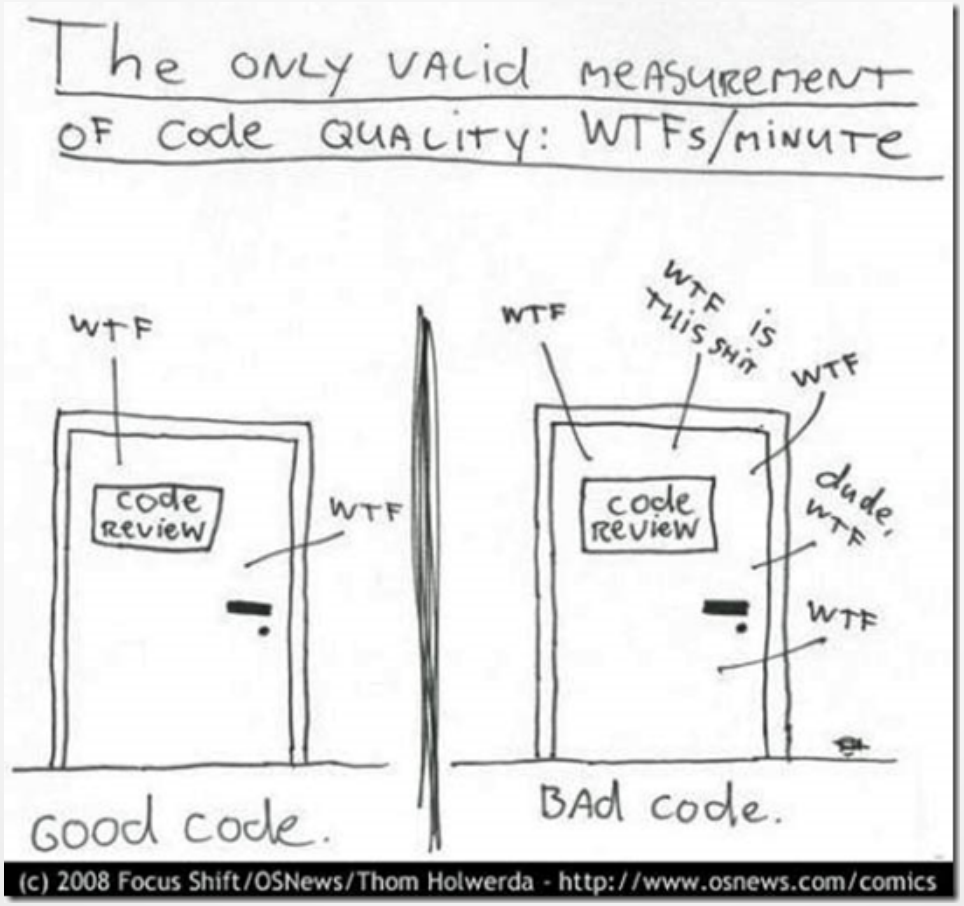
- Don't repeat yourself (DRY)
- Keep it simple, Stupid (KISS)
- One level of abstraction
- Single responsibility principle
- Separation of concern
- Avoid premature optimization
- Many others...



> Bill Mitchell View profile More options Sep 26 1991, 1:57 am In article <5...@ksr.com> j...@ksr.com (John F. Woods) writes:
>
> > [...] Always code as if the guy who ends up maintaining your code will be a violent psychopath who knows where you live. Code for readability.
>
> Damn right!

Clean Code: A Handbook of Agile Software Craftsmanship, R. C. Martin, Prentice Hall, 2008

How to measure code quality



2. Good style

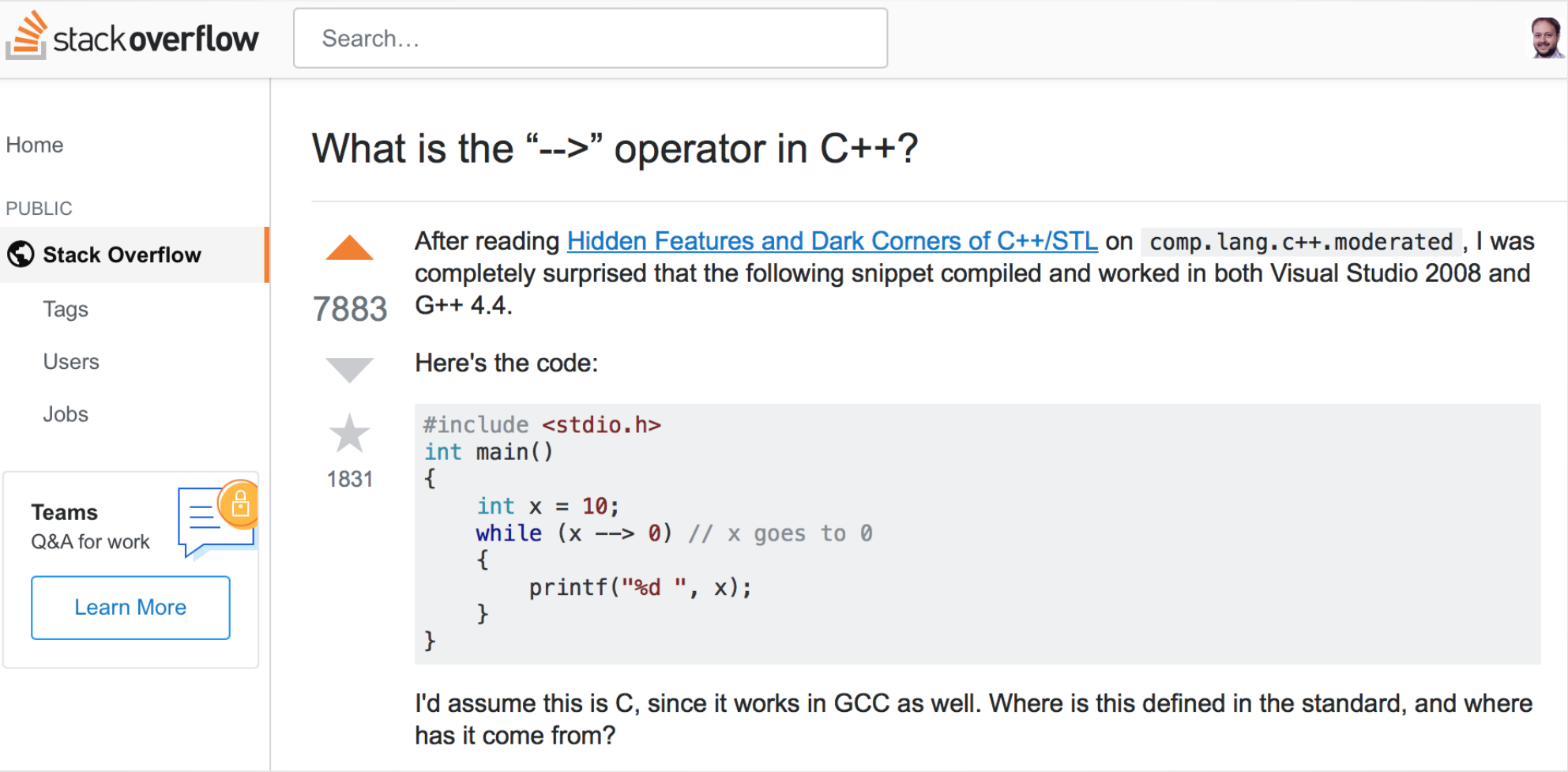
- Makes sure the code is readable by all
  - easily
  - quickly
- Depends on
  - the language
  - the project

```
if (hours < 24 && minutes < 60 && seconds < 60)
{
    return true;
}
else
{
    return false;
}
```

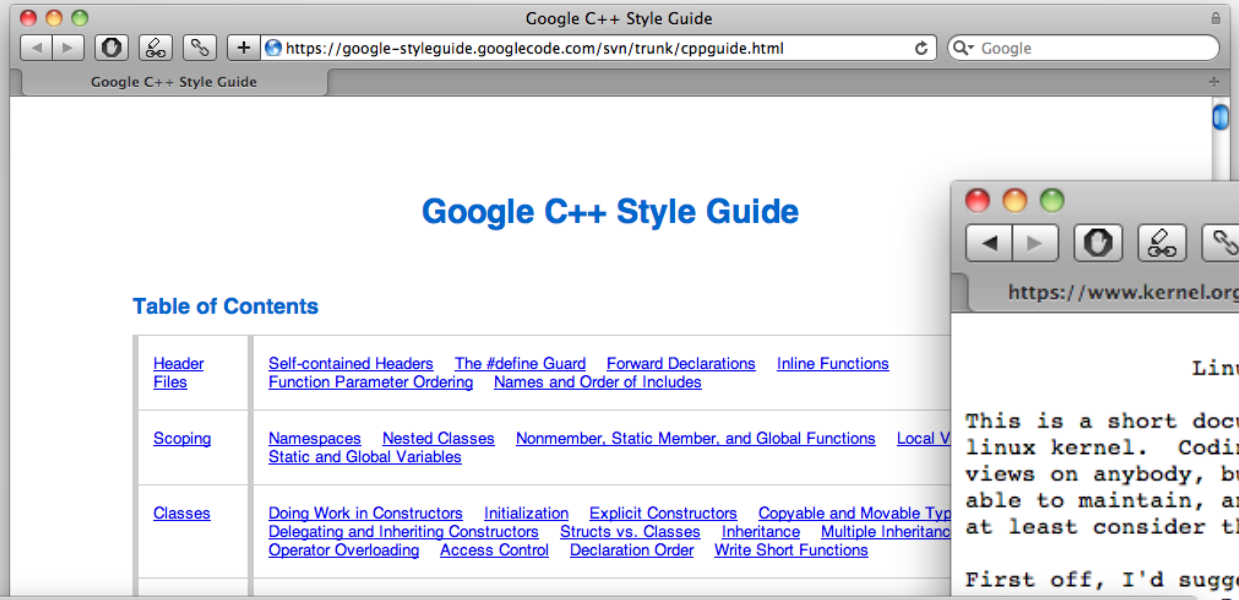
vs

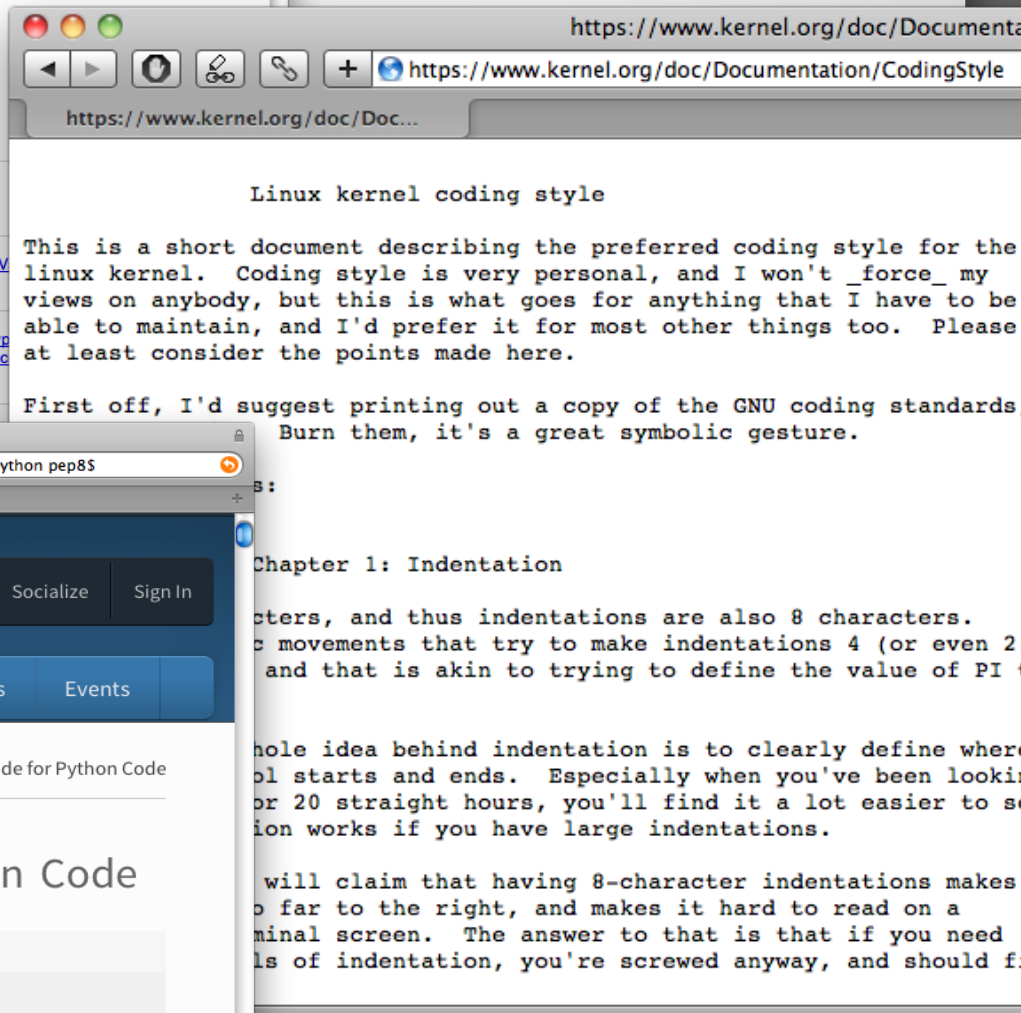
```
if  ( hours   < 24
   && minutes < 60
   && seconds < 60
)
{return    true
;}         else
{return   false
;}
```

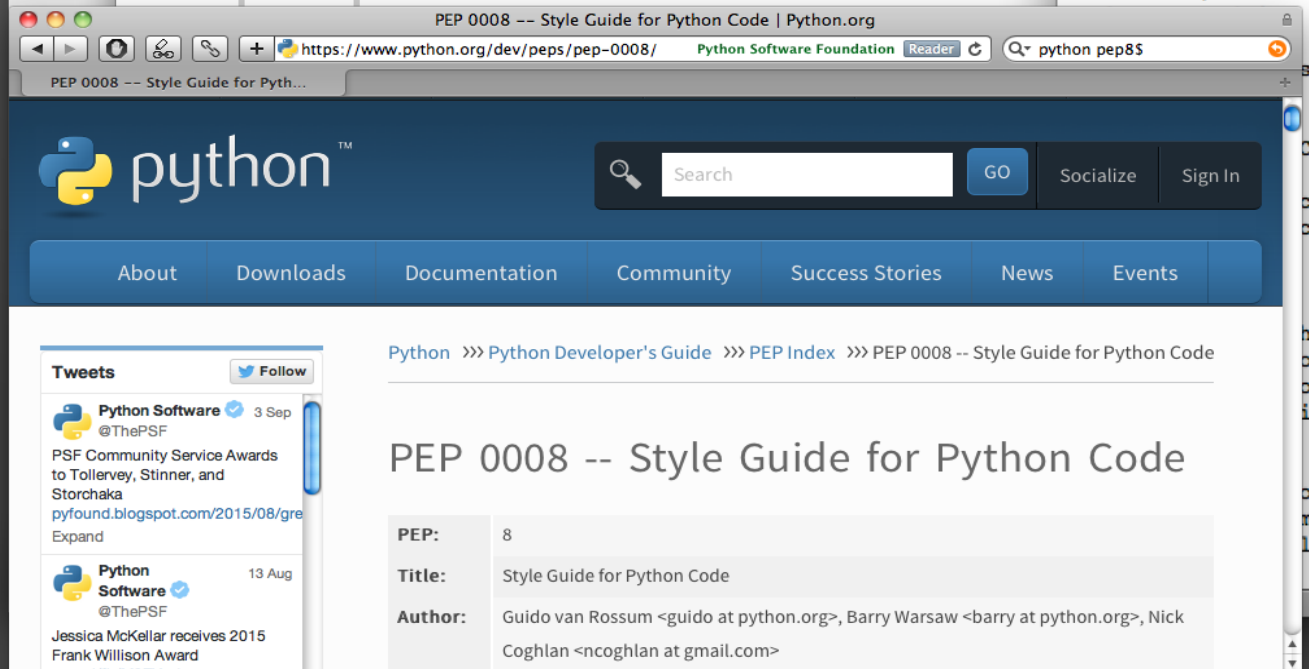
2. Good style


stackoverflow
Search…
Home
PUBLIC
Stack Overflow
Tags
Users
Jobs
Teams
Q&A for work
Learn More
What is the “-->” operator in C++?
7883
1831
After reading Hidden Features and Dark Corners of C++/STL on comp.lang.c++.moderated, I was completely surprised that the following snippet compiled and worked in both Visual Studio 2008 and G++ 4.4.
Here's the code:
#include <stdio.h>
int main()
{
int x = 10;
while (x --> 0) // x goes to 0
{
printf("%d ", x);
}
}
I'd assume this is C, since it works in GCC as well. Where is this defined in the standard, and where has it come from?

2. Good style







https://github.com/SalGnt/cscs

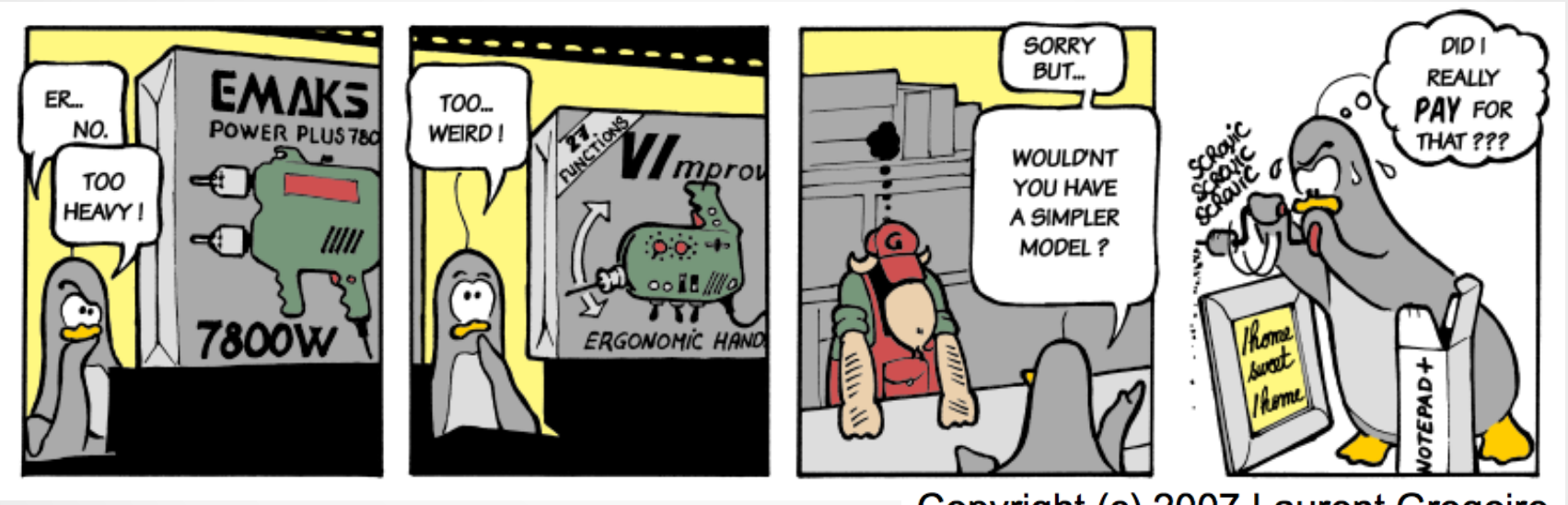
3. Text editor

- Some files are better edited directly on the clusters;



- If you prefer a graphical user interface, some good candidates are:
  - Sublime text: http://www.sublimetext.com/
  - Notepad++: https://notepad-plus-plus.org/
  - Text Wrangler: http://www.barebones.com/
  - Textmate: https://macromates.com/
  - Atom: https://atom.io/
- Choose one and learn it from inside out

3. Text editor



Dev's toolkit :

1. Programming language
2. Good practices / Code Style Guides
3. Text editor / IDE
4. **Source control management**
5. **Debuggers / Profilers**
6. **Databases**
7. Packaging / Distributing tools
8. Comments and documentation
9. Tests
10. Licensing

Own dedicated sessions

7. Packaging Fortran/C/C++ code

Making sure it compiles on your laptop is not enough

It has to compile on all the clusters...

https://cmake.org/ , https://en.wikipedia.org/wiki/GNU_build_system

8. Comments / Documentation

Lots of useless comments

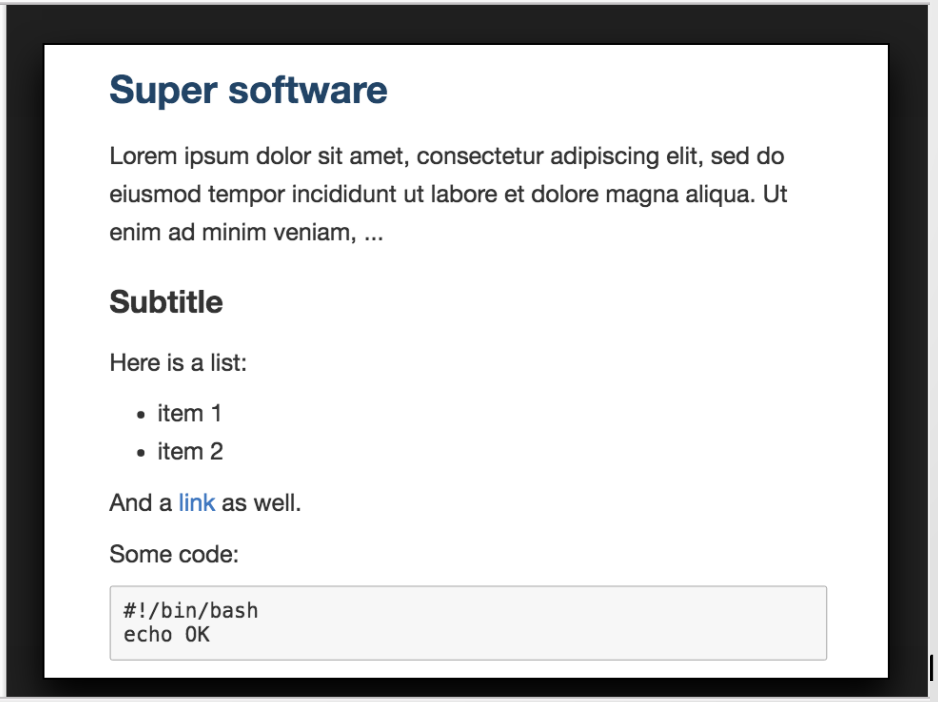
```
function res = f(base, num)
% Assign base to res
res = base
% loop from 2 to num
for i=2:num
    % multiply current res by base
    res=base*res;
end
```

Less comments but useful comments

```
function res = pow(base, num)
% compute base^num by iterative multiply for baseline check
res = base
for i=2:num
    res=res*base;
end
```

Write doc in a lightweight markup language (Markdown, rst, etc.)

```
Super software
==============

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do
eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut
enim ad minim veniam, ...

Subtitle
---------

Here is a list:

  - item 1
  - item 2

And a [link](http://www.google.com) as well.

Some code:

    #!/bin/bash
    echo OK
```



9. Tests - TDD

3 Testing methods
- 3.1 Static vs. dynamic testing
- 3.2 The box approach
  - 3.2.1 White-box testing
  - 3.2.2 Black-box testing
    - 3.2.2.1 Visual testing
  - 3.2.3 Grey-box testing

4 Testing levels
- 4.1 Unit testing
- 4.2 Integration testing
- 4.3 Component interface testing
- 4.4 System testing
- 4.5 Operational Acceptance testing

5 Testing types
- 5.1 Installation testing
- 5.2 Compatibility testing
- 5.3 Smoke and sanity testing
- 5.4 Regression testing
- 5.5 Acceptance testing
- 5.6 Alpha testing
- 5.7 Beta testing
- 5.8 Functional vs non-functional testing
- 5.9 Continuous testing
- 5.10 Destructive testing
- 5.11 Software performance testing
- 5.12 Usability testing
- 5.13 Accessibility testing
- 5.14 Security testing
- 5.15 Internationalization and localization
- 5.16 Development testing
- 5.17 A/B testing
- 5.18 Concurrent testing
- 5.19 Conformance testing or type testing

Information and Software Technology 56 (2014) 1219–1232

Contents lists available at ScienceDirect

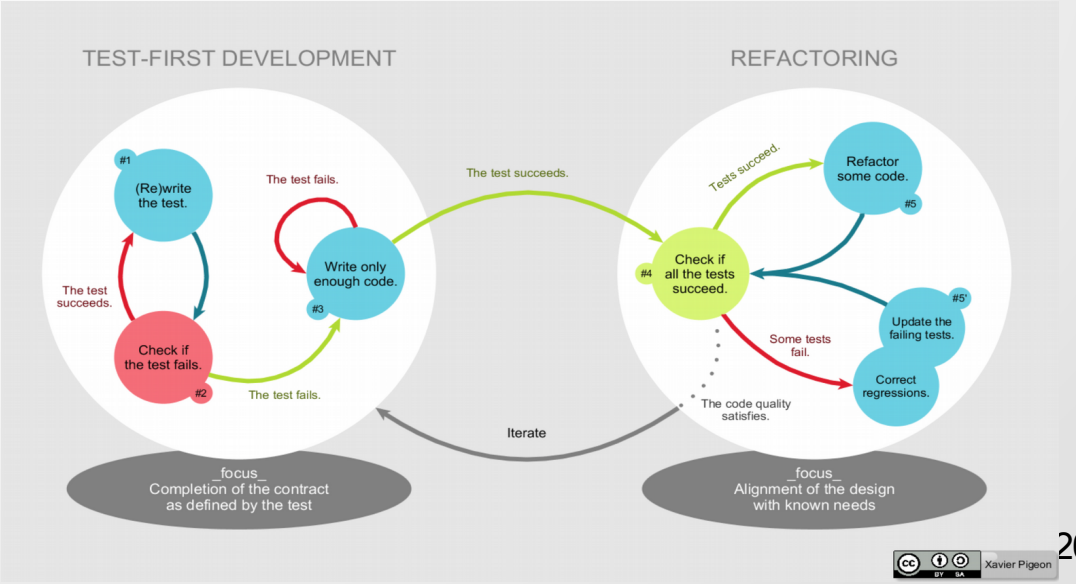
Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

Testing scientific software: A systematic literature review

Upulee Kanewala *, James M. Bieman

Computer Science Department, Colorado State University, USA



http://dx.doi.org/10.1016/j.infsof.2014.05.006

10. Licensing your code: Why?

- **Commercial reason** :
  - you want to make money out of it – forbid distribution
    – forbid reverse engineering
- **Scientific reason** :
  - you want to it to be used and get citations
    – you need to allow usage, and/or modification, etc.
    – you require others to cite your work
  - you want to protect yourself from liability claims

10. Licensing your code: How?

- Choose a license type, e.g.
  - Apache License 2.0
  - BSD 3-Clause "New" or "Revised" license
  - BSD 2-Clause "Simplified" or "FreeBSD" license
  - GNU General Public License (GPL)
  - GNU Library or "Lesser" General Public License (LGPL)
  - MIT license
  - Mozilla Public License 2.0
  - Common Development and Distribution License
  - Eclipse Public License
- Copy/adapt the text
- Distribute a LICENSE file with your code

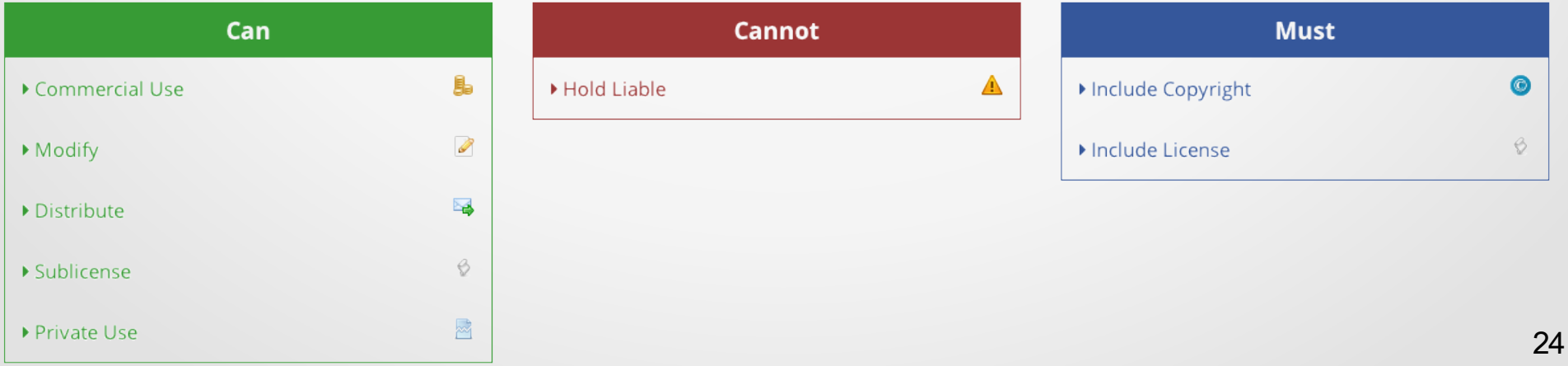
10. Licensing your code: MIT

Copyright (c) <year> <copyright holders>

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.



10. Licensing your code: BSD,GPL

BSD

Can	Cannot	Must
Commercial Use	Use Trademark	Include Copyright
Modify	Hold Liable	Include License
Distribute		
Place Warranty		

GPL

Can	Cannot	Must
Commercial Use	Sublicense	Include Original
Modify	Hold Liable	State Changes
Distribute		Disclose Source
Place Warranty		Include License
Use Patent Claims		Include Copyright
		Include Install Instructions

Ops' toolkit :

1. Virtualization platforms
2. Multi-host connexions
3. Configuration management
4. Installing
5. Automatic build tests
6. Monitoring

1. Virtualization

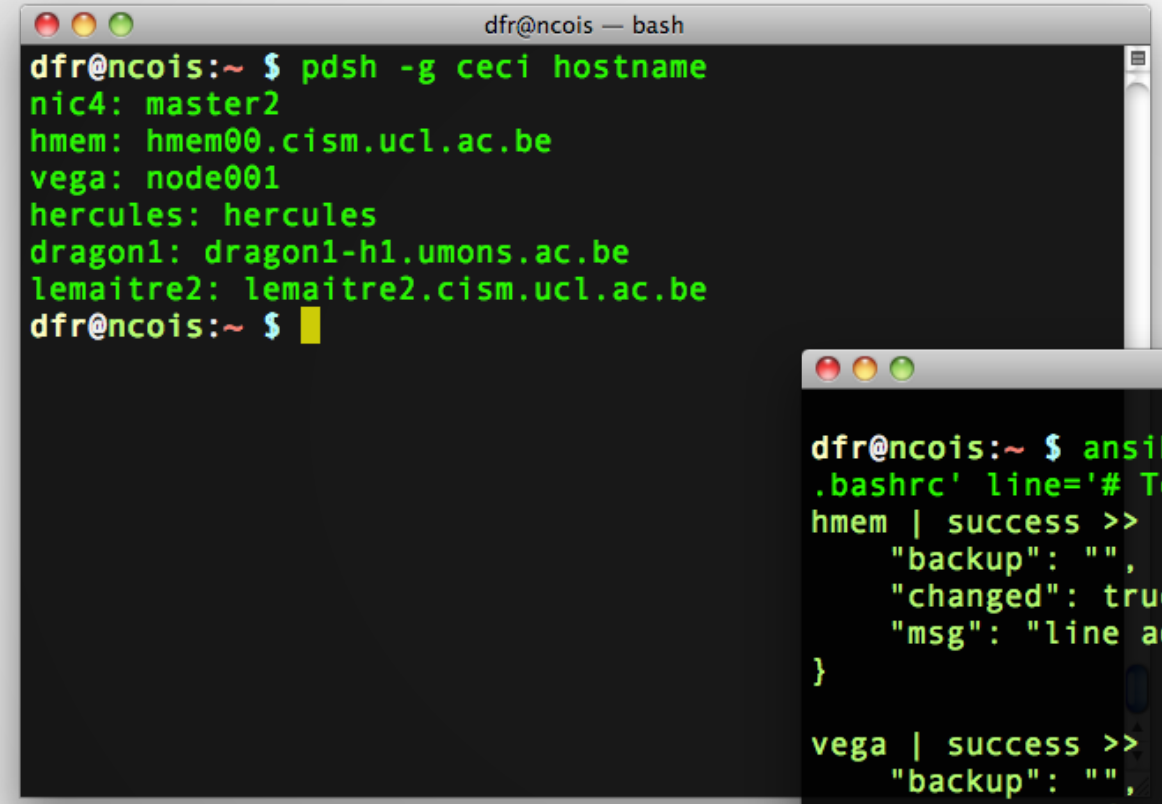
- Install on your laptop an environment similar to that of the cluster to test your workflow
- With
  - VirtualBox: https://www.virtualbox.org/
  - Vagrant: https://www.vagrantup.com/
- you can build a virtual cluster in one command:

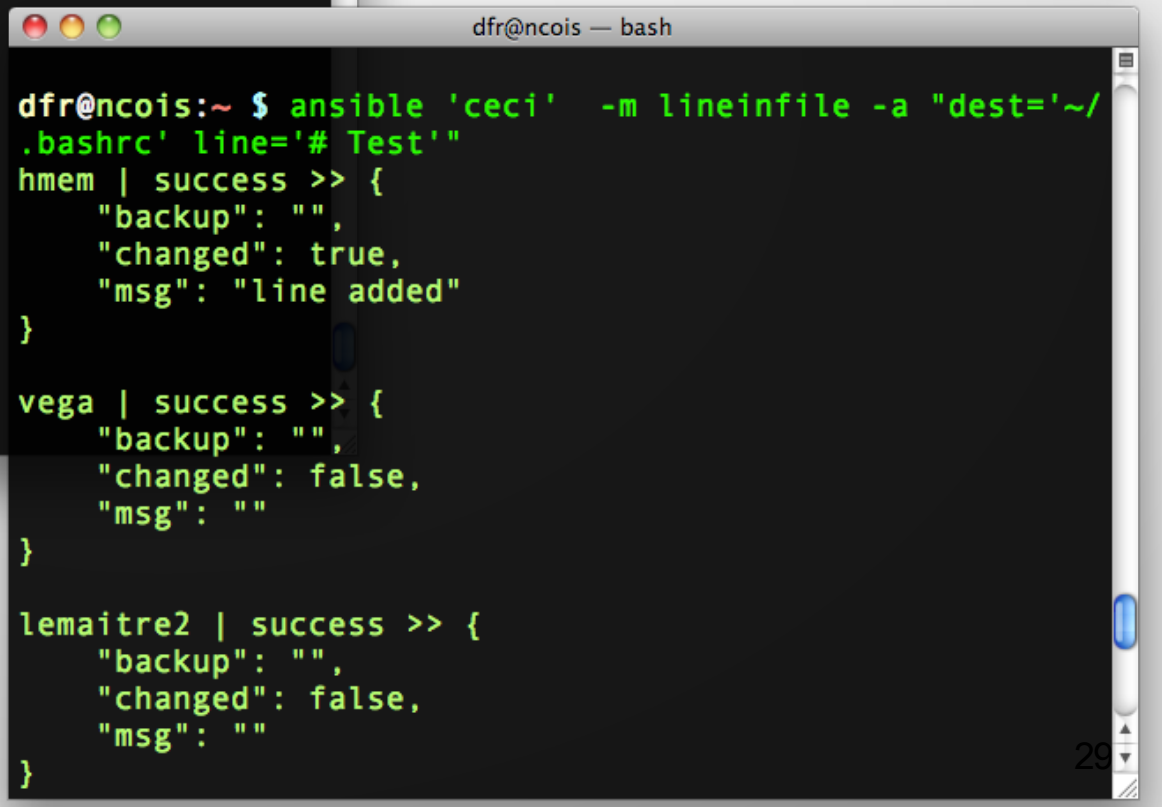
“vagrant up”

1. Virtualization


```
# -*- mode: ruby -*-
# vi: set ft=ruby :
VAGRANTFILE_API_VERSION = "2"

cluster = {
    "slave1"  => { :ip => "10.10.10.101", :cpus => 1, :mem => 512},
    "slave2"  => { :ip => "10.10.10.102", :cpus => 1, :mem => 512},
    "slave3"  => { :ip => "10.10.10.103", :cpus => 1, :mem => 512},
    "master" => { :ip => "10.10.10.10",   :cpus => 1, :mem => 1024},
}

Vagrant.configure(VAGRANTFILE_API_VERSION) do |config|

    config.vm.box = "bento/centos-6.7"

    cluster.each do |hostname, info|
        config.vm.define hostname do |cfg|
            cfg.vm.hostname = hostname
            cfg.vm.network :private_network, ip: "#{info[:ip]}", netmask: "255.255.255.0"

            cfg.vm.provider :virtualbox do |vb, override|
                vb.name = hostname
                vb.customize ["modifyvm", :id, "--memory", info[:mem]]
                vb.customize ["modifyvm", :id, "--cpus", info[:cpus]]
            end

            if hostname == 'master'
                config.vm.provision :ansible do |ansible|
                    ansible.limit = "all"
                    ansible.playbook = "bootstrap.yml"
                end
            end
        end
    end
end
```

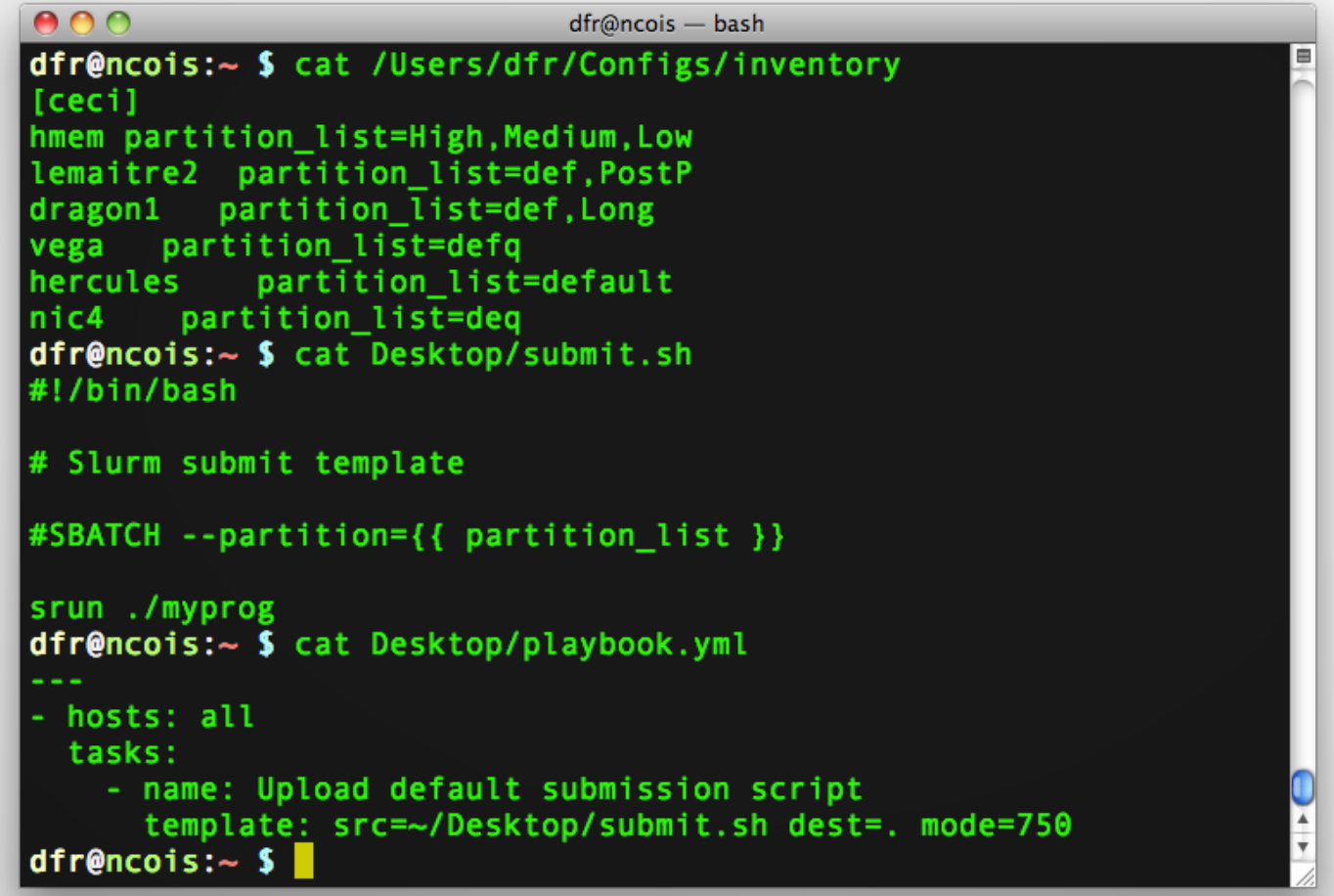
2. Multi-host SSH





https://clustershell.readthedocs.io https://code.google.com/p/pdsh/, http://www.ansible.com/

3. Configuration Management

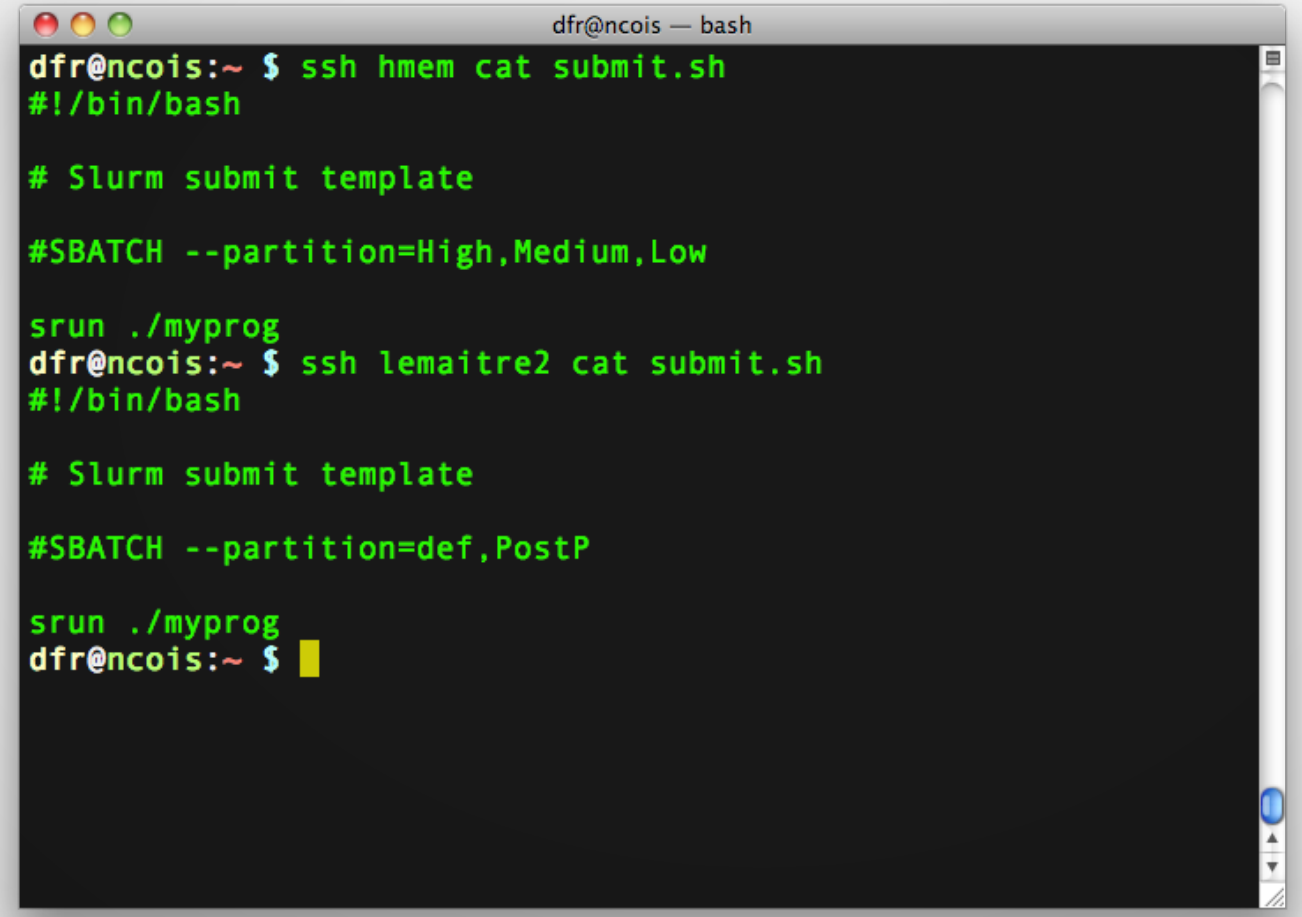


3. Configuration Management

dfr@ncois — bash

```
dfr@ncois:~ $ ansible-playbook Desktop/playbook.yml

PLAY [all] ********************************************************************

GATHERING FACTS ***************************************************************
ok: [hmem]
ok: [lemaitre2]
ok: [hercules]
ok: [vega]
ok: [dragon1]
ok: [nic4]

TASK: [Upload default submission script] **************************************
changed: [hmem]
changed: [lemaitre2]
changed: [vega]
ok: [hercules]
ok: [dragon1]
ok: [nic4]

PLAY RECAP ********************************************************************
dragon1                    : ok=2    changed=0    unreachable=0    failed=0
hercules                   : ok=2    changed=0    unreachable=0    failed=0
hmem                       : ok=2    changed=1    unreachable=0    failed=0
lemaitre2                  : ok=2    changed=1    unreachable=0    failed=0
nic4                       : ok=2    changed=0    unreachable=0    failed=0
vega                       : ok=2    changed=1    unreachable=0    failed=0

dfr@ncois:~ $
```

3. Configuration Management



4. Easy installing

EasyBuild @PyPi | EasyBuild docs | EasyBuild @GitHub

EasyBuild: building software with ease.

EasyBuild is a software build and installation framework that allows you to manage (scientific) software on High Performance Computing (HPC) systems in an efficient way.

Latest news

- *20150902* - **EasyBuild v2.3.0** is available
- *20150622* - **10th EasyBuild/Lmod hackathon** @ Austin (before SC15)
- *20150315* - **ISC'15 BoF "Getting Scientific Software Installed" accepted**
- *20141104* - **Revamped documentation @ easybuild.readthedocs.org**
- *20141020* - pre-print of **HUST-14 workshop paper** available

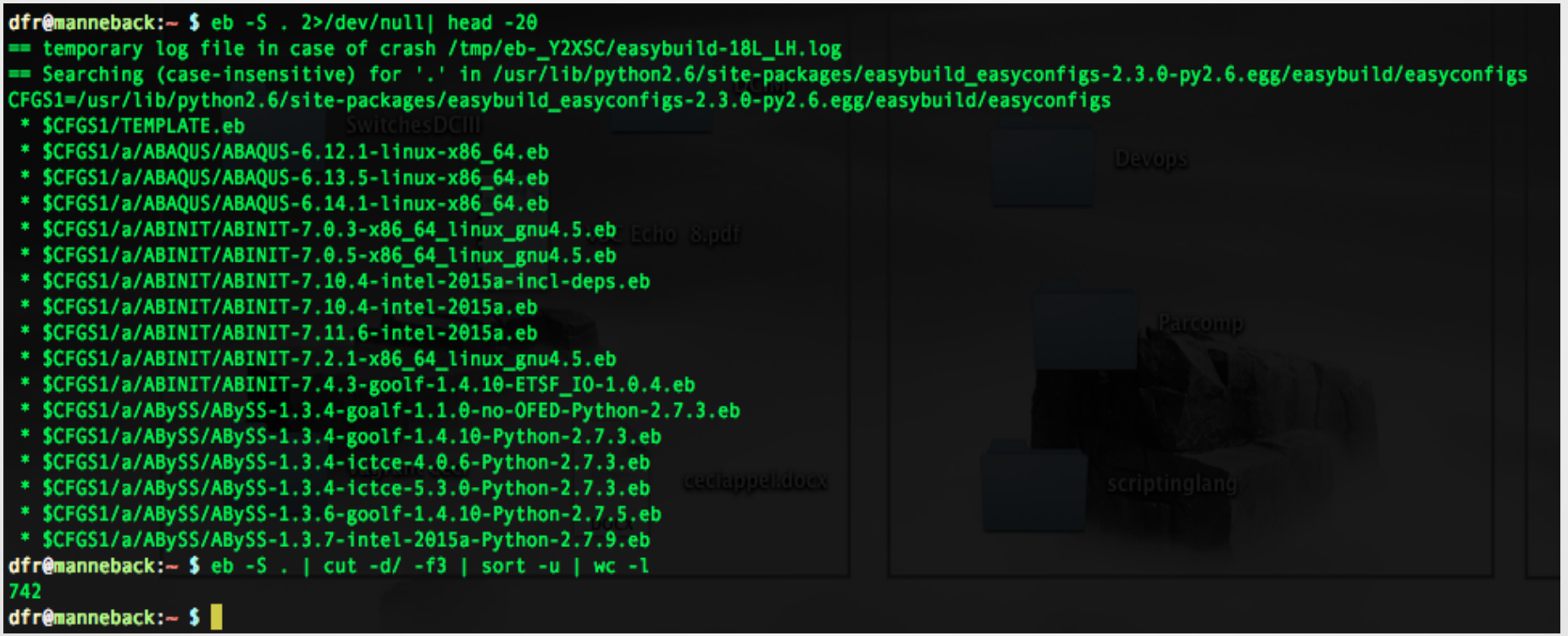
Documentation

Read the fine manual (RTFM!) at http://easybuild.readthedocs.org/.

Getting started

The recommended way of installing EasyBuild is via the documented bootstrap procedure. You should configure EasyBuild to behave as you prefer, subsequently.

4. Easy installing



4. Easy installing

CDE: Automatically create portable Linux applications

CDE (formerly known as CDEpack) automatically packages up the **C**ode, **D**ata, and **E**nvironment required to deploy and run your Linux programs on other machines without any installation or configuration. CDE is the easiest way to completely eliminate dependency hell.

To get started, download the CDE binary (32-bit or 64-bit) and follow these steps:

1. Package

Prepend any set of Linux commands with the "cde" binary, and CDE will run them and automatically package up all files (e.g., executables, libraries, plug-ins, config/data files) accessed during execution.

2. Deliver

A package is simply a directory that can be compressed and delivered to any x86-Linux machine. It contains all the files and environment variables required to run your original commands. Packages can range from 10 to 100 MB in size.

3. Run

After receiving the package, the user can now run those same commands from within the package on **any** modern x86-Linux distro. The user does not need to first compile, install, or configure anything.

http://www.pgbovine.net/cde.html

4. Easy installing

http://appimage.org

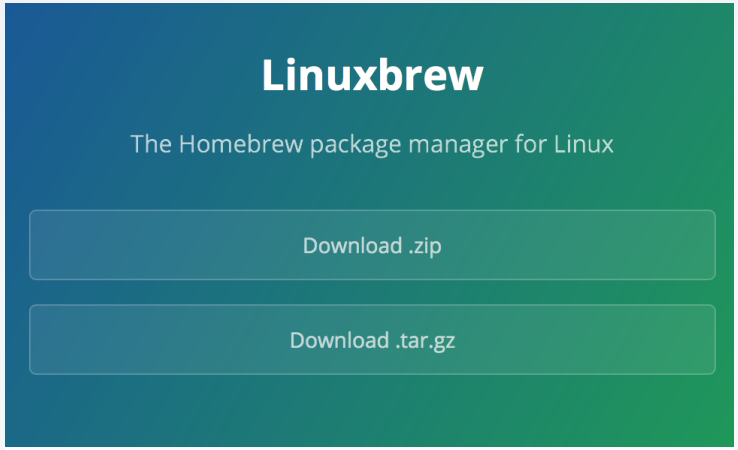
4. Easy installing

Shifter vs Charlie Cloud vs Docker vs Singularity

	Shifter	Charlie Cloud	Docker	Singularity
Privilege model	SUID	UserNS	Root Daemon	SUID/UserNS
Support current production Linux distros	Yes	No	No	Yes
Internal image build/boostrap	No*	No*	No**	Yes
No privileged or trusted daemons	Yes	Yes	No	Yes
No additional network configurations	Yes	Yes	No	Yes
No additional hardware	Maybe	Yes	Maybe	Yes
Access to host filesystem	Yes	Yes	Yes***	Yes
Native support for GPU	No	No	No	Yes
Native support for InfiniBand	Yes	Yes	No	Yes
Native support for MPI	Yes	Yes	No	Yes
Works with all schedulers	No	Yes	No	Yes
Designed for general scientific use cases	Yes	No	No	Yes
Contained environment has coorect perms	Yes	No	Yes	Yes
Containers are portable, unmodified by use	No	No	No	Yes
Trivial HPC install (one package, zero conf)	No	Yes	Yes	Yes
Admins can control and limit capabilities	Yes	No	No	Yes

* Relies on Docker
** Depends on upstream
*** With security implications

4. Easy installing



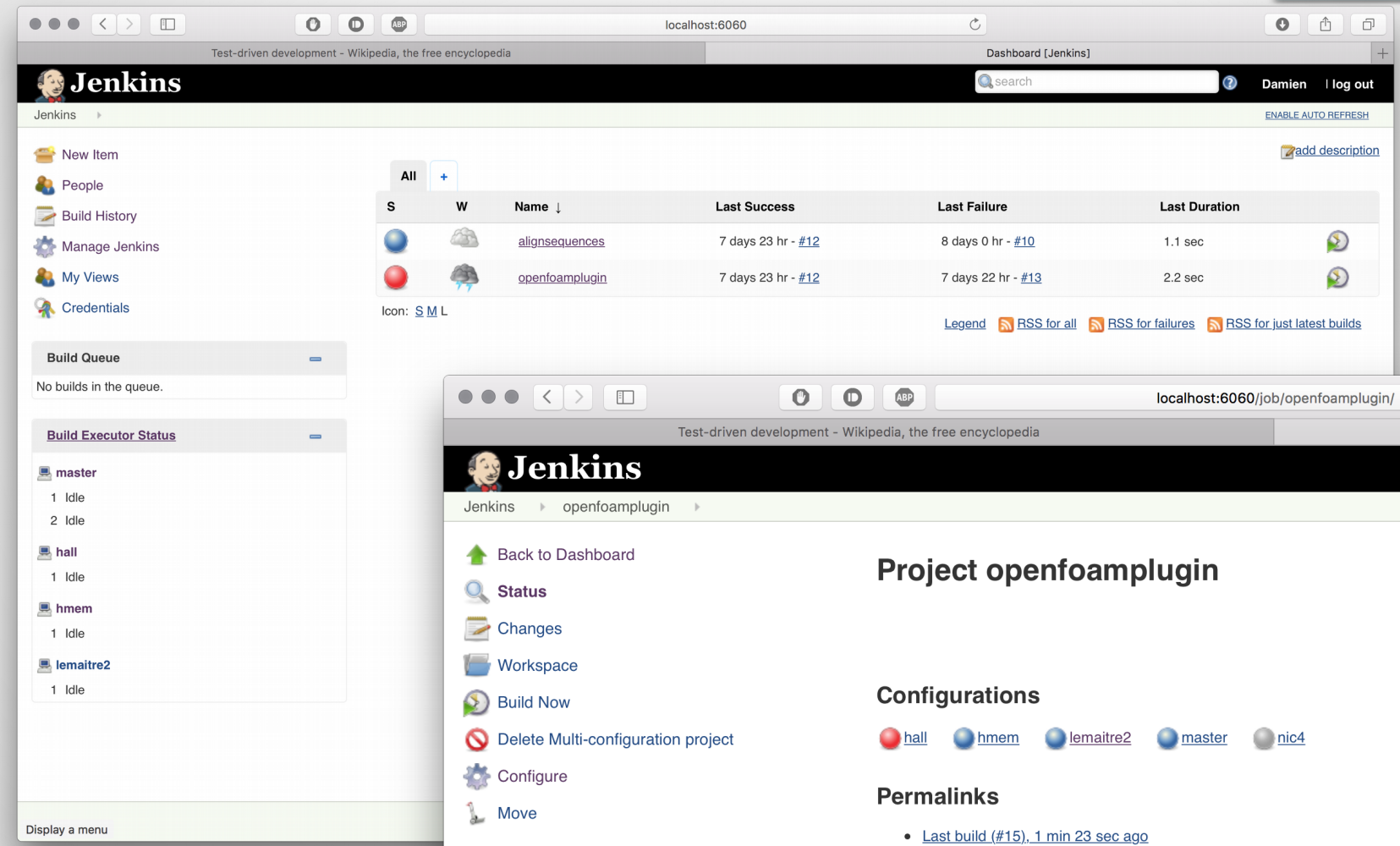
Linuxbrew

tag v1.2.2

Linuxbrew is a fork of Homebrew, the macOS package manager, for Linux.

http://linuxbrew.sh

5. Automatic build tests



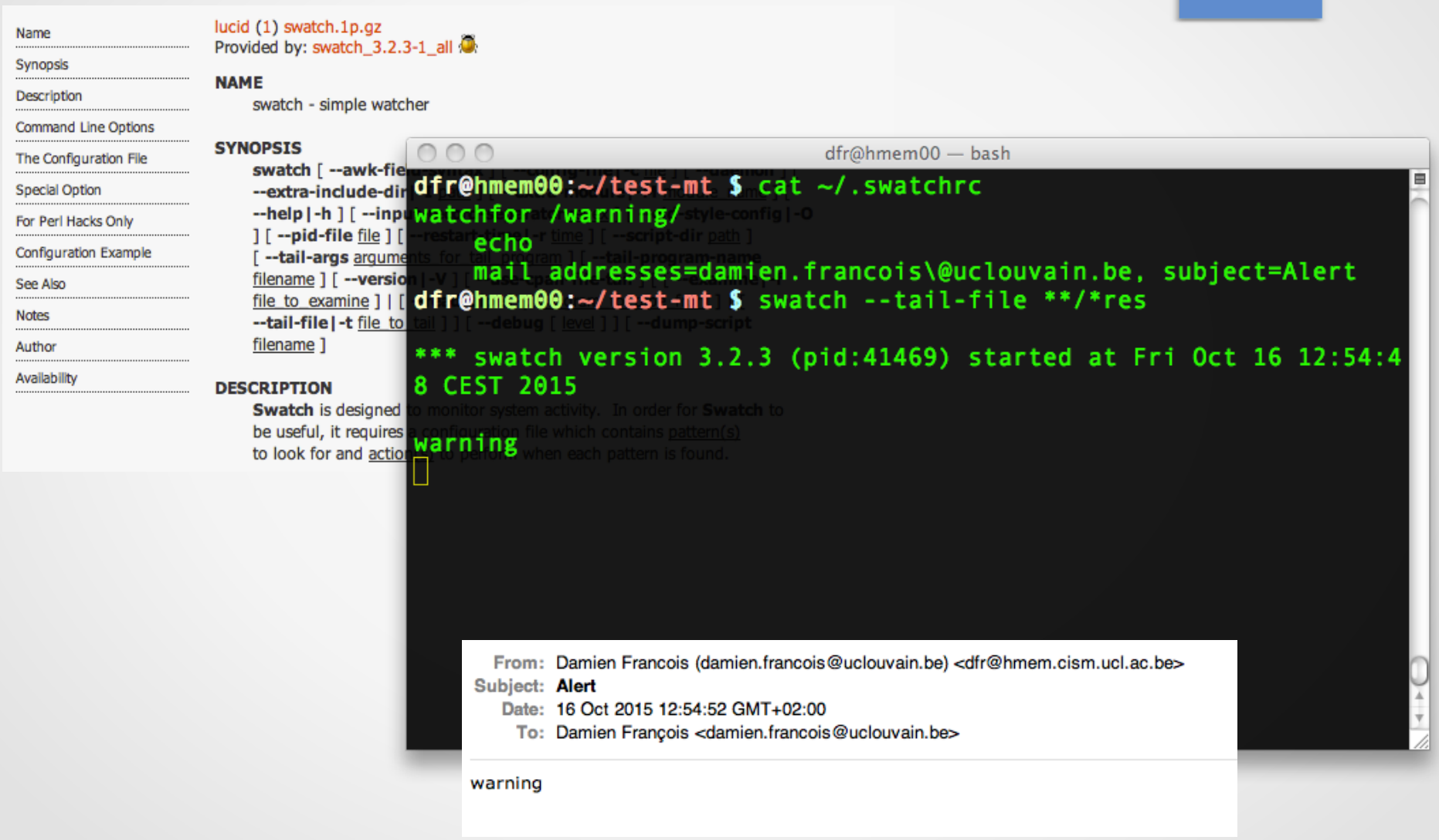
6. Monitoring

MultiTail

multitail -q 2 "/res"**

Useful with command 'screen'

6. Monitoring



Useful with command 'screen'

Dev's toolkit :

1. Programming language
2. Good practices / Code Style Guides
3. Text editor / IDE
4. Source control management
5. Debuggers / Profilers
6. Databases
7. Packaging / Distributing tools
8. Comments and documentation
9. Tests
10. Licensing

Ops' toolkit :

1. Virtualization platforms (Virtual box, Vagrant)
2. Multi-host connexions (pdsh, clustershell)
3. Configuration management/ (ansible)
4. Installing (easybuild)
5. Automatic build tests (jenkins)
6. Monitoring (multitail)

The 'Phillip' test

- 12 simple questions
- ordered by 'difficulty'
- measures quality of organization
- for research programming

If you do not score at least a 7 there is room for improvement using the tools presented here

1. Do you have reliable ways of taking, organizing, and reflecting on notes as you're working?
2. Do you have reliable to-do lists for your projects?
3. Do you write scripts to automate repetitive tasks?
4. Are your scripts, data sets, and notes backed up on another computer?
5. Can you quickly identify errors and inconsistencies in your raw data sets?
6. Can you write scripts to acquire and merge together data from different sources and in different formats?
7. Do you use version control for your scripts?
8. If you show analysis results to a colleague and they offer a suggestion for improvement, can you adjust your script, re-run it, and produce updated results within an hour?
9. Do you use `assert` statements and test cases to sanity check the outputs of your analyses?
10. Can you re-generate any intermediate data set from the original raw data by running a series of scripts?
11. Can you re-generate all of the figures and tables in your research paper by running a single command?
12. If you got hit by a bus, can one of your lab-mates resume your research where you left off with less than a week of delay?

Work quicker & more reliably

C.E.C.I