The Life of a SysAdmin

in a Research Environment

Eric G. Burgueño  |  http://ebur.co
Reproducible Research
The Role

➤ Science > Money
➤ Makefiles <3 Visio
➤ Unus pro omnibus...
The Environment

Some very big servers

Running specialised software

Heavily customised
Some very BIG servers...
...and some small ones

```latex
$ lshosts

<table>
<thead>
<tr>
<th>HOST_NAME</th>
<th>ncpus</th>
<th>maxmem</th>
<th>maxswp</th>
</tr>
</thead>
<tbody>
<tr>
<td>host01</td>
<td>4</td>
<td>15943M</td>
<td>8015M</td>
</tr>
<tr>
<td>host02</td>
<td>8</td>
<td>32102M</td>
<td>8015M</td>
</tr>
<tr>
<td>host03</td>
<td>4</td>
<td>15943M</td>
<td>8015M</td>
</tr>
<tr>
<td>host04</td>
<td>16</td>
<td>84617M</td>
<td>32767M</td>
</tr>
<tr>
<td>host05</td>
<td>8</td>
<td>32102M</td>
<td>8015M</td>
</tr>
<tr>
<td>host06</td>
<td>80</td>
<td>2068067M</td>
<td>10239M</td>
</tr>
<tr>
<td>host07</td>
<td>80</td>
<td>2068067M</td>
<td>10239M</td>
</tr>
<tr>
<td>host08</td>
<td>16</td>
<td>84617M</td>
<td>32767M</td>
</tr>
<tr>
<td>host09</td>
<td>8</td>
<td>258326M</td>
<td>8015M</td>
</tr>
<tr>
<td>host10</td>
<td>40</td>
<td>258269M</td>
<td>10239M</td>
</tr>
<tr>
<td>host11</td>
<td>32</td>
<td>129013M</td>
<td>2047M</td>
</tr>
<tr>
<td>host12</td>
<td>32</td>
<td>129013M</td>
<td>2047M</td>
</tr>
<tr>
<td>hostnn</td>
<td>32</td>
<td>129013M</td>
<td>2047M</td>
</tr>
</tbody>
</table>

```
Specialised Software

Genome/Transcriptome assembly
  Phylogenetic analysis
RNA analysis tools
  Protein annotation
Search pattern, motif, profiles domains, families
  Data visualisation
  Other tools
Find transcriptional elements
  Organism classification
Align sequences
  DNA Sequence analysis tools
    Detection of repeats, restriction sites, and patterns
    Gene finding
    Sequence alignment
    Find tRNA
Expression analysis
  Protein function assignment
    Search, computation & analysis of pathways
Protein-protein interactions
Protein sequence analysis
  Multiple sequence alignment
Customised Environment

[eric@host01]$ lsid
openlava project 2.2, Oct 19 2015

My cluster name is genomecluster
My master name is host01

[eric@host01 myapp-5.15-54.0]$ grep temp launcher.sh
TMP="$(dirname $0)/temp"

[eric@host01 myapp-5.15-54.0]$ pwd
/apps/myapp-5.15-54.0

[eric@host01 myapp-5.15-54.0]$ ls temp
lrwxrwxrwx. 1 nonpriv nonpriv 35 Jan  5 12:38 temp ->
/nfs/appscratch/myapp/temp
The Struggles

➤ Some very big servers
➤ Running specialised software
➤ Heavily customised
Some very BIG servers...
Some very BIG servers...

<table>
<thead>
<tr>
<th>Memory Type</th>
<th>last</th>
<th>min</th>
<th>avg</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free memory</td>
<td>1.3 TB</td>
<td>486.14 MB</td>
<td>310.35 GB</td>
<td>1.75 TB</td>
</tr>
<tr>
<td>Cached memory</td>
<td>671.52 GB</td>
<td>15.25 MB</td>
<td>545.23 GB</td>
<td>1.94 TB</td>
</tr>
<tr>
<td>Buffers memory</td>
<td>404.71 MB</td>
<td>1.93 MB</td>
<td>319.92 MB</td>
<td>615.35 MB</td>
</tr>
<tr>
<td>Real used memory</td>
<td>17.86 GB</td>
<td>13.33 GB</td>
<td>1.14 TB</td>
<td>1.97 TB</td>
</tr>
</tbody>
</table>
Some very BIG servers...

“With the expected growth in the size of data sets, moving to distributed-memory parallel computing seems inevitable.”

Bill - July 12, 2015 at 3:33 pm

http://bit.ly/1WCqSlQ
Specialised Software

$ ls /software/
454
AATpackage
AATpackage-r03052011
AdapterRemoval-1.4
AdapterRemoval-1.5.4
allpathslg
allpathslg-44837
allpathslg-44913
allpathslg-49722
allpathslg-50191
amos
AMOS
amos-3.0.1
amos-3.1.0
AMOS-Short
Anchor-0.3.1
antismash
anytag-2.5.2
arachne
arachne-39099
aragorn
cgal-0.9.6-beta
CGAT-0.2.4
circos
circos-0.62-1
clustal-omega-1.1.0
clustal-omega-1.2.1
clustalw-2.1
cndsrc
cndsrc-
correction
cortex_con
cortex_var
crac-1.2.3
cufflinks
cufflinks-1.3.0

cufflinks-1.3.0.Linux_x86_64
cufflinks-2.0.2.Linux_x86_64
cufflinks-2.1.1.Linux_x86_64
geneid_v1.4.4.Jan_13_2011
genomemapper0.4.3
genometools
genometools-1.3.3
genometools-1.5.6
HarrismerHMM
gmap
gmap-2014-07-04
gmap-2014-10-22
gmap-2014-12-29
gmap-2015-09-29
gmato
gmatoV1.2Build20130106
gmes
gmorse
gmorse_02mar2011
gmp-6.0.0
hagfish

Dependency Hell
Customised Environment

How Standards Proliferate:
(See: A/C chargers, character encodings, instant messaging, etc)

Situation: There are 14 competing standards.

14?! Ridiculous! We need to develop one universal standard that covers everyone's use cases. Yeah!

Soon:

Situation: There are 15 competing standards.
[eric@host01]$ aws-shell || gcloud
-bash: aws-shell: command not found
-bash: gcloud: command not found

[eric@host01]$ ./corkscrew ssh login01@eu.collaborator.edu
‘bhosts || qstat -a || qhost’
-bash: bhosts: command not found
-bash: qstat: command not found
-bash: qhost: command not found

[eric@host01]$ :(
Reproducible?
Research
The Solution?

“Put it in a container!

Dear Google: Please give us sparkles in Google Slides.
Thank you

...any questions?