Farms, Fabrics and Clouds

Steve Loughran
Julio Guijarro
HP Laboratories, Bristol, UK

December 2007

steve.loughran@hpl.hp.com
julio.guijarro@hpl.hp.com
Julio Guijarro

Researcher at HP Laboratories

Area of interest: Deployment

In charge of OSS release

http://smartfrog.org/
Steve Loughran

Researcher at HP Laboratories

Area of interest: Deployment

Author of *Ant in Action*
Our research

- How to host big applications across distributed resources
  - Automatically
  - Repeatably
  - Dynamically
  - Correctly
  - Securely
- How to manage them from installation to removal
- How to make dynamically allocated servers useful
Who had breakfast this morning?
Question

Who harvested wheat or corn, or killed an animal for that breakfast?
Farms provide food.

It is somebody else's problem
Who is wearing clothes they wove or knitted themselves?
Provisioning of clothing -fabrics- is outsourced

It is somebody else's problem
All new applications are on the Web

• Web Browser, AJAX clients
• Richer: Flash, XUL, Silverlight
• "… as a Service"

⇒ Lots of code running in the server
⇒ Data mining/analysis problems
⇒ Unpredictable demand
Old world installation: single server

Single web server,
Single DB
RAID filestore

-SPOF
-limintations of scale
yesterday: clustering

Multiple web servers, Replicated DB
RAID Network filestore
Load-balancing router

- Cost
- Complexity
- Limitations of scale

Maintains the illusion of a single server
Now: server farms

500 web servers,
Distributed filestore
Rented storage & CPU

Scales up
No capital outlay

*Agile infrastructure*
UK Grid Status at 03 Dec 2007 23:21:05

<table>
<thead>
<tr>
<th>Institute</th>
<th>CPU Tot</th>
<th>CPU Free</th>
<th>Jobs Cur</th>
<th>Jobs Wait</th>
<th>Disk Tot</th>
<th>Disk Free</th>
<th>CE</th>
<th>SE</th>
<th>SRM</th>
<th>24 Hrs</th>
<th>Week</th>
<th>ATLAS CMS</th>
<th>LHCb</th>
<th>CE</th>
<th>Release</th>
<th>Replica</th>
<th>HW</th>
<th>NP</th>
<th>UA</th>
<th>24 Hrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brunel</td>
<td>396</td>
<td>117</td>
<td>200</td>
<td>121</td>
<td>17.0</td>
<td>0.0</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>100%</td>
<td>99%</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.020</td>
</tr>
<tr>
<td>Imperial HEP</td>
<td>402</td>
<td>386</td>
<td>37</td>
<td>0</td>
<td>47.9</td>
<td>11.3</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>100%</td>
<td>99%</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>Imperial LeSc</td>
<td>200</td>
<td>148</td>
<td>273</td>
<td>3</td>
<td>0.0</td>
<td>0.0</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>100%</td>
<td>100%</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>QMUL</td>
<td>1054</td>
<td>474</td>
<td>406</td>
<td>0</td>
<td>17.3</td>
<td>10.8</td>
<td>W</td>
<td>P</td>
<td>P</td>
<td>50%</td>
<td>54%</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>RHUL</td>
<td>144</td>
<td>1</td>
<td>58</td>
<td>23</td>
<td>8.1</td>
<td>3.3</td>
<td>W</td>
<td>E</td>
<td>F</td>
<td>65%</td>
<td>73%</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>UCL CCC</td>
<td>252</td>
<td>123</td>
<td>60</td>
<td>0</td>
<td>20.4</td>
<td>17.4</td>
<td>M</td>
<td>F</td>
<td>F</td>
<td>10%</td>
<td>0%</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>UCL HEP</td>
<td>102</td>
<td>50</td>
<td>10</td>
<td>159</td>
<td>1.0</td>
<td>0.7</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>100%</td>
<td>80%</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>Lancaster</td>
<td>376</td>
<td>177</td>
<td>199</td>
<td>0</td>
<td>71.4</td>
<td>49.4</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>85%</td>
<td>54%</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>Liverpool</td>
<td>472</td>
<td>364</td>
<td>107</td>
<td>0</td>
<td>12.0</td>
<td>10.9</td>
<td>W</td>
<td>P</td>
<td>P</td>
<td>100%</td>
<td>100%</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>Manchester</td>
<td>1740</td>
<td>1298</td>
<td>412</td>
<td>0</td>
<td>1953.1</td>
<td>0.0</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>100%</td>
<td>100%</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>Sheffield</td>
<td>159</td>
<td>93</td>
<td>66</td>
<td>0</td>
<td>2.4</td>
<td>2.1</td>
<td>W</td>
<td>P</td>
<td>P</td>
<td>100%</td>
<td>100%</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>Durham</td>
<td>104</td>
<td>13</td>
<td>87</td>
<td>0</td>
<td>17.1</td>
<td>14.5</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>100%</td>
<td>100%</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>Edinburgh</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>23</td>
<td>29.0</td>
<td>21.6</td>
<td>W</td>
<td>P</td>
<td>P</td>
<td>0%</td>
<td>85%</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>Glasgow</td>
<td>636</td>
<td>406</td>
<td>130</td>
<td>0</td>
<td>82.2</td>
<td>68.9</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>100%</td>
<td>100%</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>Birmingham</td>
<td>16</td>
<td>4</td>
<td>14</td>
<td>110</td>
<td>10.2</td>
<td>8.5</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>100%</td>
<td>100%</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>Bristol</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>10.2</td>
<td>6.7</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>100%</td>
<td>100%</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>Cambridge</td>
<td>138</td>
<td>125</td>
<td>9</td>
<td>20</td>
<td>19.4</td>
<td>8.3</td>
<td>E</td>
<td>F</td>
<td>F</td>
<td>10%</td>
<td>70%</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.020</td>
</tr>
<tr>
<td>Oxford</td>
<td>72</td>
<td>71</td>
<td>53</td>
<td>11.8</td>
<td>11.8</td>
<td>8.8</td>
<td>P</td>
<td>E</td>
<td>E</td>
<td>91%</td>
<td>90%</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>RAL PPD</td>
<td>320</td>
<td>311</td>
<td>11</td>
<td>0</td>
<td>42.0</td>
<td>21.1</td>
<td>D</td>
<td>P</td>
<td>P</td>
<td>49%</td>
<td>95%</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>RAL Tier-1</td>
<td>382</td>
<td>170</td>
<td>212</td>
<td>0</td>
<td>322.2</td>
<td>212.7</td>
<td>D</td>
<td>E</td>
<td>P</td>
<td>84%</td>
<td>94%</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
<tr>
<td>Overall</td>
<td>6940</td>
<td>4274</td>
<td>2446</td>
<td>4956</td>
<td>2080.4</td>
<td>477.8</td>
<td></td>
<td></td>
<td></td>
<td>76%</td>
<td>84%</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.030</td>
</tr>
</tbody>
</table>
Application architectures and deployment problems change radically in this world.
Application architectures

- ROA/REST
- Virtualized
- Map/Reduce
- Shards
- Tuple-spaces
- Grid
Why?

- Save on hardware (and power, space)
- Dynamically move running servers
- Demand creation of new images
- Testing complex system configurations
- Redistributing entire machine image
- 'virtual appliance'
Assumptions that are now invalid

- Systems have a long lifespan
- It is slow/expensive to create a new system
- It is expensive to duplicate one
- Systems can/should be managed by hand
- Clocks proceed at the same rate
- Physical RAM doesn’t get swapped out
- Running machines can't be moved/cloned
Server Farms
Assumptions that are now invalid

- System failure is an unusual event
- 100% availability can be achieved
- Data is always near the server
- You need physical access to the servers
- Databases are the best storage form
- You need millions of $/£/€ to play
Who has the servers?

- Yahoo, Google, MSN, eBay: services
- MMORPG Game Vendors: Word of Warcraft, Second Life
- EU Grid: Scientists
- HP, IBM, Sun: rent to companies - focus on CPU performance
- Amazon: rent to anyone with an Amazon account - focus on startups
Amazon EC2

Public Internet

$\uparrow$

Host

AMI (Xen VM)

/mnt

AMI (Xen VM)

/mnt

$\downarrow$

Fast (free) network

free access; slow initial read time

S3 Storage

pay per GET; per megabyte

$\uparrow$

$\downarrow$

$\downarrow$

$\downarrow$

$\downarrow$

$\downarrow$

$\downarrow$
Amazon EC2

- Pay as you go Virtual Machine Hosting
- No persistent storage other than S3 filestore - uses HTTP GET/PUT/DELETE operations
- $0.10 per CPU/hour
- S3 Storage has own billing (by MB & by access - cheaper in bulk)
Map/Reduce

Run code near the data, then merge the results
Assumptions that are now invalid

• Terabyte datasets are hard to work with
• Code runs on a single machine
• Sequential code is better than parallel code
• RAID hardware is the best way to store data
• Databases are better than filesystems
Assumptions that are now invalid

- A single farm needs to scale to infinity
- You need to provide 100% availability to 100% of users
- You have to roll out simultaneous updates to the application, changes to the DB schema, \textit{globally}
Changes for developers

• Many classic assumptions are invalid

• Design for scale

• Rent servers from the outset —every developer can have their own set

• Cover your server costs from the outset and you are in the black from day 1
Problems for us farmers

- Power management
- Predictive disk failure management
- Load balancing for availability, power
- Data cache management
- Billing
- Security/Isolation
- How this will change server hardware
- Managing/Configuring Machine Images
- Diagnostics when things go wrong
Topic for discussion

Where is all this heading?