Towards a Green Grid'5000

Anne-Cécile Orgerie, Laurent Lefèvre
INRIA RESO – Université de Lyon
annececile.orgerie@ens-lyon.fr, laurent.lefevre@inria.fr

April 8, 2009, Grid'5000 Spring School, Nancy - France
Green-Net Project

- Power aware software frameworks for high performance data transport and computing in large scale distributed systems
- ARC GREEN-NET : (Action de Recherche Coopérative supported by INRIA)
- Partners teams :
  - IRIT (Toulouse)
  - INRIA MESCAL (Grenoble)
  - INRIA RESO (Lyon)
  - Virginia Tech (USA)

http://www.ens-lyon.fr/LIP/RESO/Projects/GREEN-NET
Plan

1. Introduction
2. Two years in the life of Grid'5000
3. The consumption measurement infrastructure
4. EARI: Energy-Aware Reservation Infrastructure
5. Greening Grid'5000
6. Conclusion and future works
Adressed challenges

- How to reduce energy usage without compromising QoE: Quality of Experiment?
- How to understand and to analyze the usage of large scale platforms?
- How to apply energy usage models on this experimental usage?
- How to monitor lively such usage (multiple views (Grids, datacenters, clusters, nodes, services, processes, threads))?
- How to design energy aware software frameworks?

Our context: Ressources always powered on / Reservation infrastructure / Large-scale distributed systems
Two years in the life of Grid'5000
Grid'5000 usage

- Two years of logs: 2007 & 2008
- Logs furnished by oarstat
- 8 sites
- Grid view, site view and node view
- 2 INRIA reports to appear
Grid statistics for 2007

job = reservation; resource = core
'real' activity = without dead and absent time

<table>
<thead>
<tr>
<th>Site</th>
<th>Number of jobs</th>
<th>Number of resources</th>
<th>Mean number of resources per job</th>
<th>Mean duration of a job in seconds</th>
<th>Percentage of ‘real’ activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bordeaux</td>
<td>45775</td>
<td>650</td>
<td>55.50</td>
<td>5224.59</td>
<td>47.40%</td>
</tr>
<tr>
<td>Grenoble</td>
<td>19211</td>
<td>72</td>
<td>4.06</td>
<td>4473.76</td>
<td>14.25%</td>
</tr>
<tr>
<td>Lille</td>
<td>330694</td>
<td>250</td>
<td>4.81</td>
<td>1446.13</td>
<td>36.08%</td>
</tr>
<tr>
<td>Lyon</td>
<td>33315</td>
<td>322</td>
<td>41.64</td>
<td>3246.15</td>
<td>45.92%</td>
</tr>
<tr>
<td>Nancy</td>
<td>63435</td>
<td>574</td>
<td>22.46</td>
<td>19480.49</td>
<td>56.21%</td>
</tr>
<tr>
<td>Orsay</td>
<td>26448</td>
<td>684</td>
<td>47.45</td>
<td>4322.54</td>
<td>18.69%</td>
</tr>
<tr>
<td>Rennes</td>
<td>36433</td>
<td>714</td>
<td>54.85</td>
<td>7973.39</td>
<td>49.42%</td>
</tr>
<tr>
<td>Sophia</td>
<td>35179</td>
<td>568</td>
<td>57.93</td>
<td>4890.28</td>
<td>51.36%</td>
</tr>
<tr>
<td>Toulouse</td>
<td>20832</td>
<td>434</td>
<td>12.89</td>
<td>7420.07</td>
<td>49.99%</td>
</tr>
</tbody>
</table>
Usage evolution

46% of activity

69% of activity
Nancy in 2008: big burst / impact of best effort jobs

Distribution in time of the different resource’s states per week for Nancy

- Jobs
- Idle
- Work
- Absent
- Suspected
- Dead

574 resources

Percentage of time

Number of jobs

Weeks
Nancy: max, min and median resources
Toulouse in 2008: 15% of absent
Conclusions about the usage

- Specific usage of an experimental Grid
- Great differences between 2007 and 2008
- Significative bursts $\rightarrow$ significative gaps
- A lot of small reservations
The consumption measurement infrastructure deployed on Grid'5000
Consumption measurement

- Autonomic wattmeter
- Furnished by Omegawatt
- One measure per second and per node
- 6 nodes at a time
- 3 different sites: Lyon, Grenoble, and Toulouse
Example: Monitoring of 6 nodes in Lyon

- Idle consumption really high: around 190 Watts
- Off consumption around 10 Watts

Idea: to switch off the unused nodes
Live energy monitoring

Energy consumption of sagittaire-30

Energy consumption of capricorne-30

Energy consumption of sagittaire-30
EARI: Energy-Aware Reservation Infrastructure
Architecture for energy efficient system
When can we switch off a node?

Consumption of the resource if it is switched off and on

Consumption of the resource if it stays idle
Role of $T_s$

$$T_s = \frac{E_s - P_{OFF}(\delta_{ON\rightarrow OFF} + \delta_{OFF\rightarrow ON}) + E_{ON\rightarrow OFF} + E_{OFF\rightarrow ON}}{P_I - P_{OFF}} + T_r$$
Submission of a reservation

Reservation submission:
\[ R = (l, n_0, t_0) \]

- \( n_0 < (N - n) \):
  - no: Compute earliest possible start time = \( t_1 \)
  - yes:
    - Compute energy costs of:
      - reservation at \( t_0 \) or \( t_1 \)
      - reservation at \( t_{\text{end}} \)
      - reservation at \( t_{\text{start}} \)
      - reservation at \( t_{\text{slack}} \)
    - Give the consumptions and the start times to the user
    - user's choice between the 4 proposals
    - Validate the reservation
    - Update the agenda
End of a reservation

Reservation end: R' frees M nodes

- Compute the energy used by R'
- Give this information to the user

per resource:
- Does the resource have an imminent reservation?

yes
- Considered as busy
- It stays awake

no
- Estimate next reservation R = (i, n, t_0)
- Look if there are awake resources
- For each of the m awake resources:
  - look if it can be used for R
  - Keep min(m, n) resources awake if
  - R is imminent
- Turn off the remaining resources

At next submission:
- compute the error
- turn off idle resources
Predictions for greater energy savings

We should predict:

- The next reservation (length, size, start time)
- The next slack period
- The consumption of a given reservation

Solutions:

- Recent history (last reservations + feedback)
- History of the days before + feedback
- History of the user + type of resource + number of resources to switch on/off.
User behavior modelling

- **user**: wished date (or the nearest which is possible);

- **fully green**: solution that costs the less in terms of energy (the one where we switch on/off the smaller number of resources);

- **green-percentage-25**: 25% of fully green taken at random and user for the other ones;

- **green-percentage-50**: 50% of fully green and user for the other ones;

- **green-percentage-75**: 75% of fully green and user for the other ones;

- **deadlined**: fully green if it does not delay the reservation for more than 24 hours, otherwise user.
Evaluation of EARI

Replay on 4 different traces of Grid'5000 in 2007

100% = present consumption (all nodes always on)

all glued = theoretical lower bound
Greening Grid'5000
Demo
Conclusion and future works
Conclusions

Our Grid'5000 experience:

- Logs collecting with *oarstat*
- Usage over 2007 and 2008
- Consumption monitoring of 18 nodes
- Replay of the logs
- Modelling user's behavior
Soon on your G5K screens

- 'Green' advices and reservations (with Dynamic Voltage Scaling)
- Availability of the energy logs
- The energy profile per job available on the web
- Full Energy Monitoring of the complete Lyon site (135 nodes)
Thank you for your attention!

Questions?