Adaptive Resource and Job Management for limited power consumption
Introduction

DVFS & Switch-off

The model

Algorithm and implementation

Experimentations

Conclusion and future works
- Introduction
- DVFS & Switch-off
- The model
- Algorithm and implementation
- Experimentations
- Conclusion and future works
Powercap: limit the power consumption during a certain amount of time
Introduction - Energy

• Why control?
  - Power peak = O(power of a city)
  - Power installations lifetime
  - Electricity providers limitations
  - Controling energy consumption = Controling cost

• How control?
  - DVFS
  - Switch-off
    • (or shutdown, or sleep mode, or hibernation...)
Introduction

DVFS & Switch-off

The model

Algorithm and implementation

Experimentations

Conclusion and future works
Switch-off

- Switch-off some resources
- switched-off has a cost
- Not possible on all clusters
- Jobs can not run on switched-off nodes!
The RJMS level – Switch-off

- **« Power Bonuses »**
  - If all components of a level are switched-off, the component of the upper level can be switched-off and provide an additional gain

- Exemples :
  - Nodes are made of processors
  - Chassis are made of nodes
  - Rack are made of Chassis
The RJMS level – Switch-off

- « Power Bonuses » on CURIE cluster:
  - 18 nodes per chassis, 5 chassis per rack
  - Power gained by switching off a **Chassis**
    \[ \sim = \text{Power(} \text{computing node} \text{)} \]
  - Power gained by switching off a **Rack**
    \[ \sim = 10 \times \text{Power(} \text{computing node} \text{)} \]

![Diagram showing power savings](attachment:image.png)

- 3400 W
- 500 W
- 344 W

© Bull, 2014
The RJMS level - DVFS

- **DVFS**
  - It's a trade-off between *performance* and *power consumption*
  - What about *performance / energy* trade-off?

\[
\int POWER \cdot dt = Energy
\]
• **DVFS**
  - It's a trade-off between **performance** and **power consumption**
  - What about **performance / energy** trade-off?
The RJMS level - DVFS

- **DVFS**
  - It's a trade-off between **performance** and **power consumption**
  - What about **performance / energy** trade-off?

---

*Energy-Performance Trade-off for IMB-MPI executions upon a 16 nodes cluster with different CPU Frequencies*
The RJMS level - DVFS

- DVFS is a trade-off between completion time and power

- No obvious performance / energy trade-off
  - Minimizing power != minimizing energy
  - The impact of DVFS is highly dependant on the job
Introduction

DVFS & Switch-off

The model

Algorithm and implementation

Experimentations

Conclusion and future works
Our Model

- We work with maximum power consumptions
- $W$ is the maximal computational work possible

\[ W = T \left( \frac{N - N_{off} - N_{dvfs}}{\sigma_{Max}} + \frac{N_{dvfs}}{\sigma_{Min}} \right) \]

- Powercap limitation

\[ N_{off} \cdot P_{off} + N_{dvfs} \cdot P_{Min} + \left( N - N_{off} - N_{dvfs} \right) \cdot P_{Max} \leq P \]

$N_X =$ number of node in state $X$
$\Sigma_Z =$ speed degradation at state $Z$
$P_Y =$ power consumption at $Y$
$P =$ powercap
Our model

- In the space 3D \((N_{dvfs}, N_{off}, W)\)

\[
W = T \left( \frac{N - N_{off} - N_{dvfs}}{\sigma_{Max}} + \frac{N_{dvfs}}{\sigma_{Min}} \right)
\]

is a plane

\[
N_{off} \cdot P_{off} + N_{dvfs} \cdot P_{Min} + (N - N_{off} - N_{dvfs}) \cdot P_{Max} \leq P
\]

is an half space

⇒ The intersection is a straight line

- Within the bound of the total number of nodes, \(W\) is maximized when:

\[
\begin{cases} 
N_{off} = \frac{P - N \cdot P_{Max}}{P_{off} - P_{Max}} \\
N_{dvfs} = 0
\end{cases}
\]

or

\[
\begin{cases} 
N_{off} = 0 \\
N_{dvfs} = \frac{P - N \cdot P_{Max}}{P_{Min} - P_{Max}}
\end{cases}
\]
• 3 cases:

– DVFS is better ⇒ we only use DVFS

– Switch-off is better ⇒ we only use Switch-off

– The powercap is so low that we should use both
Our model – switch-off or DVFS?

\[
\begin{align*}
N_{off} &= \frac{P - N \cdot P_{Max}}{P_{off} - P_{Max}} \\
N_{dvfs} &= 0
\end{align*}
\quad \text{or} \quad
\begin{align*}
N_{off} &= 0 \\
N_{dvfs} &= \frac{P - N \cdot P_{Max}}{P_{Min} - P_{Max}}
\end{align*}
\]

How to choose?

\[
\rho = 1 - \frac{\sigma_{Max}}{\sigma_{Min}} - \frac{P_{Max} - P_{dvfs}}{P_{max} - P_{off}}
\]

When \( \rho < 0 \), switch-off is preferred
Our Model – DVFS or switch-off?

- On CURIE cluster:

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Degradation</th>
<th>$\rho$</th>
<th>Best mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>2.27</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>linpack</td>
<td>2.14</td>
<td>-0.027</td>
<td>Switch-off</td>
</tr>
<tr>
<td>IMB</td>
<td>2.13</td>
<td>-0.029</td>
<td>Switch-off</td>
</tr>
<tr>
<td>SPEC Float [11]</td>
<td>1.89</td>
<td>-0.088</td>
<td>Switch-off</td>
</tr>
<tr>
<td>SPEC Integer [11]</td>
<td>1.74</td>
<td>-0.134</td>
<td>Switch-off</td>
</tr>
<tr>
<td>Common value [22]</td>
<td>1.63</td>
<td>-0.174</td>
<td>Switch-off</td>
</tr>
<tr>
<td>NAS suite [11]</td>
<td>1.5</td>
<td>-0.225</td>
<td>Switch-off</td>
</tr>
<tr>
<td>STREAM</td>
<td>1.26</td>
<td>-0.350</td>
<td>Switch-off</td>
</tr>
<tr>
<td>GROMACS</td>
<td>1.16</td>
<td>-0.422</td>
<td>Switch-off</td>
</tr>
</tbody>
</table>

Fig. 5: Comparison between DVFS and switch-off in Curie for various benchmarks.
- Introduction
- DVFS & Switch-off
- The model
- Algorithm and implementation
- Experimentations
- Conclusion and future works
The algorithm

• **When a powercap limit is set**
  • Choose between DVFS and switch-off

• **If DVFS**
  - When a job is being launched,
  - Try to schedule it at the highest frequency

• **If switch-off**
  - switch-off nodes at runtime,
  - mark these nodes as « reserved » for the scheduler
The algorithm

$scontrol create res Watts=123151 ...$
• Introduction
• DVFS & Switch-off
• The model
• Algorithm and implementation
• Experimentations
• Conclusion and future works
Experimental validation

• **Replay interesting parts of the CURIE workload**
  - 5 hours, high utilization, jobs representative of the whole workload

• **Slurm can emulate his environment**
  - 336 Slurm nodes on 1 physical node
  - *Sleep* instead of real computational jobs

• **Add a powercap**
  - Case study: 1 hour, in the middle of the trace, at different powers
Experimental validation

Fig. 7: System utilization for the IDLE, DVFS and SHUT policies in terms of cores (up) and power (bottom) during the 5 hours workload with a reservation of 60% of total powercap
Experimental validation

Idle

Switch-off

© Bull, 2014
Experimental validation

Switch-off

DVFS

Cores

0 5000 10000 15000

2.7 GHz

Watts

0 500000 1000000 1500000

2.7 GHz

DVFS

Cores

0 5000 10000 15000

2.7 GHz

Watts

0 500000 1000000 1500000

2.7 GHz

© Bull, 2014
• Introduction
• DVFS & Switch-off

• The model
• Algorithm and implementation
• Experimentations

• Conclusion and future works
Current and future works

- **Powercap on live power values**
  - Implemented using Layouts

- **Powercap on nodes**

- **DVFS**
  - What about reproducibility of jobs runs?
  - To do DVFS right, we need to know the job

- **Switch-off**
  - New scheduling algorithms
  - Switch-off (with bonuses) without powercaps
  - Switch-off particular components (cpus, gpus, network...)
References


Fig. 8: Comparison of different scenarios of policies and powercaps based on normalized values of launched jobs, accumulated cpu time and total consumed energy during the 5 hours workload interval