Server-level Power Control

Charles Lefurgy, Xiaorui Wang, Malcolm Ware
The datacenter must wire to label power supplies. However, real workloads do not use that much power. Label power is 308 W, whereas P4Max is 273 W. Real workloads use 50 W above label power.
The problem

- **Server power consumption is not well controlled.**
  - System variance (workload, configuration, process, etc.)
  - Design for worst-case power

- **Results:**
  - Power supplies are significantly over-provisioned
  - Therefore, datacenters provision for power that cannot be used
  - High cost, with no benefit in most environments
Our approach

- **Use “better-than-worst-case” design**
  - Example: Intel’s Thermal Design Power (TDP)
  - Power, like temperature, can be controlled

- **Reduce design-time power requirements**
  - Run real workloads at full performance
  - Use smaller, cost-effective power supplies

- **Enforce run-time power constraint with feedback control**
  - Slow system when running power virus
Our contributions

- Control of peak server-level power (to 0.5 W in 1 second)
- Derivation and analysis [see paper]
  - Guaranteed accuracy and stability
- Verified on real hardware
- Better application performance than previous methods
Caveats

- **Our prototype is a blade server**
  - The results of the study also apply to rack-mount servers.

- **Power controller uses clock throttling, not dynamic voltage and frequency scaling (DVFS)**
  - At the time of the study, only clock throttling was available on our prototype system.
  - DVFS is not available on all processors (lower speed grades)
  - Recently, we have built a prototype using DVFS
Rest of the talk

- **Power measurement**
- **Power control**
  - Open loop controller
  - Ad-hoc controller
  - Proportional controller
- **Experimental results**
- **Conclusions**
Power measurement

Measure 12V bulk power
0.1 W precision, 2% error

HS20 8843 (Intel Xeon blade)

controller firmware on service processor (Renesas H8 2168)

Measurement/calibration circuit
Sense resistors
Options for power control

- **Open-loop**
  - No measurement of power
  - Chooses fixed speed for a given power budget
  - Based on most power hungry workload

- **Ad-hoc**
  - Measures power and compares to power budget
  - +1/-1 adjustments to processor clock throttle register

- **Proportional Controller (“P control”)**
  - Designed using control theory
  - Guaranteed controller performance
Open loop design

- **P4MAX workload used as basis for open-loop controller**
- **Graph shows maximum 1 second power for workload**

![Graph showing power consumption vs. processor performance setting](image-url)
Proportional controller design

- **Settle to within 0.5 W of desired power in 1 second**
  - Based on BladeCenter power supply requirements

- **Every 64 ms**
  - Compare power to target power
  - Use proportional controller to select desired processor speed
    - 12.5% - 100% in units of 0.1%

- **Clock throttling**
  - Intel processor: 8 settings in units of 12.5% (12.5% - 100%)
  - Use delta-sigma modulation to achieve finer resolution
Why not use ad-hoc control?

Ad-hoc

Set point = 211.0 W

P Controller

Settles to 216.0 W  5 W Violation
CPU speed: 68.8%

Settles to 211.0 W  No violation
CPU speed: 65.8%
Steady-state error

- P controller has no steady-state error \((x=y)\)
- Ad-hoc controller has steady-state error
  - Add safety margin of 6.1 W to ad-hoc
Comparison of 3 controllers

- Run each controller with 5 power budgets
- Compare throughput of workloads

Table shows settings used for each controller

<table>
<thead>
<tr>
<th>Power budget</th>
<th>Open-loop processor performance setting</th>
<th>Ad-hoc (with safety margin) set point</th>
<th>P control set point</th>
</tr>
</thead>
<tbody>
<tr>
<td>250 W</td>
<td>75%</td>
<td>238.9 W</td>
<td>245.0 W</td>
</tr>
<tr>
<td>240 W</td>
<td>62.5%</td>
<td>229.1 W</td>
<td>235.2 W</td>
</tr>
<tr>
<td>230 W</td>
<td>62.5%</td>
<td>219.3 W</td>
<td>225.4 W</td>
</tr>
<tr>
<td>220 W</td>
<td>50%</td>
<td>209.5 W</td>
<td>215.6 W</td>
</tr>
<tr>
<td>210 W</td>
<td>37.5%</td>
<td>199.7 W</td>
<td>205.8 W</td>
</tr>
</tbody>
</table>
Application performance summary

- **P controller**
  - 31-82% higher performance than open-loop
  - 1-17% higher performance than ad-hoc
  - Quicker settling time
  - Zero steady state error

![Graph showing application throughput vs. budget](image-url)

1 W = 1.1% performance!
Power supply reduction

- 308 W: Label power of HS20 blade
- 260 W: Real workloads run at full performance
  - A reduction of 15% in supply power.

- Fit 15% more servers per circuit
Conclusions

- **Power is a 1st class resource that can be managed.**
  - Power is no longer the accidental result component configuration, manufacturing variation, and workload.

- **Reduce power supply capacity, safely.**
  - Relax design-time constraints, enforce run-time constraints.
  - Install more servers per rack.

- **Power control is a fundamental mechanism for power management in a power-constrained datacenter.**
  - Move power to critical workloads.