Power-Performance Modeling of Heterogeneous Cluster-Based Web Servers

Hiroshi Sasaki†, Takatsugu Oya†, Masaaki Kondo‡, and Hiroshi Nakamura†

† The University of Tokyo
‡ The University of Electro-Communications
Small to Large-scale internet services everywhere
- Parallel computation of a bunch of requests
- Throughput oriented computing
  We focus on web servers

Cluster-based web servers
- Heterogeneous clusters are popular

Importance of low power computing
- Computation cost, cooling cost, ...
Characteristics of web services

- Requests
  - CPU-bound
    - Dynamic files (CGI, php, Java servlets...)
  - Disk-bound
    - Static files (html web pages, jpeg photos, tar balls...)

- Response time restrictions
  - Guarantee comfortable web services
  - Web servers must be able to handle max loads
Server configuration

- Basic configuration
  - Front-end: handle and distribute the requests, reconfiguration (# of nodes, level of frequencies)
  - Back-end: execute the requests

- Back-end servers
  - Composed from several homogeneous clusters

Clients

distribution

homogeneous cluster A

homogeneous cluster B
Objective of this work

- Power reduction of heterogeneous cluster-based web servers

Requests: CPU-bound and Disk-bound
- Satisfy the response time
- Minimize the power consumption
- Dynamically select the optimal configuration (# of nodes, levels of frequencies)
Overall picture

Clients requests

Disk bound
CPU bound

Load distribution
Model

Predict the load

Decide the optimal distribution and configuration

Load distribution
Configuration management

Distribute
“70 % of the load to A,
30 % of the load to B”

Configurations are
A: mnodes
  f0, f1, ..., fm-1 GHz
B: nnodes
  f0, f1, ..., fn-1 GHz

A

B

2009/10/13
Overview of the proposed technique

1. Power-performance modeling
   - Performance model
     ✧ How much load can a certain configuration handle within the response time restriction?
   - Power model
     ✧ How much power will a certain configuration consume?
     
     Constructing a model for a single node is enough
     ✧ All the requests are parallel
     ✧ Power and performance are just a sum

2. Derive the optimal configuration
   - Homogeneous -> heterogeneous
   - Mathematically derive from the constructed model
What is a load?
- CPU-bound requests: the time to execute a page (ms)
- Disk-bound requests: the size of a page (KB)

To handle it more effectively, we define the load as a single dimensional value:
- Actual amount of requests/max amount of requests
- CPU-bound load: Load$_C$; Disk-bound load: Load$_D$
- $0 \leq \text{Load} \leq 1$
Below are two equations a CPU should satisfy to execute both Load\textsubscript{C} and Load\textsubscript{D} simultaneously:

- **CPU**
  \[
  \text{Performance\_for\_Load}_c(= f_1(\text{Load}_c)) + \text{Performance\_for\_Load}_d(= f_2(\text{Load}_d)) \leq \text{CPU performance}
  \]

- **Memory bus**
  \[
  \text{Bandwidth\_for\_Load}_c(= g_1(\text{Load}_c)) + \text{Bandwidth\_for\_Load}_d(= g_2(\text{Load}_d)) \leq \text{Memory bus bandwidth}
  \]

*Details in the paper...*
Power modeling

- Power = Base power
  + Power_for_Load_C(= F(Load_C))
  + Power_for_Load_D(= G(Load_D))
Optimization (homogenous)

- For a given amount of load, the optimal configuration is to 
  
  Distribute the load equally to every node

  All the frequencies will be the same

Details in the paper...
Challenges for optimizing heterogeneous cluster

Optimal configuration for heterogeneous cluster depends on:
- # of homogeneous clusters
- # of nodes and frequency within each homogeneous cluster
- Distribution ratio of the load

-> Derive from the model

<table>
<thead>
<tr>
<th># of nodes</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.8G, 1.8G</td>
</tr>
<tr>
<td>3</td>
<td>1.0G, 1.4G</td>
</tr>
<tr>
<td>2</td>
<td>2.0G</td>
</tr>
</tbody>
</table>

Total load: 1.0
Power distribution:
- Total

2009/10/13
Optimization (heterogenous)

- Unkowns: Distribution ratio of the load, # of nodes and frequency within each homogeneous clusters
- Known: Load

I. Within homogeneous clusters
   i. Frequency (= f(Distribution ratio, # of nodes)): substitute a load (for single node) for performance model and derive the min frequency
   ii. # of nodes (= g(Distribution ratio)): substitute the frequency for power model and derive the # of nodes which minimizes the power

II. Derive the optimal distribution ratio that minimizes the sum of the power of each homogeneous clusters
## Evaluation environment

<table>
<thead>
<tr>
<th>type</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>Intel Pentium M 760 (0.8-2.0 GHz)</td>
<td>AMD Opteron 150 (1.0-2.2 GHz)</td>
</tr>
<tr>
<td>memory</td>
<td>DDR2-SDRAM 1GB PC2-4300</td>
<td>DDR-SDRAM 1GB PC-3200</td>
</tr>
<tr>
<td>Disk</td>
<td>80GB 7200rpm SATA3.0GB/s seek time 8.8ms</td>
<td>80GB 7200rpm SATA3.0GB/s seek time 8.8ms</td>
</tr>
<tr>
<td>OS</td>
<td>Linux kernel-2.6.11</td>
<td>Linux kernel-2.6.16</td>
</tr>
<tr>
<td>ServerSW</td>
<td>Apache 2.2.3</td>
<td>Apache 2.2.3</td>
</tr>
</tbody>
</table>

- Clients: httpperf 0.8 (by HP)
- Loads: CPU-bound (cgi), Disk-bound (html)
- Response time restriction: 200ms for both types of loads
Validation: performance model

- Actual (A, max)
- Actual (A, min)
- Actual (B, max)
- Actual (B, min)
- Model (A)
- Model (B)

Load

Load

Load

Load

2009/10/13
Validation: power model

Coefficients are in the paper

\[ \text{Load}_{D} \]
Evaluation

1. Optimizing within homogeneous cluster
   - Best case vs. proposed (derived from the model)

2. Optimizing heterogeneous cluster
   - Compare the three policies below
     1. *Conventional*
        Load: distribute equally
        Configuration: all nodes are on and max frequency
     2. *Model-even*
        Load: distribute equally
        Configuration: derive from the model
     3. *Proposed*
        Load: derive from the model
        Configuration: derive from the model
Proposed mostly (94%) selected the optimal configuration
### Result 2 (A: 4 nodes B: 4 nodes)

<table>
<thead>
<tr>
<th>Policy</th>
<th>Configuration</th>
<th>Power [W]</th>
</tr>
</thead>
<tbody>
<tr>
<td>conventional</td>
<td>A(4N, 2.0G), B(4N, 2.2G)</td>
<td>410.8</td>
</tr>
<tr>
<td>Model-even</td>
<td>A(3N, 2.0G), B(3N, 2.2G)</td>
<td>326.4</td>
</tr>
<tr>
<td>proposed</td>
<td>A(4N, 2.0G), B(1N, 1.4G)</td>
<td>276.2</td>
</tr>
</tbody>
</table>

The graph shows the power [W] consumed by different load policies over time. The vertical axis represents power in watts ranging from 0 to 400, while the horizontal axis represents time from 00:00 to 18:00. The graph is divided into three segments: conventional, model-even, and proposed, each with different line styles.
Conclusions and future work

Conclusions

- Objective: power reduction of a heterogeneous cluster-based web servers
- Constructed a power-performance model
- Derived the configuration from the model
  - Showed that proposed technique can reduce significant power

Future work

- Control the power and performance of other devices (HDD, DRAM Memory, ...)
- Implement our technique in the OS (power on/off, suspend, dynamic prediction, recovery from mispredictions, ...)