Energy Management for Server Clusters

Ram Rajamony
IBM Austin Research Lab

Ricardo Bianchini
Rutgers University

Copyright 2002 by Ram Rajamony and Ricardo Bianchini
Tutorial Outline

- Motivation
- Energy Consumption in Data Centers
- Energy Conservation for Single-Node Systems
- Energy Conservation for Front-End and Compute Clusters
- Energy Conservation for Storage Clusters
- QoS and Energy Conservation
- Case Study
- Future Challenges
- Discussion
Server Cluster

Set of computers used as a single system

Load Balancer

Content Store NASD, servers

Web Server Cluster
Motivation
Motivation

- One (key) reason: Power and energy account for large fraction of data center operating cost [The Industry Standard, 2/19/01]
- Power consumption affects cooling and backup power generation requirements, as well as reliability
  - Higher power means greater investments in sophisticated racks, air conditioning and power generation infrastructure
  - Excessive heat may cause intermittent failures
  - Power draw may become problem for utilities
- Energy consumption affects electricity bills
Motivation (contd.)

- Another reason: Environmental implications
  - Most power generation technologies are harmful to the environment (e.g., coal-based, nuclear energy)
- One last reason: Regulatory or governmental pressure
  - Energy* program in the US
  - New Japanese law
Energy* Program

- US government classifies computer equipment, consumer products, homes, commercial buildings, etc.

- Key points
  - Energy* label certifies that equipment is “energy-efficient” with respect to pre-established rules
  - An Energy* PC goes to sleep mode after 30 minutes of inactivity and starts consuming less than 15W
  - An Energy* server may consume more, depending on rating of the power supply; e.g., for a power supply between 200 and 300W, maximum consumption is 20W
Japanese Energy Law

- Requires all products sold in Japan to meet energy requirements starting 2005
- Key points
  - "Top runner" highlights best model/category/year (since ‘99)
  - Weighted average based on sales volume
  - Law becomes stricter with time
Japanese Energy Law (contd.)

The achievement of the target is determined by whether the sum of the difference in energy efficiency weighted by the number of units shipped within the product division is positive or negative. In the figure above, if the sum \( F (= X_1 \times Y_1 + X_2 \times Y_2 + X_3 \times Y_3 + X_4 \times Y_4) \) is negative \((F<0)\), it is determined that the target is not achieved, while if the sum is positive \((F>0)\), it is determined that the target is achieved.

Source: http://unfccc.int/sessions/workshop/000411/jpnja.pdf
Slight Digression: Background and Terminology

- Power and energy are related but different things. In fact, energy = power x time
- Power is measured in Watts (W), whereas energy is measured in Joules (J = Ws) or Wh
- To simplify the rest of the presentation, we will refer to power and energy collectively as “energy”, unless we explicitly differentiate
Energy Consumption in Data Centers
Typical Data Center

- 30,000 square feet
  - Space rented by square feet, number of floor tiles
- Servers, networking, and storage in racks
- Cooling and backup power generation
Data Center Layout

- Equipment placed in racks
  - 24” wide x 36” deep x 43” high
  - Building codes typically require 3 feet aisles
  - A single rack consumes 12 sq. ft. of total space
- Significant confusion about energy usage per sq. ft.
  - Everybody has their favorite Watt/sq. ft. number
    - Ranges from 30W/sq. ft. to 300 W/sq. ft.
  - Comprehensive study from UC-Berkeley, LBNL
    - Average data center power = 50 W/sq. ft. (more on this later)
    - Jennifer Mitchell-Jackson’s thesis:
Where Does the Energy Go?

- **Ratings**
  - **Server: IBM 1U x330**
    - 200 Watts
    - Measured: 110W when busy, 90W when idle (excl. storage)
  - **Networking: Extreme BlackDiamond**
    - 1.2W/100Mb port
    - 11W/1Gb port
  - **Storage: EMC Clariion**
    - ~50W per drive (~70GB)
Cost of Cooling

- How to measure cooling capacity?
  - Cooling effect: BTU, Efficiency: EER or SEER
  - Cooling unit with EER = 12 will cool 12000 BTU for 1kWh

- Example: Yearly operating energy for single 200W server
  - Operation: 200 * 24 * 365 / 1000 = 1752 kWh
  - Cooling: 1752 * 3414 / 12000 = 498 kWh
    - Note: 1kWh = 3414 BTU
  - Cooling takes 22% of total energy
  - Energy costs at 8 cents/kWh ➔ $180/year

- *This analysis is simplified. Eg: Air distribution not included*
Total Data Center Energy Usage

- Difficult to break energy usage down
  - Data centers have computer space, offices, etc.
  - Cooling equipment cools whole building
- Conservative estimate of total energy usage
  - Computer room area = X sq. ft.
  - Total data center power usage = 50X Watts
  - Includes cooling, support
  - See also http://www.repp.org/articles/static/1/binaries/data_centers_report.pdf
Expense info is hard to come by, income info is easier

Take co-location prices for example
- Data center gives you space, power, cooling, network
- You install your own servers (BYOS)
- Sample co-lo prices: $700-$1000/month/rack (uses 12 sq. ft)

Average energy cost per rack per month: $35
- (50W/sq. ft * 12 sq. ft for a rack) * 24 * 30 * .08 / 1000 = $35
  - Includes cooling, support, lighting, etc.

Energy accounts for 3.5% to 5% of co-lo income
What Makes It Harder

- Data center price wars
  - Prices keep falling, lot cheaper than 2001
- Cost of provided energy capacity is high
  - Data center gives rack ~ 3000W of power, say you use it all
  - $3000W * 24 * 30 * .08 / 1000 = $173
  - This is 17% - 25% of income
- Utilities are getting smarter
  - Requiring up-front deposits from data centers
  - ComEd asks for $500K to $10 million depending on peak demand, ConEd, Nstar, SoCalEd thinking similarly
- Bottom line: It pays to reduce energy costs

Bottom line: It pays to reduce energy costs
Energy Conservation for Single-Node Systems
Single-Node Systems: Mechanisms

- Dynamic voltage scaling
  - Processor does not need to run at max speed all the time
  - Leverage non-linear relationship between power and performance
  - Must account for scaling costs

- Power components down
  - Idle components do not have to be on
  - Must account for transition costs
  - Clock gating, powering down memory, disk, network, etc.

- Other techniques (request batching/aggregation)
Dynamic Voltage Scaling

- Dynamically adjust voltage and/or frequency
- Non-linear voltage \(\leftrightarrow\) frequency relationship
Power Components Down

- Processors
  - Clock gating
- Storage
  - Spin down media
- Memory
  - Low-energy state (RDRAM), turn off memory
- Network
  - Turn network interface off
  - Turn part of switch down
Example: Powering Down Disks

3 states: idle, standby, sleep

- green: access
- green: state transition
- blue: disk ready
- red: Direct Deactivation, Pre-activation
Bladed Servers

- Blade: Board that plugs into chassis (backplane)
- Two approaches
  - High function blades (high-power CPUs)
  - High density blades (low-power CPUs)
- Advantages
  - Less cabling (this is important!)
  - Potentially lower space, power, cooling needs
  - Mix and match: server, storage, network, etc.
- Disadvantage
  - Current data centers may not be able to cope with higher energy density [esp. true for high function blades]
High Function Server Blades

- Typically perform all traditional server functions
  - Drop in replacement
- Denser component packaging; evolution of the pizza box
- IBM’s eServer BladeCenter (3Q02)
  - Multiple Pentium-4 or Power processors
    - [Tech Update](http://techupdate.zdnet.com/techupdate/stories/main/0,14179,2862624,00.html)
    - [IBM's eServer XSeries BladeCenter](http://www.pc.ibm.com/ww/eserver/xseries/blade.html)
- Dell PowerEdge 1655MC (2H02)
  - Two Pentium-III processors, 2GB memory, 146GB, 1.33Gb/s
    - [Dell PowerEdge 1655MC](http://www.dell.com/us/en/esg/topics/esg_pedge_rackmain_servers_3_pedge_1655mc.htm)
  - 1000W power supply per 3U height
- A few other examples: Mellanox, NEC, NexCom, Cubix
High Density Server Blades

- Lower power components
  - The army of turtles approach
- RLX ServerBlade
  - One Transmeta or low-voltage Pentium-III processor, 1GB memory, two IDE disks, 100Mb/s
    http://www.rlxtechnologies.com
- A few other examples: HP bc1000, HP Proliant BL e-class, Fujitsu, IBM-ARL prototype
Other Blades

- Storage blades
  - Example: HP A6781A
    - Two 30GB, 4200rpm IDE disks
- Network switch blades
  - Example: Znyx cPCI blade
    - Twenty-four 100Mb/s with two 1 Gb/s uplink ports
- “Other” blades
  - Depends on chassis
  - compactPCI has GPS, SCSI, Management, …
Energy Management for Blades?

- Does it make sense to energy manage blades?
  - Aren’t high density blades already energy conscious?
- Tutorial approach - show it both ways
  - Evaluation for bladed servers
  - Evaluation for traditional servers
- Bottom line: Energy management important for all types of servers
Energy Conservation for Front-End and Compute Clusters
Cluster of Front-End Servers

- Goal: reduce the energy consumption of the cluster, with minimal performance impact
- Two mechanisms
  - Varyon/Varyoff (VOVO) - turn off some nodes of the cluster when workload is low
    - Proposed independently by Chase et al. (SOSP '01) and Pinheiro et al. (COLP '01)
  - Voltage Scaling - reduce the operating voltage/frequency of some/all server nodes
    - Proposed for servers by Bohrer et al. (PAC’02)
Cluster Energy Mgmt Policies

- IVS - Independent Voltage Scaling
- VOVO - VaryOn/VaryOff
- CVS - Coordinated Voltage Scaling
- VOVO-IVS - VOVO combined with IVS
- VOVO-CVS - VOVO combined with CVS

Increasing implementation complexity
VOVO-CVS - Details

Based on model for system power

\[ P(f) = c_0 + c_1 f^3 \]

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>CPU frequency (0 &lt;= f &lt;= 1.0)</td>
</tr>
<tr>
<td>c_0</td>
<td>fixed system power</td>
</tr>
<tr>
<td>c_1</td>
<td>coefficient for variable system power</td>
</tr>
</tbody>
</table>

- VOVO policy targets \( c_0 \) component
- CVS policy targets \( c_1 f^3 \) component
- VOVO-CVS targets the component that offers the most benefit
VOVO-CVS - Details

Vary off a node if:

\[(n-1) \left( c_0 + c_1 \left( \frac{n}{n-1} f \right)^3 \right) < n \left( c_0 + c_1 f^3 \right)\]

or

\[f_{\text{varyoff}} < \sqrt[3]{\frac{c_0}{c_1} \frac{(n-1)^2}{2n^2 - n}}\]

Vary on a node if:

\[(n+1) \left( c_0 + c_1 \left( \frac{n}{n+1} f \right)^3 \right) < n \left( c_0 + c_1 f^3 \right)\]

or

\[f_{\text{varyon}} < \sqrt[3]{\frac{c_0}{c_1} \frac{(n+1)^2}{2n^2 + n}}\]
VOVO-CVS - Details

Varyon/Varyoff points for $c_0 = 13.5$, $c_1 = 22.7$
Evaluation

- Energy-accurate simulator (Salsa) for web server cluster
- Workloads derived from real internet server logs
- Machine configuration
  - Voltage-scaled CPU: 600MHz–1.2GHz, 5.0W-27.7W
  - Two base machine configurations
    - Bladed system 1 (today’s system): 8.5 Watts ($c_0 = 13.5W$)
    - Bladed system 2 (future, more efficient): 5.0 Watts ($c_0 = 10.0W$)
Salsa: Energy Accuracy
Salsa: Response Time Accuracy

![Graph showing response time measurements and simulations over time.](image)
Analysis

Olympics98 Workload

![Graph showing the workload over time](image)
Analysis

Olympics98 Energy Savings

% savings over base

Time (hours)

VOYO-CVS
VOYO
CVS
Analysis

Olympics98 Active Servers

![Graph showing the number of active servers over time with two lines labeled 'VOVO' and 'VOVO-CVS'.]
Analysis

Olympics98 Power Consumption

5.0W base machine configuration
Analysis

Finance Workload

![Graph showing Finance Workload over time](image-url)
Analysis

Finance Energy Savings

[Graph showing percentage savings over time with lines for VOVO-CVS, VOVO, and CVS]
Analysis

Finance Active Servers
Analysis

Finance Power Consumption

5.0W base machine configuration
Results

All policies have avg response time < 40ms
Conclusion: Front-End Servers

- Energy management using simple voltage scaling mechanisms produced energy savings of 20%-29%.
- Varyon/Varyoff achieved savings of 22%-43%.
- A complex combination of voltage scaling and Varyon/Varyoff achieved the largest energy savings of 36%-50% over no energy management.
- More results in Elnozahy, Kistler, and Rajamony (PACS’02).
Cluster of Compute/Cycle Servers

- Same idea as for front-end servers, however may need to migrate processes dynamically
- Key principles: (1) remove nodes if resulting estimated performance degradation <= pre-established threshold; (2) add nodes if current estimated degradation > threshold
- Estimate degradation based on resource utilization information collected at all nodes
- Plug estimates into a PID feedback controller:
  \[ \text{output}(t) = k_p \text{ e}(t) + k_i \sum_{0}^{t} \text{e}(t) + k_d \Delta \text{e}(t) \]
Cluster Reconfiguration Algorithm

Periodically do

if node removal is acceptable
  choose victim nodes with low demand
  if necessary, determine nodes to receive load and
  ask victims to migrate their loads out
  ask victims to turn themselves off
else
  if addition is necessary
    turn on new nodes
    if necessary, determine load for new nodes and
    ask nodes to share their load with new nodes
Evaluation

- Implementation in the Nomad operating system by Pinheiro and Bianchini (IWCC99) – provides SSI, load balancing, and intelligent application initiation

- Methodology: 8-node cluster (idle node = 70W), synthetic application mix (SPEC, MPEG encoder, I/O benchmarks), comparison against static cluster

- Parameters
  - Controller constants: $k_p = 0.7$, $k_i = 0.15$, $k_d = 0.15$
  - Acceptable degradations: 0% and 20%
Results
Results
Results
Conclusion: Compute Servers

- Simple VOVO for compute servers achieves power savings of up to 88%
- Achieves energy savings of up to 32% when degradation is unacceptable. Up to 40% energy savings when 20% degradation is acceptable
- More results in Pinheiro et al. (Kluwer Academic ‘02)
Energy Conservation for Storage Clusters
Basic Principles and Policies

- All the data NOT needed all the time
- Access-based allocation to machines/disks
  - MAID: Massive Arrays of Idle Disks [WIP FAST'02]
  - NomadFS [WIP SOSP'01]
- MAID: copy recently accessed data to a subset of the disks; keep the other disks off
- NomadFS: migrate the popular data to a subset of the disks; keep the other disks off
- When data is accessed with high temporal locality, two approaches are similar
- In both systems, unpopular files may suffer high latencies
MAID

Active set

Inactive set

Energy Management for Clusters
Rajamony & Bianchini
NomadFS: Single-Node Server

- Data requests logged and periodically analyzed
- Analysis determines where files should be stored
- Writes to disks that are off are accumulated in logs
Preliminary Results

- Simplistic approach: allocate files to disks according to popularity. Turn disks off based on inactivity threshold.
- Simulator: single-node Web server with 4 IDE disks
- Workload: 45-day World Cup ’98 trace
- Transition overheads:
  - Power: active = 2.7W, idle = 2W, sleep = 0.1W
  - Energy: spin up = 12.4J, spin down = 0.1J
  - Time: spin up = 4 seconds
Preliminary Results (contd.)

- Energy savings of 25% for 10-sec disk inactivity threshold. Savings go up to 40% for 1-sec threshold
- Overall throughput indistinguishable from energy-oblivious server with random allocation of files to disks
- Real implementation is underway; TR out soon
Next Step: Storage Cluster

- Divide files into groups and allocate them differently
- Extend ideas to cluster environment, e.g. set of NASDs or cluster of disk-enabled servers
- Creates new possibilities for energy conservation: whole nodes can be turned off, if the files they store are *really* unpopular
QoS and Energy Conservation
Quality of Service vs. Energy

- System responsiveness affects energy
  - More slack ➔ More energy savings (up to a point)
  - Slack enables system to smooth out load bursts
Internet Delivery: Advantages

- Current RTT
  - From 88ms (North America) to 350 ms (Asia)
  - Sampled on 6/14/2002

- Assume:
  - Server time is 0, one packet request/response
  - 75ms response time degradation at the server
    - Varies from 85% (NA) to 21% (Asia) CPRT degradation
      - But is not noticeable within human sensory limits
      - [CPRT: Client-Perceived Response Time]
    - Will show this provides significant energy savings
Taking Advantage of Slack

- Olympics98 workload, VOVO-CVS
  - Yields 24.3% savings over no EM case
  - 90%-ile RT relaxed to 75ms ➞ 34.1% savings
Slack in Finance

- Finance workload, VOVO-CVS
  - Yields 42.8% energy savings over no EM case
  - 90%-ile RT relaxed to 75ms ➔ 49% savings
Relaxing Response Time

- Service Level Agreements (SLAs) based on system responsiveness and availability
    - 95% of server’s responses must be delivered in less than 50ms
    - Timeframe from 3 to 6am does not affect above statistics
    - Response time greater than 10s considered a failure (treated in the availability part of the SLA and does not affect above stats)

- You can engineer an SLA to provide for slack
  - Customer agrees to specified response-time degradation
  - Similar to voluntary load shedding

- Controller design is involved
  - Ongoing research topic
Case Study

Build Your Own Cluster
Measuring Energy Consumption

- External measurement, IP-enabled power strips
  - Ex. multimeter: Yokogawa WT110
  - Ex. strip: Baytech RPC-22
- In-server monitoring
  - Design hardware to monitor
  - Extract using I2C bus, report over network
  - Example: ARL H8 implementation (see blade)
Using Performance Counters

- Use performance counter information:
  - Example: Instruction decode, MMU, FPU, …
  - Aha! Events and energy values correlate
  - Pre-compute event $\rightarrow$ energy table
  - Main disadvantage: Few counters, limited functionality

- Frank Bellosa (Univ. of Erlangen)

- CASTLE (Princeton)
Building Your Own EM Cluster

- **Ingredients**
  - Component control: CPU, disk, memory, network, server
  - Request Dispatcher + System Energy Controller
Component Control

- CPU voltage scaling
  - Transmeta, AMD PowerNow
  - Enable software control of core voltage/frequency
- Disk spin down
  - Software for IDE drives available today
  - SCSI needs some work
- Memory, Network
  - No stable methods exist today
- Server power on/off
Server Power On/Off

- Service processor based
  - IBM x330 series
  - SNMP/Java-based system to power nodes on/off
  - Power consumption reduced to ~10W when off
- I2C bus-based control
  - compactPCI blades (PICMG 2.16)
  - HP blade servers
Request Dispatcher

- Linux Virtual Server
  - See http://www.linuxvirtualserver.org
  - Uses connection-based load determination
  - Needs adaptation to enforce VOVO
    - Determine maximum server capacity (in connections)
    - Augment LVS to enforce policy [write code!]
      - See “Cluster Reconfiguration Algorithm” (VOVO)
How Easy Is It Today?

- Two choices when it comes to the servers you need
  - Servers with remote on/off: xSeries, blades, power switches
  - Voltage-scaled systems: Transmeta, AMD PowerNow, IBM 405GP

- LVS is distributed with RedHat 7.3
  - IVS: No modifications needed
  - VOVO: LVS modifications not hard, done at ARL
  - VOVO-CVS: LVS modifications + PowerNow control
    - More work involved

- Energy management for storage is harder
  - No stock solutions for SCSI (unstable Linux utilities)
Future Challenges
Future Challenges

- Reassessment of single-node server design decisions with energy conservation also in mind
- Development of more sophisticated energy conservation techniques for clusters
- Development of standards and off-the-shelf software + hardware for energy-efficient clusters
- Further research needed on storage servers, QoS vs energy, and the role of heterogeneous systems
- Energy managing clusters of “virtual servers”
- Finding other ways of exploiting non-linearity in processors
DVS: Future Prospects

- Technology causing Vdd, Vt to decrease
  - Power/performance relationship likely to remain non-linear
Discussion