Where Does the Power Go in High-Scale Data Centers?

USENIX ‘09 San Diego

James Hamilton, 2009/6/17
VP & Distinguished Engineer, Amazon Web Services

e: James@amazon.com
w: mvdirona.com/jrh/work
b: perspectives.mvdirona.com
Agenda

- High Scale Services
  - Infrastructure cost breakdown
  - Where does the power go?
- Power Distribution Efficiency
- Mechanical System Efficiency
- Server & Applications Efficiency
  - Work done per joule & per dollar
  - Resource consumption shaping
Background & Biases

• 15 years in database engine development
  – Lead architect on IBM DB2
  – Architect on SQL Server

• Past 5 years in services
  – Led Exchange Hosted Services Team
  – Architect on the Windows Live Platform
  – Architect on Amazon Web Services

• Talk does not necessarily represent positions of current or past employers
Services Different from Enterprises

• **Enterprise Approach:**
  – Largest cost is people -- scales roughly with servers (~100:1 common)
  – Enterprise interests center around consolidation & utilization
    • Consolidate workload onto fewer, larger systems
    • Large SANs for storage & large routers for networking

• **Internet-Scale Services Approach:**
  – Largest costs is server & storage H/W
    • Typically followed by cooling, power distribution, power
    • Networking varies from very low to dominant depending upon service
    • People costs under 10% & often under 5% (>1000+:1 server:admin)
  – Services interests center around work-done-per-$ (or joule)

• **Observations:**
  • People costs shift from top to nearly irrelevant.
  • Expect high-scale service techniques to spread to enterprise
  • Focus instead on work done/$ & work done/joule
Power & Related Costs Dominate

• **Assumptions:**
  - Facility: ~$200M for 15MW facility (15-year amort.)
  - Servers: ~$2k/each, roughly 50,000 (3-year amort.)
  - Average server power draw at 30% utilization: 80%
  - Commercial Power: ~$0.07/kWhr

![Monthly Costs](http://perspectives.mvdirona.com/2008/11/28/CostOfPowerInLargeScaleDataCenters.aspx)

3yr server & 15 yr infrastructure amortization

• **Observations:**
  - $2.3M/month from charges functionally related to power
  - Power related costs trending flat or up while server costs trending down

PUE & DCiE

• Measure of data center infrastructure efficiency
• Power Usage Effectiveness
  – PUE = (Total Facility Power)/(IT Equipment Power)
• Data Center Infrastructure Efficiency
  – DCiE = (IT Equipment Power)/(Total Facility Power) * 100%
• Help evangelize tPUE (power to server components)

Where Does the Power Go?

• Assuming a pretty good data center with PUE ~1.7
  – Each watt to server loses ~0.7W to power distribution losses & cooling
  – IT load (servers): 1/1.7 => 59%

• Power losses are easier to track than cooling:
  – Power transmission & switching losses: 8%
    • Detailed power distribution losses on next slide
  – Cooling losses remainder: 100-(59+8) => 33%

• Observations:
  – Server efficiency & utilization improvements highly leveraged
  – Cooling costs unreasonably high
Agenda

• High Scale Services
  – Infrastructure cost breakdown
  – Where does the power go?

• Power Distribution Efficiency

• Mechanical System Efficiency

• Server & Applications Efficiency
  – Work done per joule & per dollar
  – Resource consumption shaping

2009/6/17
http://perspectives.mvdirona.com
Power Distribution

High Voltage Utility Distribution

8% distribution loss
0.997^3 * 0.94 * 0.99 = 92.2%

2.5MW Generator (180 gal/hr)

IT Load (servers, storage, Net, ...)

~1% loss in switch gear & conductors

Transformers

115kV

13.2kV

13.2kV

13.2kV

0.3% loss
99.7% efficient

6% loss
94% efficient, ~97% available

0.3% loss
99.7% efficient

0.3% loss
99.7% efficient

0.3% loss
99.7% efficient

UPS: Rotary or Battery

UPS: Rotary or Battery

Transformers

480V

208V

2009/6/17

http://perspectives.mvdirona.com
Power Yield Management

• “Oversell” power, the most valuable resource:
  – e.g. sell more seats than airplane holds
• Overdraw penalty high:
  – Pop breaker (outage)
  – Overdraw utility (fine)
• Considerable optimization possible, if workload variation is understood
  – Workload diversity & history helpful
  – Degraded Operations Mode to shed workload

Source: Power Provisioning in a Warehouse-Sized Computer, Xiabo Fan, Wolf Weber, & Luize Borroso
Power Distribution Efficiency Summary

- Two additional conversions in server:
  1. Power Supply: often <80% at typical load
  2. On board step-down (VRM/VRD): ~80% common
     - ~95% efficient both available & affordable

- Rules to minimize power distribution losses:
  1. Oversell power (more theoretic load that power)
  2. Avoid conversions (Less transformer steps & efficient or no UPS)
  3. Increase efficiency of conversions
  4. High voltage as close to load as possible
  5. Size voltage regulators (VRM/VRDs) to load & use efficient parts
  6. DC distribution potentially a small win (regulatory issues)
Agenda

• High Scale Services
  – Infrastructure cost breakdown
  – Where does the power go?
• Power Distribution Efficiency
• Mechanical System Efficiency
• Server & Applications Efficiency
  – Work done per joule & per dollar
  – Resource consumption shaping
Conventional Mechanical Design

- Cooling Tower
- CWS Pump
- Heat Exchanger (Water-Side Economizer)
- A/C Condenser
- Primary Pump
- A/C Evaporator
- A/C Compressor
- Secondary Pump

Server fans 6 to 9W each

Overall Mechanical Losses ~33%

Diluted Hot/Cold Mix

Cold

Hot

leakage

fans

Air Impeller

Computer Room Air Handler
Cooling & Air Handling Gains

- Tighter control of air-flow increased delta-T
- Container takes one step further with very little air in motion, variable speed fans, & tight feedback between CRAC and load
- Sealed enclosure allows elimination of small, inefficient (6 to 9W each) server fans
Water!

• It’s not just about power
• Prodigious water consumption in conventional facility designs
  – Both evaporation & blow down losses
  – For example, roughly 360,000 gal/day at typical 15MW facility
ASHRAE 2008 Recommended

Most data centers run in this range.

ASHRAE 2008 Recommended Class 1

81°F
Most data center run in this range

ASHRAE Allowable Class 1

ASHRAE 2008 Recommended Class 1

90°F
Dell PowerEdge 2950 Warranty

ASHRAE Allowable Class 1

Most data center run in this range

ASHRAE 2008 Recommended Class 1

95°F
NEBS (Telco) & Rackable Systems

Psychrometric Chart

- NEBS & Rackable CloudRack C2
- Dell Servers (Ty Schmitt)
- Most data center run in this range
- ASHRAE Allowable Class 1
- ASHRAE 2008 Recommended Class 1

104°F
Air Cooling

• Allowable component temperatures higher than hottest place on earth
  – Al Aziziyah, Libya: 136F/58C (1922)

• It’s only a mechanical engineering problem
  – More air & better mechanical designs
  – Tradeoff: power to move air vs cooling savings & semi-conductor leakage current
  – Partial recirculation when external air too cold

• Currently available equipment:
  – 40C: Rackable CloudRack C2
  – 35C: Dell Servers

Memory: 3W - 20W
Temp Spec: 85C-105C

Hard Drives: 7W - 25W
Temp Spec: 50C-60C

Rackable CloudRack C2
Temp Spec: 40C

Thanks for data & discussions:
Ty Schmitt, Dell Principle Thermal/Mechanical Arch.
& Giovanni Coglitore, Rackable Systems CTO
Air-Side Economization & Evaporative Cooling

• Avoid direct expansion cooling entirely

• Ingredients for success:
  – Higher data center temperatures
  – Air side economization
  – Direct evaporative cooling

• Particulate concerns:
  – Usage of outside air during wildfires or datacenter generator operation
  – Solution: filtration & filter admin or heat wheel & related techniques

• Others: higher fan power consumption, more leakage current, higher failure rate
Mechanical Efficiency Summary

• **Mechanical System Optimizations:**
  1. Tight airflow control, short paths & large impellers
  2. Raise data center temperatures
  3. Cooling towers rather than A/C
  4. Air side economization & evaporative cooling
     • outside air rather than A/C & towers
Agenda

• High Scale Services
  – Infrastructure cost breakdown
  – Where does the power go?
• Power Distribution Efficiency
• Mechanical System Efficiency
• Server & Applications Efficiency
  – Work done per joule & per dollar
  – Resource consumption shaping
CEMS Speeds & Feeds

• CEMS: Cooperative Expendable Micro-Slice Servers
  – Correct system balance problem with less-capable CPU
    • Too many cores, running too fast, and lagging memory, bus, disk, ...
• Joint project with Rackable Systems ([http://www.rackable.com/](http://www.rackable.com/))

### System-X vs. CEMS V3 (Athlon 4850e) vs. CEMS V2 (Athlon 3400e) vs. CEMS V1 (Athlon 2000+)

<table>
<thead>
<tr>
<th></th>
<th>System-X</th>
<th>CEMS V3 (Athlon 4850e)</th>
<th>CEMS V2 (Athlon 3400e)</th>
<th>CEMS V1 (Athlon 2000+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU load%</td>
<td>56%</td>
<td>57%</td>
<td>57%</td>
<td>61%</td>
</tr>
<tr>
<td>RPS</td>
<td>95.9</td>
<td>75.3</td>
<td>54.3</td>
<td>17.0</td>
</tr>
<tr>
<td>Price</td>
<td>$2,371</td>
<td>$500</td>
<td>$685</td>
<td>$500</td>
</tr>
<tr>
<td>Power</td>
<td>295</td>
<td>60</td>
<td>39</td>
<td>33</td>
</tr>
<tr>
<td>RPS/Price</td>
<td>0.04</td>
<td>0.15</td>
<td>0.08</td>
<td>0.03</td>
</tr>
<tr>
<td>RPS/Joule</td>
<td>0.33</td>
<td>1.25</td>
<td>1.39</td>
<td>0.52</td>
</tr>
<tr>
<td>RPS/Rack</td>
<td>1918.4</td>
<td>18062.4</td>
<td>13024.8</td>
<td>4080.0</td>
</tr>
</tbody>
</table>

• CEMS V2 Comparison:
  • Work Done/$: +375%
  • Work Done/Joule +379%
  • Work Done/Rack: +942%

**Update**: New H/W SKU will likely reduce advantage by factor of 2.

S/W & Utilization

• Work done/Joule & work done/$ optimization led to CEMS
  – But, there are limits where this can be difficult to apply
  – Some workloads partition poorly(e.g. commercial DB engines)
• The technique applies well to highly partitioned workloads
  – Under 10W fail-in-place servers
  – Requires porting entire S/W stack (practical with server workloads)
• But inefficient S/W & poor utilization problems remain:
  – Inefficient software can waste more resources than savings so far
  – Average server utilization industry-wide is estimated at 15%
• We need:
  1. Improve utilization through dynamic resource management
  2. Power proportionality
    • Today zero-load server draws ~60% of fully loaded server
Resource Consumption Shaping

- Resourced optimization applied to full DC
- Network charge: base + 95th percentile
  - Push peaks to troughs
  - Fill troughs for “free”
  - Dynamic resource allocation
    - Virtual machine helpful but not needed
    - Symmetrically charged so ingress effectively free
- Power also often charged on base + peak
  - Push some workload from peak into “free” troughs
  - S3 (suspend) or S5 (off) when server not needed
- Disks come with both IOPS capability & capacity
  - Mix hot & cold data to “soak up” both resources
- Incent priority (urgency) differentiation in charge-back model

David Treadwell & James Hamilton / Treadwell Graph

http://perspectives.mvdirona.com
Summary

• It's not about application performance but performance & efficiency of a multi-server S/W system, the H/W, and hosting infrastructure

• In work at all levels, focus on:
  – Work done per dollar
  – Work done per joule

• Single dimensional performance measurements are not interesting at scale unless balanced against cost

• Measure data center efficiency using tPUE

• Big opportunity to improve overall system efficiency
More Information

- **This Slide Deck:**
  - I will post these slides to [http://mvdirona.com/jrh/work](http://mvdirona.com/jrh/work) later this week

- **Power and Total Power Usage Effectiveness (tPUE):**

- **Berkeley Above the Clouds**

- **Degraded Operations Mode**

- **Cost of Power**

- **Power Optimization:**

- **Cooperative, Expendable, Microslice Servers**

- **Power Proportionality**

- **Resource Consumption Shaping:**

- **Email**
  - James@amazon.com