Data Center Efficiency
Best Practices

Data Center Efficiency Summit

James Hamilton, 2009/4/1
VP & Distinguished Engineer, Amazon Web Services
e: James@amazon.com
w: mvdirona.com/jrh/work
b: perspectives.mvdirona.com
Agenda

• Where does the power go?
• Power distribution optimization
• Mechanical systems optimization
• Server & other optimization
  – Cooperative, Expendable, Micro-Slice Servers
  – Improving existing builds
• Summary

http://perspectives.mvdirona.com
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

• This talk focuses on industry best practices
  – Not about Amazon (or past employers)
  – specialized data center design techniques
  – 2x gain over current averages easily attainable without advanced techniques
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%
- I’m looking for help defining tPUE (pwr to chip rather than server)

http://www.thegreengrid.org/gg_content/TGG_Data_Center_Power_Efficiency_Metrics_PUE_and_DCiE.pdf
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**

- Servers: $2,997,090
- Power & Cooling Infrastructure: $1,296,902
- Power: $1,042,440
- Other Infrastructure: $284,686

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

Fully Burdened Cost of Power

- **Infrastructure cost/watt:**
  - 15 year amortization & 5% money cost
  - \( \text{PMT}(5\%, 15, 2\text{MM}, 0)/(15\text{MW}) \) => $1.28/W/yr
- **Cost per watt using $0.07 Kw*hr:**
  - \(-0.07\times1.7/1000\times0.8\times24\times365\) => $0.83/W/yr (@80% power utilization)

- **Annually fully burdened cost of power:**
  - $1.28 + $0.83 => $2.11/W/yr

Details at: [http://perspectives.mvdirona.com/2008/12/06/AnnualFullyBurdenedCostOfPower.aspx](http://perspectives.mvdirona.com/2008/12/06/AnnualFullyBurdenedCostOfPower.aspx)
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%
Power Distribution

High Voltage Utility Distribution

8% distribution loss
\[0.997^3 \times 0.94 \times 0.99 = 92.2\%\]

2.5MW Generator (180 gal/hr)

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

~1% loss in switch gear & conductors

0.3% loss
99.7% efficient

9.15 kv

Transformer

UPS: Rotary or Battery

6% loss
94% efficient, ~97% available

13.2 kv

Transformer

0.3% loss
99.7% efficient

480V

Transformer

0.3% loss
99.7% efficient

208V
Power Redundancy to Geo-Level

- Roughly 20% of DC capital costs is power redundancy
- Instead use more, smaller, cheaper, commodity data centers
- Non-bypass, battery-based UPS in the 94% efficiency range
  - ~900kW wasted in 15MW facility (4,500 200W servers)
  - 97% available (still 450kW loss in 15MW facility)
Power Distribution Optimization

• Two additional conversions in server:
  – Power Supply: often <80% at typical load
  – Voltage Regulation Module: ~80% common
  – ~95% efficient available & affordable

• Rules to minimize power distribution losses:
  1. Avoid conversions (Less transformer steps & efficient or no UPS)
  2. Increase efficiency of conversions
  3. High voltage as close to load as possible
  4. Size voltage regulators (VRM/VRDs) to load & use efficient parts
  5. DC distribution potentially a small win (regulatory issues)
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
  – Graceful Degradation Mode to shed workload

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

• Where does the power go?
• Power distribution optimization
• Mechanical systems optimization
• Server & other optimization
  – Cooperative, Expendable, Micro-Slice Servers
  – Improving non-new builds
• Summary
Conventional Mechanical Design

- **Cooling Tower**
- **CWS Pump**
- **A/C Condenser**
- **Heat Exchanger (Water-Side Economizer)**
- **Primary Pump**
- **A/C Evaporator**
- **A/C Compressor**
- **Secondary Pump**
- **Diluted Hot/Cold Mix**
- **Computer Room Air Handler**
- **Air Impeller**
- **Overall Mechanical Losses ~33%**
- **Server fans 6 to 9W each**

- **leakage**
- **cold**
- **Hot**
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 gallons/day at fairly typical 15MW facility
ASHRAE 2008 Recommended

Most data center 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

Dell Servers (Ty Schmitt)

Most data center run in this range

ASHRAE 2008 Recommended Class 1
NEBS (Telco) & Rackable Systems

Psychrometric Chart

St (metric) units
Barometric Pressure 101.325 kPa (Sea level)
based on data from
Carrier Corporation Cat. No. 794-001, dated 1975

Dell Servers (Ty Schmitt)

Most data center run in this range

ASHRAE Allowable Class 1

ASHRAE 2008 Recommended Class 1

NEBS & Rackable CloudRack C2

104°F
Air Cooling

• Allowable component temperatures higher than hottest place on earth
  – Al Aziziya, Libya: 136F/58C (1922)
• It’s only a mechanical engineering problem
  – More air and better mechanical designs
  – Tradeoff: power to move air vs cooling savings
  – Partial recirculation when external air too cold
• Currently available equipment:
  – 40C: Rackable CloudRack C2
  – 35C: Dell Servers

I/O: 5W - 25W
Temp Spec: 50C-60C

Processors/Chipset: 40W - 200W
Temp Spec: 60C-70C

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 Optimization Summary

• Simple rules to minimize cooling costs:
  1. Raise data center temperatures
  2. Tight airflow control, short paths & large impellers
  3. Cooling towers rather than A/C
  4. Air side economization & evaporative cooling
     • outside air rather than A/C & towers
Agenda

• Where does the power go?
• Power distribution optimization
• Mechanical systems optimization
• Server & other optimization
  – Cooperative, Expendable, Micro-Slice Servers
  – Improving non-new builds
• Summary
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/)

<table>
<thead>
<tr>
<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>
</tr>
<tr>
<td>RPS</td>
<td>95.9</td>
<td>75.3</td>
<td>54.3</td>
</tr>
<tr>
<td>Price</td>
<td>$2,371</td>
<td>$500</td>
<td>$685</td>
</tr>
<tr>
<td>Power</td>
<td>295</td>
<td>60</td>
<td>39</td>
</tr>
<tr>
<td>RPS/Price</td>
<td>0.04</td>
<td>0.15</td>
<td>0.08</td>
</tr>
<tr>
<td>RPS/Joule</td>
<td>0.33</td>
<td>1.25</td>
<td>1.39</td>
</tr>
<tr>
<td>RPS/Rack</td>
<td>1918.4</td>
<td>18062.4</td>
<td>13024.8</td>
</tr>
</tbody>
</table>

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

**Update:** New H/W SKU likely will improve numbers by factor of 2. CEMS still a win.

Resource Consumption Shaping

- Essentially yield mgmt 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
  - Server idle to full-load range: ~65% (e.g. 158W to 230W)
  - S3 (suspend) or S5 (off) when server not needed
- Disks come with both IOPS capability & capacity
  - Mix hot & cold data to “soak up” both
- Encourage priority (urgency) differentiation in charge-back model
Existing Builds: Containers

• Existing enterprise deployments often:
  – Very inefficient with PUE in 2 to 3 range
  – Out of cooling, out of power & out of space

• Rather than continue to grow bad facility
  – Drop container on roof or parking lot
  – Convert existing data center to offices or other high value use

• Easy way to get PUE to 1.35 range

http://perspectives.mvdirona.com
Existing Builds: Cloud Services

• Deploy new or non-differentiated workloads to cloud
  – Focus the on-premise facility to differentiated computing that adds value to the business
  – Focus people resources on revenue generating, differentiated IT work

• No upfront capital outlay

• Very high scale, cloud service deployments offer lower costs and can be more efficient
  – Better for environment & lower cost
Summary

- Average DCs have considerable room to improve
- Use tPUE rather PUE to track improvement
- Power & related costs drive infrastructure expenses
  - Don’t use floor space or rack positions as metric
- Server costs still (barely) dominate power
- What to do with existing, inefficient infrastructure
  - Modular data center designs
  - Utility computing
More Information

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

- **Berkeley Above the Clouds**

- **Designing & Deploying Internet-Scale Services**

- **Architecture for Modular Data Centers**

- **Perspectives Blog**
  - [http://perspectives.mvdirona.com](http://perspectives.mvdirona.com)

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