Internet-Scale Service Infrastructure Efficiency

International Symposium on System Architecture

James Hamilton, 2009/6/23
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
  – Hot I/O workloads & NAND flash
  – Resource consumption shaping
  – Work done per joule & per dollar
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

**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
  – Hot I/O workloads & NAND flash
  – Resource consumption shaping
  – Work done per joule & per dollar
Power Distribution

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

2.5MW Generator (180 gal/hr)

IT Load (servers, storage, Net, …)

~1% loss in switch gear & conductors

High Voltage Utility Distribution

0.3% loss
99.7% efficient

13.2kv

115kv

Transformers

13.2kv

13.2kv

Transformers

Transformers

13.2kv

13.2kv

Transformers

480V

208V

0.3% loss
99.7% efficient

6% loss
94% efficient, ~97% available

0.3% loss
99.7% efficient

0.3% loss
99.7% efficient

UPS:
Rotary or Battery

0.3% loss
99.7% efficient

220V

99.7% efficient

8% distribution loss
\[0.997^3 \cdot 0.94 \cdot 0.99 = 92.2\% \]
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 (fewer transformer steps & efficient UPS)
  3. Increase efficiency of conversions
  4. High voltage as close to load as possible
  5. Size VRMs & 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
  – Hot I/O workloads & NAND flash
  – Resource consumption shaping
  – Work done per joule & per dollar
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
- Diluted Hot/Cold Mix
- Overall Mechanical Losses ~33%
- Computer Room Air Handler
- Air Impeller

Blow down & Evaporative Loss for 15MW facility: ~360,000 gal/day

2009/6/16

http://perspectives.mvdirona.com
Most data centers run in this range.
ASHRAE Allowable

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

Dell Servers (Ty Schmitt)
NEBS (Telco) & Rackable Systems

Most data center run in this range

ASHRAE Allowable Class 1

ASHRAE 2008 Recommended Class 1

Dell Servers (Ty Schmitt)

NEBS & Rackable CloudRack C2

104F
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

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
  – Hot I/O workloads & NAND flash
  – Resource consumption shaping
  – Work done per joule & per dollar
Disk sequential BW lagging DRAM and CPU
Disk random access BW growth ~10% of sequential
Conclusion: Storage Chasm widening requiring larger memories & more disks

Source: Dave Patterson with James Hamilton updates

http://perspectives.mvdirona.com
Memory to Disk Chasm

• Disk I/O rates grow slowly while CPU data consumption grows near Moore pace
  – Random read 1TB disk: 15 to 150 days*

• Sequentialize workloads
  – Essentially the storage version of cache conscious algorithms
    – e.g. map/reduce
  – Disks arrays can produce acceptable aggregate sequential bandwidth

• Redundant data: materialized views & indexes
  – Asynchronous maintenance
  – Delta or stacked indexes (from IR world)

• Distributed memory cache (remote memory “closer” than disk)

• I/O Cooling: Blend hot & cold data (using HDD)

• I/O concentration: partition hot & cold (SSD & HDD mix)

* Tape is Dead, Disk is Tape, Flash is Disk, Ram Locality is King (Jim Gray)

2009/6/23
http://perspectives.mvdirona.com
# Case Study: TPC-C with SSD

<table>
<thead>
<tr>
<th>Slot</th>
<th>Controller</th>
<th>Disks</th>
<th>Capacity</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Dell PERC5i</td>
<td>6x73GB, 15K, SAS</td>
<td>RAID 10</td>
<td>Disk 6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>279.99GB</td>
<td>260GB</td>
</tr>
<tr>
<td>3</td>
<td>Dell PERC6/E</td>
<td>15x36GB, 15K, SAS</td>
<td>RAID 0</td>
<td>Disk 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15x36GB, 15K, SAS</td>
<td>RAID 0</td>
<td>Disk 3</td>
</tr>
<tr>
<td>4</td>
<td>Dell PERC6/E</td>
<td>15x36GB, 15K, SAS</td>
<td>RAID 0</td>
<td>Disk 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15x36GB, 15K, SAS</td>
<td>RAID 0</td>
<td>Disk 5</td>
</tr>
<tr>
<td>6</td>
<td>Dell PERC6/E</td>
<td>15x73GB, 15K, SAS</td>
<td>RAID 0</td>
<td>Disk 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15x73GB, 15K, SAS</td>
<td>RAID 0</td>
<td>Disk 1</td>
</tr>
</tbody>
</table>

- **98 HDD total**
  - 90 data disks (primarily random access)
  - 8 log & O/S disks (primarily sequential access)

- **Compute HDD/SSD cross-over using fictitious SSD**
  - 128GB SSD @ 7k IOPS

- **90 HDD to store 2,464GB (short stroked)**
  - 106GB static & 2,357GB dynamic (60 day rule)
  - 90 disk HDD budget: $26,910 (disks $299 each)
  - Requires **20 SSDs to support @ up to $1,346 each**

- **Static content only (drop 60 day rule)**
  - Conservatively estimate 45k IOPS
  - Used 90 short stroked disks at 500 IOPS each
  - Requires **7 SSDs at up to $3,844 (easy)**

- **Very hot I/O workloads a win on SSD**

---


2009/6/23

http://perspectives.mvdirona.com
Summary

• CPU optimizations are always welcome but the biggest design & optimization problems today are at the datacenter level

• 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
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