The “Device” Layer of Energy Systems: Datacenters and Microgrids

Randy H. Katz
University of California, Berkeley
Berkeley, CA 94720-1776

Datalink Layer: The Network Analogy

- DATA LINK
  - means of connecting one location to another to transmit/receive digital information
  - can also refer to a set of electronics assemblies, consisting of a transmitter and a receiver and the interconnecting data telecommunication circuit. Governed by a link protocol enabling digital data to be transferred from a data source to a data sink
- By analogy, the interface between the aggregated load, i.e., the machine room, the building, and the energy system
- Network interface device (NID): device that serves as the demarcation point between the carrier's local loop and the customer's premises wiring
  - Weatherproof box
  - Protection circuits
  - Metering and rate limiting (on the service provider side in telecomms, on the consumer side for power)
Energy Proportional Computing


It is surprisingly hard to achieve high levels of utilization of typical servers (and your home PC or laptop is even worse)

Figure 1. Average CPU utilization of more than 5,000 servers during a six-month period. Servers are rarely completely idle and seldom operate near their maximum utilization, instead operating most of the time at between 10 and 50 percent of their maximum

Energy Efficiency = Utilization/Power


Doing nothing well … NOT!

Figure 2. Server power usage and energy efficiency at varying utilization levels, from idle to peak performance. Even an energy-efficient server still consumes about half its full power when doing virtually no work.
Energy Proportional Computing


Energy Efficiency = Utilization/Power

Figure 4. Power usage and energy efficiency in a more energy-proportional server. This server has a power efficiency of more than 80 percent of its peak value for utilizations of 30 percent and above, with efficiency remaining above 50 percent for utilization levels as low as 10 percent.
Datacenter Arms Race

• Amazon, Google, Microsoft, Yahoo!, … race to build next-gen mega-datacenters
  – Industrial-scale Information Technology
  – 100,000+ servers
  – Located where land, water, fiber-optic connectivity, and cheap power are available

• E.g., Microsoft Quincy
  – 43600 sq. ft. (10 football fields), sized for 48 MW
  – Also Chicago, San Antonio, Dublin @ $500M each

• E.g., Google:
  – The Dalles OR, Pryor OK, Council Bluffs, IW, Lenoir, NC, Goose Creek, SC
Google Oregon Datacenter

Performance vs. Distance in the Datacenter

[Graph showing performance metrics against distance]
Energy Use in Datacenters

- UPS: 18%
- Chiller: 33%
- IT Equipment: 30%
- CRAC: 9%
- PDU: 5%
- Lighting: 1%
- Transformers / Switchgear: 1%

Datacenter Energy Overheads

Ideal Machine Room Cooling
Hot and Cold Aisles

- Interstitial Ceiling Space
- Air Barrier (Plastic Sheet)
- CRAH
- Supply Air: 70-75°F
- Return Air: 95-100°F
- Cold Aisle
- Raised Floor
- Air Barrier (Melamine Board)
- Hot Aisle
Real Machine Rooms
More Complicated

DC Infrastructure Energy Efficiencies
Cooling (Air + Water movement) + Power Distribution
Containerized Datacenter
Mechanical-Electrical Design

Microsoft Chicago Datacenter

Modular Datacenters

• Just add power, chilled water, & network
• Drivers of move to modular
  — Faster pace of infrastructure innovation
  — Efficient scale-down
    • Driven by latency & jurisdictional restrictions
  — Service-free, fail-in-place model
    • 20-50% of system outages caused by admin error
    • Recycle as a unit
  — Incremental data center growth
    • Transfer fixed to variable cost
• Microsoft Chicago deployment: entire first floor with ½ MW containers

James Hamilton, Amazon
Containerized Datacenters

- Sun Modular Data Center
  - Power/cooling for 200 KW of racked HW
  - External taps for electricity, network, water
  - 7.5 racks: ~250 Servers, 7 TB DRAM, 1.5 PB disk
Google Container

Power Usage Effectiveness Rapidly Approaching 1!

Bottom-line: the frontier of DC energy efficiency IS the IT equipment

Doing nothing well becomes incredibly important

Figure 4. Our San 1.28, which trans datacenter built
• Typical structure 1MW Tier-2 datacenter
• Reliable Power
  – Mains + Generator
  – Dual UPS
• Units of Aggregation
  – Rack (10-80 nodes)
  – PDU (20-60 racks)
  – Facility/Datacenter

Nameplate vs. Actual Peak

<table>
<thead>
<tr>
<th>Component</th>
<th>Peak Power</th>
<th>Count</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>40 W</td>
<td>2</td>
<td>80 W</td>
</tr>
<tr>
<td>Memory</td>
<td>9 W</td>
<td>4</td>
<td>36 W</td>
</tr>
<tr>
<td>Disk</td>
<td>12 W</td>
<td>1</td>
<td>12 W</td>
</tr>
<tr>
<td>PCI Slots</td>
<td>25 W</td>
<td>2</td>
<td>50 W</td>
</tr>
<tr>
<td>Mother Board</td>
<td>25 W</td>
<td>1</td>
<td>25 W</td>
</tr>
<tr>
<td>Fan</td>
<td>10 W</td>
<td>1</td>
<td>10 W</td>
</tr>
<tr>
<td>System Total</td>
<td></td>
<td></td>
<td>213 W</td>
</tr>
</tbody>
</table>

Nameplate peak
Measured Peak
(Power-intensive workload) 145 W

In Google’s world, for given DC power budget, deploy as many machines as possible


Typical Datacenter Power

Clusters driven to modest utilization/67% power
Racks can be driven to high utilization/95% power

Power-aware allocation of resources can achieve higher levels of utilization – harder to drive a cluster to high levels of utilization than an individual rack

Better to have one computer at 50% utilization than five computers at 10% utilization: Save $ via Consolidation (& Save Power)

**SPECpower:**
- Two 3.0-GHz Xeons, 16 GB DRAM, 1 Disk
- One 2.4-GHz Xeon, 8 GB DRAM, 1 Disk

- 50% utilization → 85% Peak Power
- 10% → 65% Peak Power
- Save 75% power if consolidate & turn off
  - 1 computer @ 50% = 225 W
  - v. 5 computers @ 10% = 870 W

**Atoms are Quite Better at Doing Nothing Well**

<table>
<thead>
<tr>
<th>Name</th>
<th>Xeon 1.54GHz</th>
<th>Atom 1.33GHz</th>
<th>Atom 1.27GHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>2.30GHz</td>
<td>1.60GHz</td>
<td>1.60GHz</td>
</tr>
<tr>
<td>Cache</td>
<td>2x6MB</td>
<td>2x512KB</td>
<td>512KB</td>
</tr>
<tr>
<td>CPU</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Threads/Core</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>RAM</td>
<td>16GB</td>
<td>2GB</td>
<td>1GB</td>
</tr>
<tr>
<td>Storage</td>
<td>15k SAS</td>
<td>5.4k SATA</td>
<td>SSD (low-end)</td>
</tr>
</tbody>
</table>

Measured Power in Soda Hall Machine Rooms
Climate Savers Initiative

• Improving the efficiency of power delivery to computers as well as usage of power by computers
  – Transmission: 9% of energy is lost before it even gets to the datacenter
  – Distribution: 5-20% efficiency improvements possible using high voltage DC rather than low voltage AC
  – Chill air to mid 50s vs. low 70s to deal with the unpredictability of hot spots

Datacenter Power Efficiencies

• 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)

• Two interesting approaches:
  – 480VAC (or higher) to rack & 48VDC (or 12VDC) within
  – 480VAC to PDU and 277VAC to load
    • 1 leg of 480VAC 3-phase distribution
Microslice Servers

- CEMS: Cooperative Expendable Micro-Slice Servers
  - Correct system balance problem with less-capable CPU
  - Too many cores, running too fast, for memory, bus, I/O, ...
- Joint project with Rackable Systems

<table>
<thead>
<tr>
<th></th>
<th>CPU Idle%</th>
<th>RPS</th>
<th>Price</th>
<th>Power</th>
<th>RPS/Price</th>
<th>RPS/Joole</th>
<th>RPS/RU</th>
</tr>
</thead>
<tbody>
<tr>
<td>System-X</td>
<td>56%</td>
<td>96.92</td>
<td>$2,371</td>
<td>205</td>
<td>0.04</td>
<td>0.32535254</td>
<td>1918.4</td>
</tr>
<tr>
<td>Athlon 4850e</td>
<td>57%</td>
<td>75.28</td>
<td>$5000</td>
<td>60</td>
<td>0.12</td>
<td>1.254333333</td>
<td>18062.4</td>
</tr>
<tr>
<td>Athlon 3400e</td>
<td>57%</td>
<td>54.27</td>
<td>$5685</td>
<td>39</td>
<td>0.08</td>
<td>1.391538462</td>
<td>13024.8</td>
</tr>
<tr>
<td>Athlon 2000+</td>
<td>61%</td>
<td>17</td>
<td>$5900</td>
<td>33</td>
<td>0.03</td>
<td>0.535353535</td>
<td>4080</td>
</tr>
</tbody>
</table>

- CEMS V2 Comparison:
  - Work Done/$: +372%
  - Work Done/Joule: +385%
  - Work Done/RU: +941%

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

James Hamilton, Amazon

Processor-Storage Balance

Rack Unit: Proper Speed/Cost given utilization challenges

James Hamilton, Amazon
Rethinking Rack Power Distribution

- Replace AC power supply in servers with DC-DC converters to generate required voltages
- Battery capacity per rack to simplify design of the DC-DC converter, centralizing the charge controller and energy sharing function in the IPS
- Distributed DC-DC converters provide regulation at the load

Google Computing Node
Google Computing Node

Power/Cooling Issues
Thermal Image of Typical Cluster Rack

DC Networking and Power

- Within DC racks, network equipment often the “hottest” components in the hot spot
- Network opportunities for power reduction
  - Transition to higher speed interconnects (10 Gbs) at DC scales and densities
  - High function/high power assists embedded in network element (e.g., TCAMs)
- Recent Work:
DC Networking and Power

- 96 x 1 Gbit port Cisco datacenter switch consumes around 15 kW -- approximately 100x a typical dual processor Google server @ 145 W
- High port density drives network element design, but such high power density makes it difficult to tightly pack them with servers
- Alternative distributed processing/communications topology under investigation by various research groups

Datacenter Energy Efficiency Roadmap

A "research roadmap" developed for the California Energy Commission outlines key areas for energy efficiency research, development, and demonstration
Conventional Datacenter Mechanical Design

Utilization and Efficiency

- **PUE: Power Utilization Efficiency**
  - Total facility power / Critical load
  - Good conventional data centers ~1.7 (a few are better)
  - Poorly designed enterprise data centers as bad as 3.0
- Assume a PUE of 1.7 and see where it goes:
  - 0.3 (18%): Power distribution
  - 0.4 (24%): Mechanical (cooling)
  - 1.0 (58%): Critical Load (server efficiency & utilization)
- Low efficiency DCs spend proportionally more on cooling
  - 2 to 3x efficiency improvements possible by applying modern techniques
  - Getting to 4x and above requires server design and workload management techniques
Power Distribution Optimization

• Simple rules to minimize power distribution losses in priority order
  1. Avoid conversions (indirect UPS or no UPS)
  2. Increase efficiency of conversions
  3. High voltage as close to load as possible
  4. Size board voltage regulators to load and use high quality
  5. Direct Current small potential win (but regulatory issues)

• Two interesting approaches:
  – 480VAC to rack and 48VDC (or 12VDC) within rack
  – 480VAC to PDU and 277VAC (1 leg of 480VAC 3-phase distribution) to each server

Energy Proportionality?

Inherent proportionality inefficiencies throughout the energy distribution system
Typical AC Distribution Today

- 480 Volt AC
- DC/AC
- UPS
- PDU

380 V DC after first stage conversion

Facility-level DC Distribution

- 480 Volt AC
- DC/DC
- DC UPS or Rectifier

380 V DC delivered directly to the server at the same point as in AC powered server
- Eliminates DC-AC conversion at the UPS and the AC-DC conversion in the server
- Less equipment needed
Rack-level DC Distribution

480 Volt AC

AC/DC

DC/AC

UPS

PDU

380 VDC

AC/DC

DC/DC

PSU

12 V

VRM 5 V

VRM 3.3 V

VRM 12 V

VRM 1.8 V

VRM 0.9 V

Loads using Legacy Voltages

Loads using Silicon Voltages

Rack

Server

AC System Loss Compared to DC

7-7.3% measured improvement

2-5% measured improvement

480 VAC Bulk Power Supply

UPS

PDU

DC/DC

PSU

VRM 5 V

VRM 3.3 V

VRM 12 V

VRM 1.8 V

VRM 0.9 V

Loads using Legacy Voltages

Loads using Silicon Voltages

Server

LBNL
Energy Distributions Savings for Typical Datacenter

- 20% or more facility level energy savings:
  - Redundant UPS and server power supplies operate at reduced efficiency
  - Cooling loads would be reduced
  - Demonstration comparisons were against “best in class” systems which performed better than typical systems we benchmarked
- Further optimization of conversion devices/voltages is possible

Mechanical Optimization

- Simple rules to minimize cooling costs:
  1. Raise data center temperatures
  2. Tight control of airflow with short paths
     ~1.4 to perhaps 1.3 PUE with the first two alone
  3. Air side economization (essentially, open the window)
  4. Water side economization (don’t run A/C)
  5. Low grade, waste heat energy reclamation
- Best current designs have water cooling close to the load but don’t use direct water cooling
  - Lower heat densities could be 100% air cooled but density trends suggest this won’t happen
More Efficient Air Flow

Critical Load Optimization

- Power proportionality is great, but “off” still wins by large margin
  - Today: Idle server ~60% power of full load
  - Off required changing workload location
  - Industry secret: “good” data center server utilization around ~30%
    (many much lower)

- What limits 100% dynamic workload distribution?
  - Networking constraints (e.g. VIPs can’t span L2 nets, manual config, etc.)
  - Data Locality
    - Hard to move several TB and workload needs to be close to data
  - Workload management:
    - Scheduling work over resources optimizing power with SLA constraint

- Server power management still interesting
  - Most workloads don’t fully utilize all server resources
  - Very low power states likely better than off (faster)
Bringing Resources On-/Off-line

- Save power by taking DC “slices” off-line
  - Resource footprint of applications hard to model
  - Dynamic environment, complex cost functions require measurement-driven decisions -- opportunity for statistical machine learning
  - Must maintain Service Level Agreements, no negative impacts on hardware reliability
  - Pervasive use of virtualization (VMs, VLANs, VStor) makes feasible rapid shutdown/migration/restart

- Recent results suggest that conserving energy may actually improve reliability
  - MTTF: stress of on/off cycle vs. benefits of off-hours

Datacenter Optimization Summary

- Some low-scale DCs as poor as 3.0 PUE
- Workload management has great potential:
  - Over-subscribe servers and use scheduler to manage
  - Optimize workload placement and shut servers off
    - Network, storage, & mgmt system issues need work
- 4x efficiency improvement from current generation high-scale DCs (PUE ~1.7) is within reach without technology breakthrough
Microsoft’s Chicago Modular Datacenter

- 24000 sq. m housing 400 containers
  - Each container contains 2500 servers
  - Integrated computing, networking, power, cooling systems
- 300 MW supplied from two power substations situated on opposite sides of the datacenter
- Dual water-based cooling systems circulate cold water to containers, eliminating need for air conditioned rooms

The Million Server Datacenter
Microgrids

- Distributed generation generates electricity from many small energy sources (3-10000 kW)
  - on-site generation, dispersed generation, embedded generation, decentralized generation, decentralized energy or distributed energy
  - reduces amount of energy lost in transmitting electricity—generated close to where it is used. Also reduces size and number of power lines that must be constructed

http://en.wikipedia.org/wiki/Microgrid

Microgrids

- Single self-contained entity to the distribution grid
- During utility grid disturbance:
  - separate and isolate from the utility seamlessly, with little or no disruption to the loads within the microgrid (e.g., no impacts on power quality)
  - when the utility grid returns to normal, the microgrid automatically resynchronizes and reconnects itself to the grid, in an equally seamless fashion

Microgrids

- DER: When deployed in large numbers, affect on electricity grid reliability? Systems approach needed!
  - Autonomous generation, storage, and load control technologies, located at customer premises and operated for the customer’s benefit. E.g., microturbines, fuel cells, photovoltaic systems, and traditional internal combustion engines.
  - Control and dispatch for DER; ensure safely and protection of the grid; role of power electronic interfaces in connecting DER to the grid.
Microgrids

More Resources

- http://www.datacenterknowledge.com/
- http://perspectives.mvdirona.com/
- http://www.youtube.com/results?search_query=google%27s+green+datacenter&search_type=&aq=f
Summary and Conclusions

• Energy Consumption in IT Equipment
  – Energy Proportional Computing and “Doing Nothing Well”
  – Management of Processor, Memory, I/O, Network to maximize performance subject to power constraints
  – Internet Datacenters and Containerized Datacenters: New packaging opportunities for better optimization of computing + communicating + power + mechanical