“The Big Switch,” Redux

“A hundred years ago, companies stopped generating their own power with steam engines and dynamos and plugged into the newly built electric grid. The cheap power pumped out by electric utilities didn’t just change how businesses operate. It set off a chain reaction of economic and social transformations that brought the modern world into existence. Today, a similar revolution is under way. Hooked up to the Internet’s global computing grid, massive information-processing plants have begun pumping data and software code into our homes and businesses. This time, it’s computing that’s turning into a utility.”
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
  - 43,600 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

2020 IT Carbon Footprint

- 2007 Worldwide IT carbon footprint: 2% = 830 m tons CO2
- Comparable to the global aviation industry
- Expected to grow to 4% by 2020

- Total emissions: 1.4bn tonnes CO2 equivalent

- 2020 IT Carbon Footprint

"SMART 2020: Enabling the Low Carbon Economy in the Information Age", The Climate Group
2020 IT Carbon Footprint

"SMART 2020: Enabling the Low Carbon Economy in the Information Age", The Climate Group

Fig. 2.3 The global footprint by subsector
Emissions by geography
% of US$GDP

Computers + Net + Storage + Power + Cooling

Energy Expense Dominates

Energy Use In Datacenters
Energy Use In Datacenters

- 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

2020 IT Carbon Footprint

- Fig. 4.2 Composition of data centre footprint

Utilization and Efficiency

Where do the $$$'s go?

- Assumptions:
  - Facility: $200M for 15MW facility (5-year amort.
  - Servers: $24/each, roughly 50,000 (3-year amort.
  - Average server power draw at 30% utilization: 80W
  - Commercial Power: 50.07$/kWh

- Monthly Costs

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

James Hamilton, Amazon
Overview

- Data Center Overview
- Per-node Energy
- Power Distribution
- Cooling and Mechanical Design

Nameplate vs. Actual Peak

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

Energy Proportional Computing


CPU energy improves, but what about the rest of the server architecture?

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 3. CPU contribution to total server power for two generations of Google servers at peak performance (the first two bars) and for the later generation at idle (the rightmost bar).

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 Proportional Computing


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


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

“Power” of Cloud Computing

- SPECpower: two best systems
  - 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
  - 5 computers @ 10% = 870 W

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

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
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.


**Aside: Disk Power**

**IBM Microdrive (1inch)**
- writing 300mA (3.3V) 1W
- standby 65mA (3.3V) .2W

**IBM TravelStar (2.5inch)**
- read/write 2W
- spinning 1.8W
- low power idle .65W
- standby .25W
- sleep .1W
- startup 4.7 W
- seek 2.3W

**Spin-down Disk Model**

**Disk Spindown**

- Disk Power Management – Oracle (off-line)
- Disk Power Management – Practical scheme (on-line)
Spin-Down Policies

- Fixed Thresholds
  - $T_{out} = \text{spin-down cost s.t. } 2 \cdot E_{\text{transition}} = P_{\text{spin}} \cdot T_{out}$
- Adaptive Thresholds: $T_{out} = f(\text{recent accesses})$
  - Exploit burstiness in $T_{\text{idle}}$
- Minimizing Bumps (user annoyance/latency)
  - Predictive spin-ups
- Changing access patterns (making burstiness)
  - Caching
  - Prefetching

Dynamic Spindown
Helmbold, Long, Sherrod (MOBICOM96)

- Dynamically choose a timeout value as function of recent disk activity
- Based on machine learning techniques (for all you AI students!)
- Exploits bursty nature of disk activity
- Compares to (related previous work)
  - best fixed timeout with knowledge of entire sequence of accesses
  - optimal - per access best decision of what to do
  - competitive algorithms - fixed timeout based on disk characteristics
  - commonly used fixed timeouts

Spindown and Servers

- The idle periods in server workloads are too short to justify high spinup/down cost of server disks [ISCA’03][ISPASS’03] [ICS’03]
  - IBM Ultrastar 36Z15 -- 135J/10.9s
- Multi-speed disk model [ISCA’03]
  - RPMs: multiple intermediate power modes
  - Smaller spinup/down costs
  - Be able to save energy for server workloads
- BUT... many energy/load optimizations have similar tradeoffs/algorithms

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)
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).

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.
Overview

- Data Center Overview
- Per-node Energy
- Power Distribution
- Cooling and Mechanical Design

Datacenter Power

- 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

Datacenter Power Efficiencies

- Power conversions in server
  - Power supply (<80% efficiency)
  - Voltage regulation modules (80% common)
  - Better available (95%) and inexpensive
- 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

480 Volt AC Distribution Today

- 380 V DC after first stage conversion

Facility-level DC Distribution

- 380V 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

- 380V 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

AC System Loss Compared to DC

- 7-7.3% measured improvement
- 2-5% measured improvement
Power Redundancy

- 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)

Why built-in batteries?

- Building the power supply into the server is cheaper and means costs are matched directly to the number of servers
- Large UPSs can reach 92 to 95 percent efficiency vs. 99.9 percent efficiency for server mounted batteries

Overview

- Data Center Overview
- Per-node Energy
- Power Distribution
- Cooling and Mechanical Design
Mechanical Optimization

- Simple rules to minimize cooling costs:
  - Raise data center temperatures
  - Tight control of airflow with short paths
  - ~1.4 to perhaps 1.3 PUE with the first two alone
  - Air side economization (essentially, open the window)
  - Water side economization (don’t run A/C)
  - 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

Ideal Machine Room Cooling
Hot and Cold Aisles

Real Machine Rooms
More Complicated
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

Modular Datacenters

- Just add power, chilled water, & network
- Drivers of move to modular
  - Faster pace of infrastructure innovation
  - Power & mechanical innovation to 3 year cycles
  - 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

Containerized Datacenter Mechanical-Electrical Design

James Hamilton, Amazon
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

Google

- Since 2005, its data centers have been composed of standard shipping containers--each with 1,160 servers and a power consumption that can reach 250 kilowatts
- Google server was 3.5 inches thick--2U, or 2 rack units, in data center parlance. It had two processors, two hard drives, and eight memory slots mounted on a motherboard built by Gigabyte

Google's PUE

- In the third quarter of 2008, Google's PUE was 1.21, but it dropped to 1.20 for the fourth quarter and to 1.19 for the first quarter of 2009 through March 15
- Newest facilities have 1.12
Summary and Conclusions

- Energy Consumption in IT Equipment
  - Energy Proportional Computing
  - Inherent inefficiencies in electrical energy distribution
- Energy Consumption in Internet Datacenters
  - Backend to billions of network capable devices
  - Enormous processing, storage, and bandwidth supporting applications for huge user communities
  - Resource Management: Processor, Memory, I/O, Network to maximize performance subject to power constraints: "Do Nothing Well"
  - New packaging opportunities for better optimization of computing + communicating + power + mechanical

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
- The Uptime Institute reports that the average data center Power Usage Effectiveness is 2.0 (smaller is better). What this number means is that for every 1W of power that goes to a server in an enterprise data center, a matching watt is lost to power distribution and cooling overhead. Microsoft reports that its newer designs are achieving a PUE of 1.22 (Out of the box paradox...). All high scale services are well under 1.7 and most, including Amazon, are under 1.5.