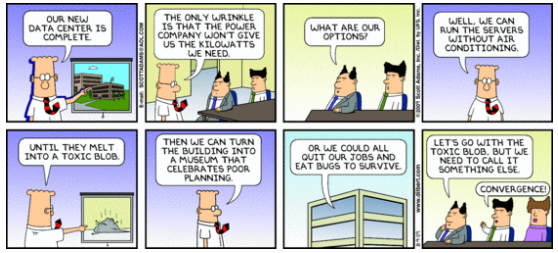
Advanced Computer Architecture

Datacenters

Fall 2016



Pejman Lotfi-Kamran

Adapted from slides originally developed by Profs. Hill, Hoe, Falsafi and Wenisch of CMU, EPFL, Michigan, Wisconsin

Where Are We?

Fr	Sa	Su	Mo	Tu
	27-Shahrivar		29-Shahrivar	
	3-Mehr		5-Mehr	
	10-Mehr		12-Mehr	
	17-Mehr		19-Mehr	
	24-Mehr		26-Mehr	
	1-Aban		3-Aban	
	8-Aban		10-Aban	
	15-Aban		17-Aban	
	22-Aban		24-Aban	
	29-Aban		1-Azar	
	6-Azar		8-Azar	
	13-Azar		15-Azar	
	20-Azar		22-Azar	
	27-Azar		29-Azar	
	4-Dey		6-Dey	

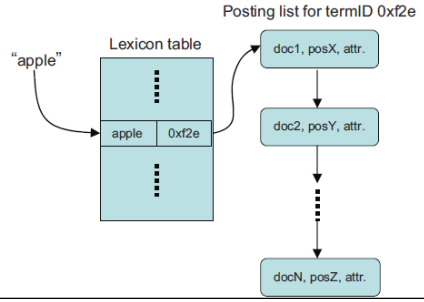
- This Lecture
  - Data Center (2)

Software-based fault tolerance

- Fault tolerance enables the service to continue to operate properly in the event of failure in some of its components.
- Key idea: hardware faults are detected and reported to software in a timely manner, software takes appropriate action to manage the fault.
- Hides complexity from application-level software.
- Servers need not be expected to run at all costs
  - Can customize reliability level to maximize overall cost efficiency.
  - Hardware and software upgrades can be easily made.
  - More generally, allows for flexibility in hardware choices.

Online workload: Web Search

- Search algorithm traverse posting lists for each term in the query until it finds all documents contained in all three posting lists.
- Ranks the documents (e.g. by computing PageRank score).



Online workload: Web Search

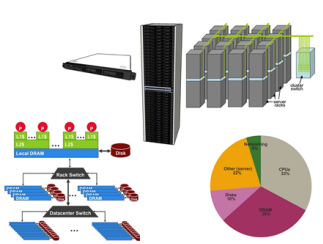
- Algorithm may need to run across a few thousand machines since index is huge.
- Index can be split into load-balanced subfiles and distributed across the machines.
- Total user-perceived latency needs to be short.
- High throughput is also important to support many queries.
- Relatively small network bandwidth requirements because size of queries are small.

Offline workload: Scholar Article Similarity

- Similarity analysis by co-citation: count every article that cites articles A and B as a vote for the similarity between A and B.
- Input data (a citation graph) is divided into hundreds of files.
- Program is distributed over hundreds of servers.
- Traffic is streaming.
- Latency is less important than overall parallel efficiency.

Data Center Power Consumption

Computing/IT equipment



Only 30% of energy overhead

What else consumes power in a datacenter?

Power distribution units, cooling system, etc.

The Components of a Typical Datacenter

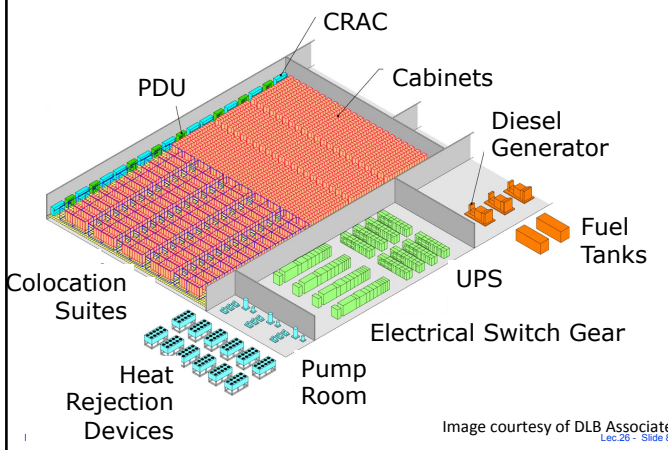
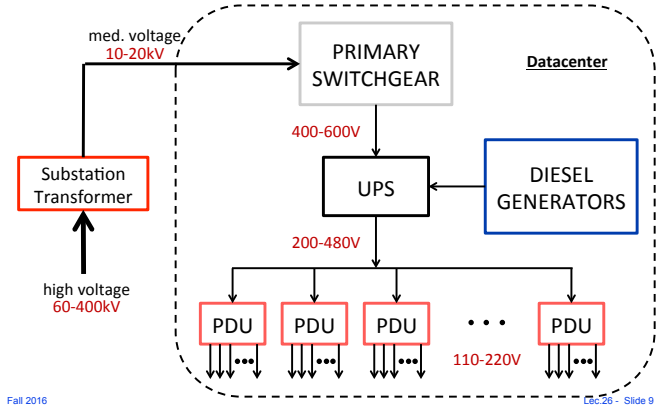


Image courtesy of DLB Associates

Datacenter Power Network



Datacenter Power Network

PRIMARY SWITCHGEAR
- Breakers for fault protection
- Transformers for voltage scaling

UPS
- Transfer switch for selecting power source
- Batteries/flywheels (AC-DC-AC double conversion)
- Conditions incoming power feed (removes spikes, distortions)

PDU
- Similar to breaker panels in residential houses
- Break a higher-voltage feed into several 110- or 220-V circuits that feed servers
- Typically handles 75-225kW of load (each circuit handles up to 6kW)
- Often provide additional redundancy

Old School Datacenter Cooling System

- Datacenter floor is raised 2-4ft above the concrete floor
- Primary use of under-floor area is to distribute cool air to the server racks
- Also used for routing power/network cables to the racks

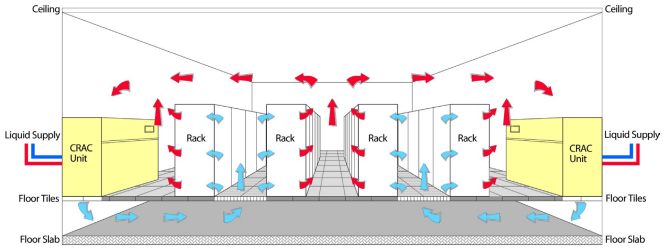


Image courtesy of DLB Associates

Old School:
CRAC (Computer Room Air Conditioning) Units

- Consist of coils pumping a liquid coolant and fans pushing air through these coils
- Redundant pumps circulate cold coolant to CRACs and warm coolant back to chiller or cooling tower
- Incoming coolant is at 12-14°C
- Air exits CRACs at 16-20°C
- Cold aisle (server intake) temperature of 18-22°C

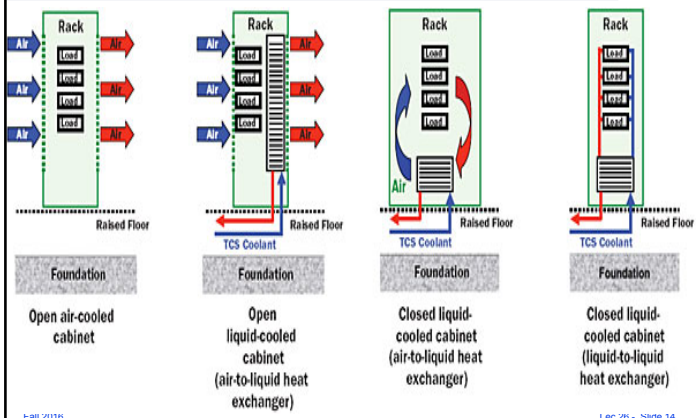
Airflow Considerations

- Air flow through an aisle can be regulated by adjusting the number of perforated tiles along that aisle
- Cold airflow out of tiles in an aisle has to match the horizontal airflow through servers in the racks (avoiding recirculation)

In-Rack Cooling

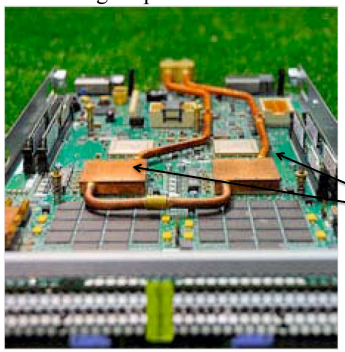
- In-rack cooler adds air-to-water heat exchanger at the back of a rack
- Can replace or supplement CRAC units
- Disadvantage: Chilled water has to be brought to the racks
  - Concerns: plumbing cost, leaky couplings

Chilled-Water Cooling

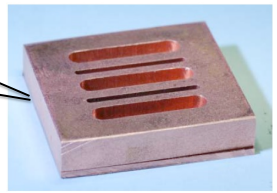


AQUASAR
Water-Cooled Blades

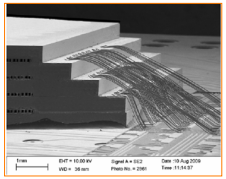
Cooling loop with two CPU's

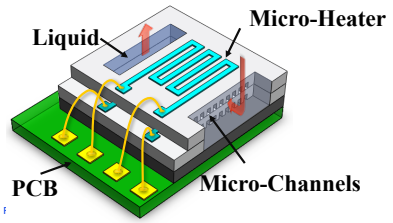


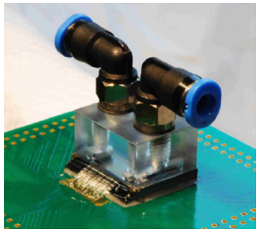
Cooling element without its cover plate and piping ports



CMOSAIC: Integrated 3D IC Computing/Cooling

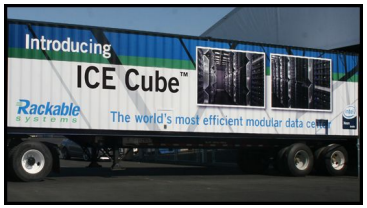






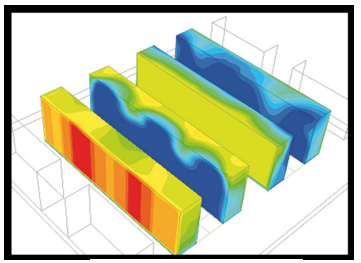
Container-based Datacenters

- Place server racks into a standard shipping container and integrate heat exchange and power distribution
- Typically achieve very high energy efficiency ratings



Holistic Datacenter Power Management

- Synergistic power/load balancing
- Real-time monitoring of 5K servers
- Fine-grain power/thermal sensors
- Achieved 30% overall reductions in power



CREDIT SUISSE

Datacenter Energy Efficiency

- Energy efficiency (DCPE):
  Amount of computational work performed divided by the total energy used in the process
- No actual standard metric exists
- Could run a standard datacenter-wide workload (SPEC/TPC benchmark) and measure total power consumption
- Hard to measure in practice

How to Measure Datacenter Energy Efficiency

- Factor DCPE into three components that can be independently measured and optimized by the appropriate engineering disciplines:
  a) A facility term
  b) A server energy conversion term
  c) The energy efficiency of computation

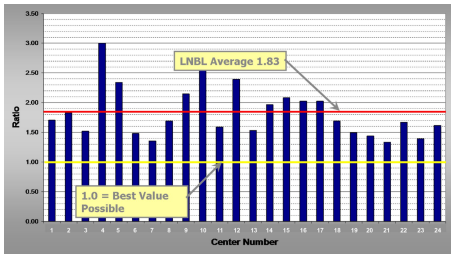
$$\text{Efficiency} = \frac{\text{Computation}}{\text{Total Energy}} = \left(\frac{1}{\text{PUE}}\right) \times \left(\frac{1}{\text{SPUE}}\right) \times \left(\frac{\text{Computation}}{\text{Total Energy to Electronic Components}}\right)$$

Power Usage Effectiveness (PUE)

- PUE: Ratio of total DC power to IT power
  - Power consumed by actual computing equipment
  - PUE=2: for each watt of power used to power IT equipment, one watt used for cooling, power distribution, etc.
- PUE decreases towards 1 as DC gets more efficient.

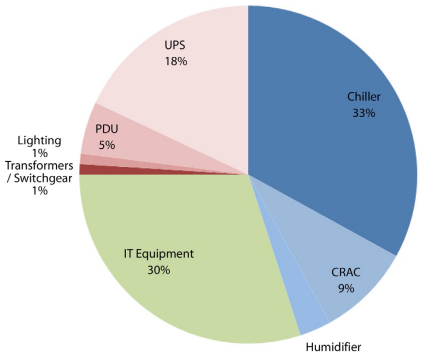
PUE Statistics

- [2006] About 85% of datacenters estimated to have PUE > 3.0
- [2006] Only about 5% of datacenters estimated to have PUE ≤ 2.0
- [2007] Average PUE of approximately 2.0 for 22 datacenters



Update: LBNL survey of PUE of 24 datacenters, 2007 [Greenberg et al.]

Breakdown of Datacenter Energy Overheads



Sources of Efficiency Loss

- Transformer/Switchgear very efficient (<0.5% loss)
- AC-DC-AC double conversion in UPS is mostly responsible for 6-12% loss (more if lightly loaded)
- Long power cables to racks cause 1-3% loss
- Cool/warm air circulation consumes lot of fan power
- Common practice of keeping datacenters at temperatures much colder than necessary (20°C) requires chilled water at 10°C

Improving PUE

- Raise cold aisle temperature to 25-27°C instead of the traditional setting of 20°C
  - No server/network equipment actually needs 20°C intake
- Effective management of warm exhaust heat
  - Separate hot and cold aisles
  - Container-based data centers (compact volume)
    - Per server 12V DC UPS (99.99% efficient)
- Google's container-based datacenter achieved state-of-the-art PUE of 1.24 in 2008.

SPUE

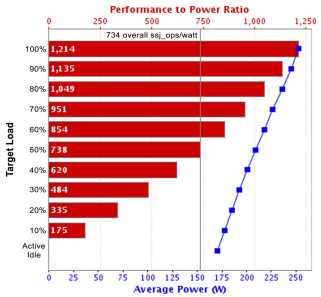
- SPUE: Ratio of total server input power to its useful power
- Useful power is power consumed by components involved directly in the computation:
  - Includes motherboard, disks, CPUs, DRAM, I/O cards, etc.
  - Excludes all losses in power supplies, VRMs, fans
- No standard metric exists, but is being worked on
- Most SPUE ratios are 1.6-1.8 (state of the art should be < 1.2)
- TPUE = PUE * SPUE
  - Average today is > 3.2!
  - State of the art ≈ 1.2*1.2 = 1.44 (70% efficiency)

Efficiency of Computing

- Hardest to measure objectively (given the general purpose nature of computing systems)
- Measuring the value obtained from the energy spent in computing can be useful for
  a) Comparing relative efficiencies of two WSCs
  b) Guide design choices for new systems
- If the goal is (a), then benchmarking for this purpose is hard due to application diversity and workload churn.
- No cluster-level benchmarks currently exist
  - Existing server-level benchmarks (Joulesort, SPECpower) can be used if meaningful extrapolations are possible
- No benchmarks for switches and storage subsystems exist, but are being worked on

Key Energy Usage Feature of Current Servers

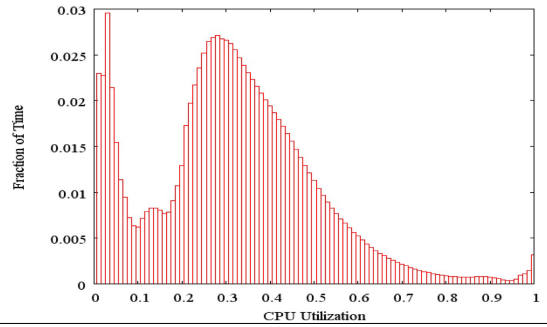
- Under low utilization, the inefficiency is significantly higher
  - Cause: Idle power consumption more than half of peak!!!
- Individual servers spend negligible time completely idle



An example benchmark result for SPECpower_ssj2008; energy efficiency is indicated by bars, whereas power consumption is indicated by the line. Both are plotted for a range of utilization levels, with the average metric corresponding to the vertical dark line. The system has a single-chip 2.83 GHz quad-core Intel Xeon processor, 4 GB of DRAM, and one 7.2 k RPM 3.5" SATA disk drive.

Load vs. Efficiency

- Unfortunately, most datacenter servers operate at 10-50% utilization most of the time ☹



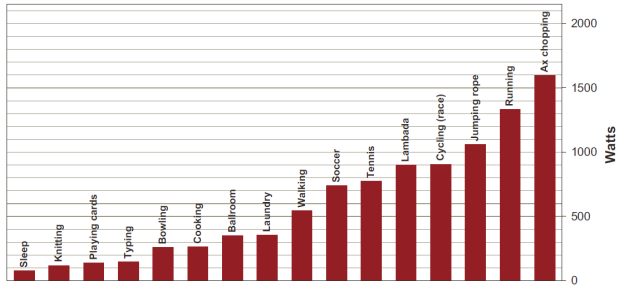
Activity profile of a sample of 5,000 Google servers over a period of 6 months.

Absence of Idle Intervals

- Due to the application of sound design principles to high-performance, robust distributed systems software
  - Lower overall load leads to lower load in multiple servers (e.g. Web search queries)
- Idleness can be manufactured by moving workloads
  - Downside: significant overhead, complex software/data distribution models
  - Difficult when resilient distributed storage is required (e.g. GFS)

Dynamic Power Range

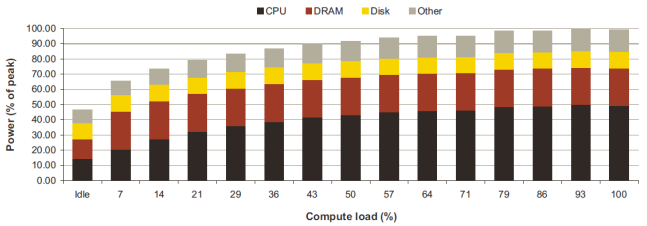
- Energy-proportional machines would exhibit a wide dynamic power range – rare in computing equipment (merely 2x), but not unprecedented in other domains (e.g. Humans have a 20x factor)



Human energy usage vs. activity levels (adult male)

Causes of Poor Energy Proportionality

- CPUs are not necessarily the main culprit!
- Over the years, CPU designers have been more attentive to energy efficiency than their counterparts for other subsystems (e.g., switching to multicore vs. higher clock frequencies)
- Server-class CPUs have dynamic power range of 3.0x or more (compare with: 2.0x for memory, 1.2x for disks, less than 1.2x for networking switches)



Improving Energy Proportionality

- Added focus on energy proportionality across all system components
- More innovation is required in some cases. For example, disk drives spend a large fraction (up to 70%) of energy simply keeping the platters spinning!
  - Smaller rotational speeds, smaller platters, multiple independent head assemblies
  - [Sankar et al.] Head movements are relatively energy-proportional, so a disk with lower rotational speed and multiple heads might achieve similar performance and lower power when compared with a single-head, high RPM disk
- Energy-proportional behavior is not only a target for electronic components, but to the entire system, including power distribution and cooling infrastructure

Relative Effectiveness of Low-Power Modes

- Inactive low-power modes are successful for mobile/embedded systems, but a poor fit for datacenter systems
  - Latency penalty is usually high and our workload pattern would trigger it often (e.g., spun-down disks)
- Active low-power modes (e.g. CPU voltage-frequency scaling) save energy at a performance cost while not requiring inactivity
  - Useful even when the latency/energy penalty to transition to a high-performance mode are significant (since overheads amortize more effectively)

Role of Software

- Clever software strategies can enhance energy-proportionality of the underlying hardware:
  - Intelligent use of power management features in existing hardware
  - Using low-overhead inactive or active low-power modes
  - Power-friendly scheduling of tasks
- Challenges
  - Encapsulation (avoid exposure to developers)
  - Robustness (avoid side-effects like increased variability of response-time)

Summary

- Datacenters basics
- Software
  - Parallel
  - Fault-tolerant
- Energy efficiency
  - Decrease PUE*SPUE
  - Energy proportional computing