Ingredients for Building Energy Efficient Computing Systems: Hardware, Software, and Tools

John D. Davis
Researcher
Microsoft Research Silicon Valley

Why energy efficiency matters?

- Power & Related Costs Dominate
  - 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

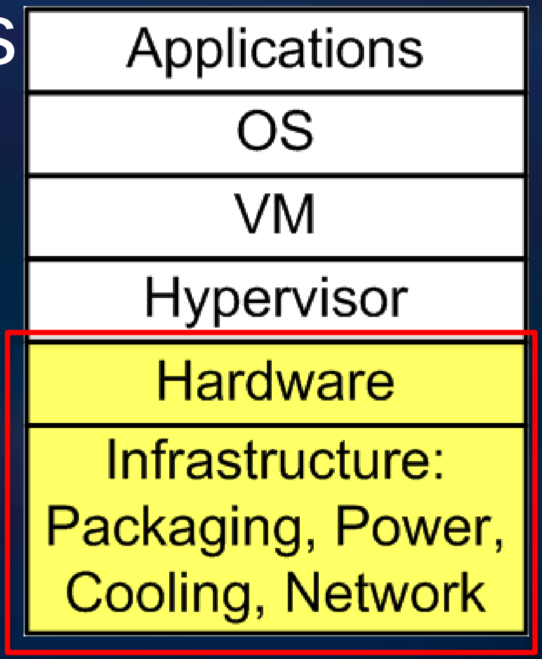
- Observations:
  - $2.3M/month from charges functionally related to power
  - Power related costs trending flat or up while server costs trending down

Details at: http://perspectives.mvdirona.com/2008/11/28/CostOfPowerInLargeScaleDataCenters.aspx

Courtesy: James Hamilton, ISCA 2009

Agenda

- Understanding the applications
  - Application-driven design
  - Software re-engineering
- Discovering what matters
  - So many metrics, so little time
- Energy efficient hardware
  - Low-power processors are NOT the solution



- Data Center design requires full system engineering

Understanding the applications

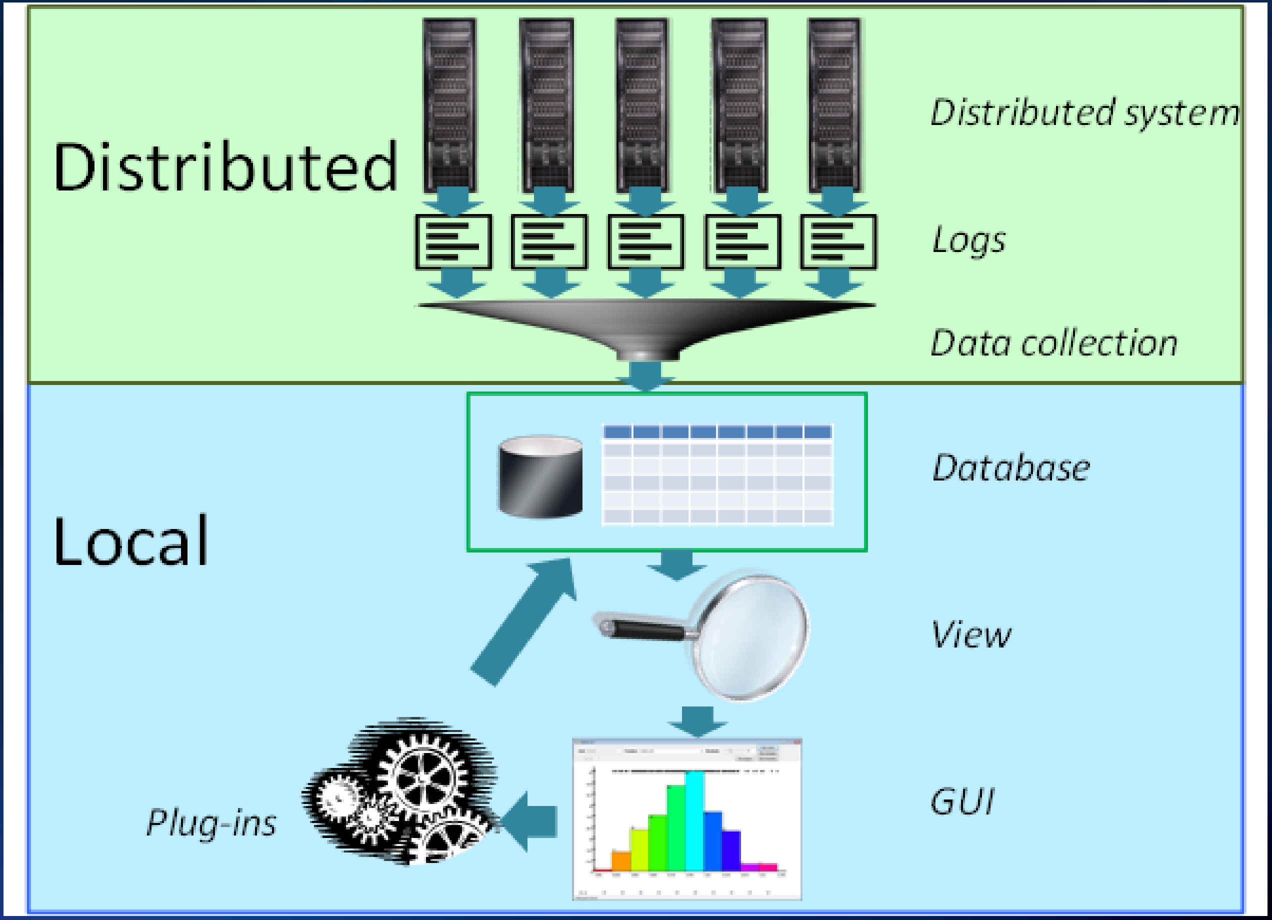
- Taking an application-driven approach to DC design
  - Internal: Search, web server, file system, etc.
  - SPEC: CPU2006, Power, and many others
  - Dryad/DryadLinq applications
  - Joulesort
  - TPC-*
- Re-engineer software
  - Remove/Reduce/Consolidate heartbeats
- How do we understand what is important?

Discovering what matters

- ETW Framework
  - 100's of metrics
- Performance Counters
  - VTUNE (Intel)
  - Code Analyst (AMD)
- Need visualization tools
  - Find the needle in the haystack
- Machine learning and other techniques
  - Identify correlations and significant metrics

Artemis

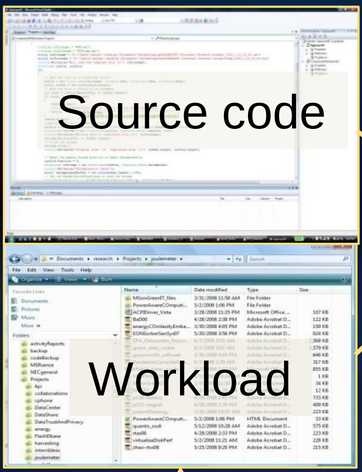
- Performance analysis of distributed systems
- Modular, extensible, and interactive



Courtesy: Mihai Budiu

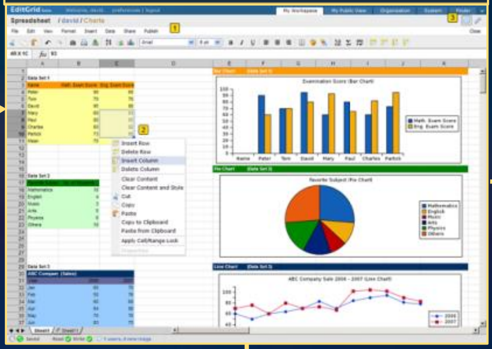
JouleMeter

- Measure application energy usage using performance events



JouleMeter

Detailed Energy Profile



Run-time control
(Eg. Tune QoS)

Auto-optimize
(Eg. Compiler warning: Change line 136 in Program.c to save …)

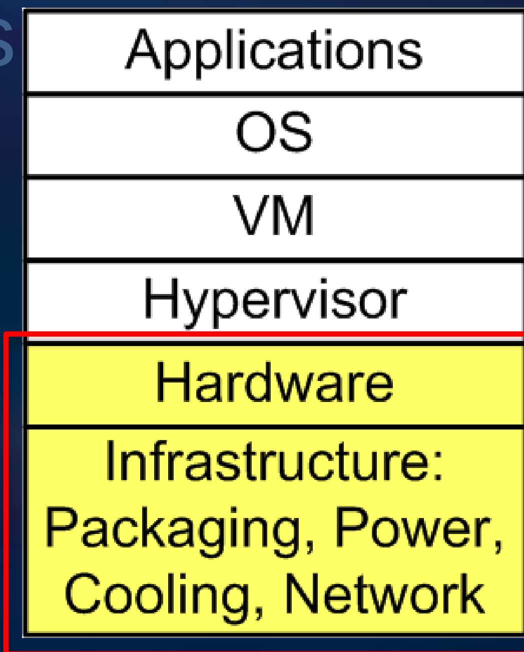
Designer Insights
(Eg. Network is holding up the CPU in method X …)

Optimizations

Courtesy: Aman Kansal

Agenda

- Understanding the applications
  - Application-driven design
  - Software re-engineering
- Discovering what matters
  - So many metrics, so little time
- Energy efficient hardware
  - Low-power processors are NOT the solution



Energy Efficient Hardware

- Building a DC
  - Servers
  - Power distribution
  - Cooling
  - Packaging
  - Networking
- All of these can be improved to reduce both capital and operating cost.

The Computers

- We currently use commodity servers designed by HP, Rackable, others.
- Higher quality and reliability than the average PC, but they still operate in the PC ecosystem.
  - IBM doesn't.
- Why not roll our own?

Designing our own

- Minimize SKUs
  - One for computing. Lots of CPU, Lots of memory, relatively few disks.
  - One for storage. Modest CPU, memory, lots of disks.
  - Maybe Flash memory has a home here.
- What do they look like?
  - Custom motherboards.
  - Redesign the power supply.
  - Commodity disks.
  - Cabling exits the front panel.
  - Error correction where possible …
  - Processor dictated by workload data.

Power Distribution

- Need to minimize conversion steps to minimize losses.
- Deliver 3-phase AC to the rack
  - Must balance the phases anyway
  - Lower ripple after rectification
- What voltage?
  - TBD, but probably 12-20 VAC.
  - Select to maximize overall efficiency

Cooling

- Once-through air cooling is possible in some locations.
  - Unfortunately, data centers tend to be built in inhospitable places.
  - Air must be filtered.
  - Designs are not compatible with side-to-side airflow.
- Cooling towers are well understood technology.
  - And need not be used all the time.
- Once-through water cooling is attractive.
  - Pump water from a river, use it once, sell the output to farms.

Packaging: Another way

- Use a shipping container – and build a parking lot instead of a building.
- Doesn't need to be human-friendly.
  - Might never open it.
- Assembled at one location, computers and all.
  - A global shipping infrastructure already exists.
- Sun's version uses a 20-foot box. 40 would be better.
- Requires only networking, power, and cooled water.
- Expands as needed, in sensible increments.
- Rackable has a similar system. So does Google.

Container Advantages

- Side-to-side airflow is not impeded by the server case. There is no case.
  - With bottom-to-top, servers at the top are hotter.
  - With front-to-back, must provide hot and cold plenums.
- The server packaging is simplified, since they are not shipped separately. Can incorporate shock mounting at the server, not the rack level.
- Cables exit at the front, simplifying assembly and service.
- Most of this also applies to conventional data centers.

Container DC

- A 40’ container holds two rows of 16 racks.
- Each rack holds 40 “1U” servers, plus network switch. Total container: 1280 servers.
- If each server draws 200W, the rack is 8KW, the container is 256 KW.
- A 64-container data center is 16 MW, plus cooling. Contains 82K computers.
- Each container has independent fire suppression. Reduces insurance cost.

Conclusions

- By treating data centers as *<u>systems</u>*, and doing full-system optimization, we can achieve:
  - More energy efficient systems.
  - Lower cost, both opex and capex.
  - Higher reliability.
  - Incremental scale-out.
  - More rapid innovation as technology improves.

Microsoft Research

Questions?

Microsoft®

Your potential. Our passion.™

BACK-UP SLIDES

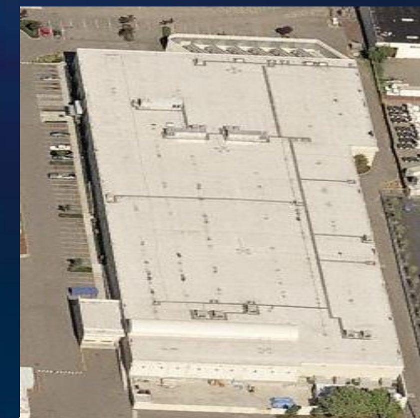
Objections

- "Commodity hardware is cheaper"
  - This *is* commodity hardware. Even for one center (and we build many).
  - And in the case of the network, it's *not* cheaper. Large switches command very large margins.
- "Standards are better"
  - Yes, but only if they do what you need, at acceptable cost.
- "It requires too many different skills"
  - Not as many as you might think.
  - And we would work with engineering/manufacturing partners who would be the ultimate manufacturers. This model has worked before.
- "If this stuff is so great, why aren't others doing it"?
  - They are.

More Information

- **James Hamilton ISCA 2009 Keynote: Internet-Scale Service Infrastructure Efficiency**
- • http://mvdirona.com/jrh/TalksAndPapers/JamesHamilton_ISCA2009.pdf
- **Power and Total Power Usage Effectiveness (tPUE)**
- •http://perspectives.mvdirona.com/2009/06/15/PUEAndTotalPowerUsageEfficiencyTPUE.aspx
- **•Berkeley Above the Clouds**
- •http://perspectives.mvdirona.com/2009/02/13/BerkeleyAboveTheClouds.aspx
- **•Degraded Operations Mode**
- –http://perspectives.mvdirona.com/2008/08/31/DegradedOperationsMode.aspx
- **•Cost of Power**
- –http://perspectives.mvdirona.com/2008/11/28/CostOfPowerInLargeScaleDataCenters.aspx
- –http://perspectives.mvdirona.com/2008/12/06/AnnualFullyBurdenedCostOfPower.aspx
- **•Power Optimization:**
- –http://labs.google.com/papers/power_provisioning.pdf
- **•Cooperative, Expendable, Microslice Servers**
- –http://perspectives.mvdirona.com/2009/01/15/TheCaseForLowCostLowPowerServers.aspx
- **•Power Proportionality**
- –http://www.barroso.org/publications/ieee_computer07.pdf
- **•Resource Consumption Shaping:**
- –http://perspectives.mvdirona.com/2008/12/17/ResourceConsumptionShaping.aspx

Is Energy a Problem?



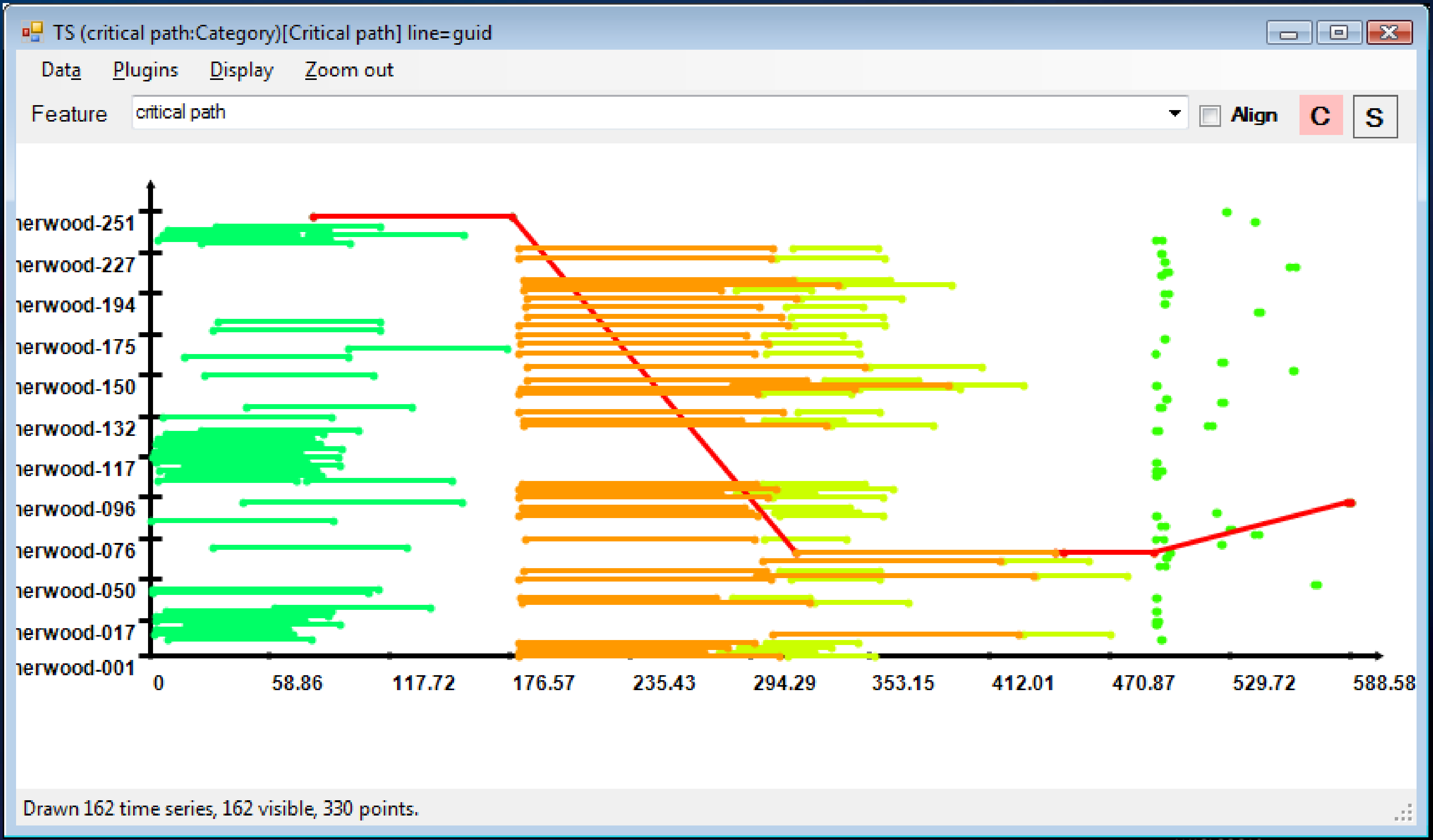
E. Cost: $4.5b

Energy usage growing at 14% yearly

- Data Center energy (excluding small DC's, office IT equip.) equals
  - Electricity used by the entire U.S transportation manufacturing industry (manufacture of automobiles, aircraft, trucks, and ships)

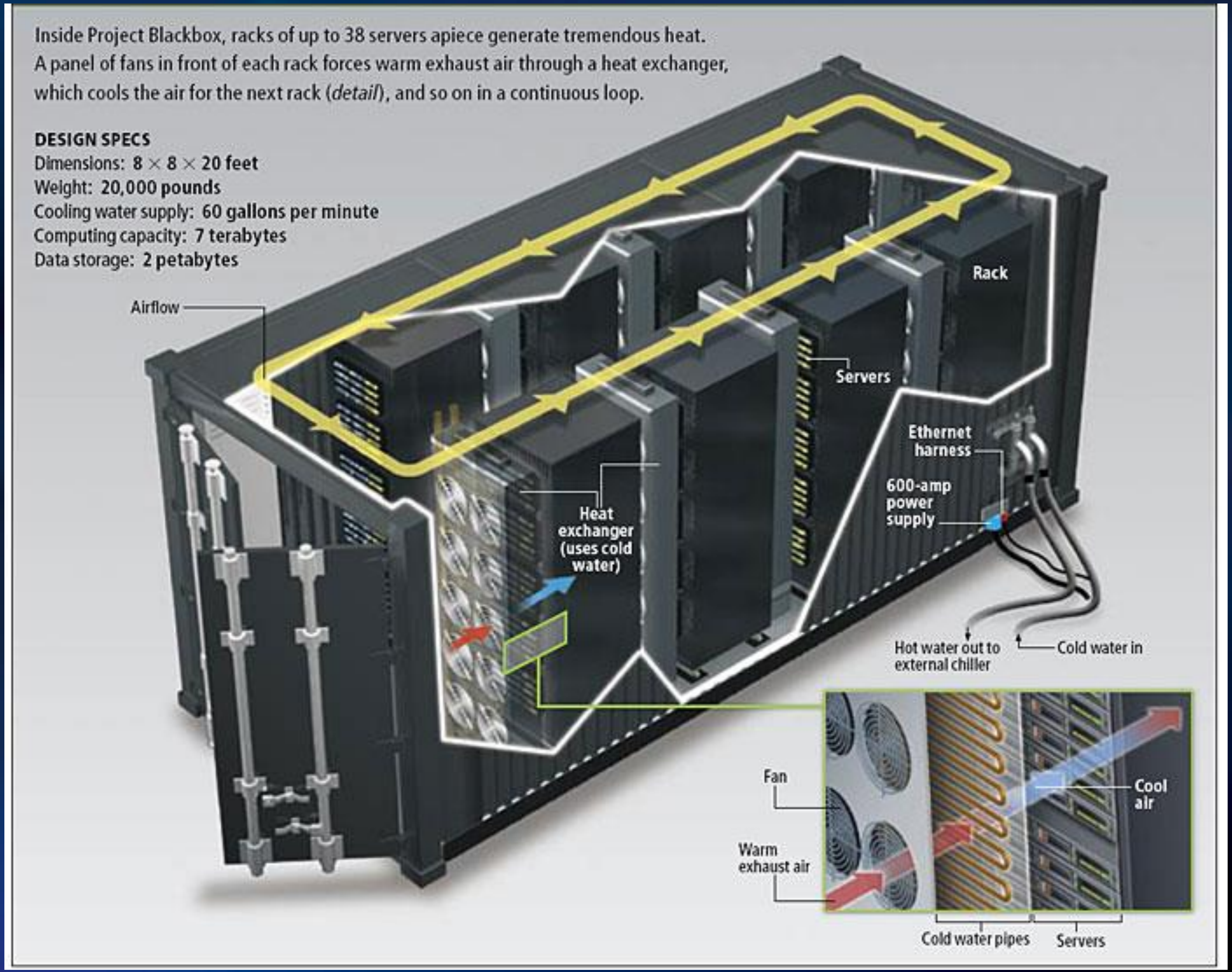
Microsoft Research

Artemis: Scheduling and Critical Path



Courtesy: Mihai Budiu

The Black Box



Power Distribution

http:// perspectives.mvdirona.com

System Design: Power distribution

- Need to minimize conversion steps to minimize losses.
- Power supplies aren't very efficient:
  - 12VDC -> 1VDC point-of-load regulators are ~90%.
  - AC -> 12VDC converters are now 2-stage (power factor correction, inverter). 85% efficient at full load, lower at low load. Can do better.
- AC transformers are 98% efficient. Two steps needed.
- Final efficiency, grid to chips/disks: ~80%.
- UPS and backup generators aren't part of the picture until the grid fails.