

Where Does the Power Go & What to do About it?

James Hamilton

Conference on Innovative Data Systems Research

2009/1/7

Amazon Web Services

e: James@amazon.com

w: mvdirona.com/jrh/work

b: perspectives.mvdirona.com



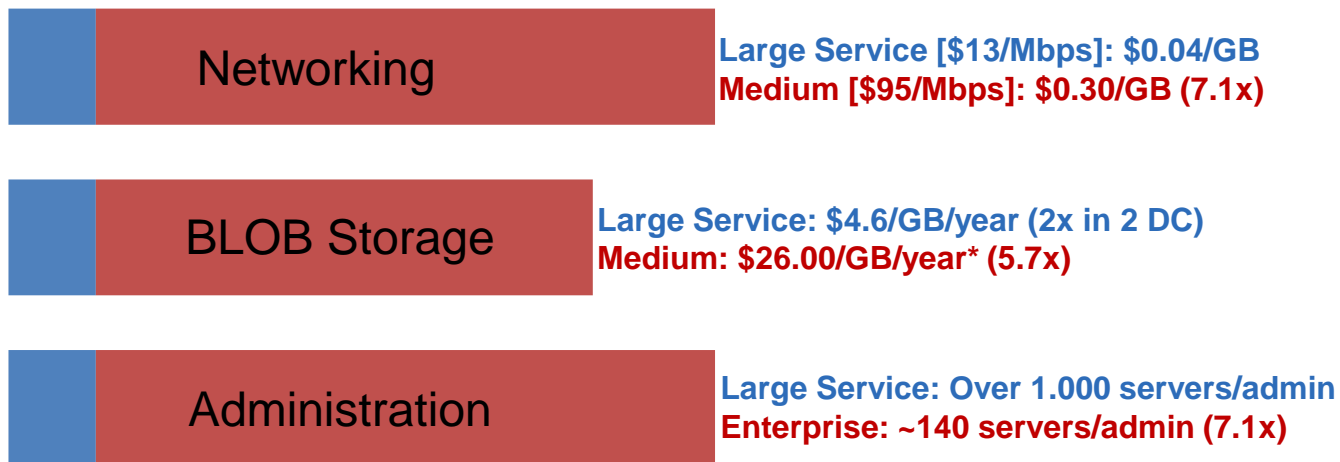
Agenda

- Services Inevitable
- Where does the power go?
 - Power distribution systems & optimizations
 - Mechanical systems & optimizations
 - Cooperative, Expendable, Micro-Slice Servers
 - Critical Load Optimizations
- Summary



Services Economies of Scale

- Substantial economies of scale possible
- Compare a very large service with a small/mid-sized: (~1000 servers):



2006 Chart data

- High cost of entry
 - Physical plant expensive: 15MW ~\$200M (infrastructure only)
- Summary: significant economies of scale but at high cost of entry
 - Small number of large players likely outcome

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

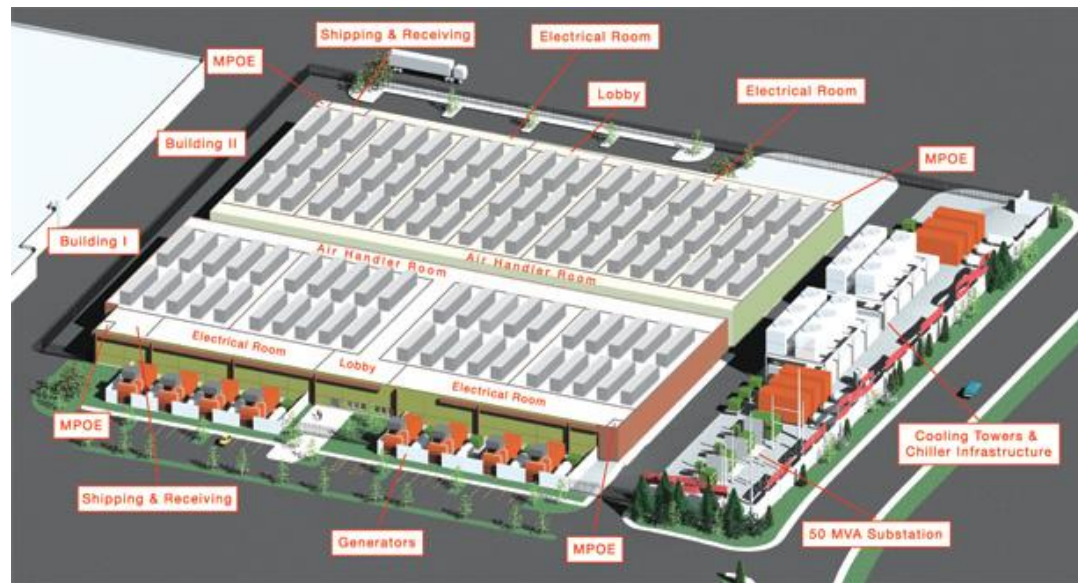
Agenda

- Services Inevitable
- Where does the power go?
 - Power distribution systems & optimizations
 - Mechanical systems & optimizations
 - Cooperative, Expendable, Micro-Slice Servers
 - Critical Load Optimizations
- Summary



PUE & DCiE

- Measure of data center infrastructure efficiency
- Power Usage Effectiveness
 - $PUE = (\text{Total Facility Power}) / (\text{IT Equipment Power})$
- Data Center Infrastructure Efficiency
 - $DCiE = (\text{IT Equipment Power}) / (\text{Total Facility Power}) * 100\%$



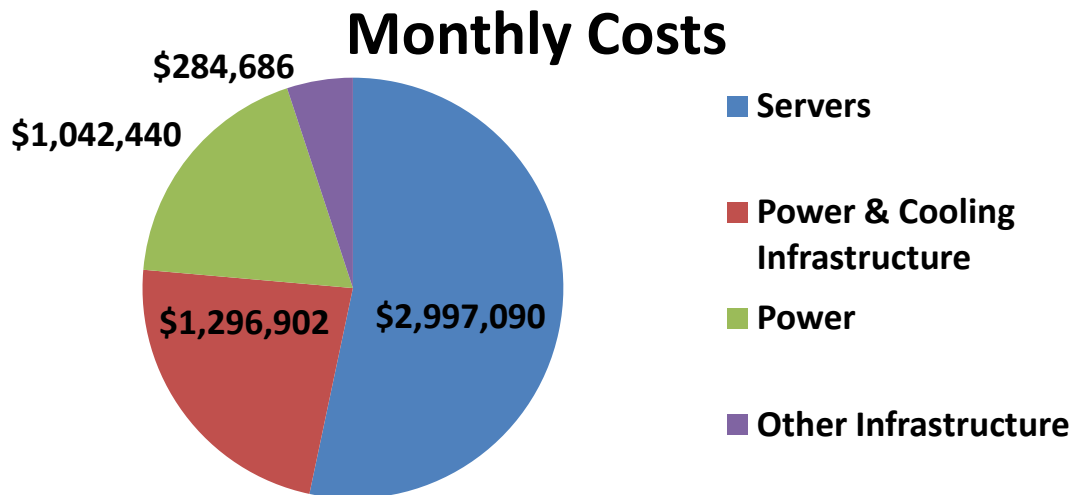
Advanced Data Centers

http://www.thegreengrid.org/gg_content/TGG_Data_Center_Power_Efficiency_Metrics_PUE_and_DCiE.pdf

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



3yr server & 15 yr infrastructure amortization



- **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>

Fully Burdened Cost of Power

- **Infrastructure cost/watt:**
 - 15 year amortization & 5% money cost
 - $=\text{PMT}(5\%,15,200000000,0)/(15,000,000)$
 - $\Rightarrow \$1.28/\text{W}/\text{yr}$
 - **Cost per watt using \$0.07 Kw*hr:**
 - $=-0.07*1.7/1000*0.8*24*365\Rightarrow$
 - $\$0.83/\text{W}/\text{yr}$ (@80% power utilization)
-

- **Fully burdened cost of power:**
 - $\$1.28 + \$0.83 \Rightarrow \$2.11$



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
- **Power losses are easier to track than cooling:**
 - Power transmission & switching losses: 8%
 - Detailed power distribution losses on next slide
 - Cooling losses remainder: $100 - (59 + 8) \Rightarrow 33\%$
- **Data center power consumption:**
 - IT load (servers): $1/1.7 \Rightarrow 59\%$
 - Distribution Losses: 8%
 - Mechanical load (cooling): 33%

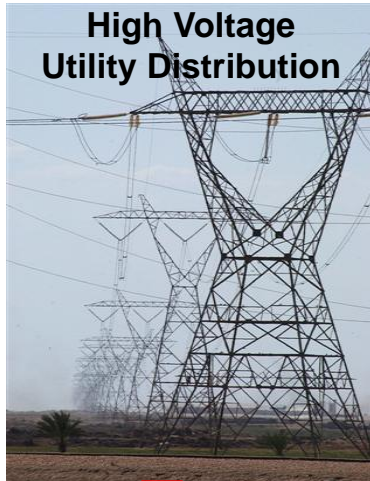


Agenda

- Services Inevitable
- Where does the power go?
 - Power distribution systems & optimizations
 - Mechanical systems & optimizations
 - Cooperative, Expendable, Micro-Slice Servers
 - Critical Load Optimizations
- Summary



Power Distribution



8% distribution loss
 $.997^3 * .94 * .99 = 92.2\%$

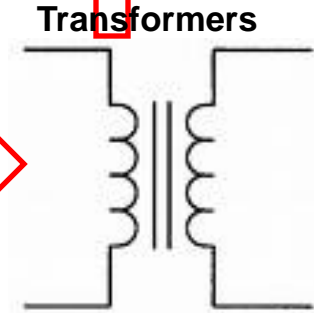
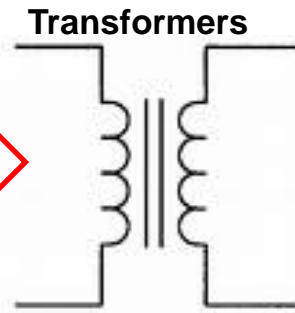
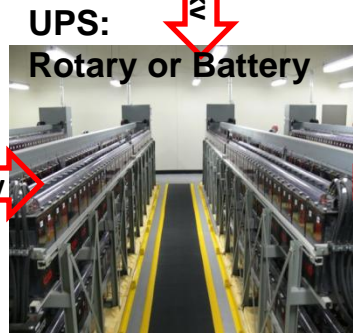


115kv

13.2kv

480V

~1% loss in switch gear & conductors



13.2kv

13.2kv

480V

0.3% loss
99.7% efficient

6% loss
94% efficient, ~97% available

0.3% loss
99.7% efficient

0.3% loss
99.7% efficient

Power Redundancy to Geo-Level

- Over 20% of entire DC costs is in power redundancy
 - Batteries supply over 10 min at some facilities (~2 min sufficient)
 - N+2 generation (2.5MW) at over \$2M each
- Instead use more, smaller, cheaper 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)



2009/1/7

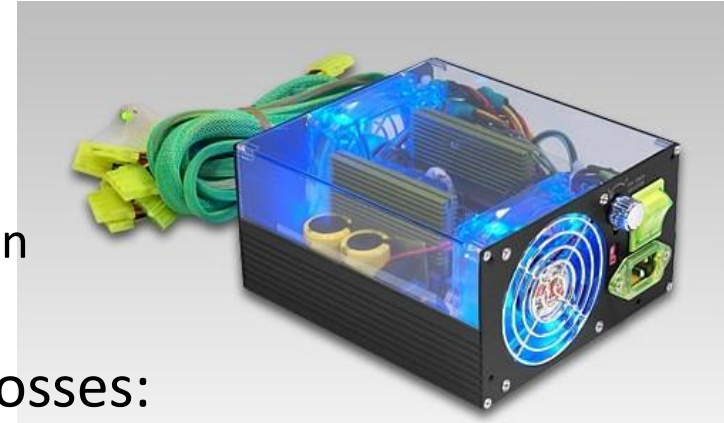
<http://perspectives.mvdirona.com>



12

Power Distribution Optimization

- 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

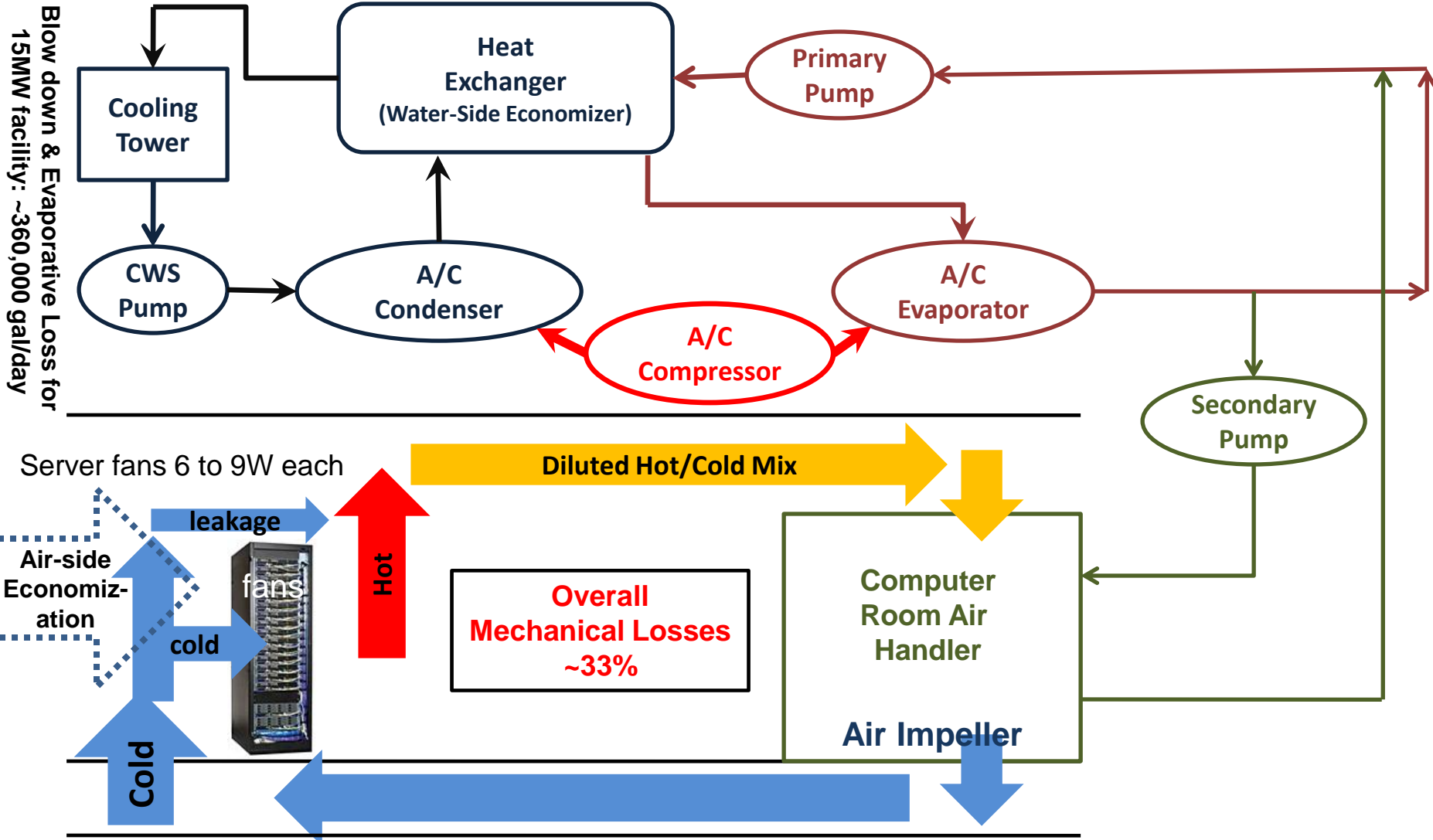


Agenda

- Services Inevitable
- Where does the power go?
 - Power distribution systems & optimizations
 - Mechanical systems & optimizations
 - Cooperative, Expendable, Micro-Slice Servers
 - Critical Load Optimizations
- Summary



Conventional Mechanical Design



Mechanical Optimization

- Simple rules to minimize cooling costs:
 1. Raise data center temperatures
 2. Tight control of airflow with short paths
 3. Cooling towers rather than A/C
 4. Air side economization (open the window)
 5. Low-grade, waste heat energy reclamation
- Best current designs still use air but bring water near servers
 - Lower heat densities could be 100% air cooled
- Common mechanical designs: 33% lost in cooling
- PUE under 1.0 within reach with some innovation
 - Waste heat reclamation in excess of power distribution & cooling overhead (~30% effective reclamation sufficient for <1.0 operation)

Agenda

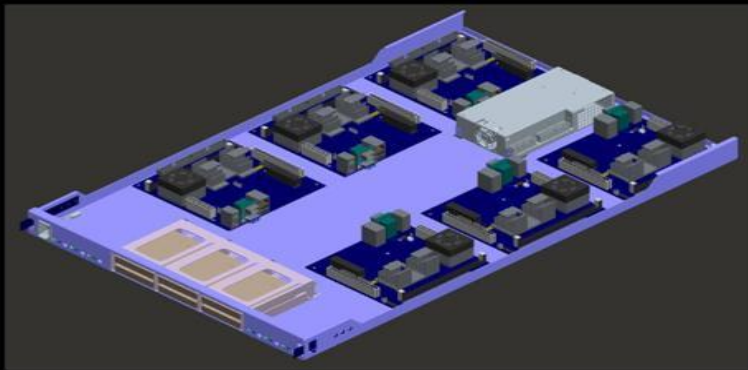
- Services Inevitable
- Where does the power go?
 - Power distribution systems & optimizations
 - Mechanical systems & optimizations
 - Cooperative, Expendable, Micro-Slice Servers
 - Critical Load Optimizations
- Summary



CEMS Speeds & Feeds

- CEMS: Cooperative Expendable Micro-Slice Servers
 - Correct system balance problem with less-capable CPU
 - Too many cores, running too fast, for memory, bus, disk, ...
- Joint project with Rackable Systems (<http://www.rackable.com/>)

	System-X	CEMS V3 (Athlon 4850e)	CEMS V2 Athlon 3400e)	CEMS V1 (Athlon 2000+)
CPU load%	56%	57%	57%	61%
RPS	95.9	75.3	54.3	17.0
Price	\$2,371	\$500	\$685	\$500
Power	295	60	39	33
RPS/Price	0.04	0.15	0.08	0.03
RPS/Joule	0.33	1.25	1.39	0.52
RPS/Rack	1918.4	18062.4	13024.8	4080.0

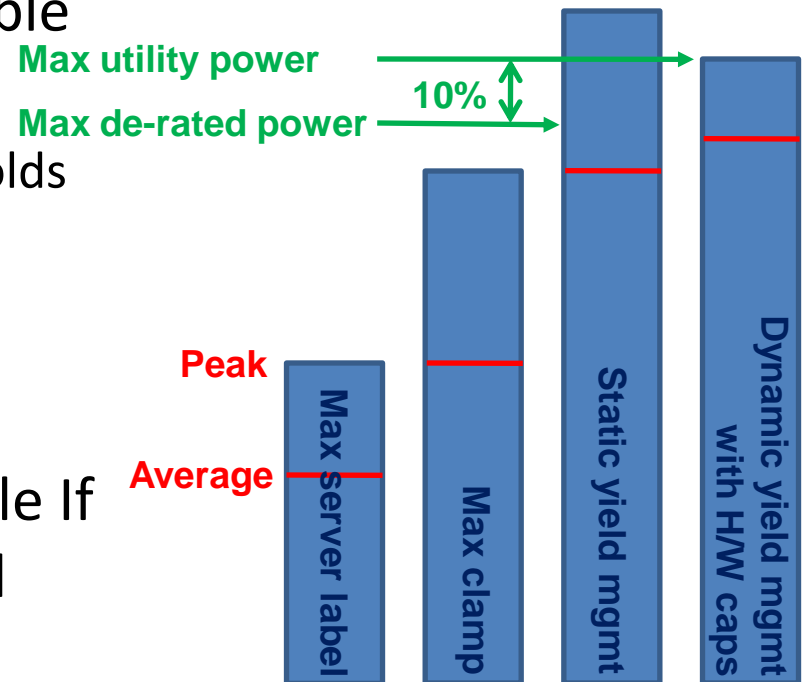


- **CEMS V2 Comparison:**
 - **Work Done/\$: +375%**
 - **Work Done/Joule +379%**
 - **Work Done/Rack: +942%**

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

Power Yield Management

- “Oversell” power, the most valuable resource:
 - e.g. sell more seats than airplane holds
- Overdraw penalty high:
 - Pop breaker (outage)
 - Overdraw utility (fine)
- Considerable optimization possible if workload variation is understood
 - Workload diversity & history helpful
 - Graceful Degradation Mode to shed workload



Power Provisioning in a Warehouse-Sized Computer, Xiabo Fan, Wolf Weber, Luize Borroso

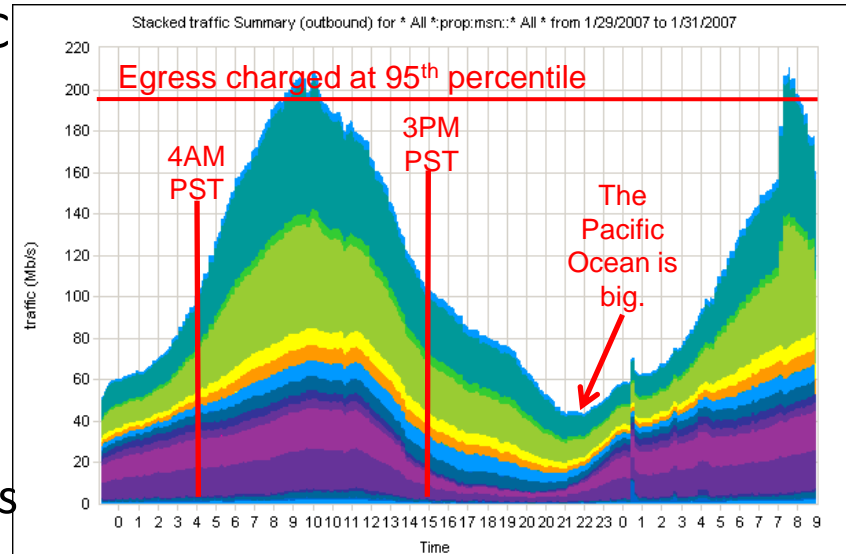
Critical Load Optimization

- Power proportionality is great but full load or shut off is even better
 - Idle server consumes ~60% power of full load
 - Industry secret: “great” data center utilization around ~30%
 - All solutions require changing where a workload is executed
- What limits dynamic workload migration?
 - Networking constraints: VIPs can’t span L2 nets, ACLs static, manual config, etc.
 - Data Locality: Hard to efficiently move several TB
 - Workload management: Scheduling work over available resources
- Critical load optimizations, in order:
 - Use the servers: any workload with marginal value over power
 - Shut them off if you can’t use them
 - Those servers on should be fully loaded but not all resources fully consumed so use power management (e.g. dynamic voltage & frequency scaling)
- Efficient S/W algorithms as important as H/W



Resource Consumption Shaping

- Essentially yield mgmt applied to full DC
- Network charged at 95th percentile:
 - Push peaks to troughs
 - Fill troughs for “free”
 - e.g. Amazon S3 replication
 - Dynamic resource allocation
 - Virtual machine helpful but not needed
 - Charged for symmetrically so ingress effectively free
- Power also charged at 95th percentile
 - Server idle to full-load range: ~65% (e.g. 158W to 230W)
 - S3 (suspend) or S5 (off) when server not needed
- Disks come with both IOPS capability & capacity
 - Mix hot & cold data to “soak up” both
- Encourage priority (urgency) differentiation in charge-back model



David Treadwell & James Hamilton / Treadwell Graph

Agenda

- Services Inevitable
- Where does the power go?
 - Power distribution systems & optimizations
 - Mechanical systems & optimizations
 - Cooperative, Expendable, Micro-Slice Servers
 - Critical Load Optimizations
- Summary



Summary

- Current “good” data centers have considerable room for improvement
- Where do the power go?
 - 58% Servers and other IT equipment
 - 33% mechanical systems
 - 8% power distribution
- Lowest hanging fruit in servers & mechanical systems
- Server system optimizations
 - Utilization levels and general scheduling optimizations
 - Servers optimized for work done per joule & watt rather than raw performance
- CEMS takes only a small step forward but achieves better than 3x improvement in work done/\$ and work done/joule

More Information

- **This Slide Deck:**
 - I will post these slides to <http://perspectives.mvdirona.com> later this week
- **Designing & Deploying Internet-Scale Services**
 - http://mvdirona.com/jrh/talksAndPapers/JamesRH_Lisa.pdf
- **Architecture for Modular Data Centers**
 - http://mvdirona.com/jrh/talksAndPapers/JamesRH_CIDR.doc
- **Increasing DC Efficiency by 4x**
 - http://mvdirona.com/jrh/talksAndPapers/JamesRH_PowerSavings20080604.ppt
- **Perspectives Blog**
 - <http://perspectives.mvdirona.com>
- **Email**
 - James@amazon.com