Data Center Efficiency Best Practices

Data Center Efficiency Summit

James Hamilton, 2009/4/1

VP & Distinguished Engineer, Amazon Web Services

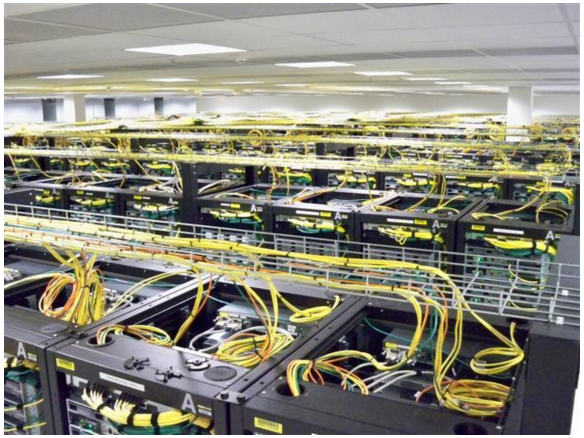
e: James@amazon.com

w: mvdirona.com/jrh/work

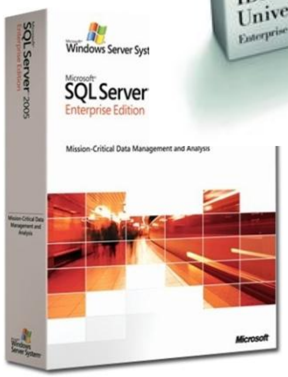
b: perspectives.mvdirona.com

Agenda

- Where does the power go?
- Power distribution optimization
- Mechanical systems optimization
- Server & other optimization
  - Cooperative, Expendable, Micro-Slice Servers
  - Improving existing builds
- Summary



Background & biases



- 15 years in database engine development
  - Lead architect on IBM DB2
  - Architect on SQL Server
- Past 5 years in services
  - Led Exchange Hosted Services Team
  - Architect on the Windows Live Platform
  - Architect on Amazon Web Services
- This talk focuses on industry best practices
  - Not about Amazon (or past employers) specialized data center design techniques
  - 2x gain over current averages easily attainable without advanced techniques

Microsoft Exchange Hosted Services

Windows Live

amazon webservices™

PUE & DCiE

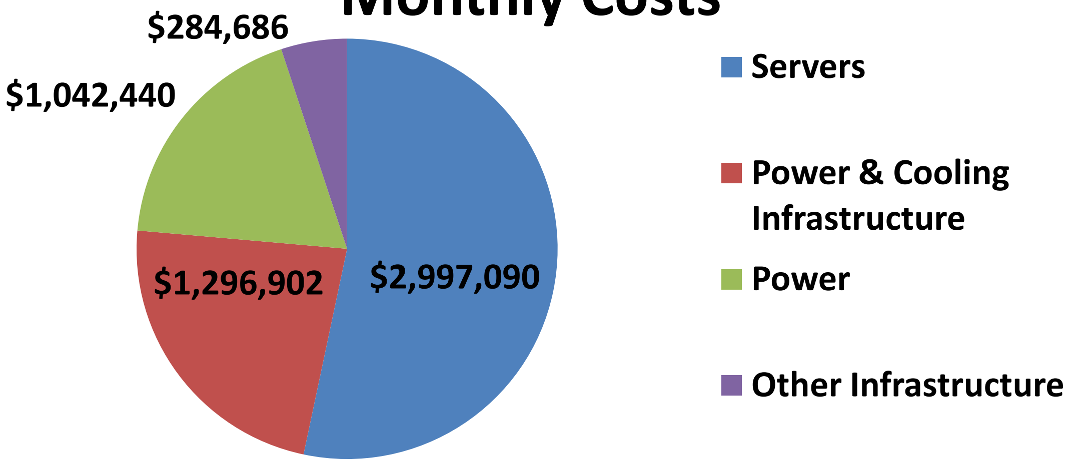
- Measure of data center infrastructure efficiency
- Power Usage Effectiveness
  - PUE = (Total Facility Power)/(IT Equipment Power)
- Data Center Infrastructure Efficiency
  - DCiE = (IT Equipment Power)/(Total Facility Power) * 100%
- I'm looking for help defining **tPUE** (pwr to chip rather than server)

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

Monthly Costs



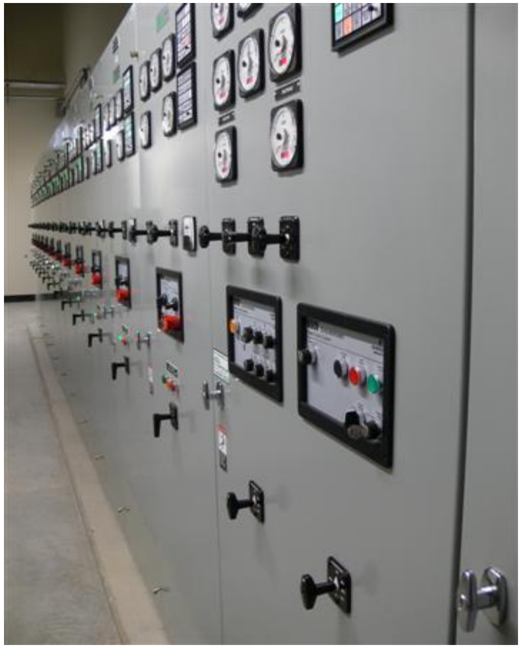
3yr server & 15 yr infrastructure amortization

- **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
  - =PMT(5%,15,2MM,0)/(15MW) => $1.28/W/yr
- **Cost per watt using $0.07 Kw*hr:**
  - =-0.07*1.7/1000*0.8*24*365=> $0.83/W/yr (@80% power utilization)



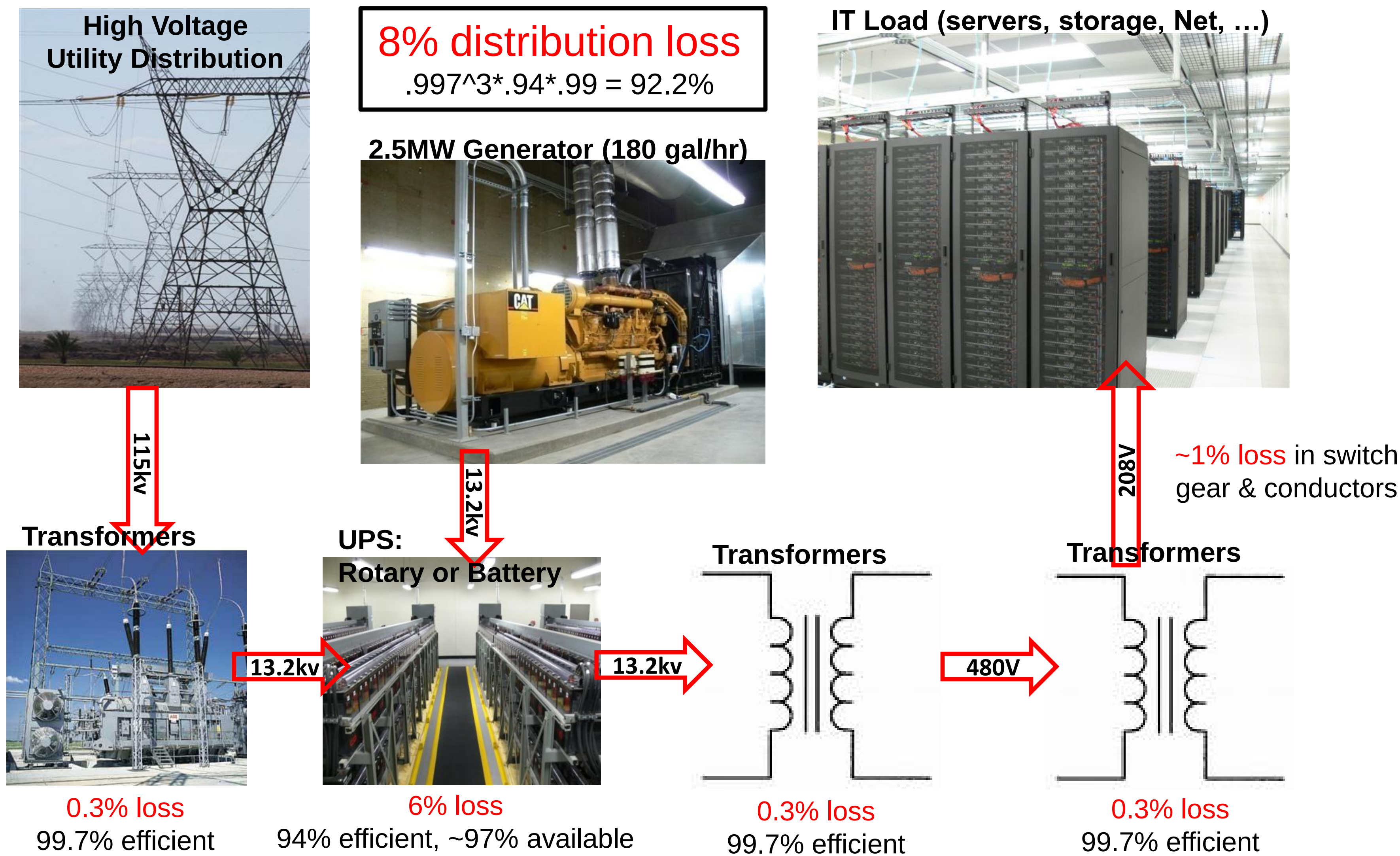
- **Annually fully burdened cost of power:**
  - $1.28 + $0.83 => $2.11/W/yr

Details at: http://perspectives.mvdirona.com/2008/12/06/AnnualFullyBurdenedCostOfPower.aspx

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
  - IT load (servers): 1/1.7=> 59%
- **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) => 33%

Power Distribution



Power Redundancy to Geo-Level

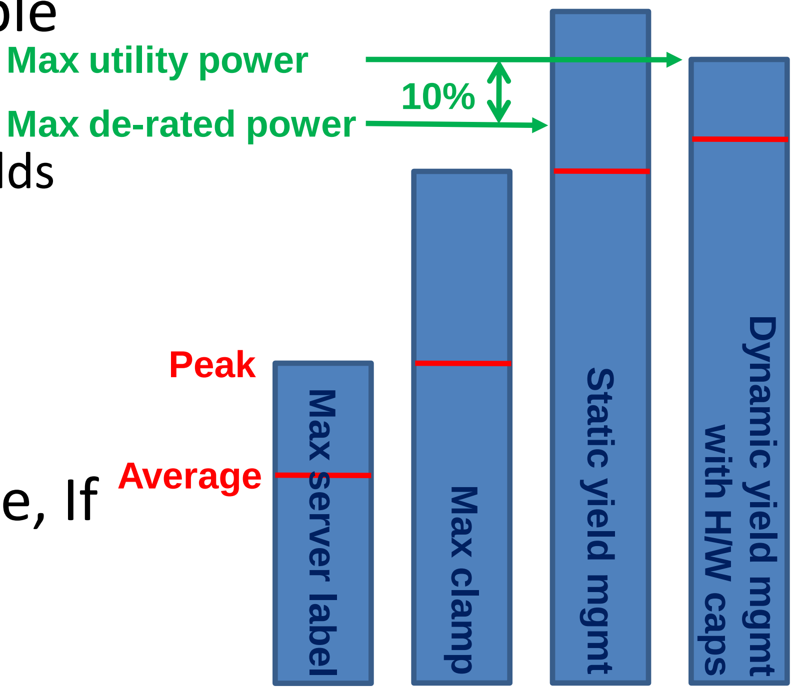
- Roughly 20% of DC capital costs is power redundancy
- Instead use more, smaller, cheaper, commodity data centers
- Non-bypass, battery-based UPS in the 94% efficiency range
  - ~900kW wasted in 15MW facility (4,500 200W servers)
  - 97% available (still 450kW loss in 15MW facility)

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)

Power Yield Management

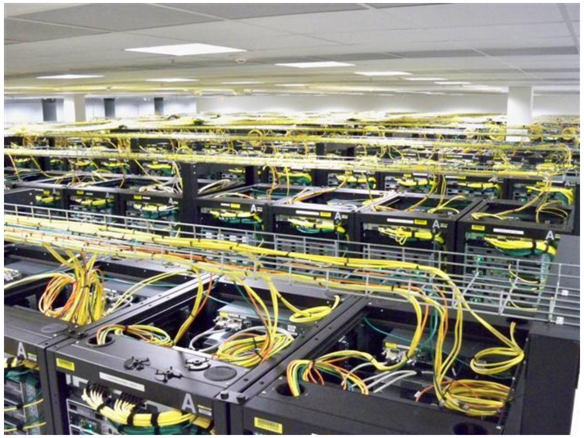
- "Oversell" power, the most valuable resource:
  - e.g. sell more seats than airplane holds
- Overdraw penalty high:
  - Pop breaker (outage)
  - Overdraw utility (fine)
- Considerable optimization possible, If workload variation is understood
  - Workload diversity & history helpful
  - Graceful Degradation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



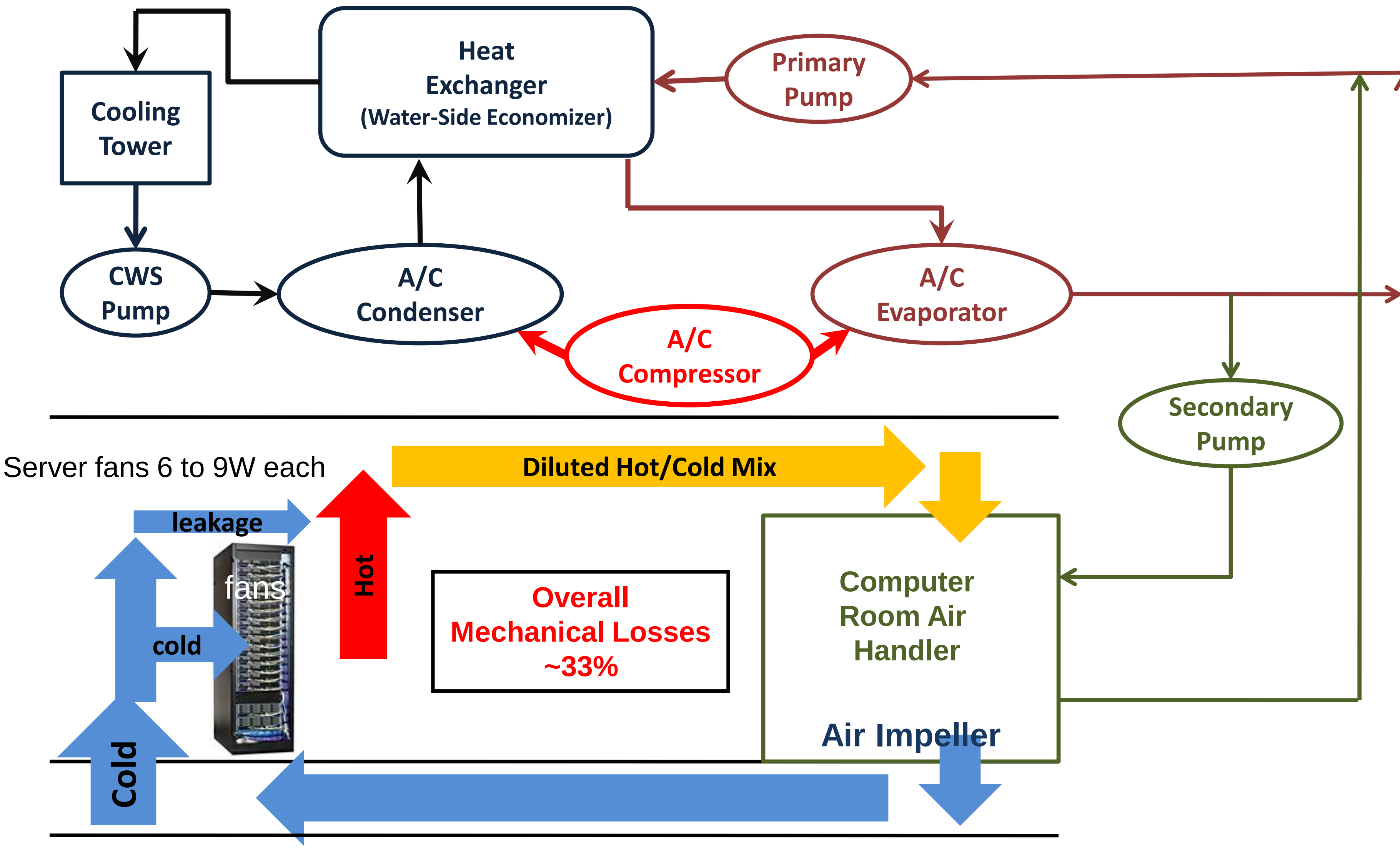
Source: Power Provisioning in a Warehouse-Sized Computer, Xiabo Fan, Wolf Weber, & Luize Borroso

Agenda

- Where does the power go?
- Power distribution optimization
- Mechanical systems optimization
- Server & other optimization
  - Cooperative, Expendable, Micro-Slice Servers
  - Improving non-new builds
- Summary

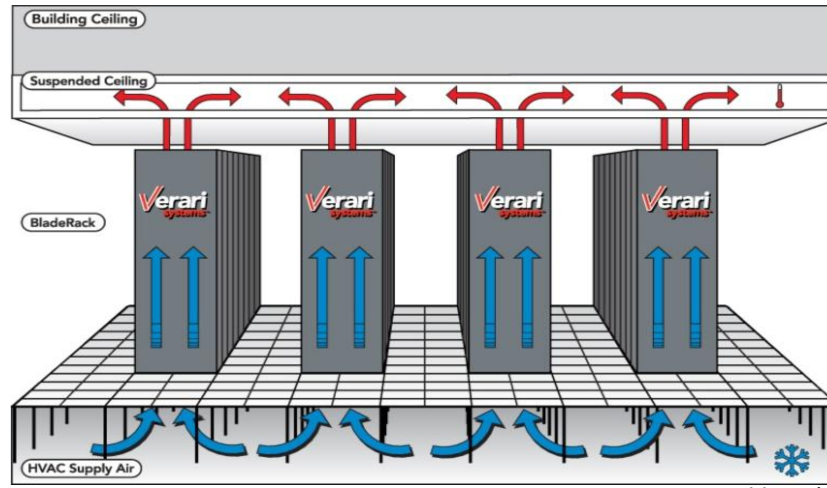


Conventional Mechanical Design

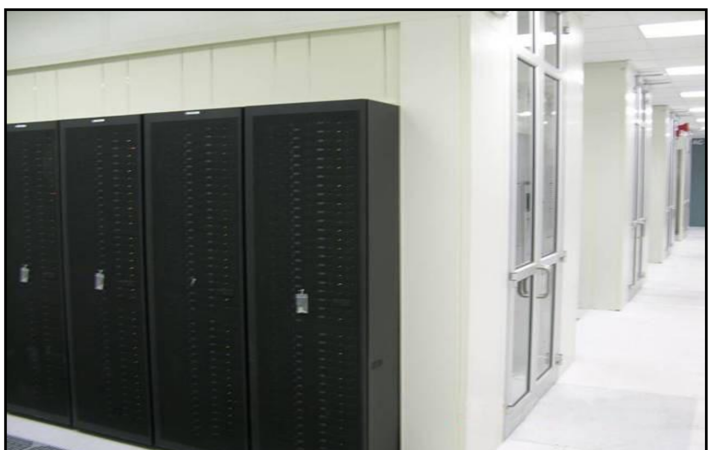


Cooling & Air Handling Gains

Intel



Verari



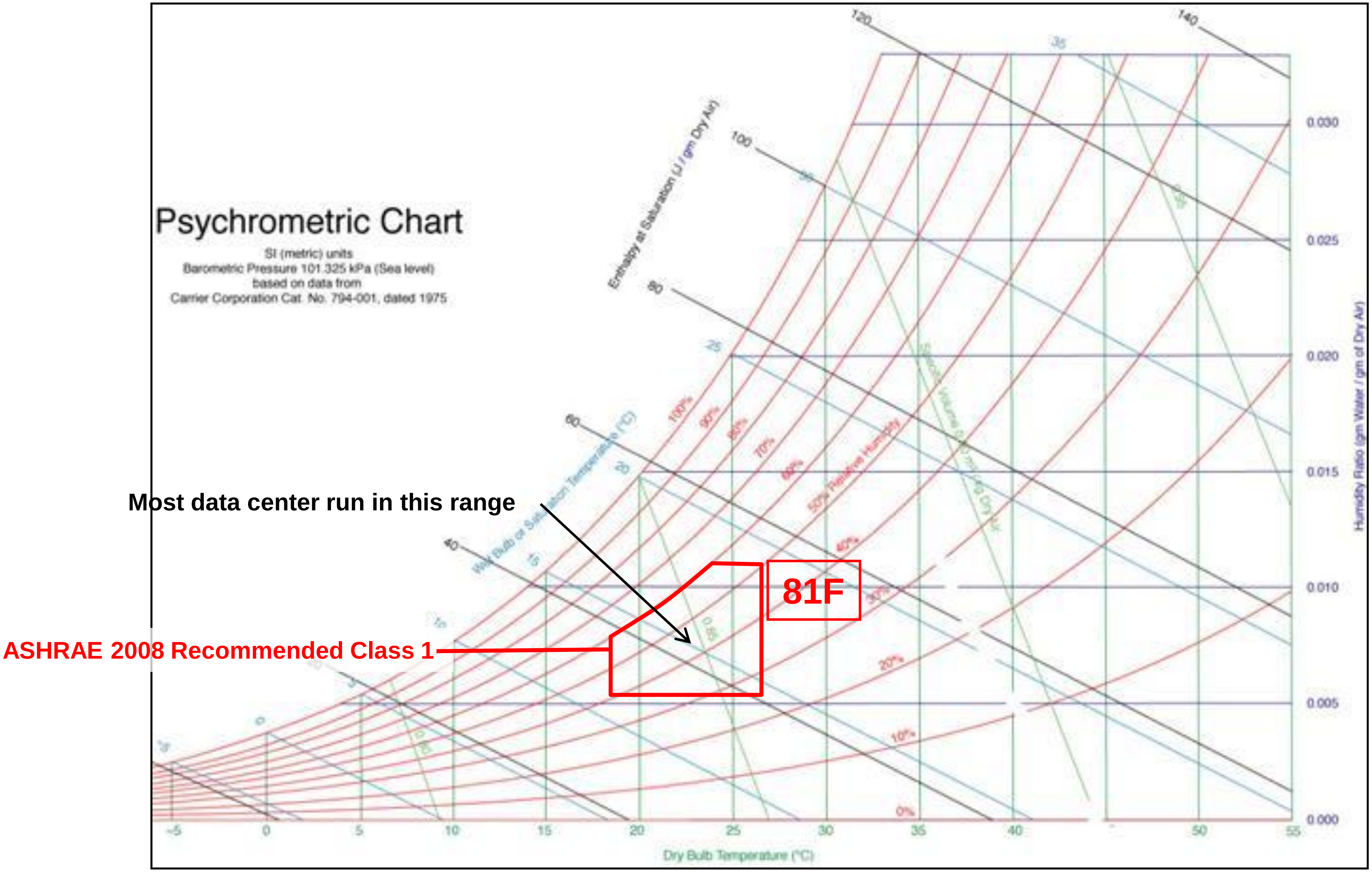
Intel

- Tighter control of air-flow increased delta-T
- Container takes one step further with very little air in motion, variable speed fans, & tight feedback between CRAC and load
- Sealed enclosure allows elimination of small, inefficient (6 to 9W each) server fans

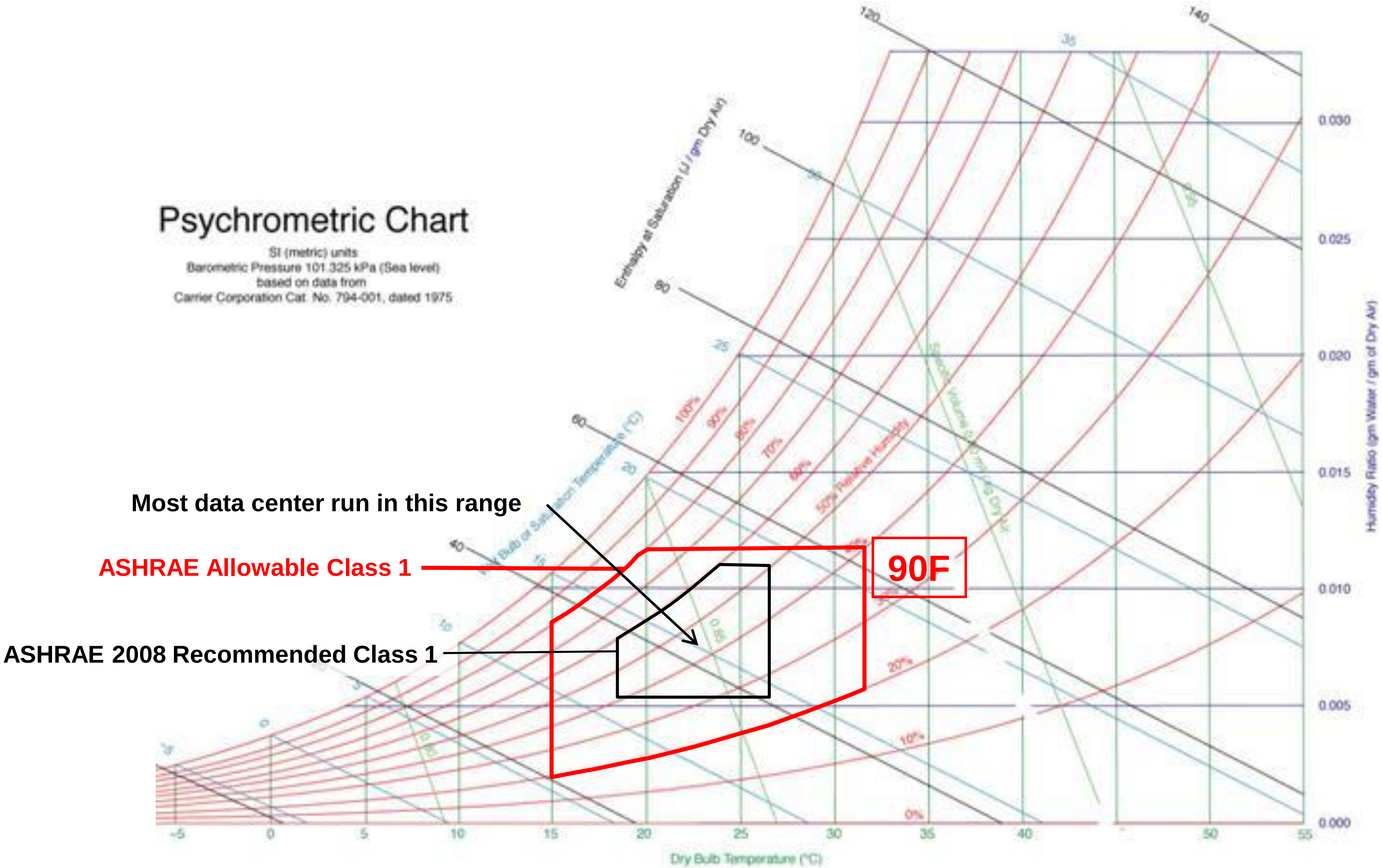
Water!

- It’s not just about power
- Prodigious water consumption in conventional facility designs
  - Both evaporation & blow down losses
  - For example, roughly 360,000 gallons/day at fairly typical 15MW facility

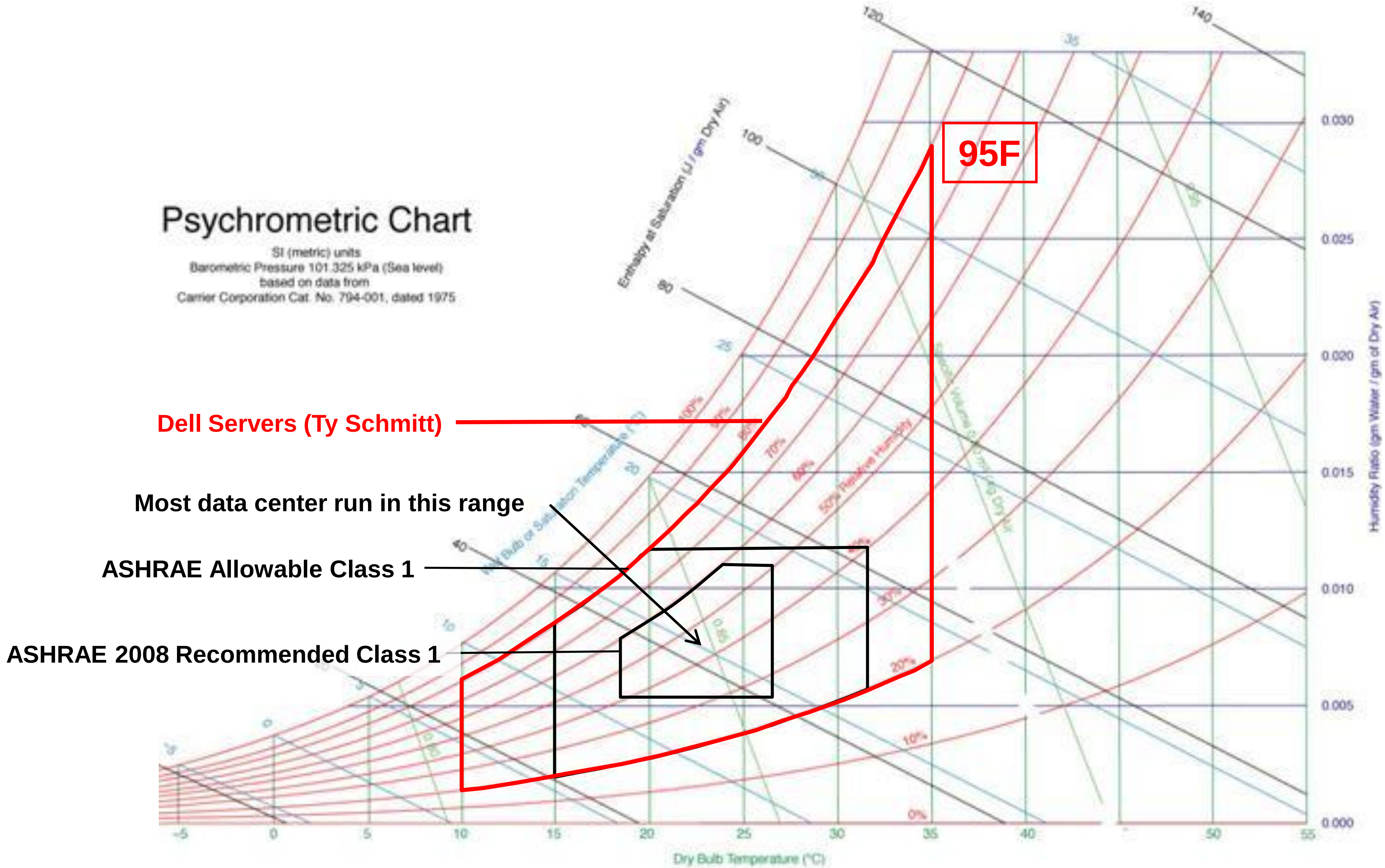
ASHRAE 2008 Recommended



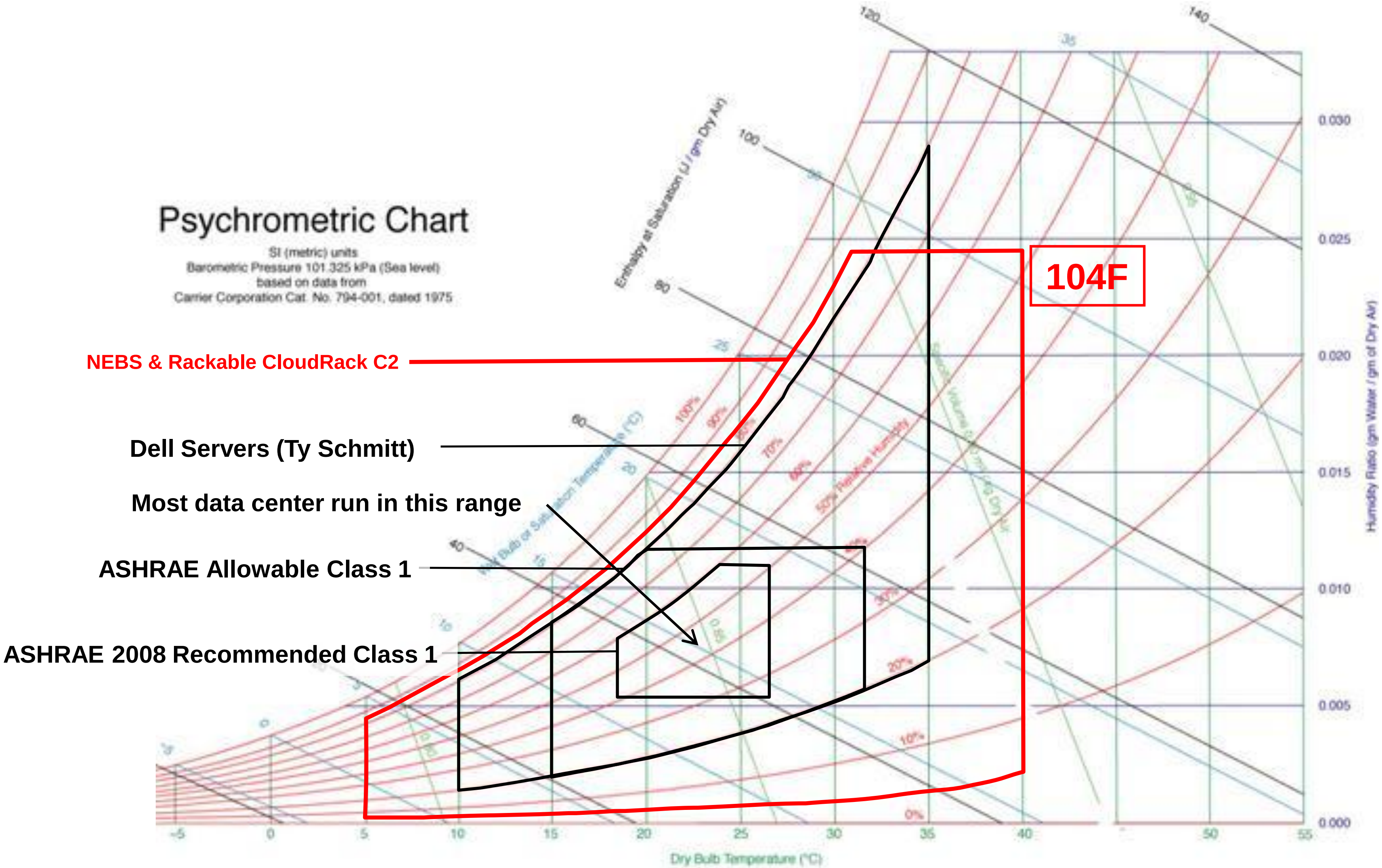
ASHRAE Allowable



Dell PowerEdge 2950 Warranty

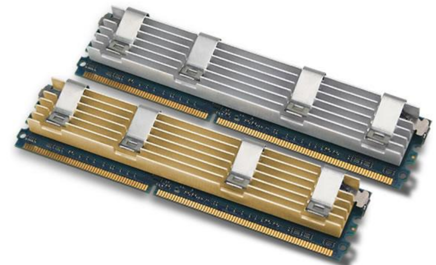


NEBS (Telco) & Rackable Systems



Air Cooling

- Allowable component temperatures higher than hottest place on earth
  - Al Aziziyah, Libya: 136F/58C (1922)
- It's only a mechanical engineering problem
  - More air and better mechanical designs
  - Tradeoff: power to move air vs cooling savings
  - Partial recirculation when external air too cold
- Currently available equipment:
  - 40C: Rackable CloudRack C2
  - 35C: Dell Servers



Memory: 3W - 20W
Temp Spec: 85C-105C

Hard Drives: 7W- 25W
Temp Spec: 50C-60C

Rackable CloudRack C2
Temp Spec: 40C

I/O: 5W - 25W
Temp Spec: 50C-60C

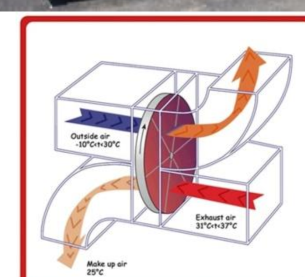
Processors/Chipset: 40W - 200W
Temp Spec: 60C-70C

Thanks for data & discussions:
Ty Schmitt, Dell Principle Thermal/Mechanical Arch.
& Giovanni Coglitore, Rackable Systems CTO

Air-Side Economization & Evaporative Cooling

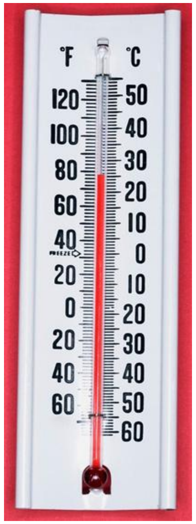
- Avoid direct expansion cooling entirely
- Ingredients for success:
  - Higher data center temperatures
  - Air side economization
  - Direct evaporative cooling

- Particulate concerns:
  - Usage of outside air during wildfires or datacenter generator operation
  - Solution: filtration & filter admin or heat wheel & related techniques
- Others: higher fan power consumption, more leakage current, higher failure rate



A diagram of a rotary heat exchanger, or "heat wheel" (From Uptime Technology BV)

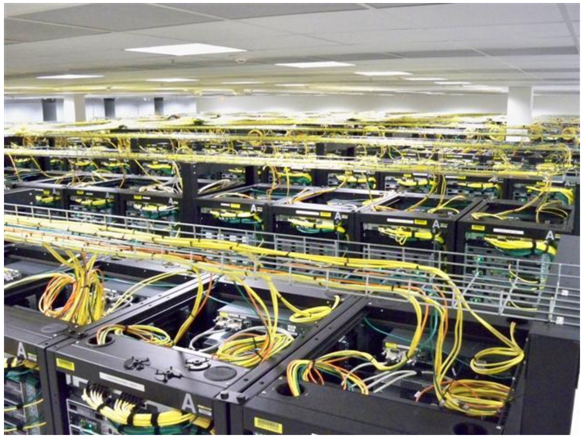
Mechanical Optimization Summary



- Simple rules to minimize cooling costs:
  1. Raise data center temperatures
  2. Tight airflow control, short paths & large impellers
  3. Cooling towers rather than A/C
  4. Air side economization & evaporative cooling
     - outside air rather than A/C & towers

Agenda

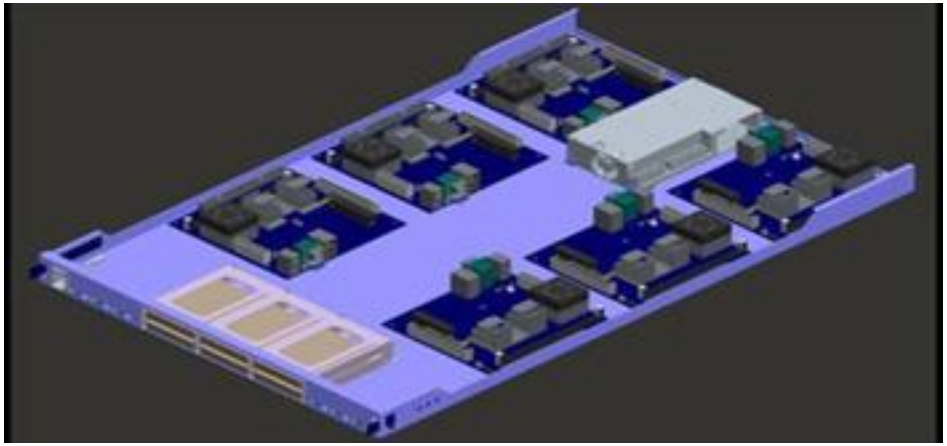
- Where does the power go?
- Power distribution optimization
- Mechanical systems optimization
- Server & other optimization
  - Cooperative, Expendable, Micro-Slice Servers
  - Improving non-new builds
- Summary



CEMS Speeds & Feeds

- CEMS: Cooperative Expendable Micro-Slice Servers
  - Correct system balance problem with less-capable CPU
    - Too many cores, running too fast, and lagging 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



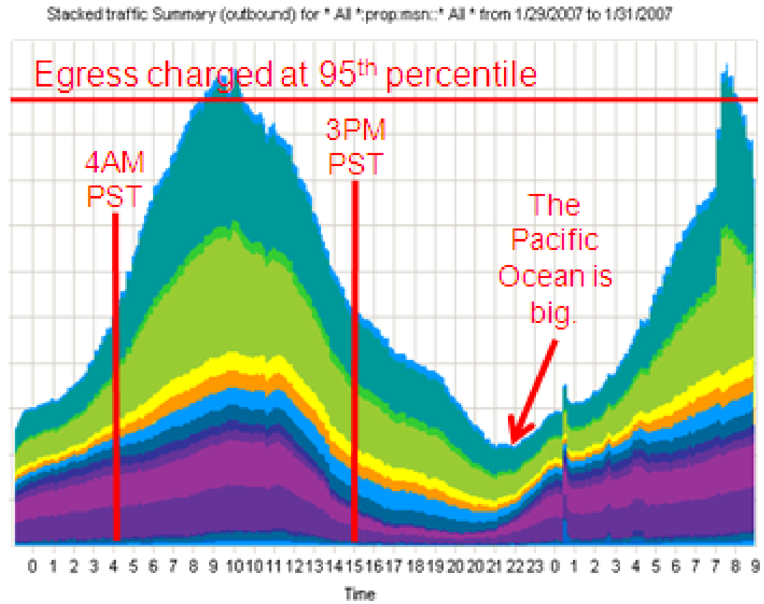
- **CEMS V2 Comparison:**
  - **Work Done/$: +375%**
  - **Work Done/Joule +379%**
  - **Work Done/Rack: +942%**

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

Details at: http://perspectives.mvdirona.com/2009/01/23/MicrosliceServers.aspx

Resource Consumption Shaping

- Essentially yield mgmt applied to full DC
- Network charge: base + 95th percentile
  - Push peaks to troughs
  - Fill troughs for “free”
  - Dynamic resource allocation
    - Virtual machine helpful but not needed
  - Symmetrically charged so ingress effectively free

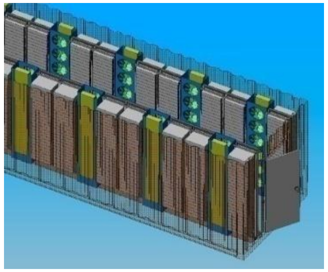


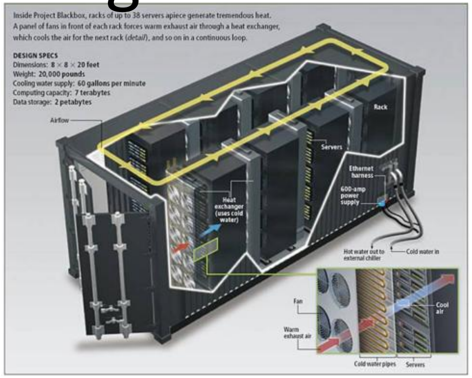
David Treadwell & James Hamilton / Treadwell Graph

- Power also often charged on base + peak
  - 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

Existing Builds: Containers

- Existing enterprise deployments often:
  - Very inefficient with PUE in 2 to 3 range
  - Out of cooling, out of power & out of space
- Rather than continue to grow bad facility
  - Drop container on roof or parking lot
  - Convert existing data center to offices or other high value use
- Easy way to get PUE to 1.35 range





Existing Builds: Cloud Services

- Deploy new or non-differentiated workloads to cloud
  - Focus the on-premise facility to differentiated computing that adds value to the business
  - Focus people resources on revenue generating, differentiated IT work
- No upfront capital outlay
- Very high scale, cloud service deployments offer lower costs and can be more efficient
  - Better for environment & lower cost

Summary

- Average DCs have considerable room to improve
- Use tPUE rather PUE to track improvement
- Power & related costs drive infrastructure expenses
  - Don't use floor space or rack positions as metric
- Server costs still (barely) dominate power
- What to do with existing, inefficient infrastructure
  - Modular data center designs
  - Utility computing

More Information

- **This Slide Deck:**
  - I will post these slides to http://mvdirona.com/jrh/work later this week
- **Berkeley Above the Clouds**
  - http://perspectives.mvdirona.com/2009/02/13/BerkeleyAboveTheClouds.aspx
- **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
- **Perspectives Blog**
  - http://perspectives.mvdirona.com
- **Email**
  - James@amazon.com