Internet-Scale Service Infrastructure Efficiency

International Symposium on System Architecture

James Hamilton, 2009/6/23

VP & Distinguished Engineer, Amazon Web Services

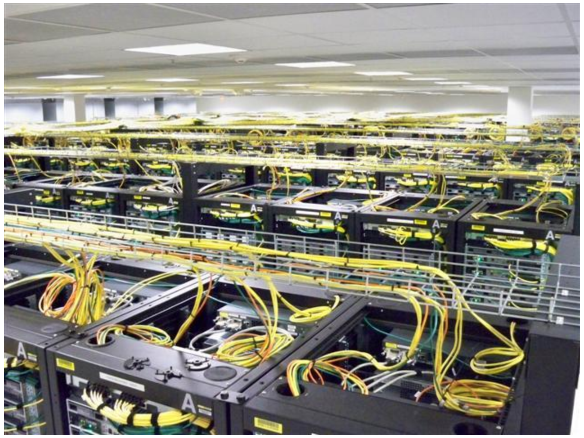
e: James@amazon.com

w: mvdirona.com/jrh/work

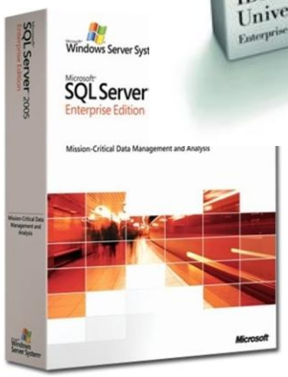
b: perspectives.mvdirona.com

Agenda

- High Scale Services
  - Infrastructure cost breakdown
  - Where does the power go?
- Power Distribution Efficiency
- Mechanical System Efficiency
- Server & Applications Efficiency
  - Hot I/O workloads & NAND flash
  - Resource consumption shaping
  - Work done per joule & per dollar



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
- Talk does not necessarily represent positions of current or past employers

Services Different from Enterprises

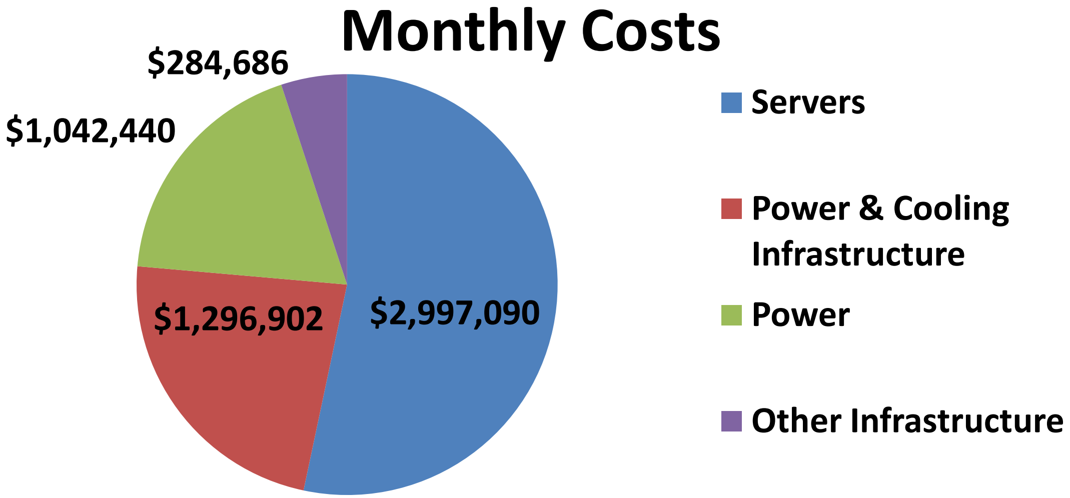
- **Enterprise Approach:**
  - Largest cost is people -- scales roughly with servers (~100:1 common)
  - Enterprise interests center around consolidation & utilization
    - Consolidate workload onto fewer, larger systems
    - Large SANs for storage & large routers for networking

- **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

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

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%
- Help evangelize **tPUE** (power to server components)
  - http://perspectives.mvdirona.com/2009/06/15/PUEAndTotalPowerUsageEfficiencyTPUE.aspx

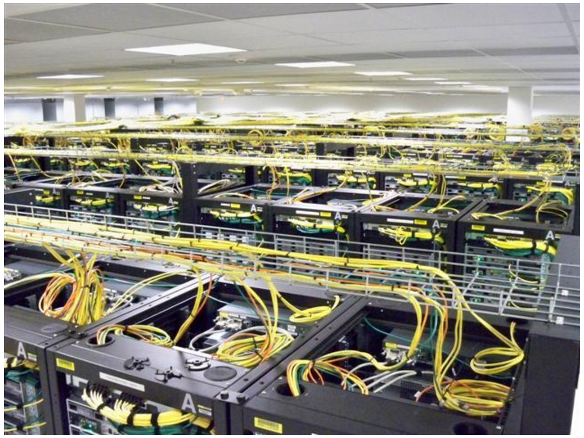
http://www.thegreengrid.org/en/Global/Content/white-papers/The-Green-Grid-Data-Center-Power-Efficiency-Metrics-PUE-and-DCiE

Where Does the Power Go?

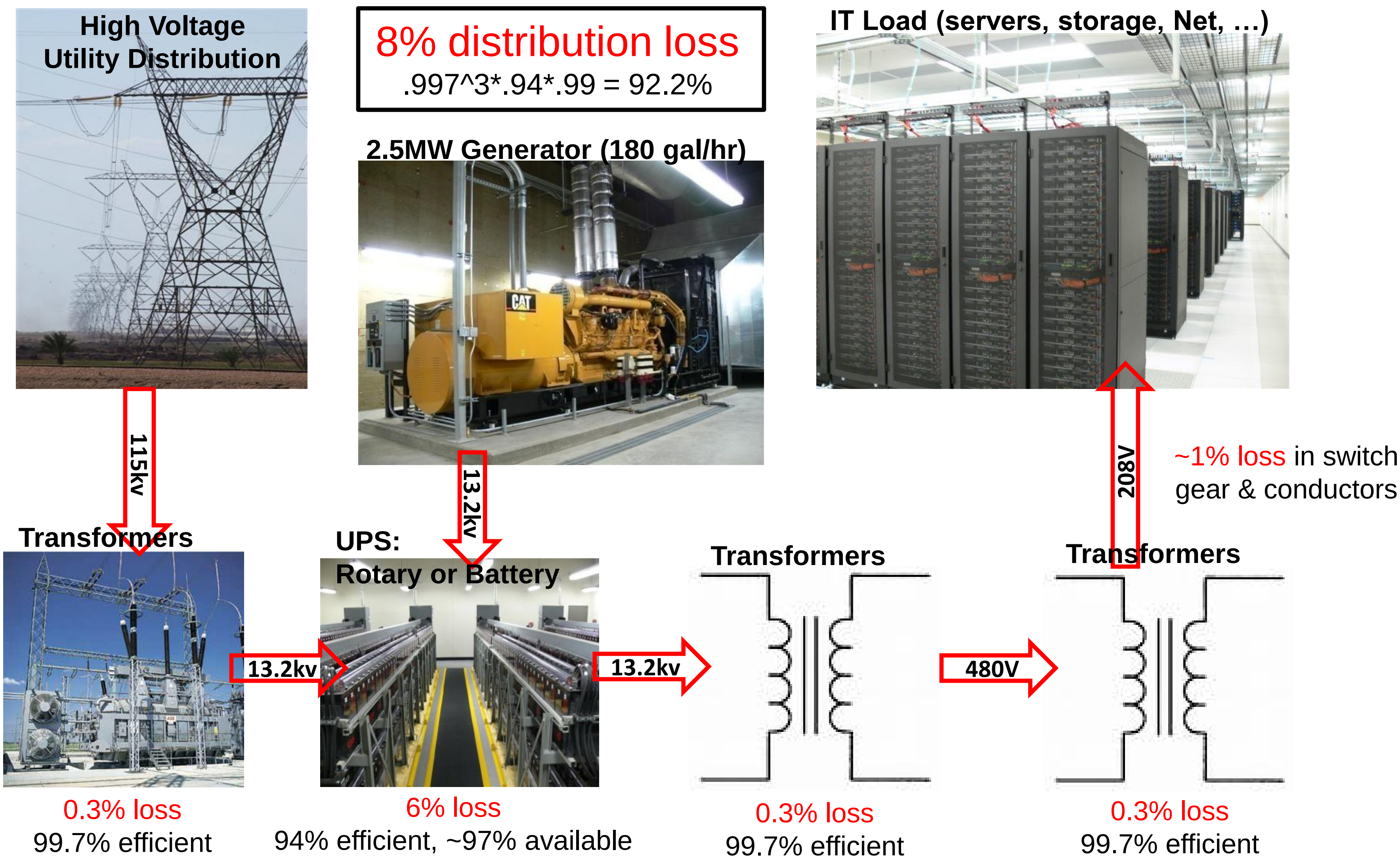
- **Assuming a pretty good data center with PUE ~1.7**
  - Each watt to server loses ~0.7W to power distribution losses & cooling
  - IT load (servers): 1/1.7=> 59%
- **Power losses are easier to track than cooling:**
  - Power transmission & switching losses: 8%
    - Detailed power distribution losses on next slide
  - Cooling losses remainder:100-(59+8) => 33%
- **Observations:**
  - **Server efficiency & utilization improvements highly leveraged**
  - **Cooling costs unreasonably high**

Agenda

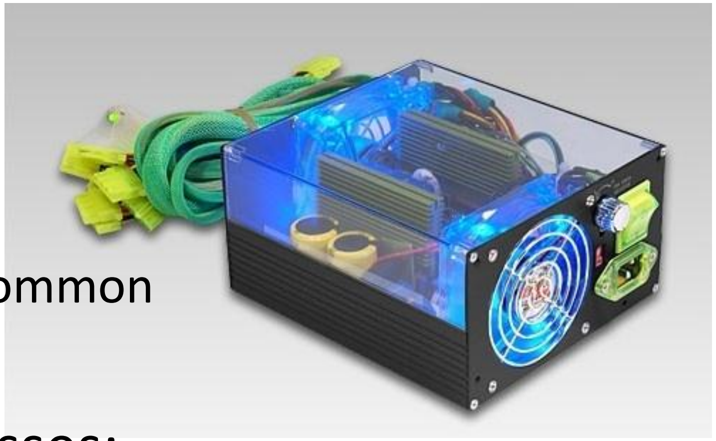
- High Scale Services
  - Infrastructure cost breakdown
  - Where does the power go?
- Power Distribution Efficiency
- Mechanical System Efficiency
- Server & Applications Efficiency
  - Hot I/O workloads & NAND flash
  - Resource consumption shaping
  - Work done per joule & per dollar



Power Distribution



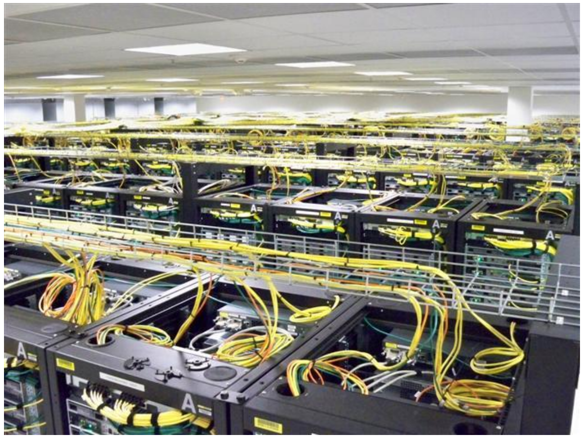
Power Distribution Efficiency Summary



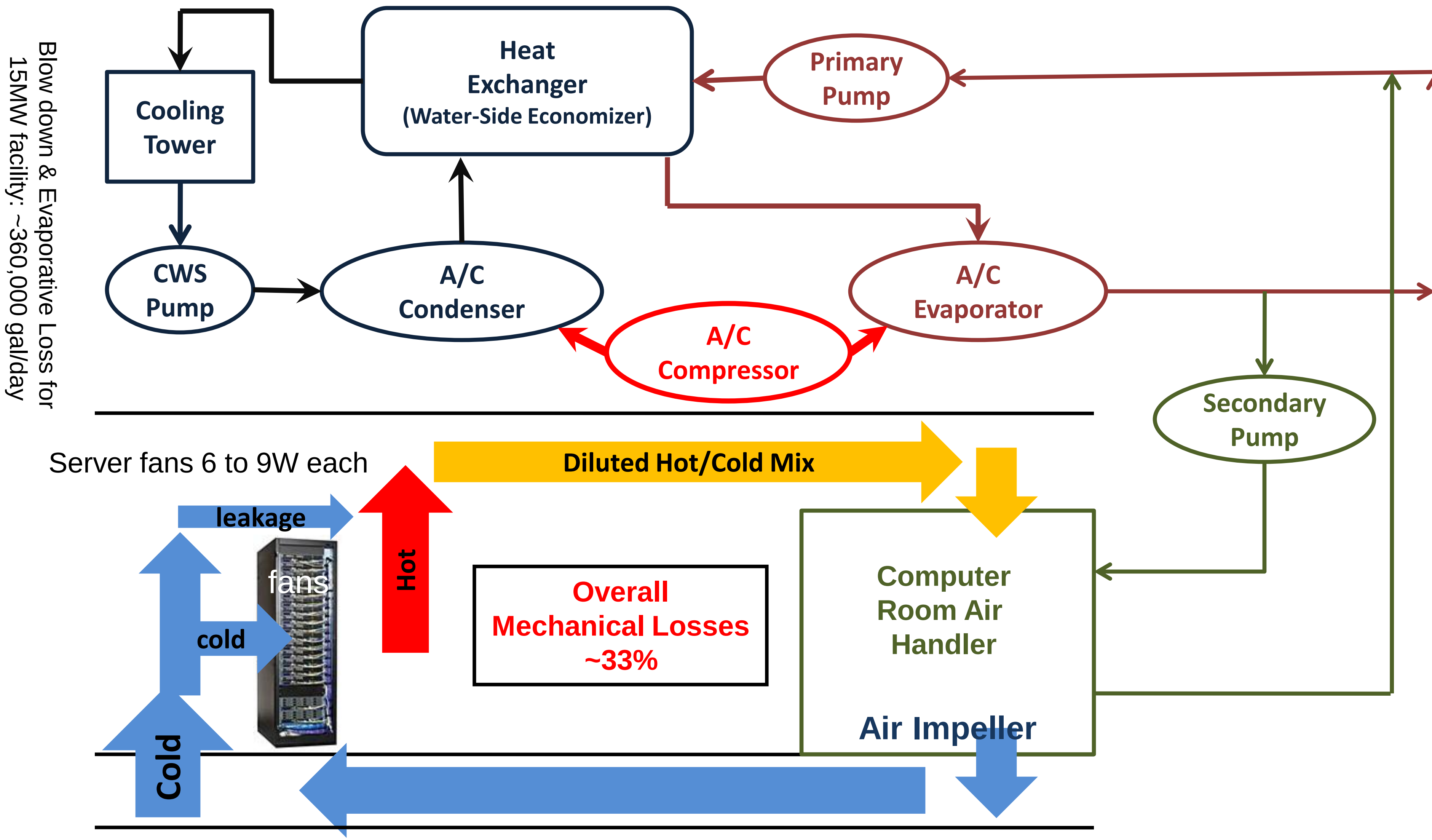
- Two additional conversions in server:
  1. Power Supply: often <80% at typical load
  2. On board step-down (VRM/VRD): <80% common
     - ~95% efficient both available & affordable
- Rules to minimize power distribution losses:
  1. Oversell power (more theoretic load that power)
  2. Avoid conversions (fewer transformer steps & efficient UPS)
  3. Increase efficiency of conversions
  4. High voltage as close to load as possible
  5. Size VRMs & VRDs to load & use efficient parts
  6. DC distribution potentially a small win (regulatory issues)

Agenda

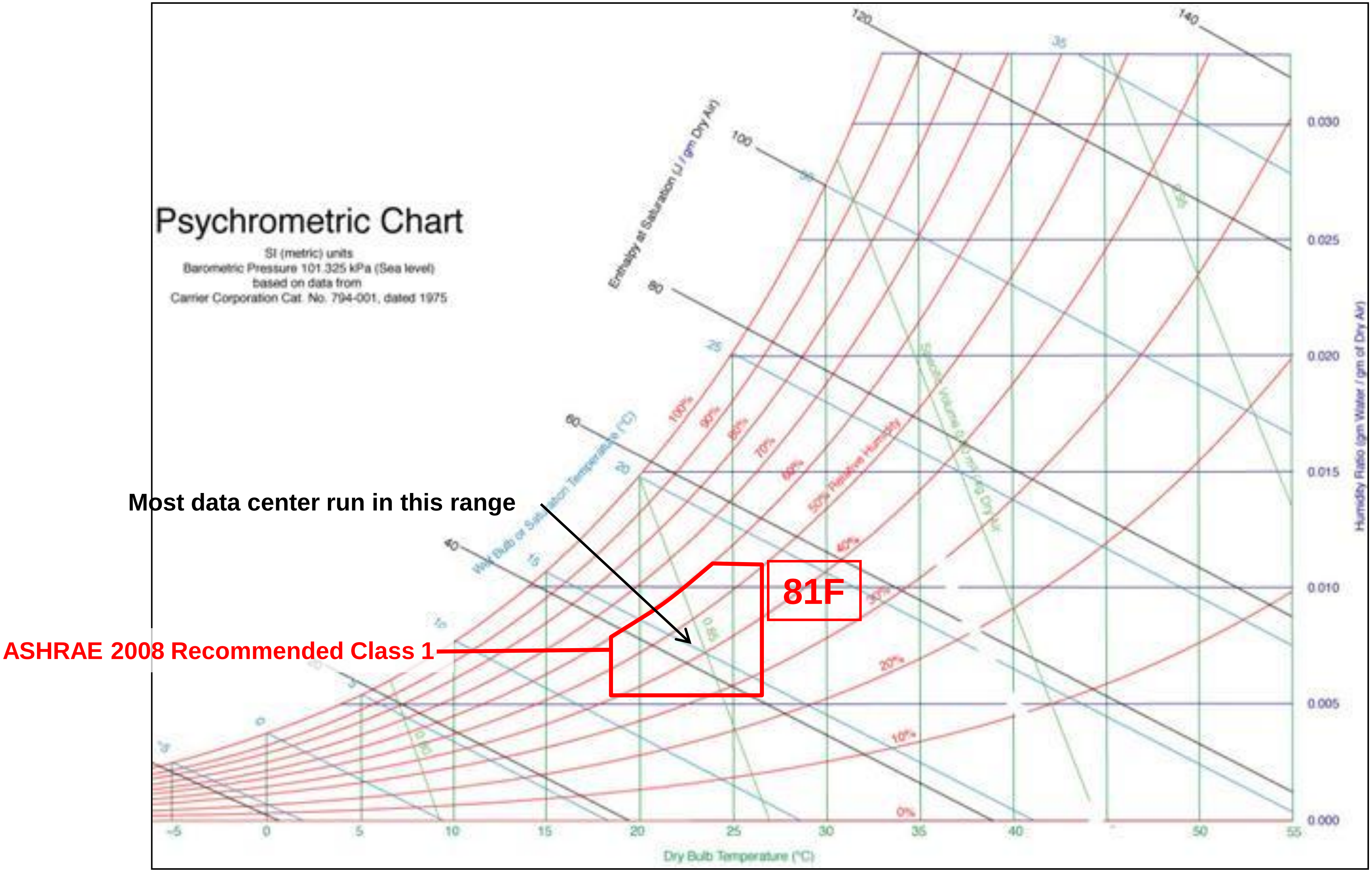
- High Scale Services
  - Infrastructure cost breakdown
  - Where does the power go?
- Power Distribution Efficiency
- Mechanical System Efficiency
- Server & Applications Efficiency
  - Hot I/O workloads & NAND flash
  - Resource consumption shaping
  - Work done per joule & per dollar



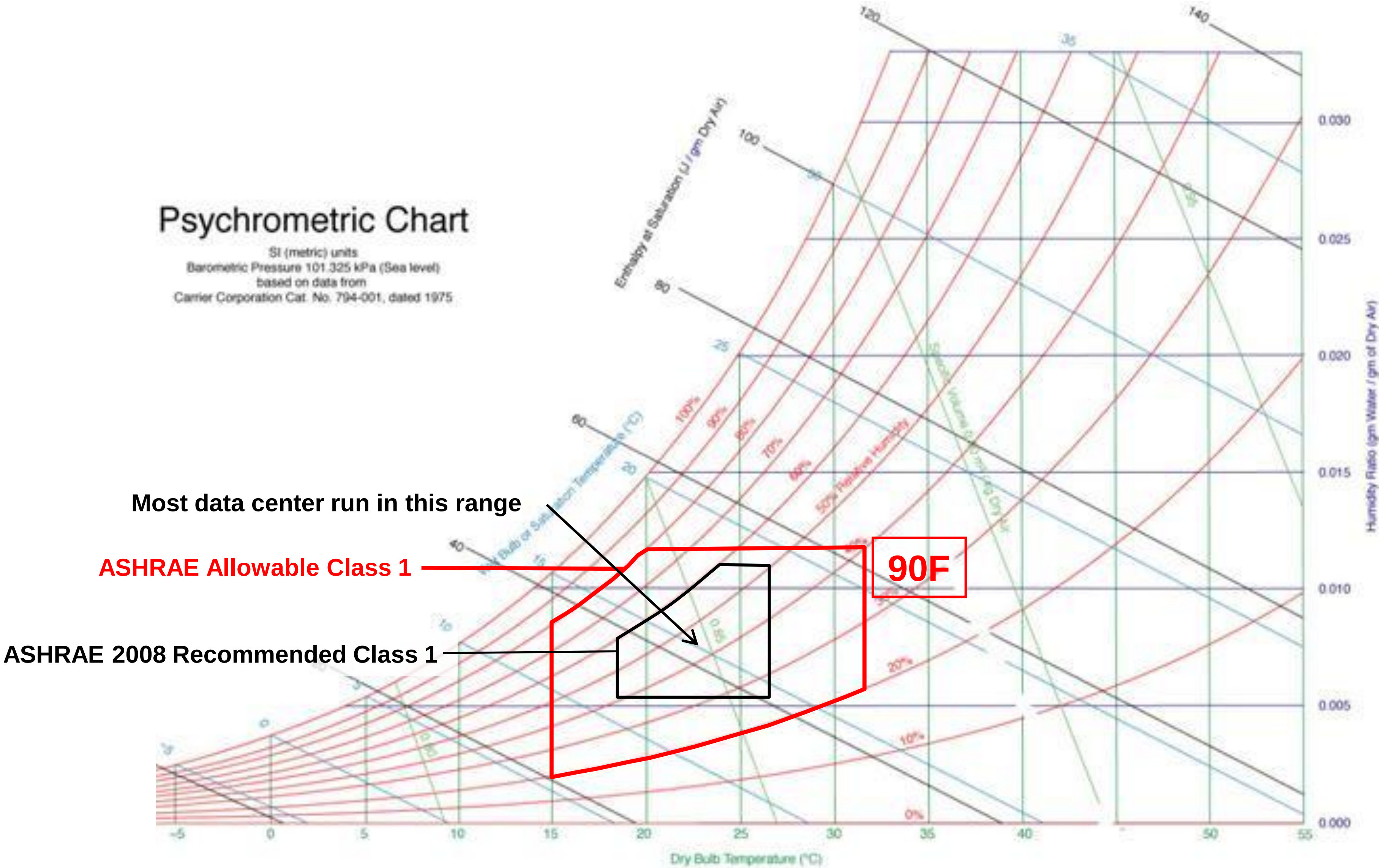
Conventional Mechanical Design



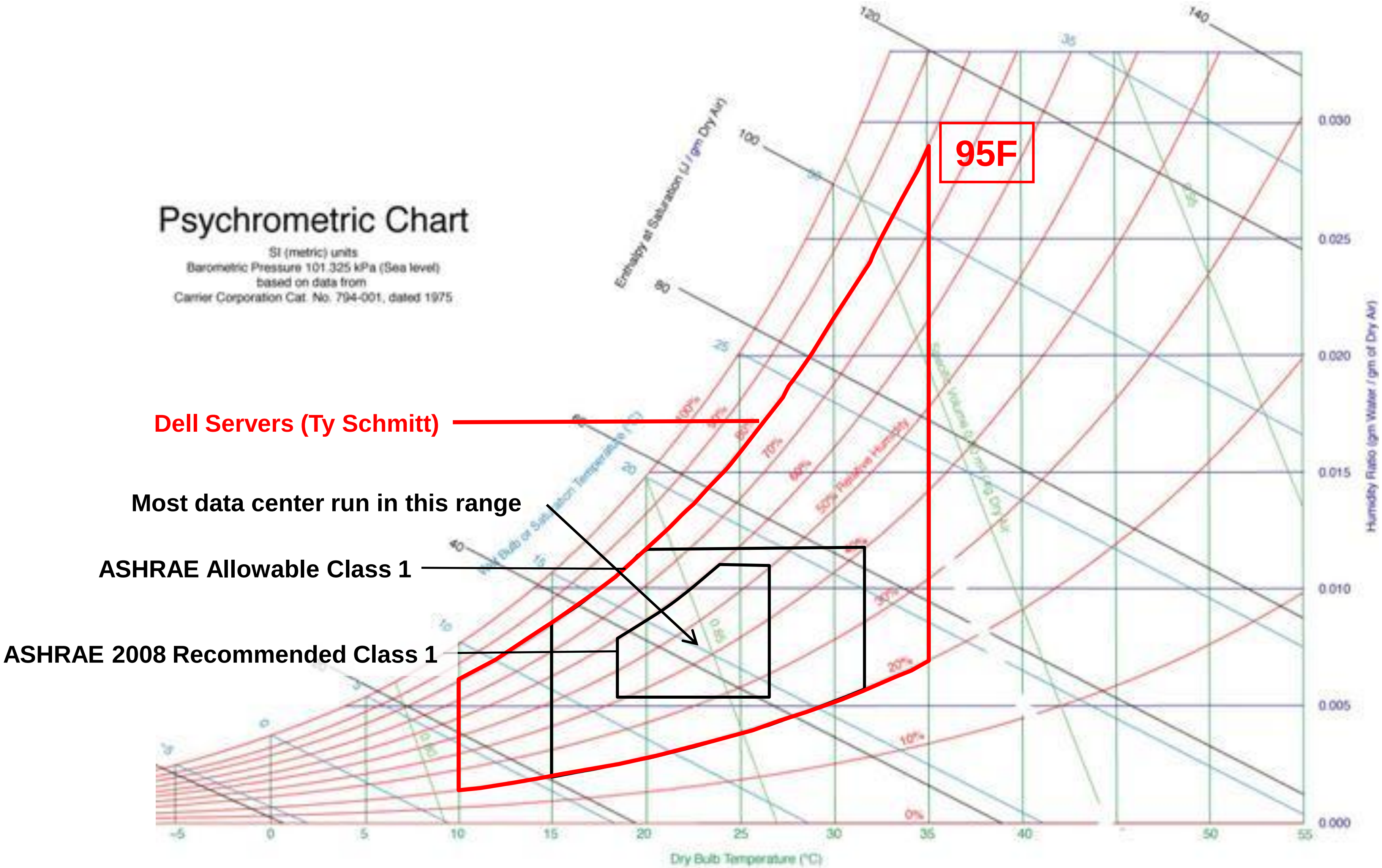
ASHRAE 2008 Recommended



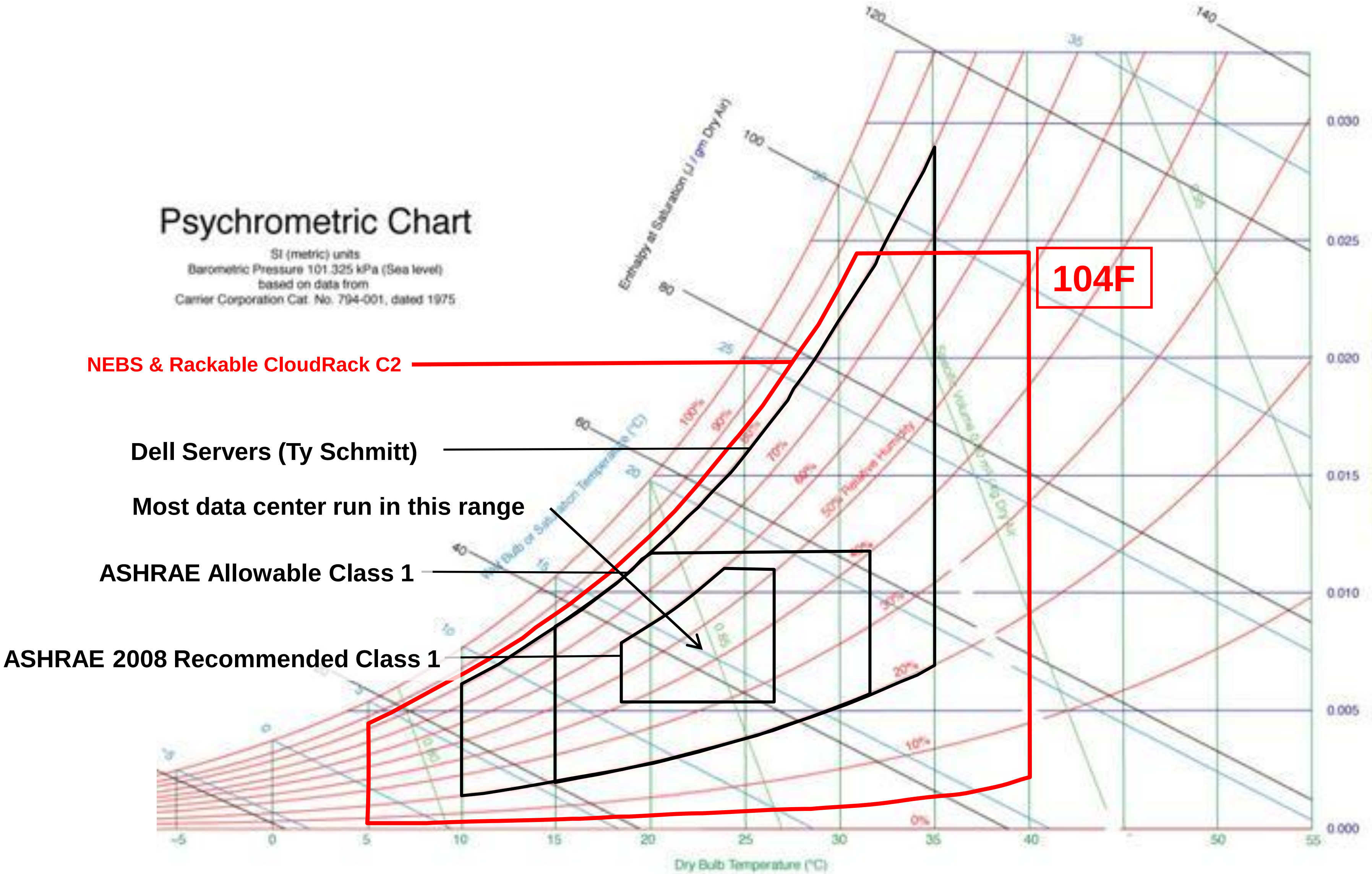
ASHRAE Allowable



Dell PowerEdge 2950 Warranty

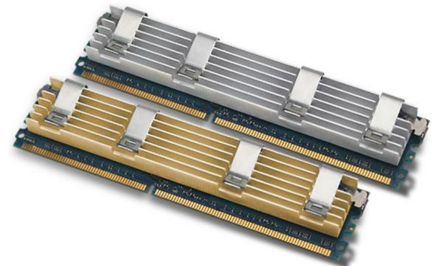


NEBS (Telco) & Rackable Systems



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 & better mechanical designs
  - Tradeoff: power to move air vs cooling savings & semi-conductor leakage current
  - 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

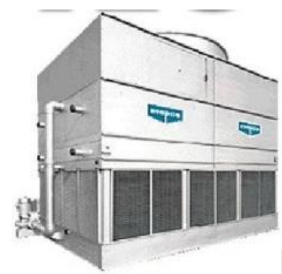
I/O: 5W - 25W
Temp Spec: 50C-60C

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

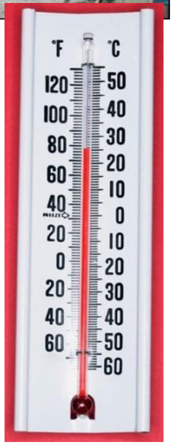


- Particulate concerns:
  - Usage of outside air during wildfires or datacenter generator operation
  - Solution: filtration & filter admin or heat wheel & related techniques

- Others: higher fan power consumption, more leakage current, higher failure rate

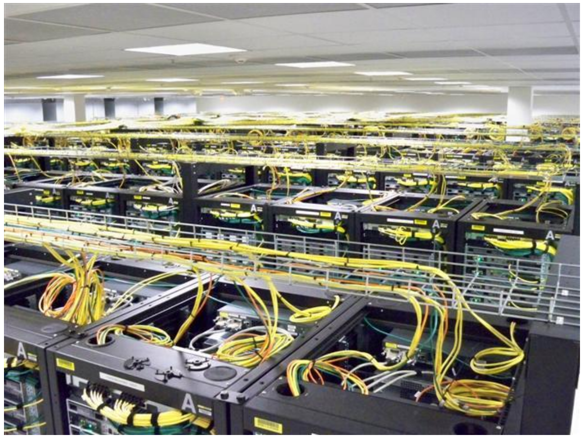
Mechanical Efficiency Summary

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

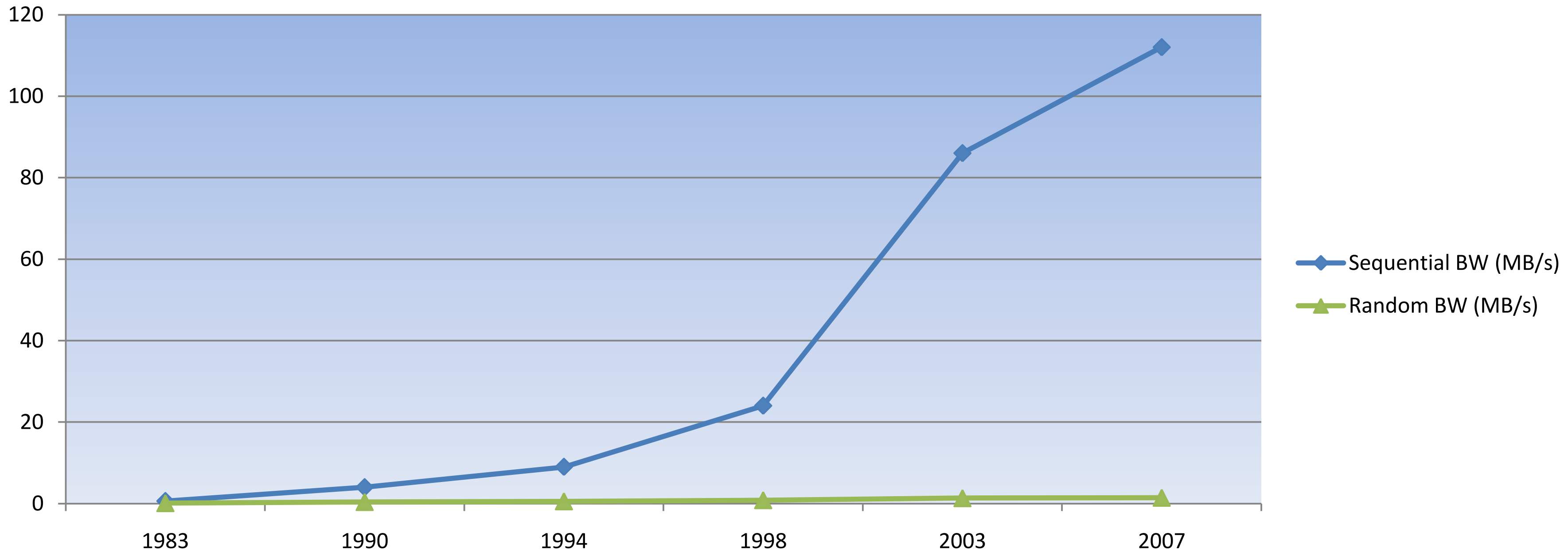


Agenda

- High Scale Services
  - Infrastructure cost breakdown
  - Where does the power go?
- Power Distribution Efficiency
- Mechanical System Efficiency
- Server & Applications Efficiency
  - Hot I/O workloads & NAND flash
  - Resource consumption shaping
  - Work done per joule & per dollar



Disk Random BW vs Sequential BW



Source: Dave Patterson with James Hamilton updates

- Disk sequential BW lagging DRAM and CPU
- Disk random access BW growth ~10% of sequential
- **Conclusion**: Storage Chasm widening requiring larger memories & more disks

Memory to Disk Chasm

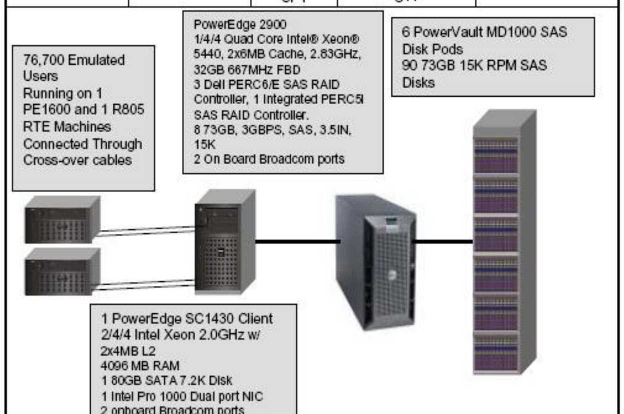
- Disk I/O rates grow slowly while CPU data consumption grows near Moore pace
  - Random read 1TB disk: 15 to 150 days*
- Sequentialize workloads
  - Essentially the storage version of cache conscious algorithms
    - e.g. map/reduce
  - Disks arrays can produce acceptable aggregate sequential bandwidth
- Redundant data: materialized views & indexes
  - Asynchronous maintenance
  - Delta or stacked indexes (from IR world)
- Distributed memory cache (remote memory "closer" than disk)
- I/O Cooling: Blend hot & cold data (using HDD)
- I/O concentration: partition hot & cold (SSD & HDD mix)

**** Tape is Dead, Disk is Tape, Flash is Disk, Ram Locality is King (Jim Gray)***

Case Study: TPC-C with SSD

DELL	PowerEdge 2900 Server with Oracle Database 11g Standard Edition One		TPC-C Rev 5.9 Original Report Date June 16 ,2008
Total System Cost	TPC-C Throughput	Price/Performance	Availability Date
$65,910	**97,083 tpmC**	**$.68 / tpmC**	June 16, 2008

Processors	Database Manager	OS	Other Software	Number of Users
1/4/4 Quad Core Intel® Xeon® 5440,2X6MB Cache, 2.83GHZ 1333MHZ FSB	Oracle Database 11g Standard Edition One	Microsoft Windows Server 2003 Standard x64 Edition SP1	Windows Server 2003 Standard Edition w/ COM+ Internet Information Server 6.0 Microsoft Visual C++	76,700



System Component	**Server**		**Each Client**	
Processor/Core/Cache	1	1/4/4 Quad Core Intel® Xeon® 5440, 2x6MB Cache, 2.83GHz, 1333	2	2/4/4 Intel® Xeon® w/ 2x4MB L2, 2.0 GHz
Memory		32GB 667 FB-DIMM		4 GB
Disk Controllers	3 1	Dell PERC6/E RAID Integrated PERC5i RAID	1	Onboard SATA
Disk Drives	38 60	73GB SAS 15K 36GB SAS 15K	1	80GB 7.2K SATA
Total Storage	98	4011 GB SAS	1	80GB SATA
Other	2 1	Broadcom NetXtreme II GigE ports CD-ROM	2 1 1	Broadcom on-board ports Dual port Intel Pro 1000 NIC CD-ROM

Slot	Controller	Disks		Capacity		Usage
0	Dell PERC5i	8x73GB,15K,SAS	RAID10	Disk 6	15GB	OS
				279.99GB	260GB	Logs
3	Dell PERC6/E	15x36GB,15K,SAS	RAID0	Disk 2	488.92GB	DB data
		15x36GB,15K,SAS	RAID0	Disk 3	488.92GB	DB data
4	Dell PERC6/E	15x36GB,15K,SAS	RAID0	Disk 4	488.92GB	DB data
		15x36GB,15K,SAS	RAID0	Disk 5	488.92GB	DB data
6	Dell PERC6/E	15x73GB,15K,SAS	RAID0	Disk 0	1016.23GB	DB data
		15x73GB,15K,SAS	RAID0	Disk 1	1016.23GB	DB data

O/S & Log

Data

- 98 HDD total
  - 90 data disks (primarily random access)
  - 8 log & O/S disks (primarily sequential access)
- Compute HDD/SSD cross-over using fictitious SSD
  - 128GB SSD @ 7k IOPS
- 90 HDD to store 2,464GB (short stroked)
  - 106GB static & 2,357GB dynamic (60 day rule)
  - 90 disk HDD budget: $26,910 (disks $299 each)
  - ***Requires 20 SSDs to support @ up to $1,346 each***
- Static content only (drop 60 day rule)
  - Conservatively estimate 45k IOPS
    - Used 90 short stroked disks at 500 IOPS each
  - ***Requires 7 SSDs at up to $3,844 (easy)***
- **Very hot I/O workloads a win on SSD**

http://www.tpc.org/results/FDR/TPCC/Dell_2900_061608_fdr.pdf

Summary

- CPU optimizations are always welcome but the biggest design & optimization problems today are at the datacenter level
- In work at all levels, focus on:
  - Work done per dollar
  - Work done per joule
- Single dimensional performance measurements are not interesting at scale unless balanced against cost

More Information

- **This Slide Deck:**
  - I will post these slides to http://mvdirona.com/jrh/work later this week
- **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
- **Email**
  - James@amazon.com