Datacenter Power Efficiency:
Separating Fact from Fiction

Kushagra Vaid
Principal Architect, Datacenter Infrastructure
Microsoft Online Services Division
kvaid@microsoft.com

Overview – Microsoft Online Services

Microsoft Datacenters: Providing services 24x7

More than 1 billion authentications per day

More than 2 billion queries per month

Microsoft Exchange Hosted Services

Processes 2–4 billion e-mails per day

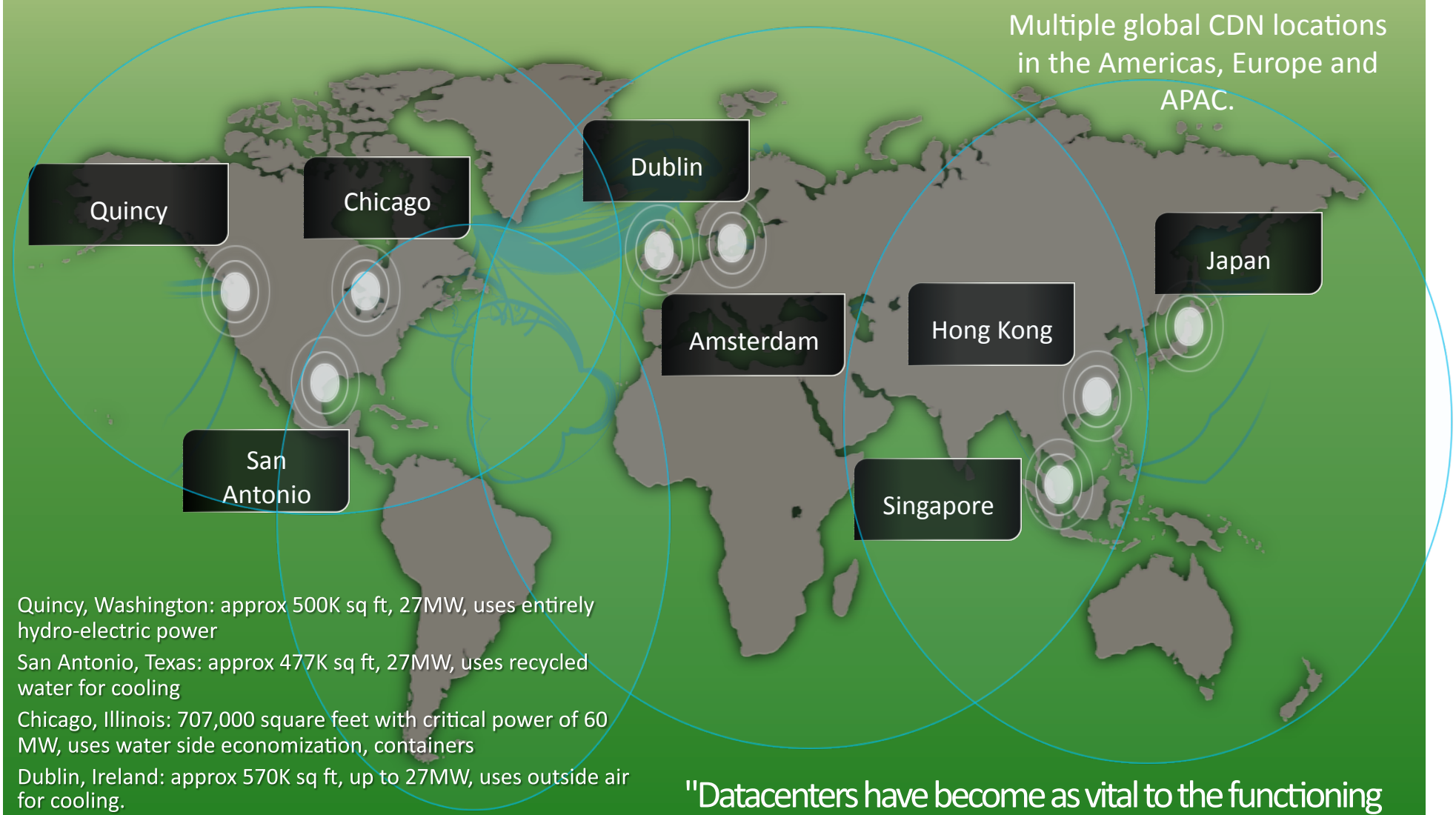
320 million active accounts

550 million unique visitors monthly

400 million active accounts

Microsoft Datacenters – Global scaling!

Multiple global CDN locations in the Americas, Europe and APAC.



Quincy, Washington: approx 500K sq ft, 27MW, uses entirely hydro-electric power

San Antonio, Texas: approx 477K sq ft, 27MW, uses recycled water for cooling

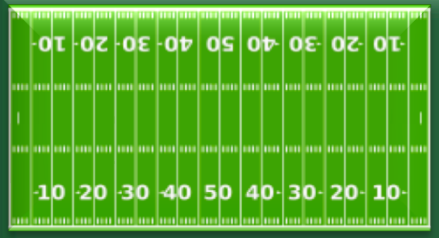
Chicago, Illinois: 707,000 square feet with critical power of 60 MW, uses water side economization, containers

Dublin, Ireland: approx 570K sq ft, up to 27MW, uses outside air for cooling.

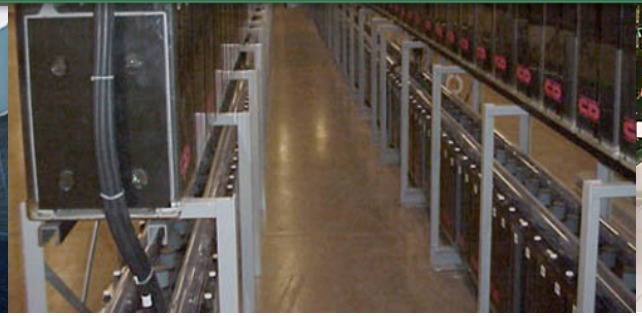
"Datacenters have become as vital to the functioning of society as power stations." – *The Economist*










A large mega-datacenter is

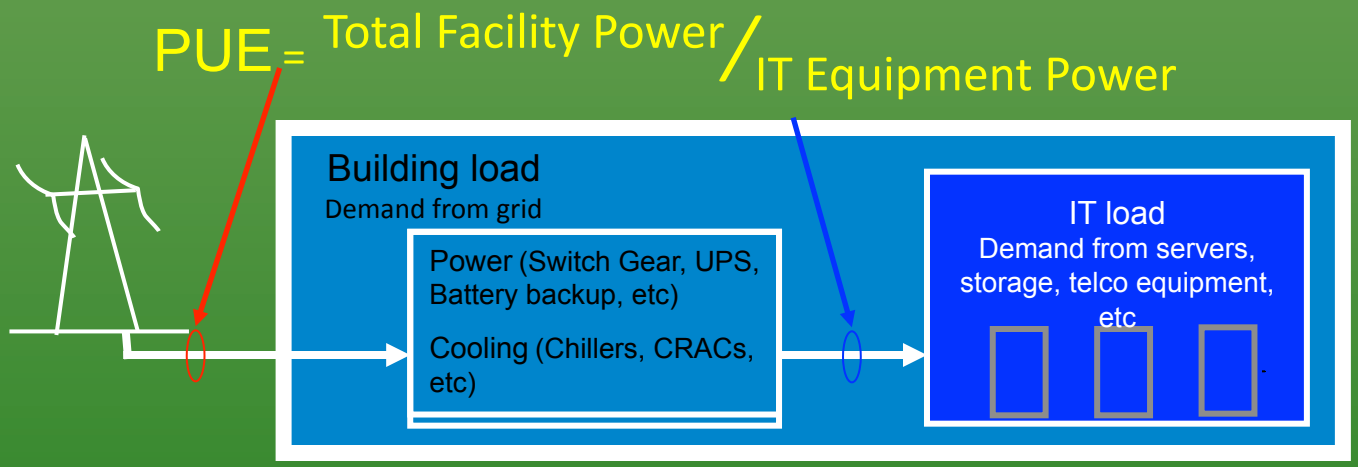
11 times

the size of a football field

Quick cost and efficiency facts

For a typical large mega-datacenter…

Building costs are between $10M to $15M per MegaWatt



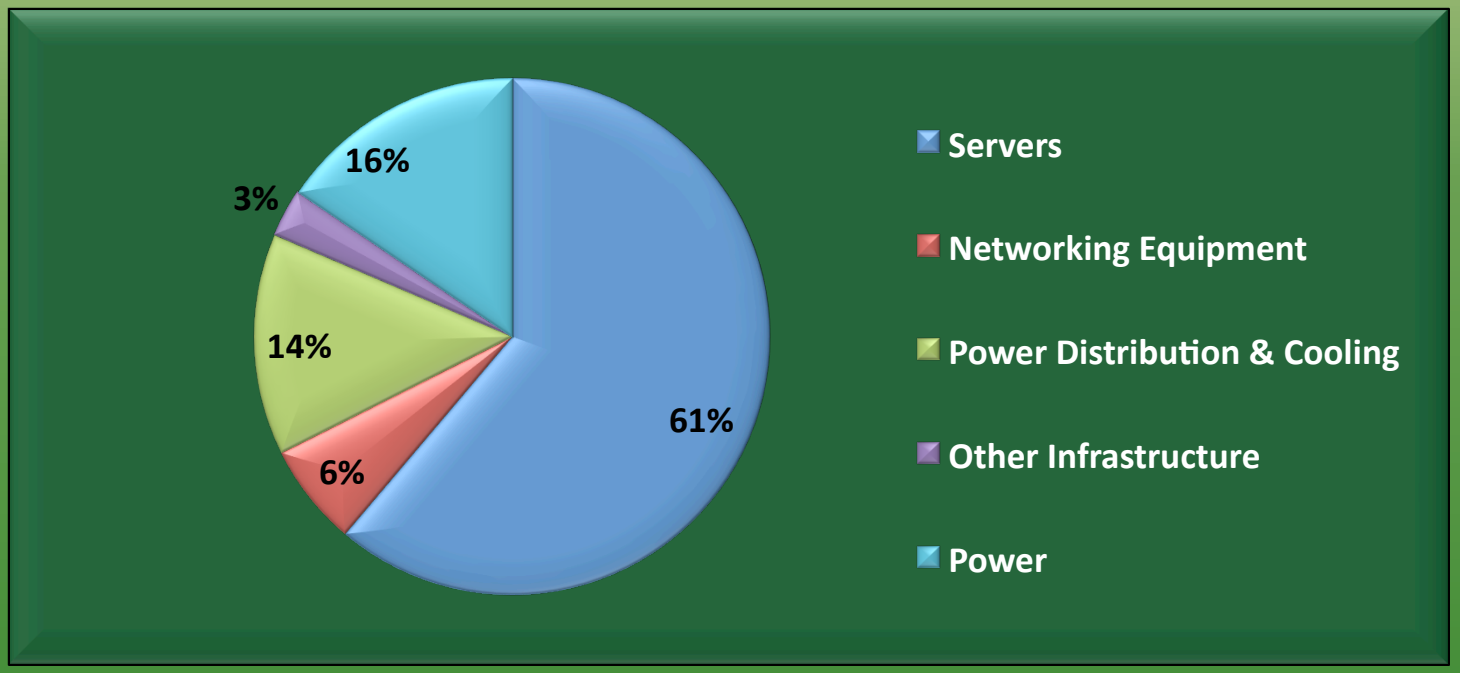
More about PUE: http://thegreengrid.org/en/Global/Content/white-papers/The-Green-Grid-Data-Center-Power-Efficiency-Metrics-PUE-and-DCiE

Typical industry PUE ranges from 1.5-2.0

Energy Consumption: US power rate 10.27 cents per Kilowatt hour) according to DOE/eia (http://www.eia.doe.gov/cneaf/electricity/epm/table5_3.html)

Datacenter TCO breakdown

Assumptions:
10MW facility
PUE 1.25
$10/W construction costs
$0.10c/KWhr power costs
Server: $2000, 200W
3yr server amortization
15yr datacenter amortization



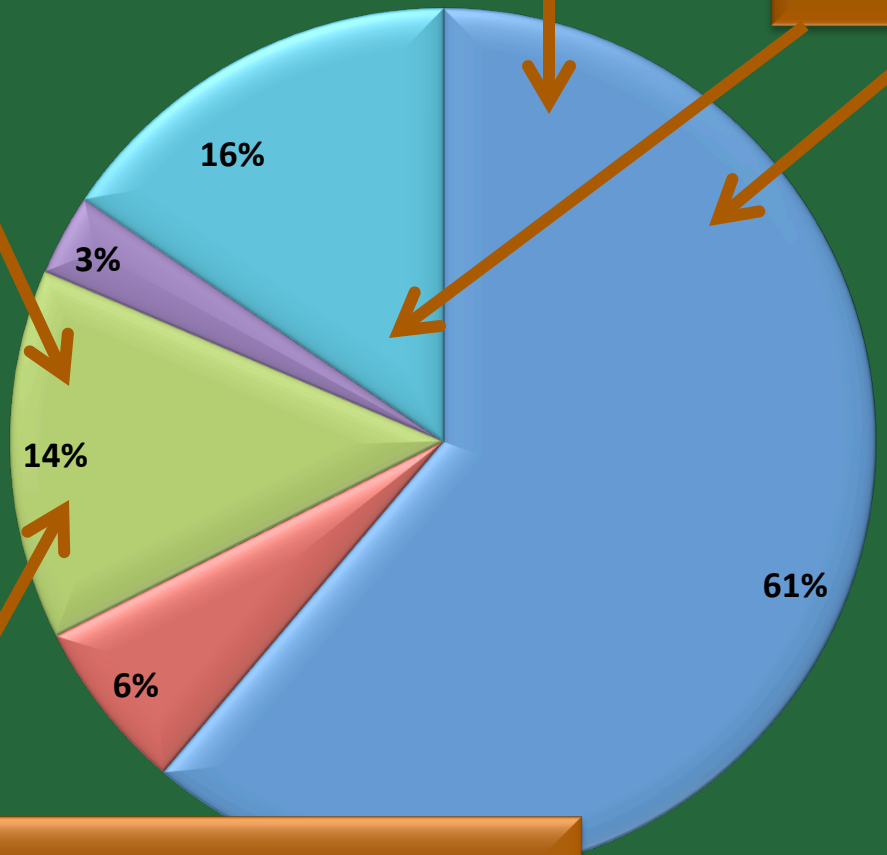
Source: James Hamilton (http://mvdirona.com/jrh/TalksAndPapers/PerspectivesDataCenterCostAndPower.xls)

- Power consumption cost is only 16%!
  - Reduction in dynamic power usage and energy proportional computing are important but have limited benefits
- Server capex accounts for largest portion of TCO (61%)
  - Invest in mechanisms to improve work done per watt
  - Improve performance and reduce power at cluster and datacenter level (not just single server)

Perspective on Datacenter power efficiency

Maximize power to IT load (improved PUE)
- Minimize losses and thermal/cooling overheads

Maximize work done per Watt
- Maximize compute density in power budget
- Maximize perf, minimize power consumption



Servers
Networking Equipment
Power Distribution & Cooling
Other Infrastructure
Power

Minimize cost of provisioning power
- Construction and cooling infrastructure

Datacenter Power Efficiency

Minimize cost of provisioning power

- **construction and cooling infrastructure**

Maximize power available to IT-load (improved PUE)

- Minimize losses and thermal/cooling overheads

Maximize work done per watt

- Maximize compute density in power budget
- Maximize performance, minimize power consumption

Microsoft's Chicago Data Center

Chicago Datacenter

- Two-story datacenter
  - First floor: container bay medium reliability
  - Second floor: high-reliability traditional co-location rooms
- Can deploy at scale
- Optimized energy and cooling efficiency
- Allows OEMs to build customized solutions based on Microsoft specifications

Microsoft's Datacenter Evolution

Datacenter Colocation
Generation 1

2005 2006

Server

Capacity
~2 PUE

San Antonio & Quincy
Generation 2

2007

Rack

Density and
Deployment
1.4 – 1.6 PUE

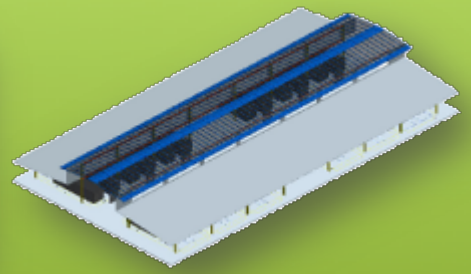
Chicago & Dublin
Generation 3

2008

Containers & Pods

Scalability and
Sustainability
1.2-1.5 PUE

Modular Datacenter
Generation 4



2009 2010+

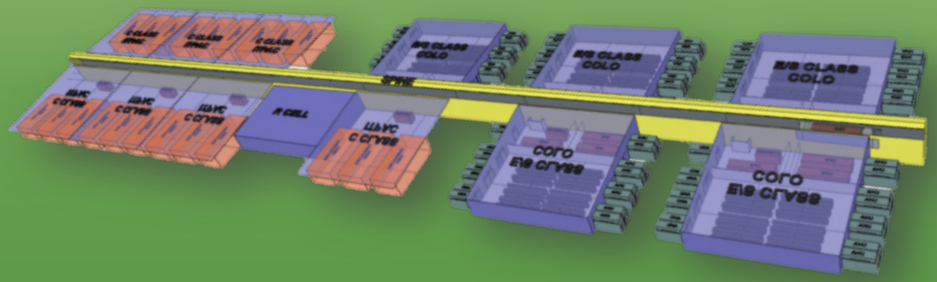
ITPAC

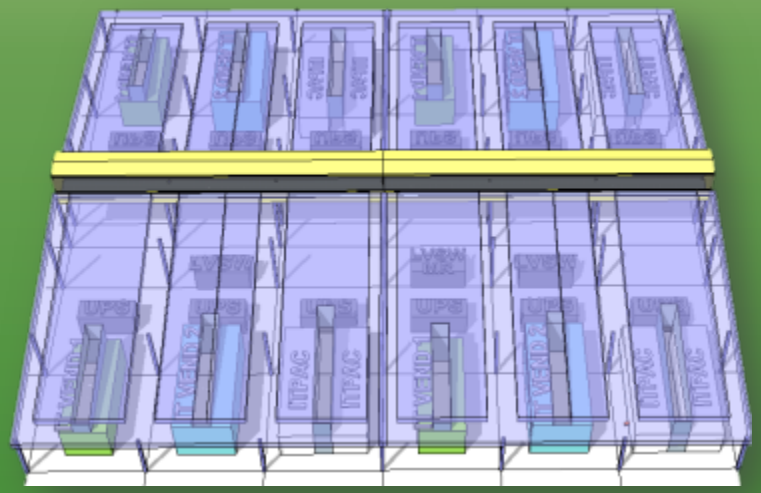
Faster Time to Market
Reduced Carbon
1.05-1.15 PUE

DEPLOYMENT SCALE UNIT

EFFICIENT RESOURCE USAGE

Microsoft's Modular Datacenters





- ***No mechanical cooling!***
- Ultra-efficient water utilization
- Focus on renewable materials
- 30-50 % more cost effective
- 1.05 – 1.15 PUE
- ITPAC is the "datacenter-in-a-box"

ITPAC design overview

video


Microsoft
IT Pre-Assembled-Components
(ITPAC)

Datacenter Power Efficiency

Minimize cost of provisioning power

- construction and cooling infrastructure

Maximize power available to IT-load (improved PUE)

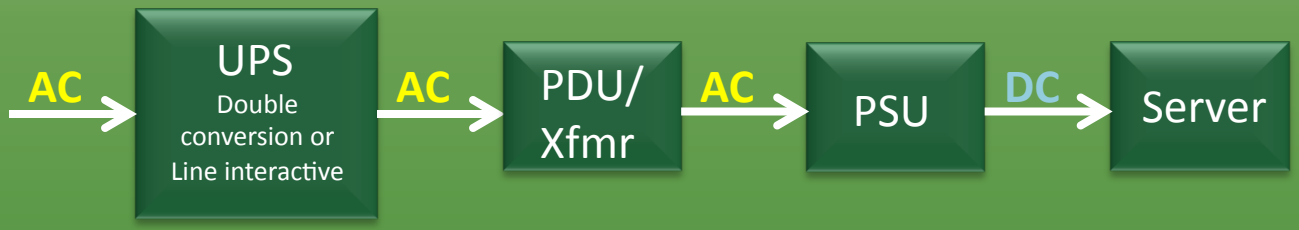
- **Minimize transmission/conversion losses** and thermal/cooling overheads

Maximize work done per watt

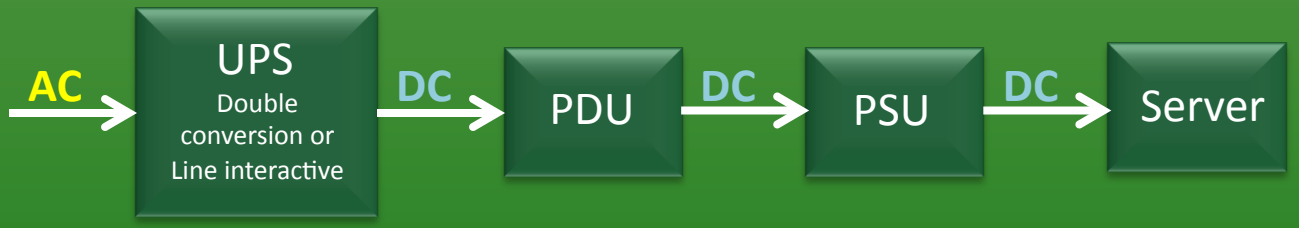
- Maximize compute density in power budget
- Maximize performance, minimize power consumption

Facility level distribution: AC or DC?

Consider the following common topologies …



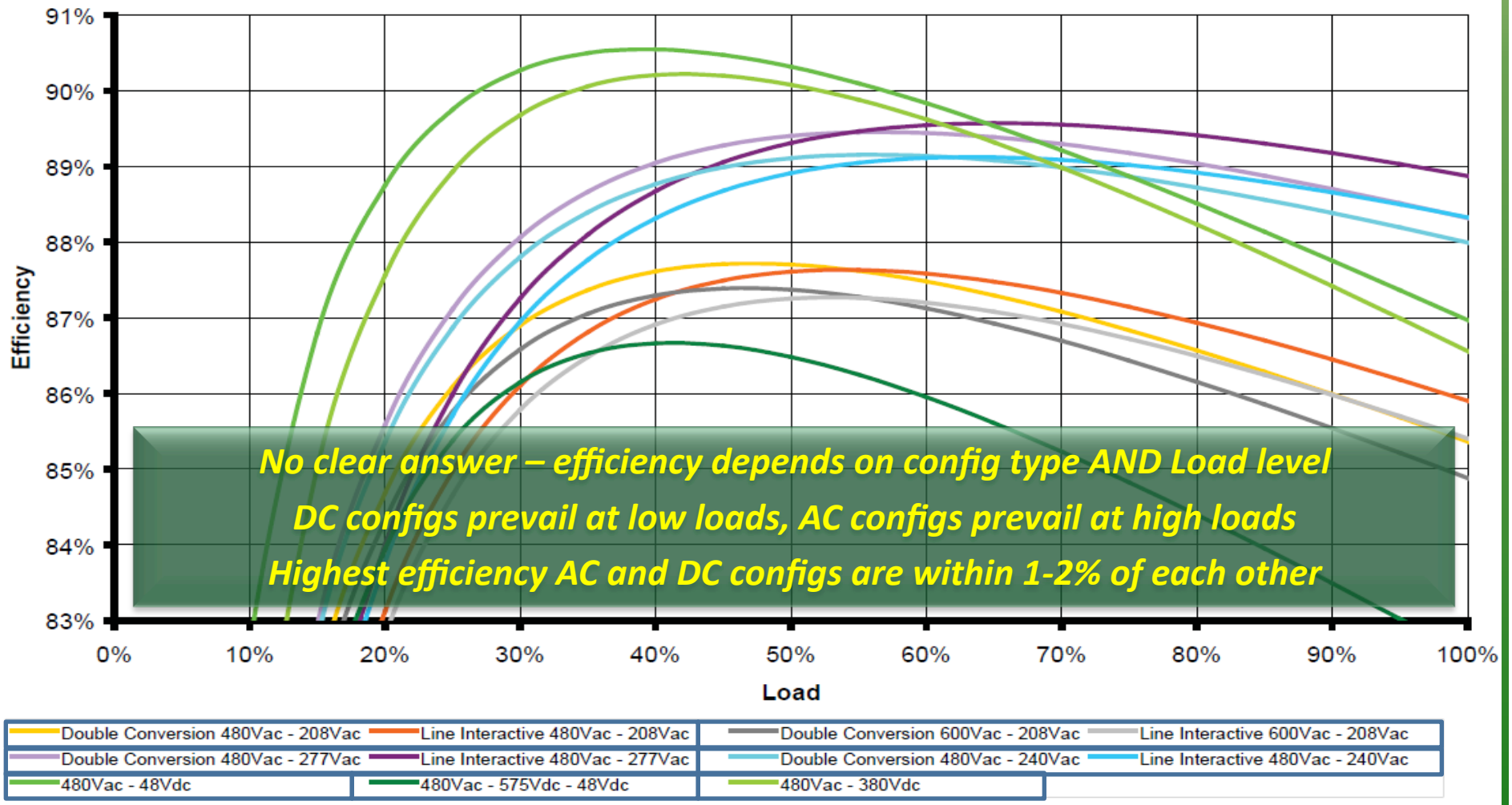
480Vac	480Vac	208Vac	12Vdc
480/277Vac	415/240Vac	240Vac	12Vdc
480/277Vac	480/277Vac	277Vac	12Vdc
600Vac	600Vac	208Vac	12Vdc



480Vac	575Vdc	48Vdc	12Vdc
480Vac	380Vdc	380Vdc	12Vdc
480/277Vac	48Vdc	48Vdc	12Vdc

Facility level distribution: AC or DC?

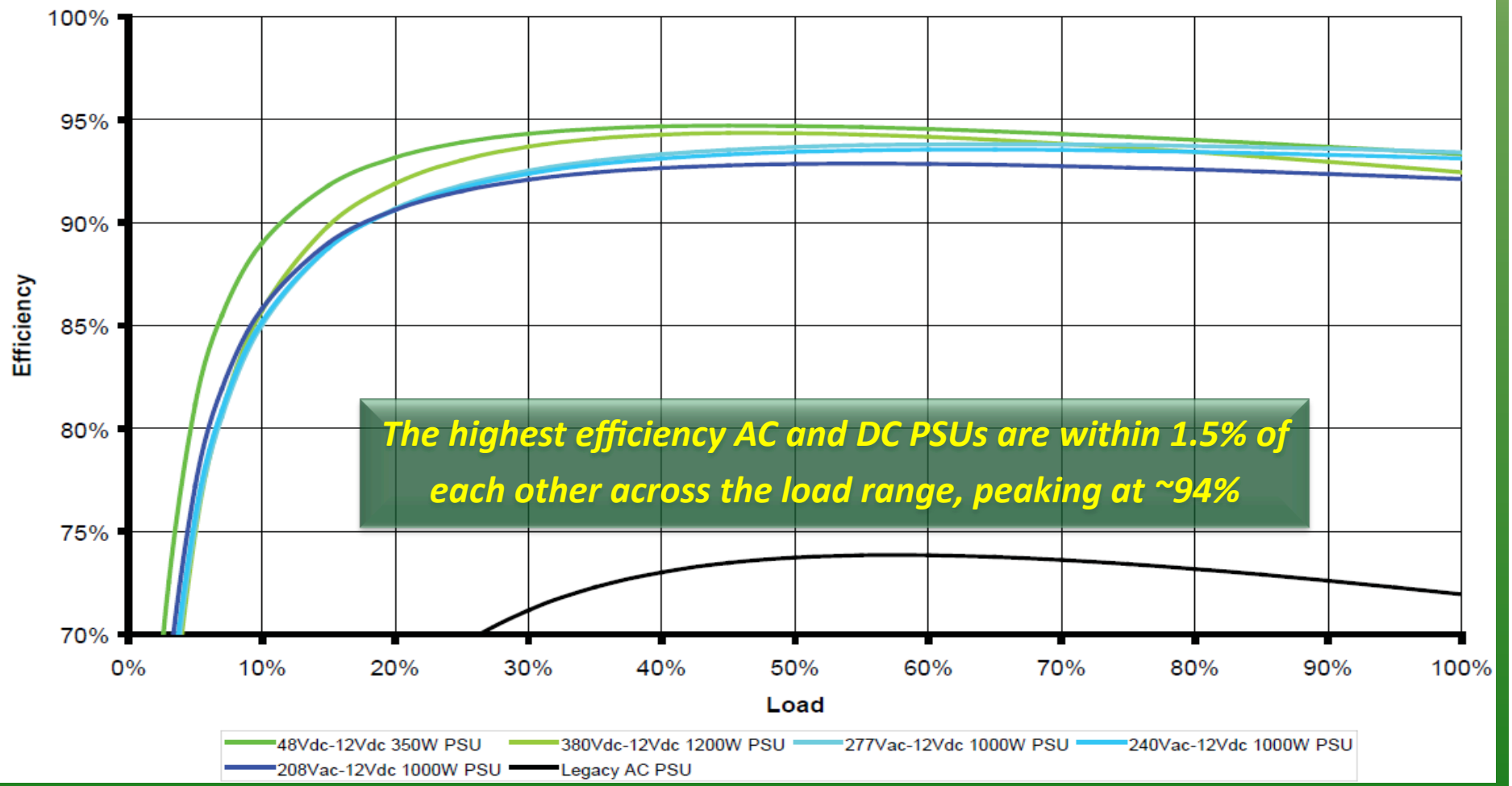
Which topology is the most efficient (lowest PUE)?



Source: Green Grid (http://www.thegreengrid.org/~/media/WhitePapers/White_Paper_16_-_Quantitative_Efficiency_Analysis_30DEC08.ashx)

AC vs DC PSU efficiency

Which PSU is the most efficient?



Source: Green Grid (http://www.thegreengrid.org/~/media/WhitePapers/White_Paper_16_-_Quantitative_Efficiency_Analysis_30DEC08.ashx)

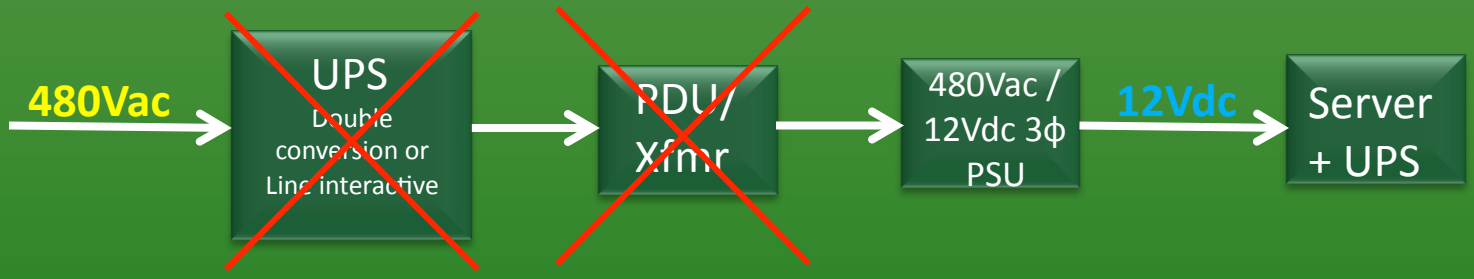
Other ways to remove losses

Phase balance

- 3ϕ power supplies → minimize losses from phase imbalance

Improve efficiency by eliminating conversions

- 480Vac-12Vdc direct conversion (no Xfmr) topologies
- Move UPS closer to IT load, preferably within the server/rack



Datacenter Power Efficiency

Minimize cost of provisioning power

- construction and cooling infrastructure

Maximize power available to IT-load (improved PUE)

- **Minimize** losses and **thermal/cooling overheads**

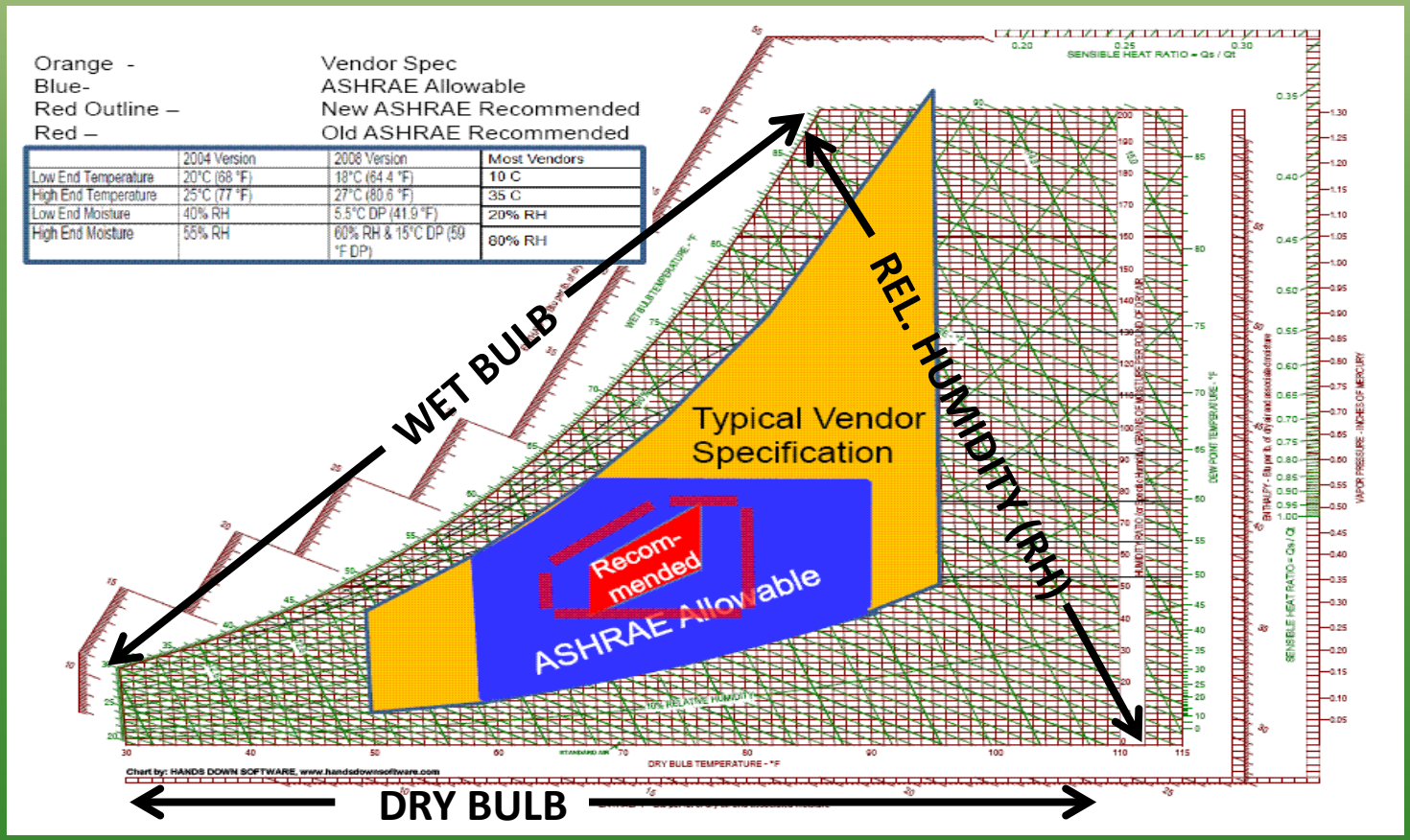
Maximize work done per watt

- Maximize compute density in power budget
- Maximize performance, minimize power consumption

Minimize thermal/cooling overheads

- For legacy datacenters …
  - Ensure better hot/cold aisle isolation → lower PUE
- Improved server chassis designs
  - Component placement for streamlined airflow
  - Shared fans across multiple chassis
- Expanded environmental range – High temp operation (more on next slide)
  - Allows for servers to operate with reduced airflow and still deal with temperature hotspots

High Temperature Server Operation



Source: http://media.techtarget.com/digitalguide/images/Misc/temp_hr.gif

ASHRAE recommended range: 64F-81F, max 60% RH

However, most Server vendors specify 50F-95F, max 90%RH

Higher temps → Lower fan speeds to cool servers → Improved PUE

Datacenter Power Efficiency

Minimize cost of provisioning power

- construction and cooling infrastructure

Maximize power available to IT-load (improved PUE)

- Minimize losses and thermal/cooling overheads

Maximize work done per watt

- **Maximize compute density in power budget**
- Maximize performance, minimize power consumption

Power capping for improving server density

- Power capping uses duty cycling based on P-states for ensuring target power level

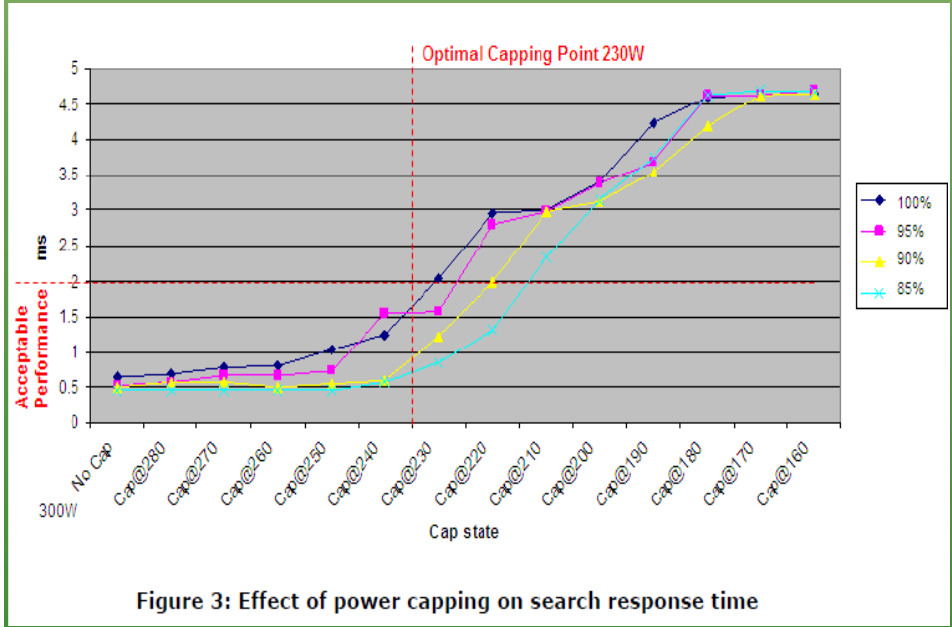
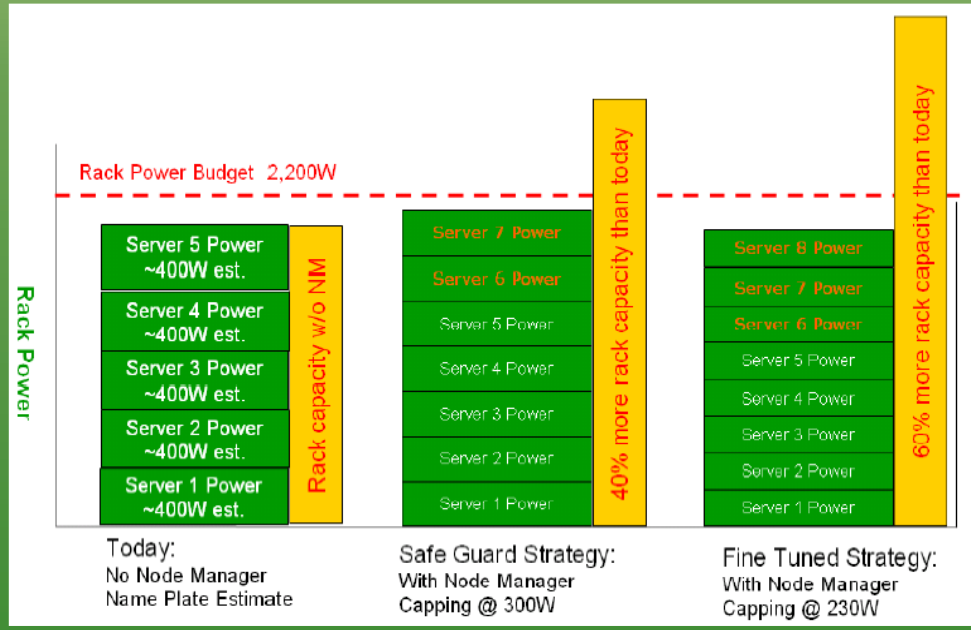


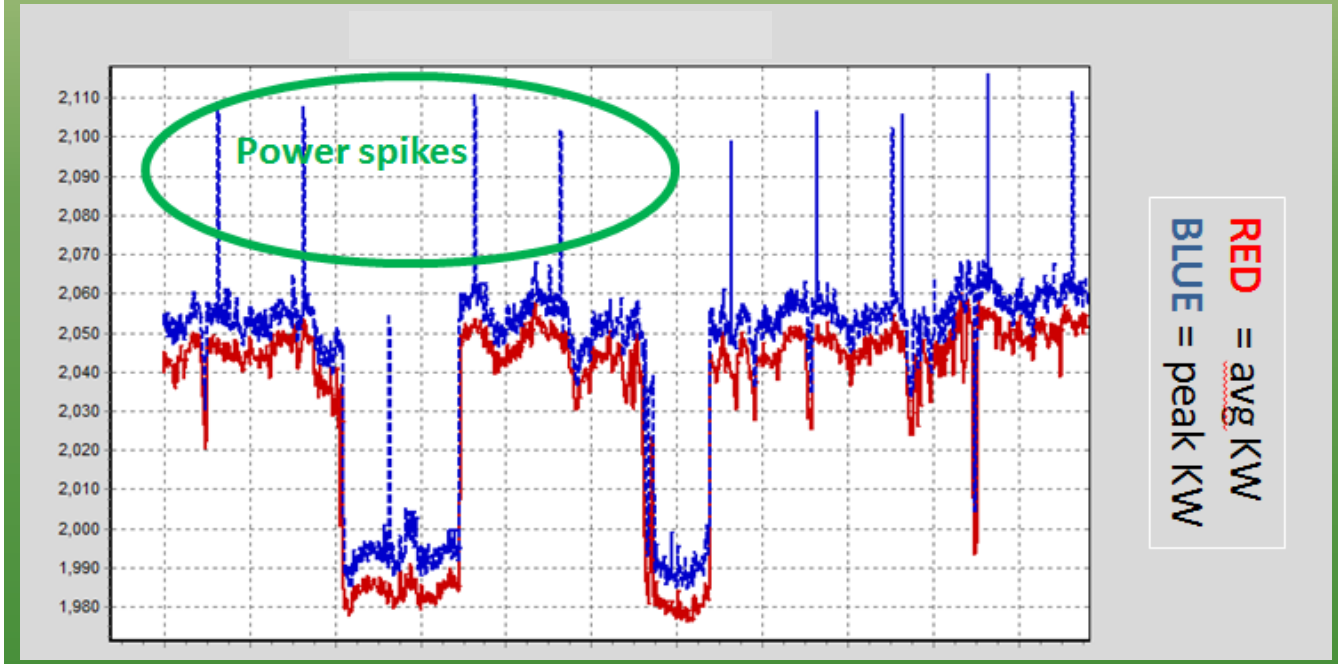
Figure 3: Effect of power capping on search response time



Source: Intel/Baidu whitepaper (http://communities.intel.com/docs/DOC-1492)

- Study shows that for a server that operates at 300W max consumed power, a setting of 230W provides acceptable query response latencies
- Allows for 15% higher server density in same power envelope
- Study doesn't account for QPS (bandwidth) maximization in given power budget – that aspect may disallow latency flexibility and affect SLAs

But does power capping scale to datacenter level?

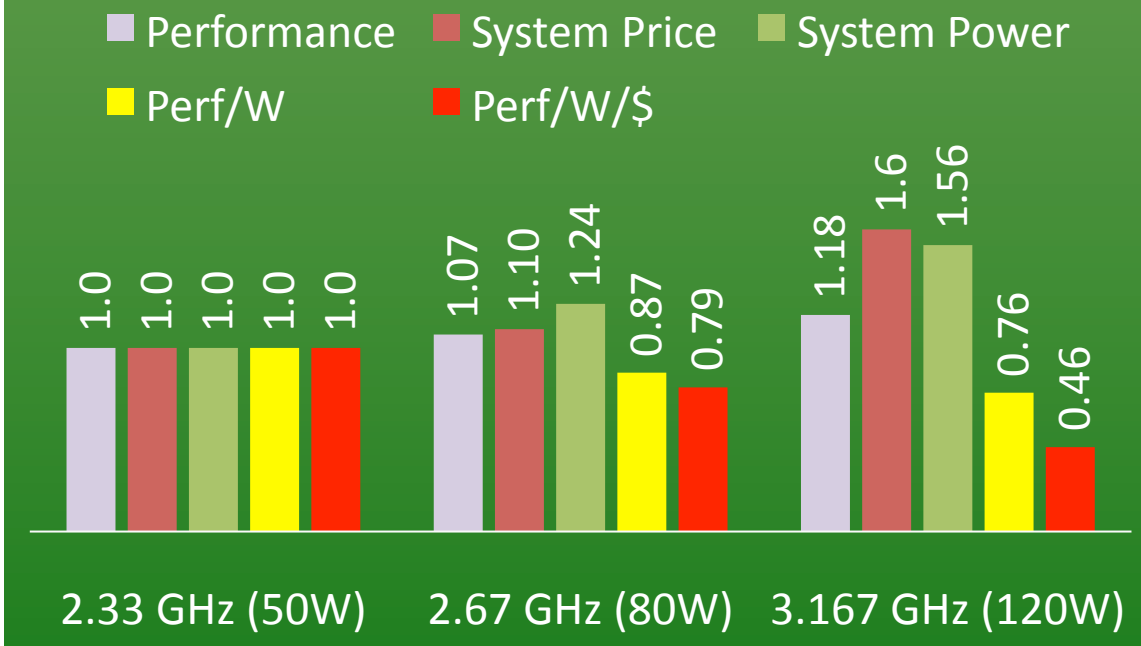


Provision servers in datacenter for peak load or nominal load?

- Short duration power spikes from sudden load increases occur in reality
- Server heterogeneity and app power profile changes also need to be considered
- *Power Capping may be used, but requires understanding tradeoff to performance SLAs and may require sophisticated policy management*

Server rightsizing for density improvements

- Characterize workloads for performance/power sensitivity
- Select power efficient components either by selecting low-voltage variants (e.g. LV CPUs/DIMMs) or lower performance variants (e.g. 800Mhz DIMMs or 5400rpm HDDs)
- E.g. shown below for CPU criteria for Internet Search app

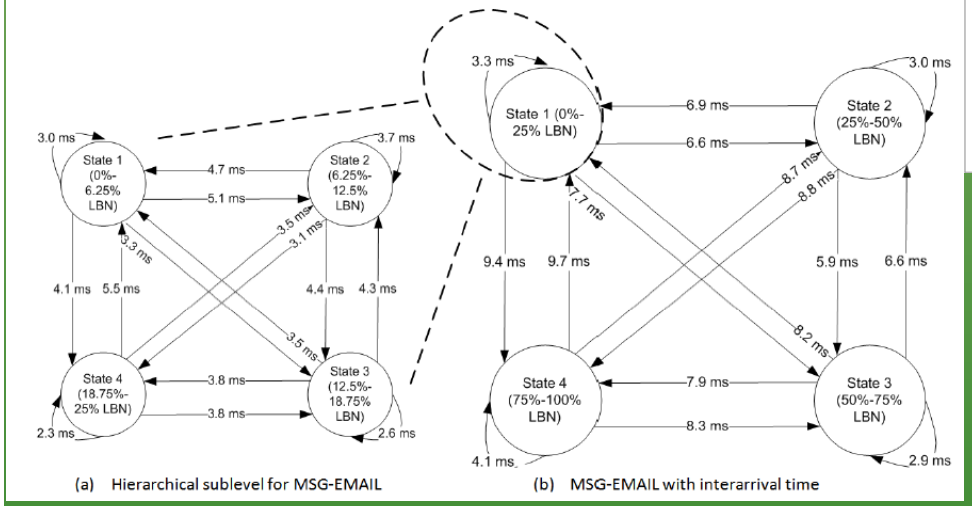


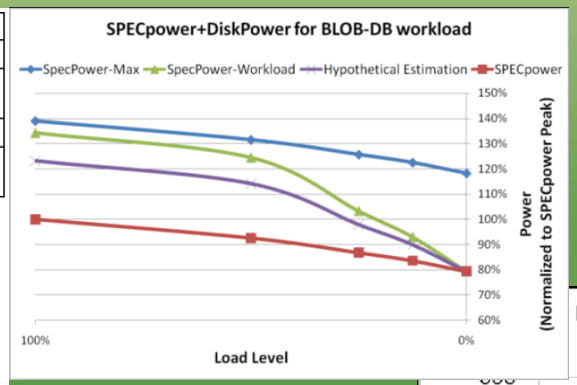
Assumption:
Server Price = $2000 + CPUs,
Server Power = 150W + CPUs

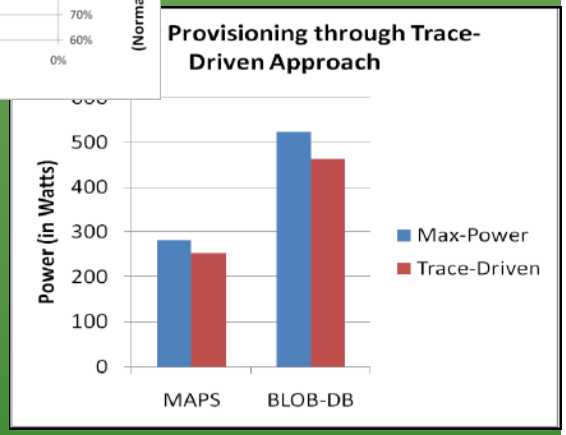
The sweet spot is often at low power processors, especially when system price and power are considered.

Storage rightsizing for power provisioning efficiency

Property	Randomness			RD:WR	Read					Write				
	Total	RD	WR	Ratio	4K	8K	16K	32K	64K	4K	8K	16K	32K	64K
MSG-EMAIL	93%	93%	94%	4.7	57%	5%			8%	14%				
MAPS	27%	24%	96%	19.8	14%				65%	4%				
USER CONTENT	91%	90%	98%	7.0	70%					10%				







Sankar et al, "Storage Characterization for Unstructured Data in Online Services Applications", IISWC 2009
Sankar et al, "Addressing the Stranded Power Problem in Datacenters Using Storage Workload Characterization, WOSP-SIPEW 2010

- E.g. above shows how in-depth storage trace analysis can be used to understand I/O workload patterns and IOPS rates
- HDD power models can then be used to determine optimal power provisioning values – minimizing stranded power and allowing higher density

Datacenter Power Efficiency

Minimize cost of provisioning power

- construction and cooling infrastructure

Maximize power available to IT-load (improved PUE)

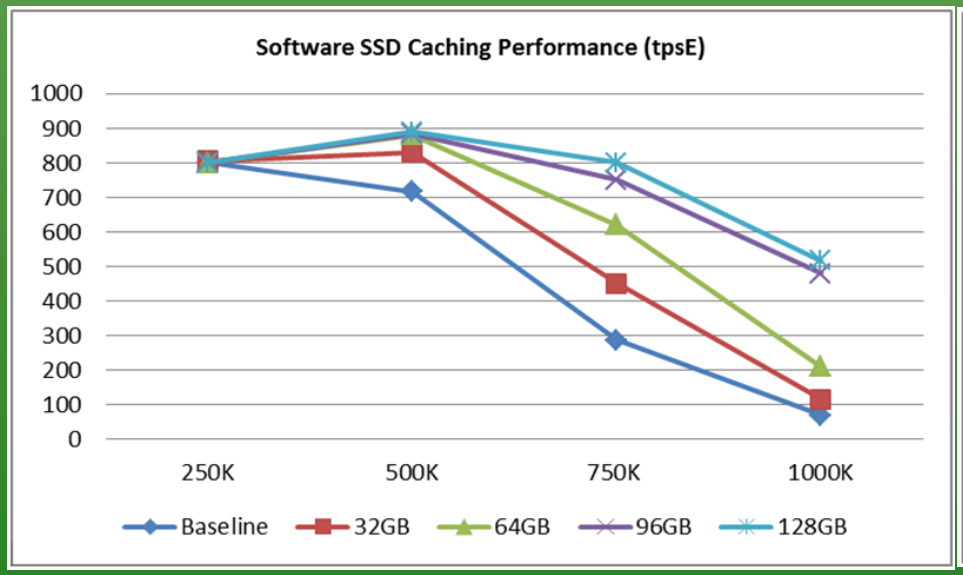
- Minimize losses and thermal/cooling overheads

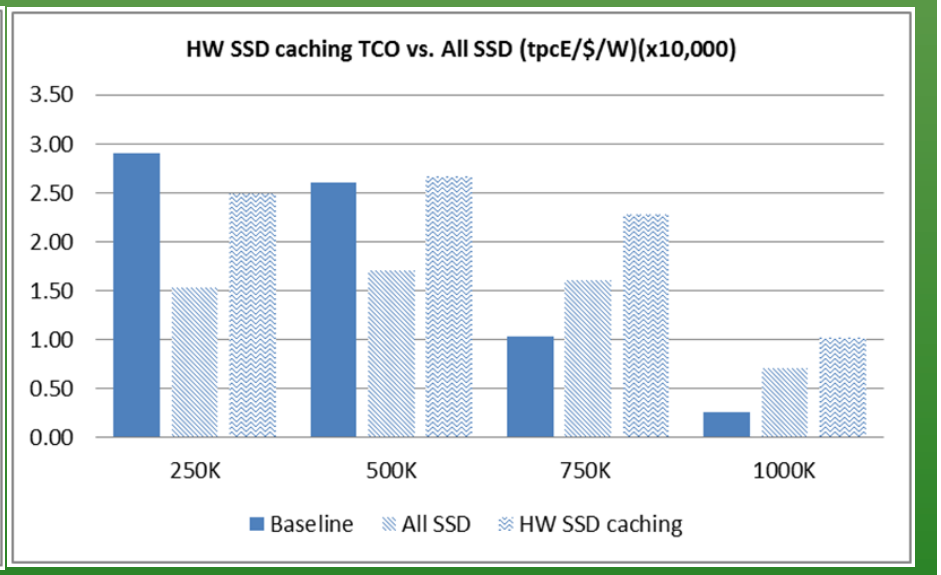
Maximize work done per watt

- Maximize compute density in power budget
- **Maximize performance, minimize power consumption**

Improving Perf/W

- Seek architectural opportunities for ***dramatic*** Perf/W improvements
- E.g. shown below for SSD caching solutions implemented in HW RAID controllers and in SW buffer pool management algorithms
- Upto ~3x server consolidation opportunity for DB workloads





Khessib et al, "Using Solid State Drives as a Mid-Tier Cache in Enterprise OLTP applications", TPC Technology Conference on Performance Evaluation and Benchmarking (TPC-TC), Sept 2010

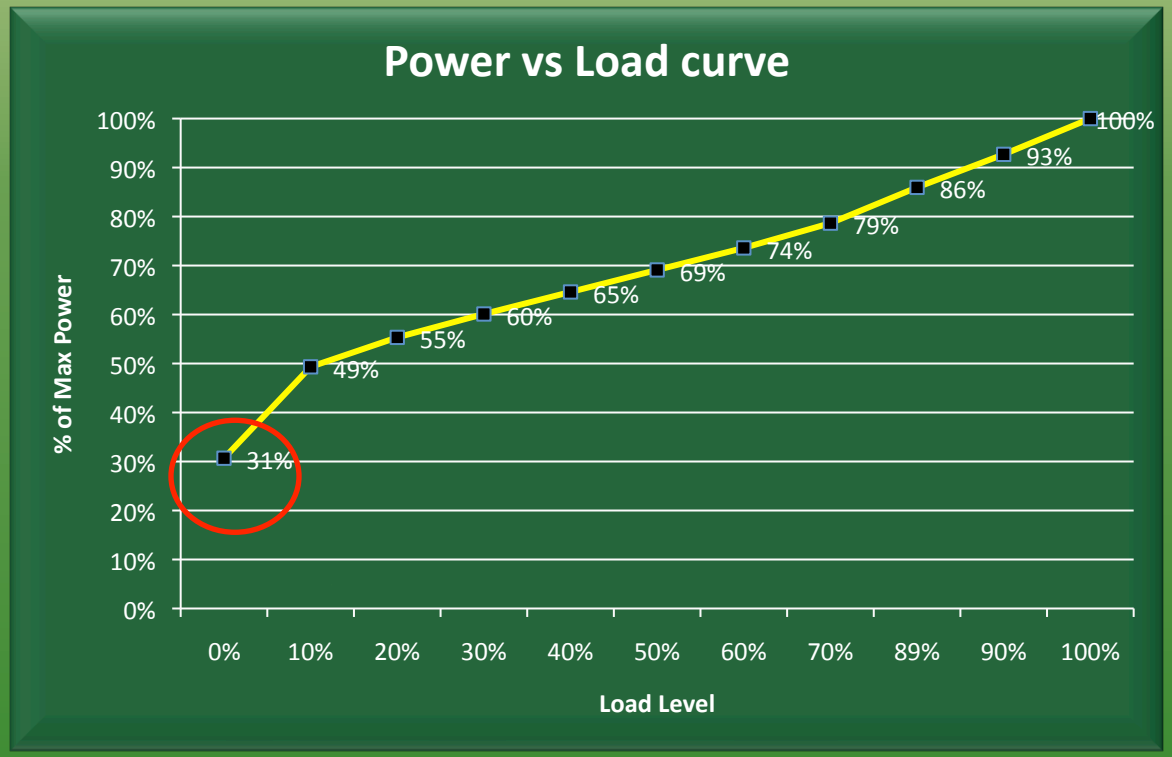
Are C and P-states really effective?

System
2S L5520/4c CPUs
LV DIMMs (32GB)

- Workload power profile shown with different C/P state combinations
- When C-states are enabled, ***P-states have no impact on power consumption***
- C-state savings vary from 0-40% of max power; C6 state helpful for idle load
- Effectiveness of P-state depends on CPU SKU selection and workload profiles

Improvements in the power load curve

SPECpower2008 benchmark
2 x Intel L5640/6c/2.26Ghz
16GB DDR3
1x160GB SSD
Windows Server 2008 R2



http://www.spec.org/power_ssj2008/results/res2010q3/power_ssj2008-20100714-00275.html

- With C6 power savings and OS feature support, idle power is now ~30% of max power
  - Not 50-60% as typically quoted in publications
- Expect significant future improvements in idle power
  - Windows Core Parking, Deeper CPU sleep states, Memory power states

Research areas

- Optimal Power provisioning for high dynamic range workloads (idle at night, peak load during day)
- Addressing Energy proportionality via system architecture innovations
- Power aware task scheduling on large clusters
- Energy conscious programming using controlled approximation (ref: http://research.microsoft.com/en-us/um/people/trishulc/papers/green.pdf)
- Several others…

Summary

- Take a holistic view when considering power efficiency opportunities
  - Minimize cost of provisioning power (datacenter design)
  - Maximize power available to IT-load (datacenter/server)
  - Maximize work done per watt (platform/OS/apps)
- Conventional wisdom on power efficiency is often incorrect (several examples in this talk)
- Workloads…workloads…workloads… Optimize for application specific scenarios

Q & A