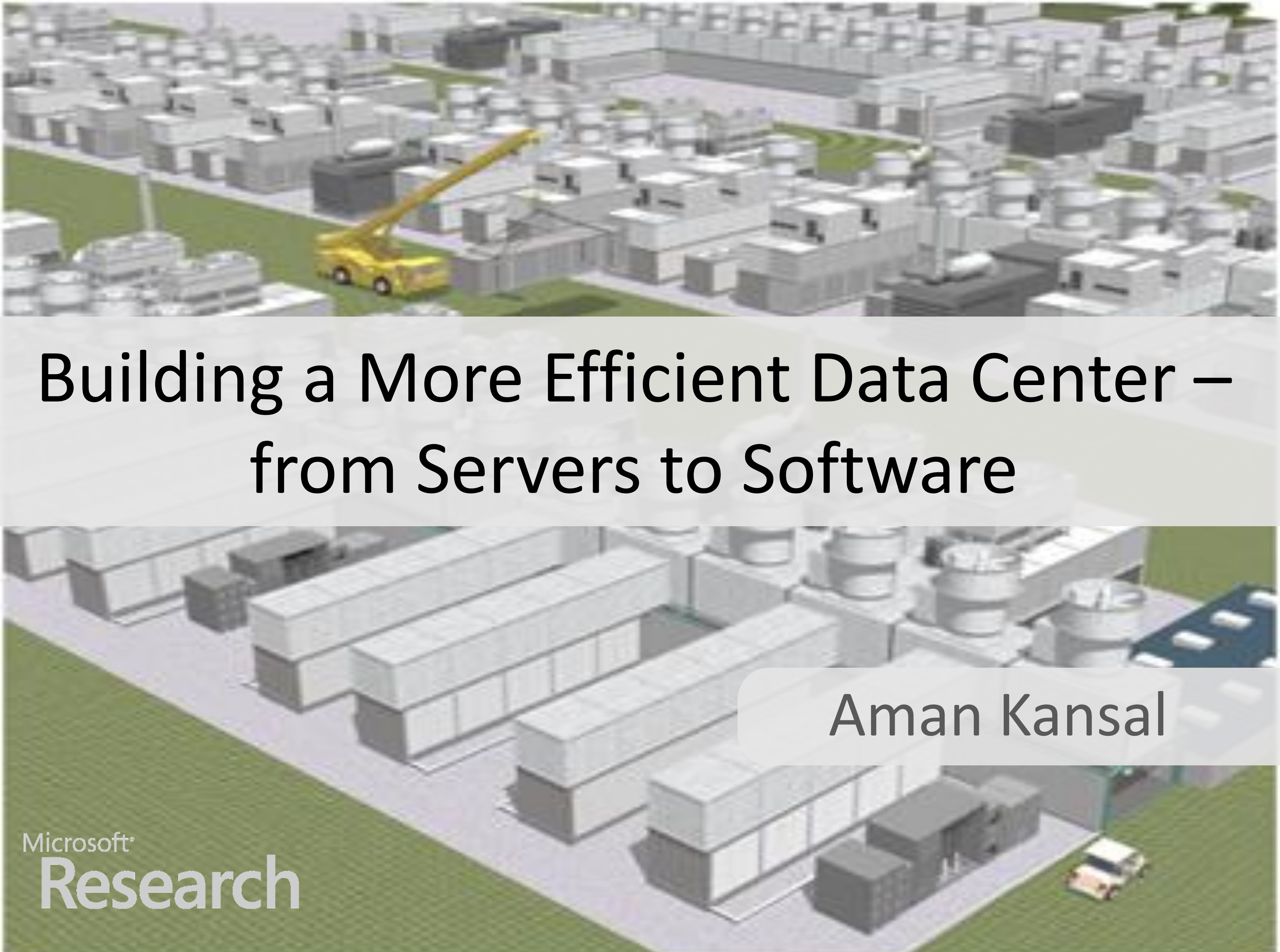


Building a More Efficient Data Center – from Servers to Software

Aman Kansal

Data centers growing in number,

Microsoft has more than 10 and less than 100 DCs worldwide

"Data Centers have become as vital to the functioning of society as power stations."

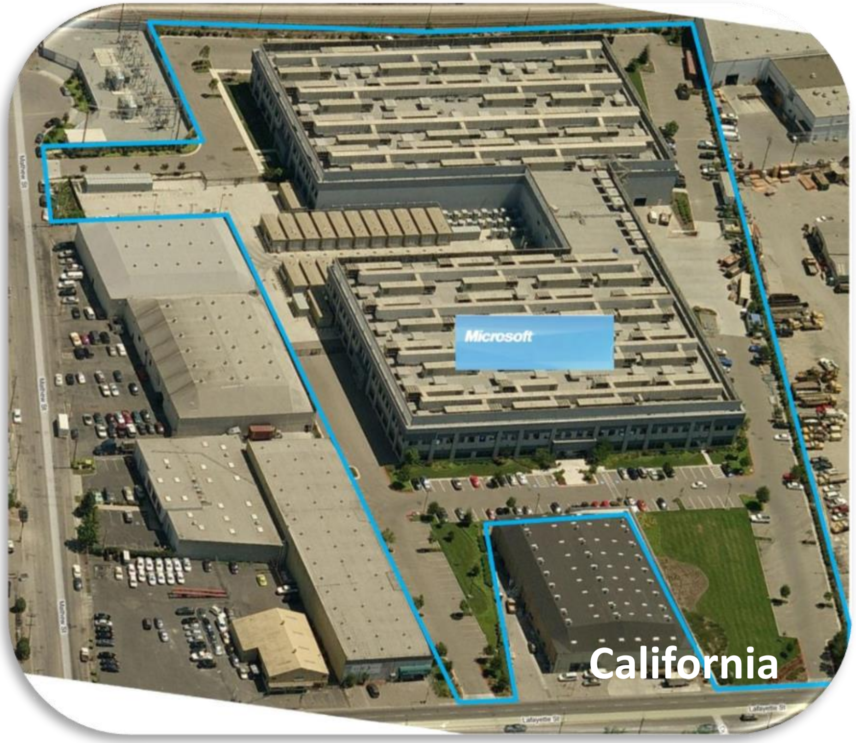
The Economist

… size,


Washington


Texas


Ireland
Microsoft
ORANGE CASTLE

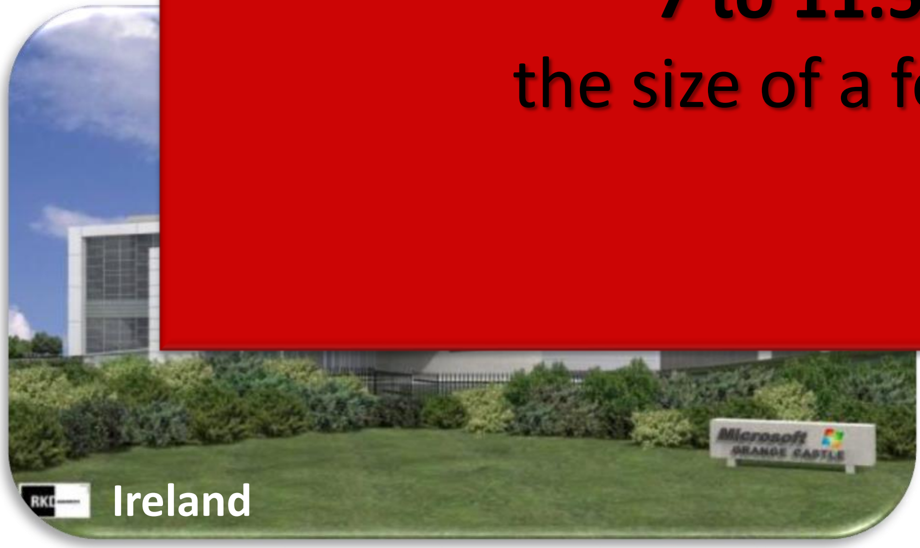

Microsoft
California

... size,

Each data center is

7 to 11.5 times

the size of a football field



…, and efficiency!

More apps have online components

– Music, office s/w, …

Lower cost DC ⇒ new scenarios

– Improved speech recognition

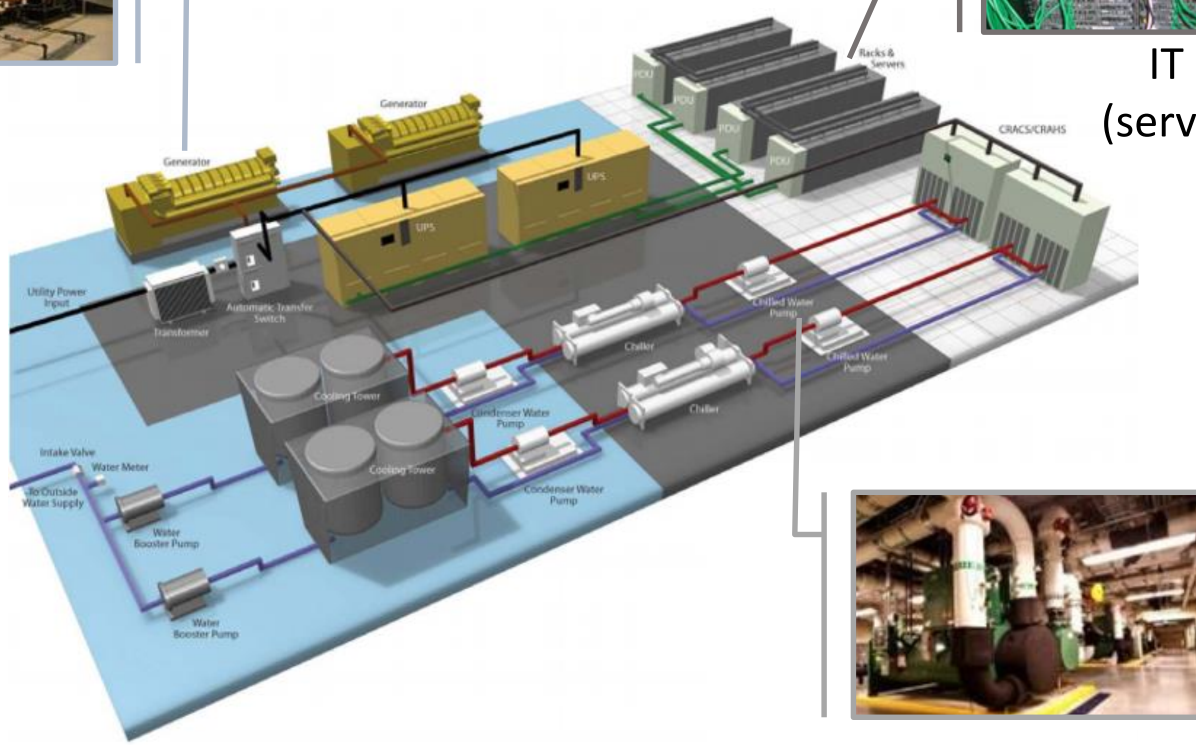
– Video on wireless HD/retina display tablets

- Better encoding needs more compute: HEVC reduces bitrate by 50%

Please speak or enter your flight number…I’m sorry I did not get that, please speak or enter your flight number…

Inside a Data Center

Power Distribution

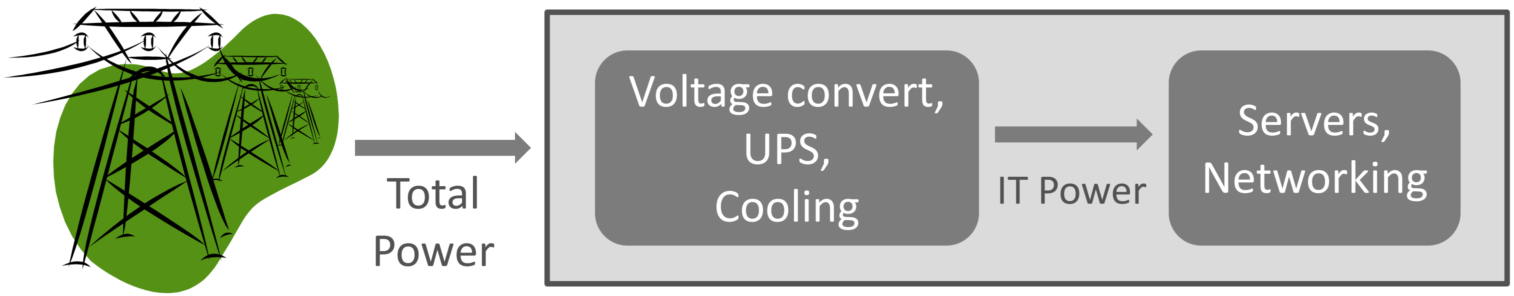


IT Equipment
(servers, network)

Cooling

Video tours: http://www.GlobalFoundationServices.com/

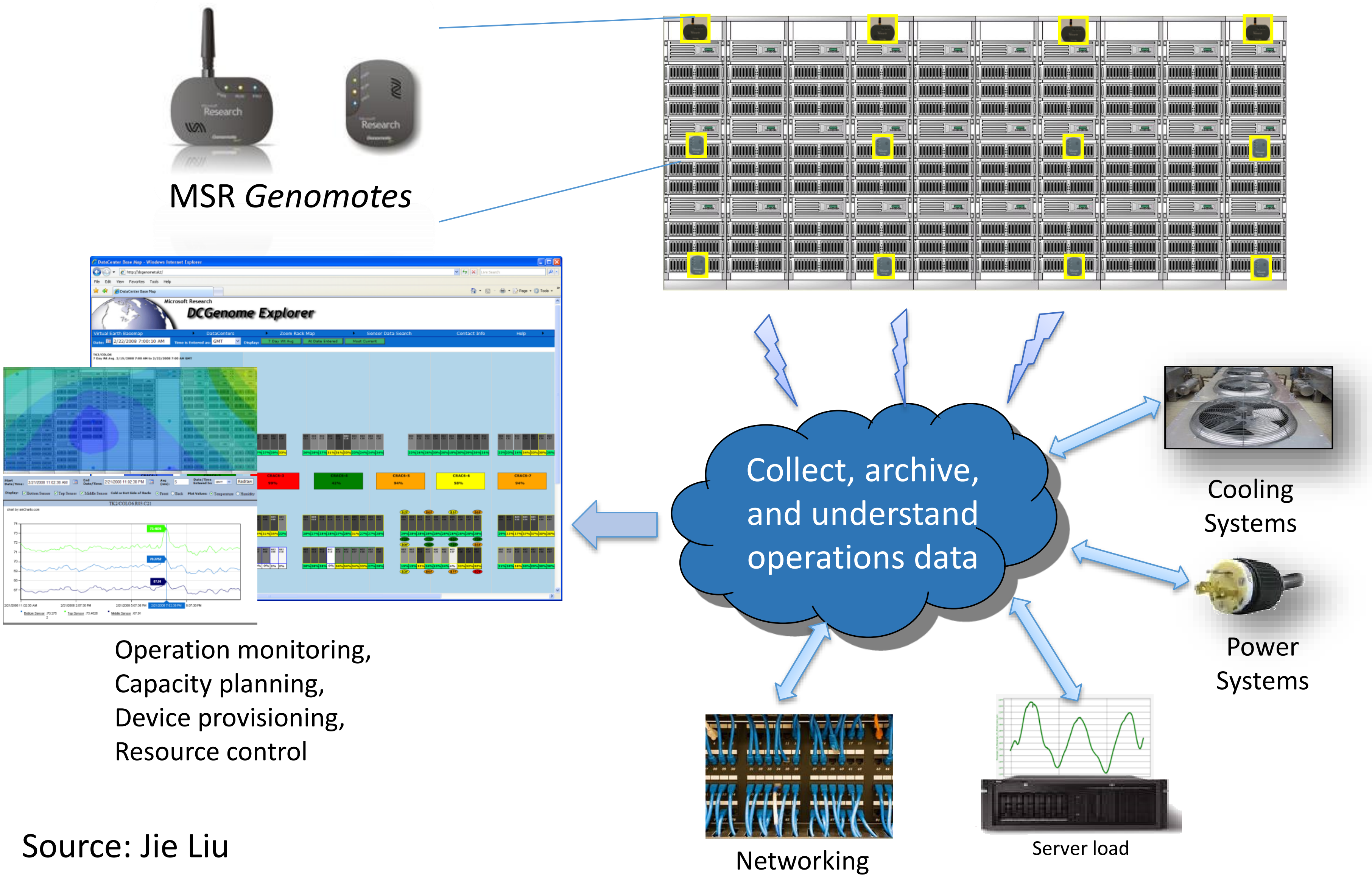
Power Usage Effectiveness (PUE)



$$PUE = \frac{Total\ Facility\ Power\ Usage}{IT\ Equipment\ Power\ Usage}$$

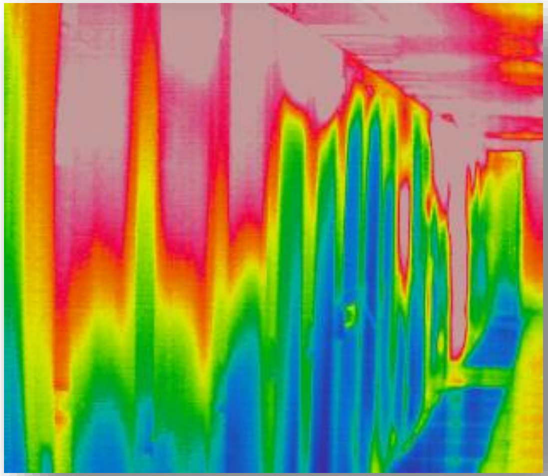
How did Microsoft improve PUE from near 2.0 to 1.05 in five years?

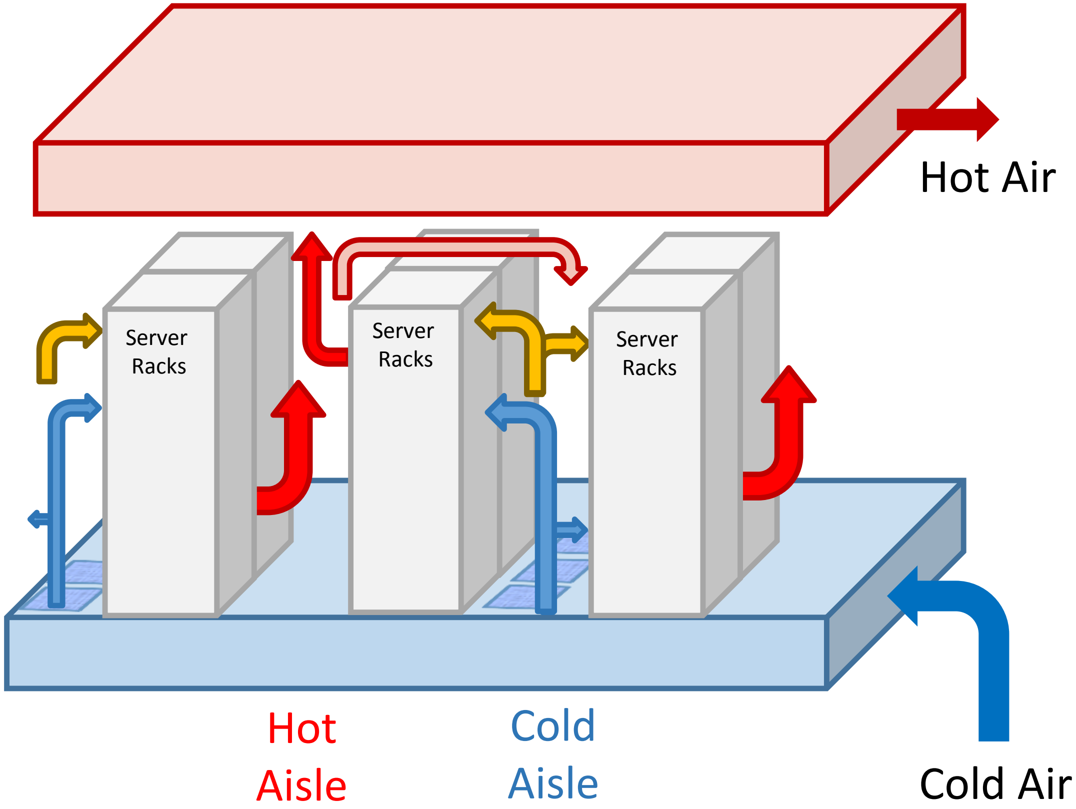
Measuring Data Centers



Source: Jie Liu

Older Cooling Design





Hot air is not contained

New and Improved

Containment: tightly guide air-flow

Use outside air: locate in cooler region

Operate servers hotter

1989-2005

Cold + hot aisles
PUE = 1.5 - 2

2008

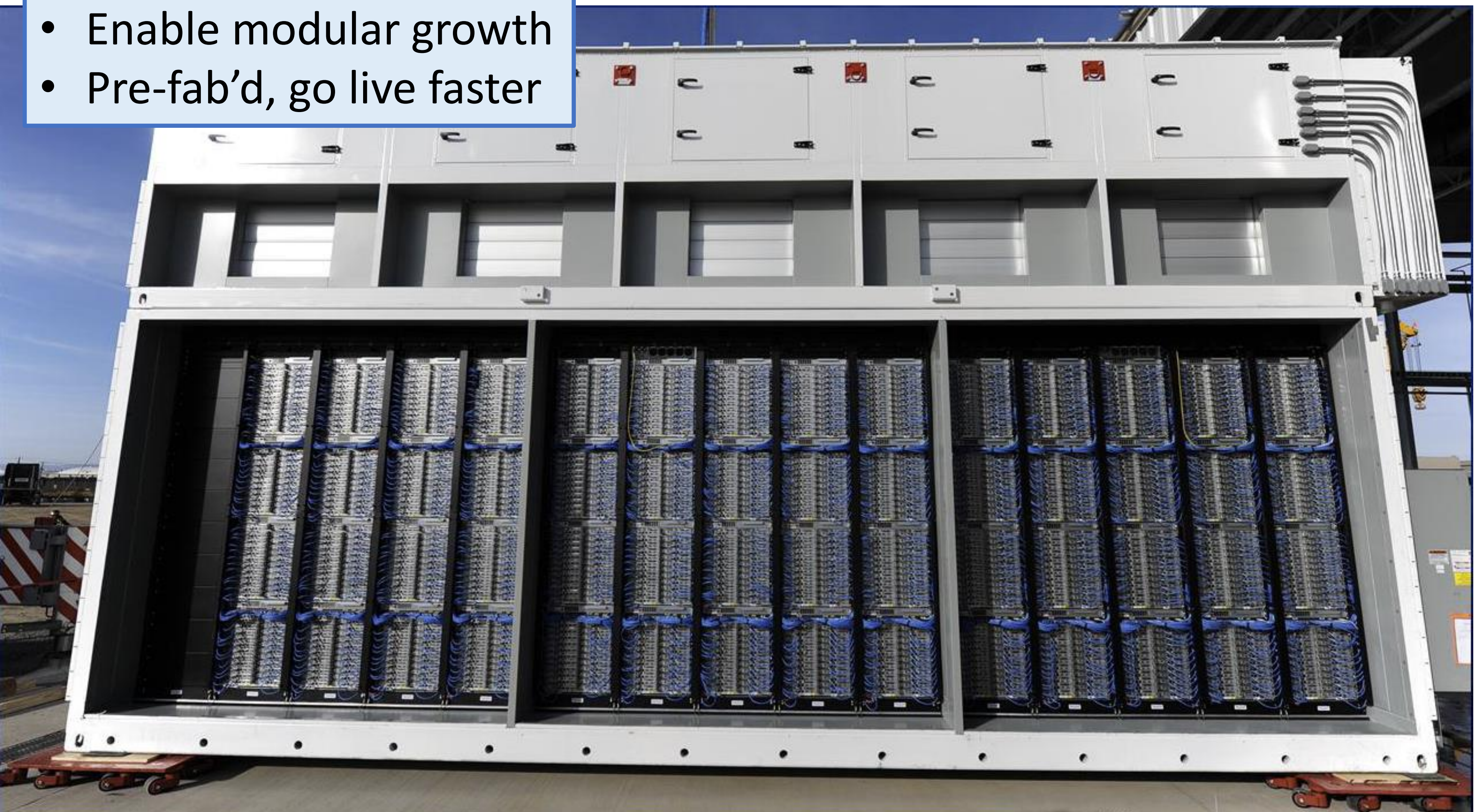
Containers
PUE = 1.2-1.5

2011

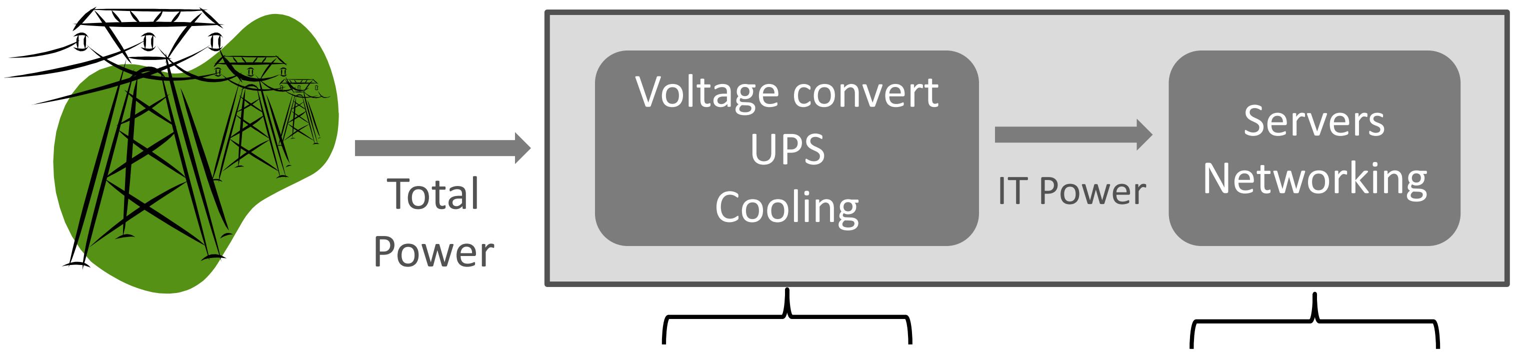
Custom Module
PUE = 1.05- 1.15

Inside a Module

- Reduce building cost
- Enable modular growth
- Pre-fab'd, go live faster



PUE ≈ 1, are we done?



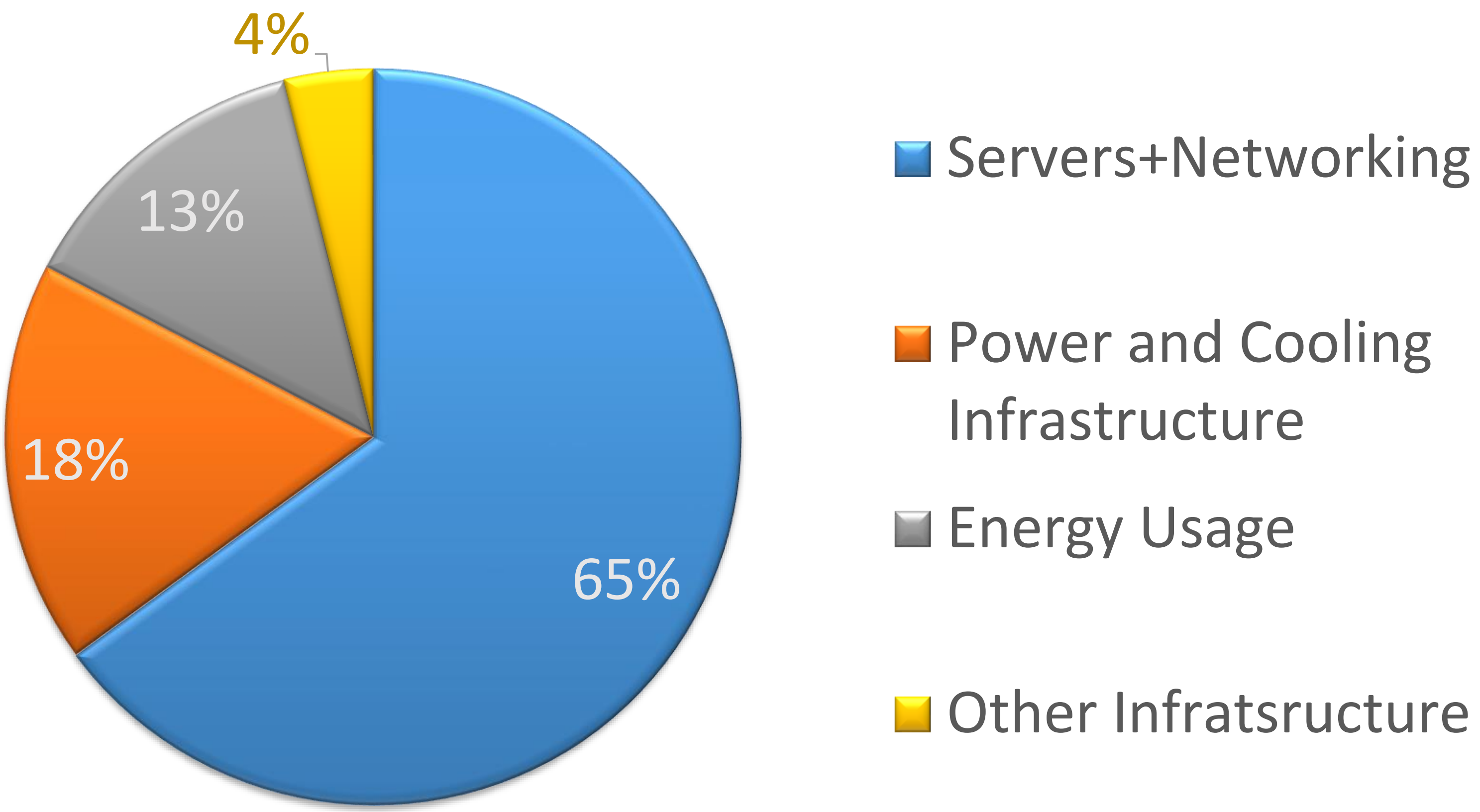
Energy overhead
Power/Cooling Infra.

IT energy cost
IT Infra. cost

For given IT load, not wasting excess energy, but we can reduce

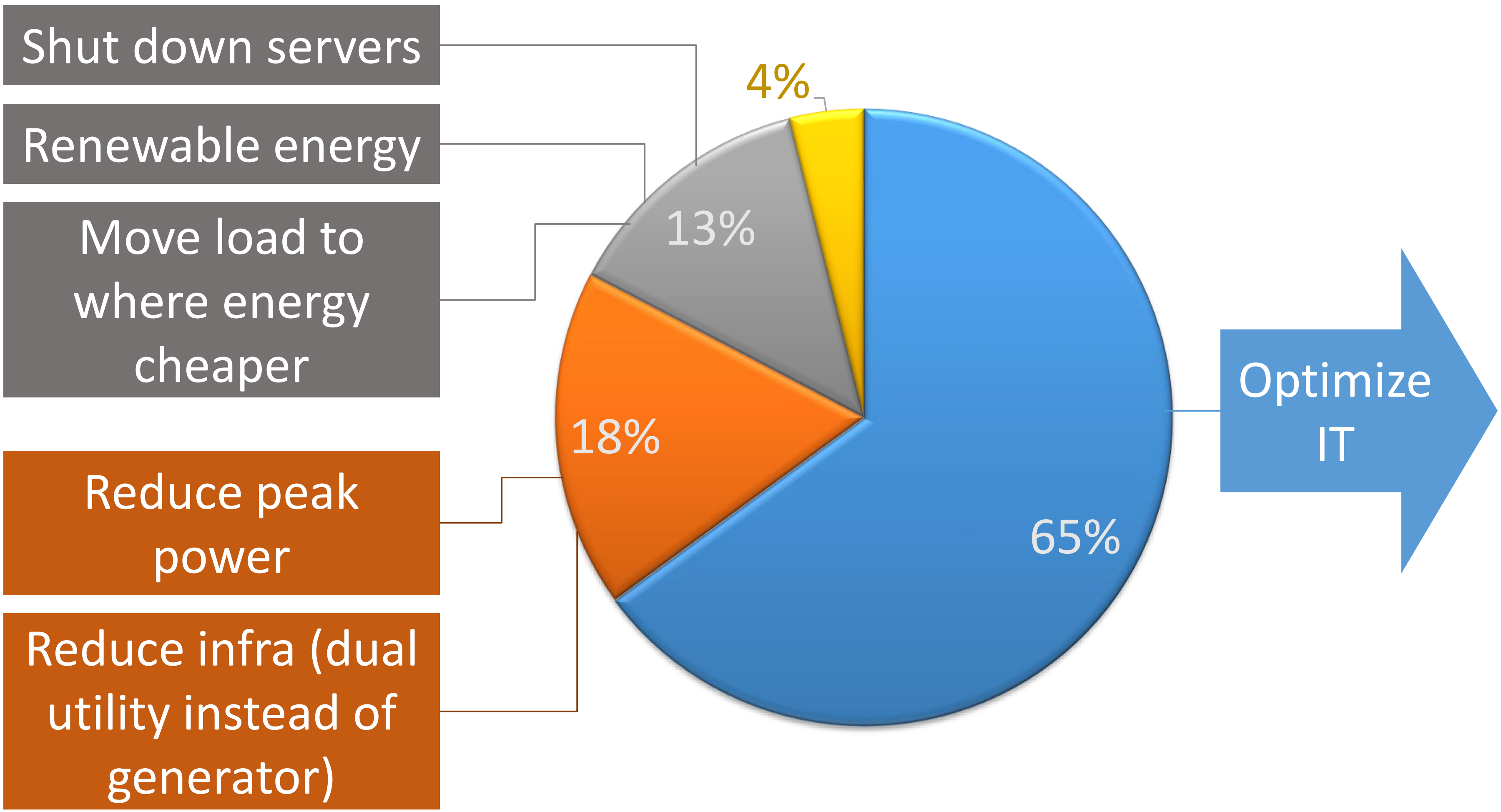
- Power required for same app
- Infrastructure

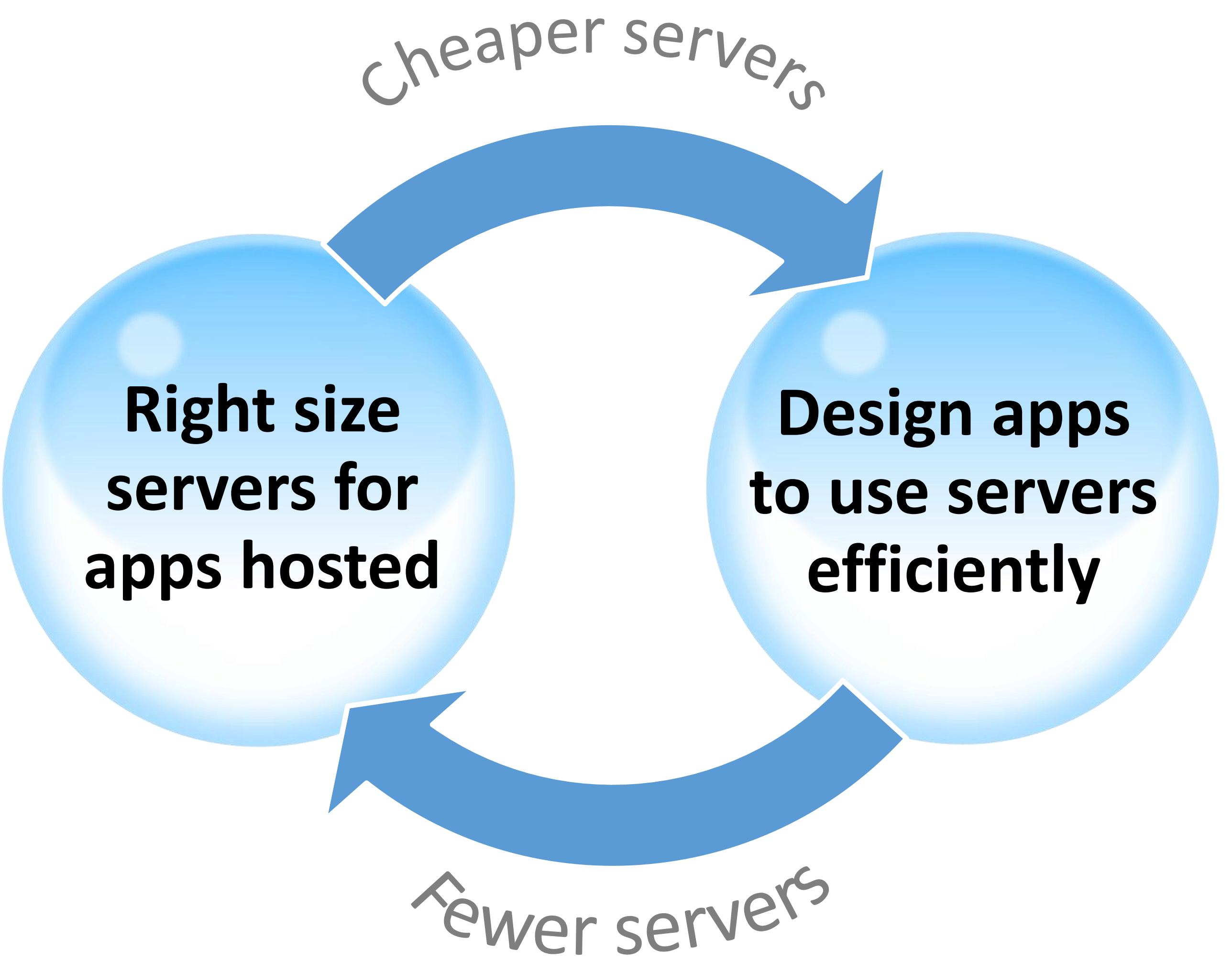
Beyond PUE



Data from: James Hamilton

[http://perspectives.mvdirona.com/2010/09/18/OverallDataCenterCosts.aspx]


Shut down servers
Renewable energy
Move load to where energy cheaper
Reduce peak power
Reduce infra (dual utility instead of generator)
4%
13%
18%
65%
Optimize IT


Cheaper servers
Right size servers for apps hosted
Design apps to use servers efficiently
Fewer servers

Cheaper Servers

Obvious

No one installs s/w from a CD on 1000s of servers: *remove the optical drive*

Use blades: share fans, power supplies

High Cost Components

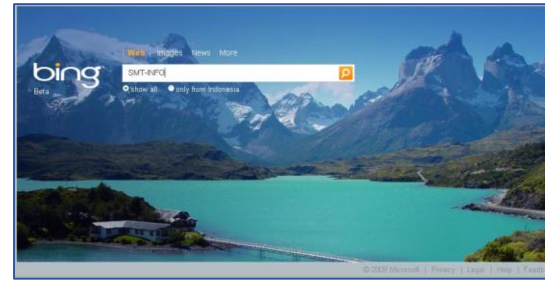
CPU	$300-1500/socket	Eg. Intel Xeon E5
Memory	$20-30/GB	64GB = $1280+
Hard disk	$100-300/TB	SATA vs. SAS, 3 - 6Gbps, 7.2 – 15 RPM
SSD	$1000-5000/TB	Vary by brand/perf.

Prices in January 2013

Right-size server to app needs

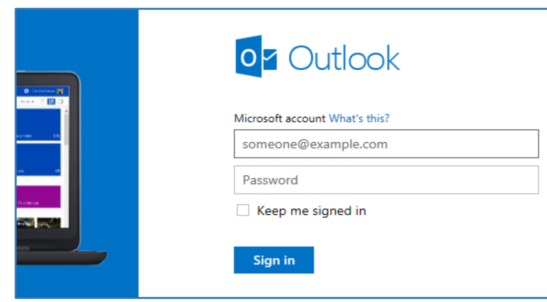
Bing

- Web crawling, index management, query lookup
- Major load: *Index lookup*
- Highly latency critical



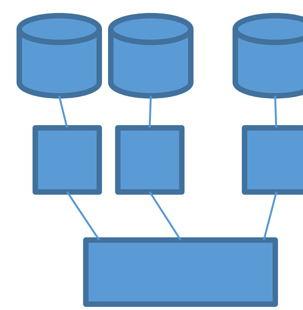
Hotmail

- UI, mail protocols, spam filtering, storage
- Major load: *Retrieve data from mailboxes*
- Stores several petabytes of data, IOPS intensive



Cosmos

- Highly parallelized data storage and analysis
- Major load: *distributed storage and batched compute*
- Throughput intensive



App Resource Usage

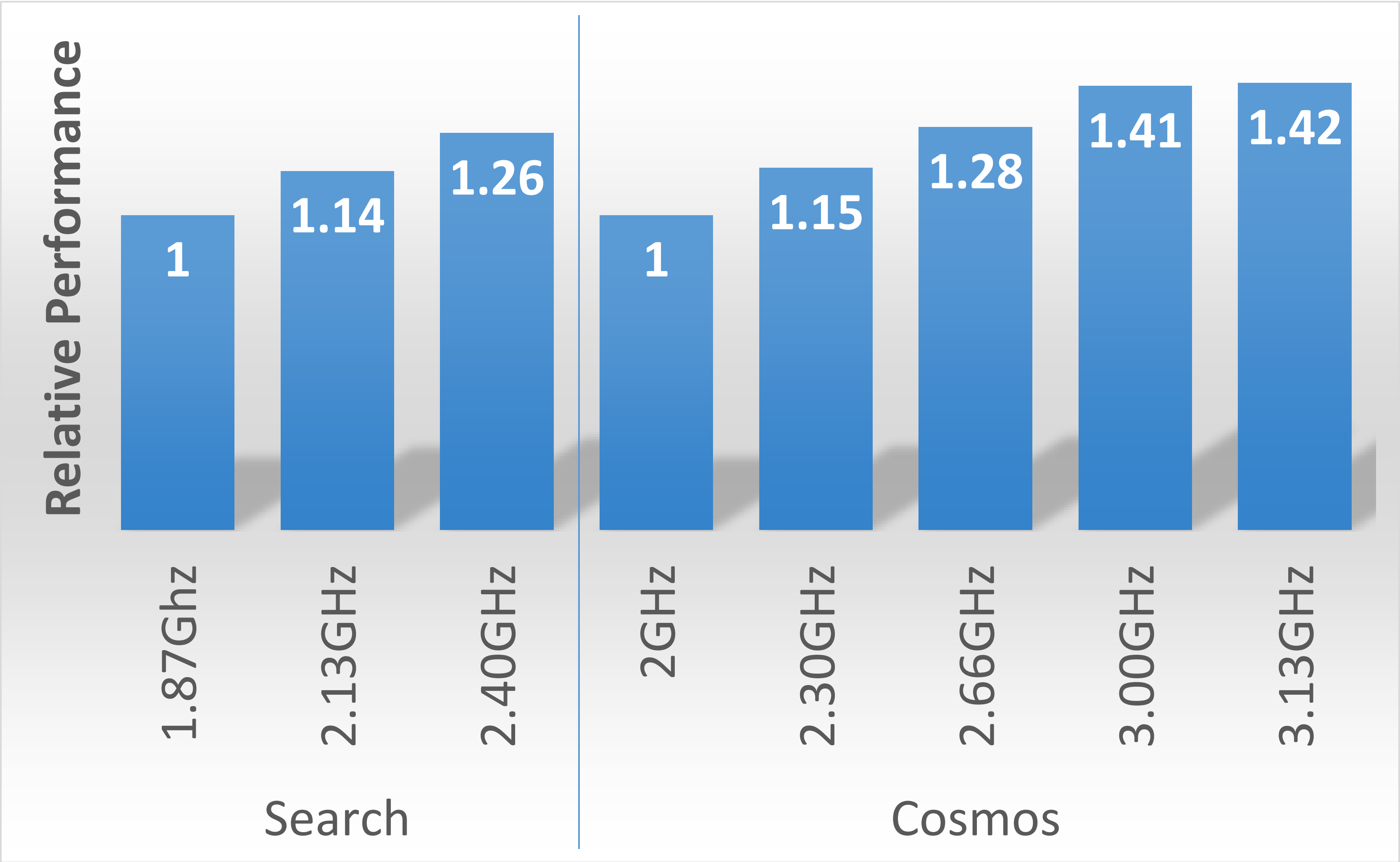
Production

App	Memory Capacity	Memory BW	Disk Capacity	Disk BW	Network BW
Hotmail	92%	NA	75%	0.91%	27%
Cosmos	39%	1.1%	52%	0.68%	9%
Bing	88%	1.8%	30%	1.10%	10%

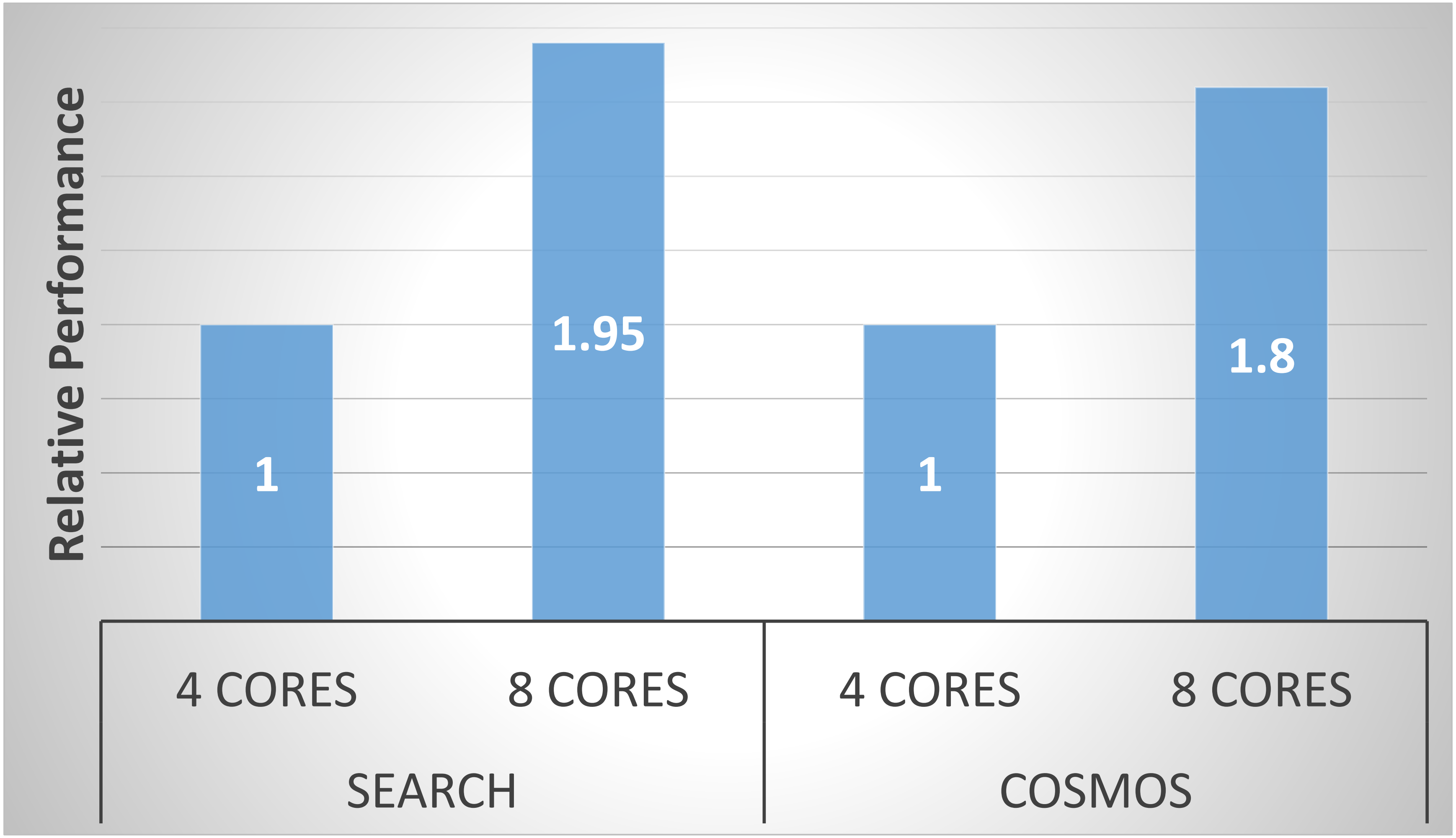
Stress

App	CPU Utilization	Memory Bandwidth	Disk Bandwidth
Hotmail	67%	NA	**71%**
Cosmos	**88%**	1.6%	8%
Bing	**97%**	5.8%	36%

CPU: Frequency



CPU: Number of Cores



Cost

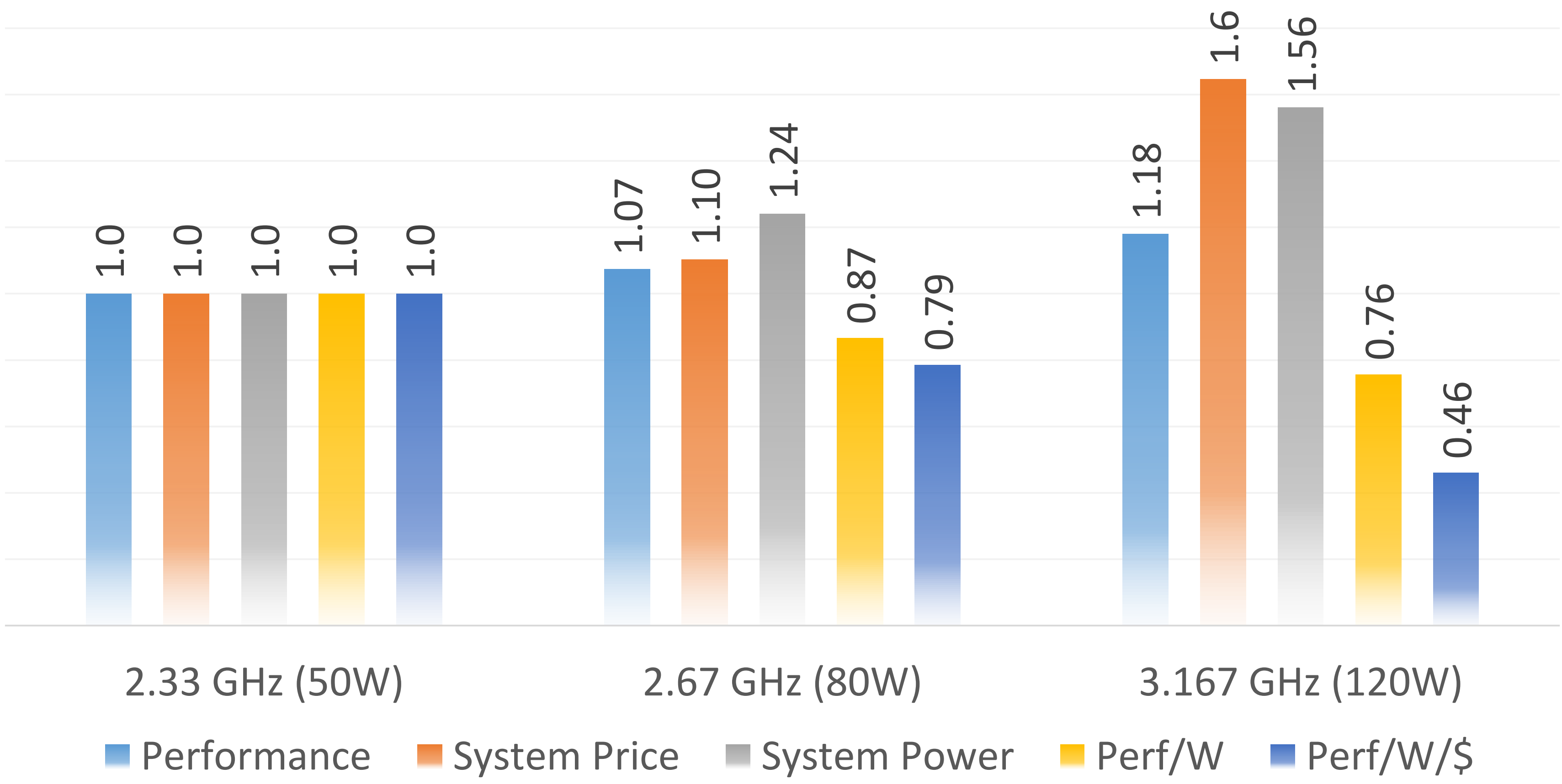
But **power** and **cost** also increase with frequency and number of cores

Figure of merit:

$$\frac{Performance}{Power\ (W) * Cost(\$)}$$

Performance / Watt / $

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



App Resource Usage: Disk

Production

App	Memory Capacity	Memory BW	Disk Capacity	Disk BW	Network BW
Hotmail	92%	NA	**75%**	**0.91%**	27%
Cosmos	39%	1.1%	52%	0.68%	9%
Bing	88%	1.8%	30%	1.10%	10%

Stress

App	CPU Utilization	Memory Bandwidth	Disk Bandwidth
Hotmail	67%	NA	**71%**
Cosmos	88%	1.6%	8%
Bing	97%	5.8%	36%

Disk

Bandwidth optimizations

- Hotmail: Mix hot and cold data to spread bandwidth
- Striping/mirroring instead of RAID

Latency

- *Use memory to cache data*

Flash storage

- *Expensive per byte stored but cheaper in bandwidth*
  - *Bandwidth is not a bottleneck for above apps*
- *Flash may potentially enhance memory*

Memory

Low latency for interactive apps demands high memory capacity

- Bing is memory bound
- Hotmail: SQL index uses available memory for caching
- Cosmos: disk bound, smaller memory sufficient
- Rising popularity of memcached

Halving the processor cache did not degrade performance for Bing and Cosmos

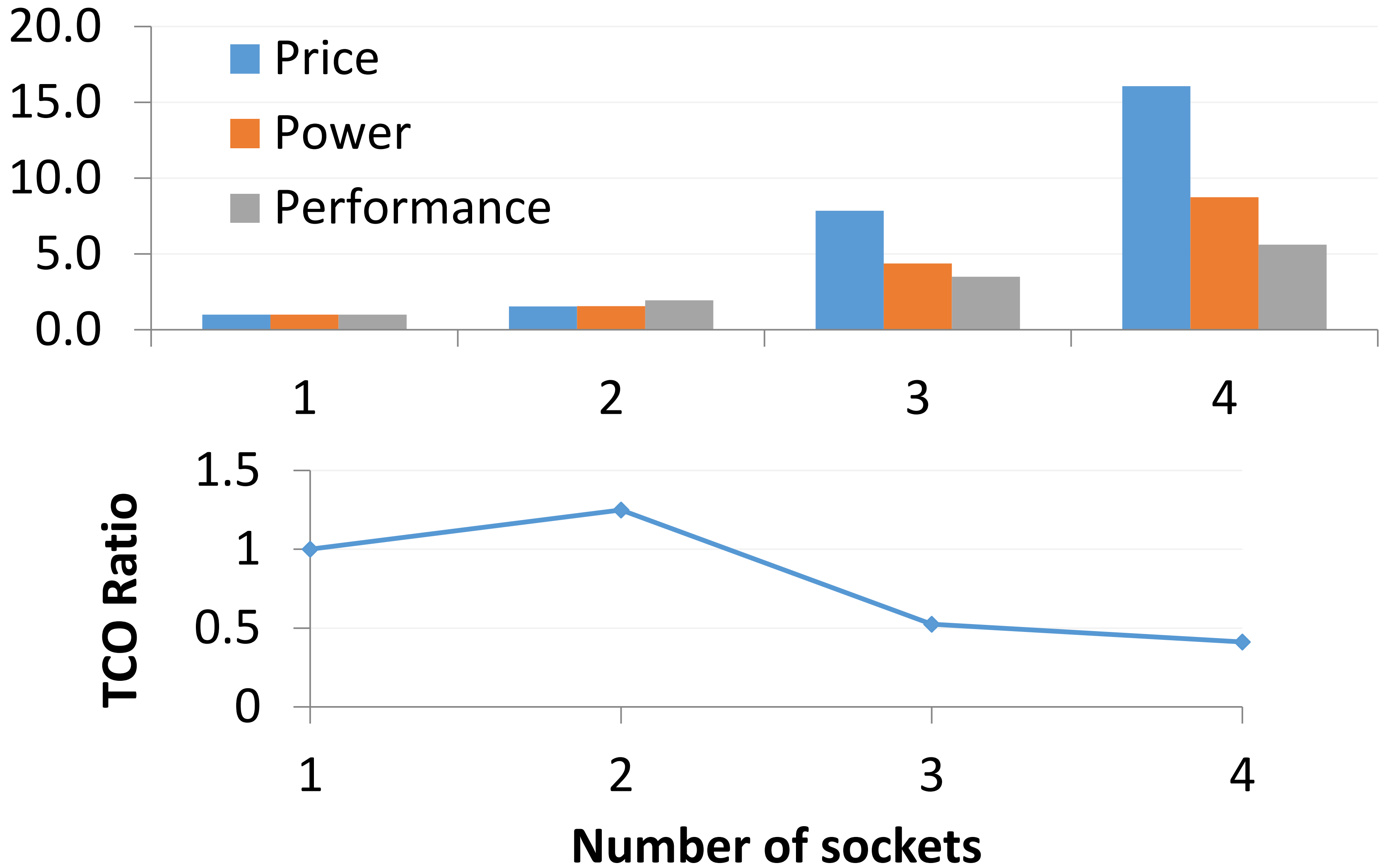
- Cache does not significantly reduce memory access

Scale Up or Scale Out

Are two cheaper servers better than one higher capability server?

	UP (1S)	DP (2S)	MP (4S)	MP (8S)
CPUs	1	2	4	8
Cores per CPU	4	8	24	48
Memory	8	16	48	96
Drives	2	3	8	16

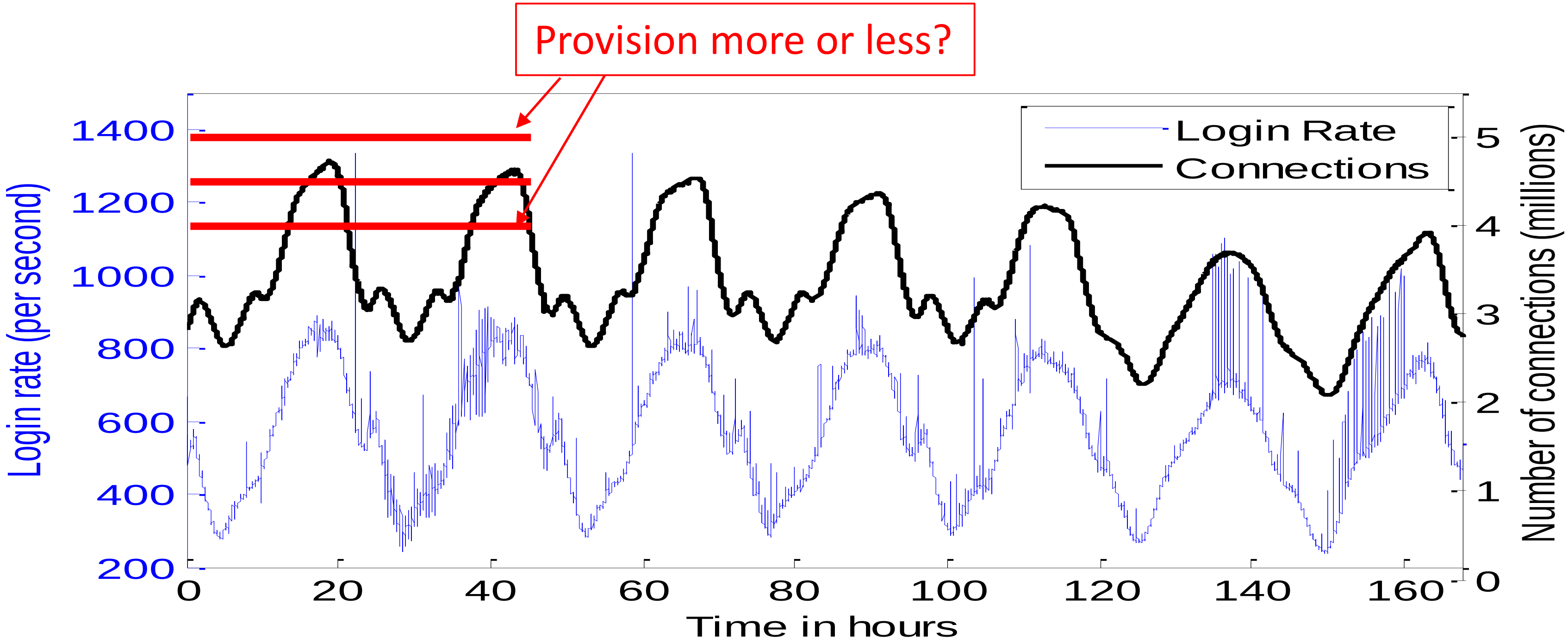
Performance/W/$



Fewer Servers

Over-provisioning Dilemma

Load varies with time



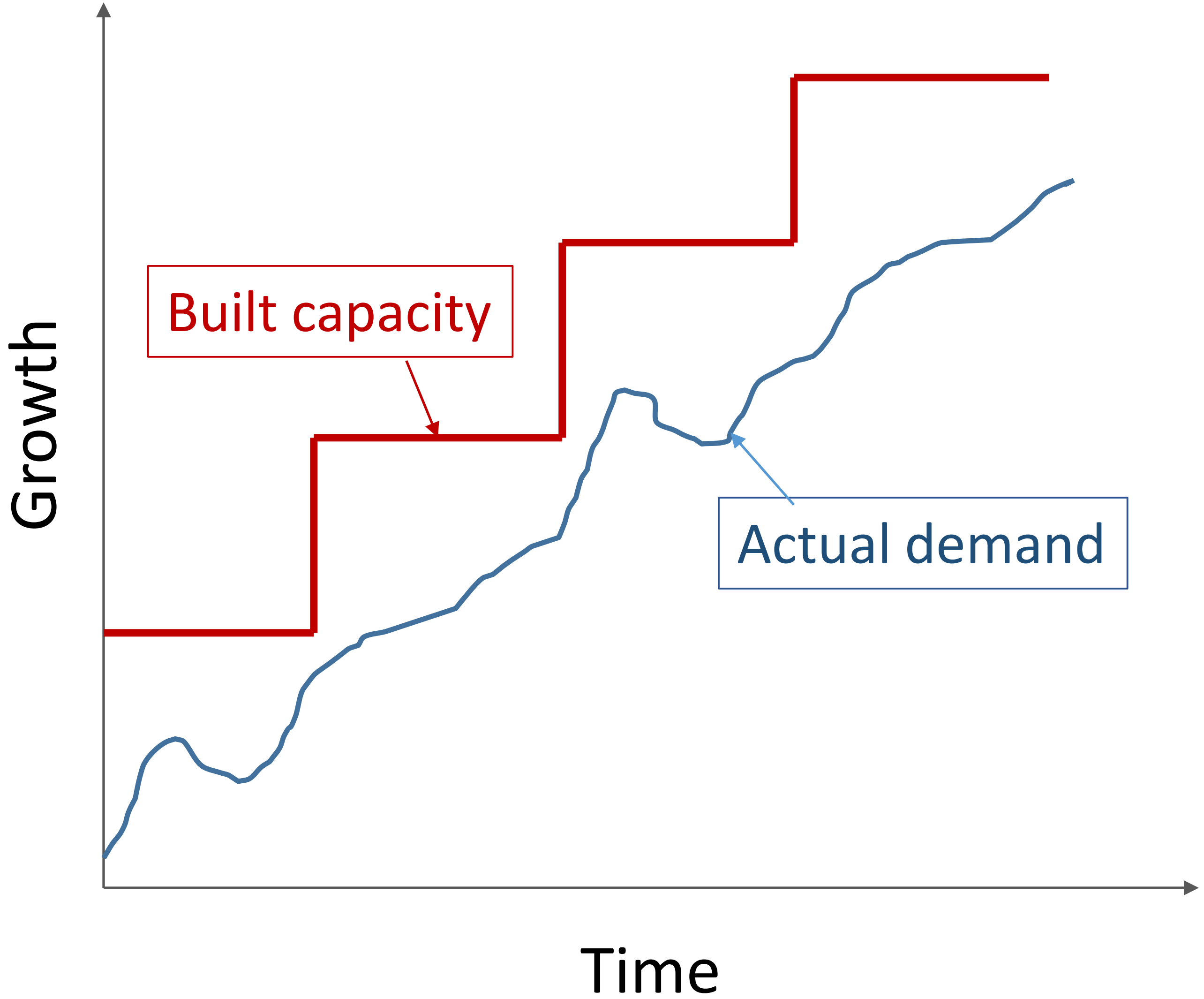
Messenger load with time

Over-provisioning (contd.)

Large difference between peak and typical

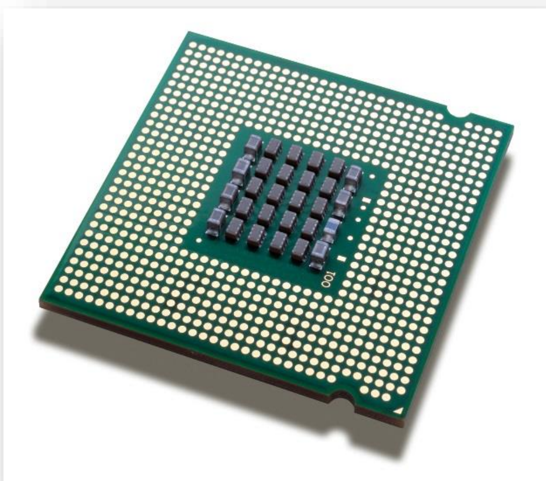
	MIPS/ Core	Disk MBps/Core		MIPS/Disk MBps	
		Avg+2Sigma	Max	Avg+2Sigma	Max
Amdahl	1			8	8
Hotmail	1059	0.32	25.22	3271	42
Cosmos	3698	0.24	2.73	15173	1357
Bing	1849	0.17	5.73	10643	323

Growth Granularity



Consolidate in a Shared Cloud

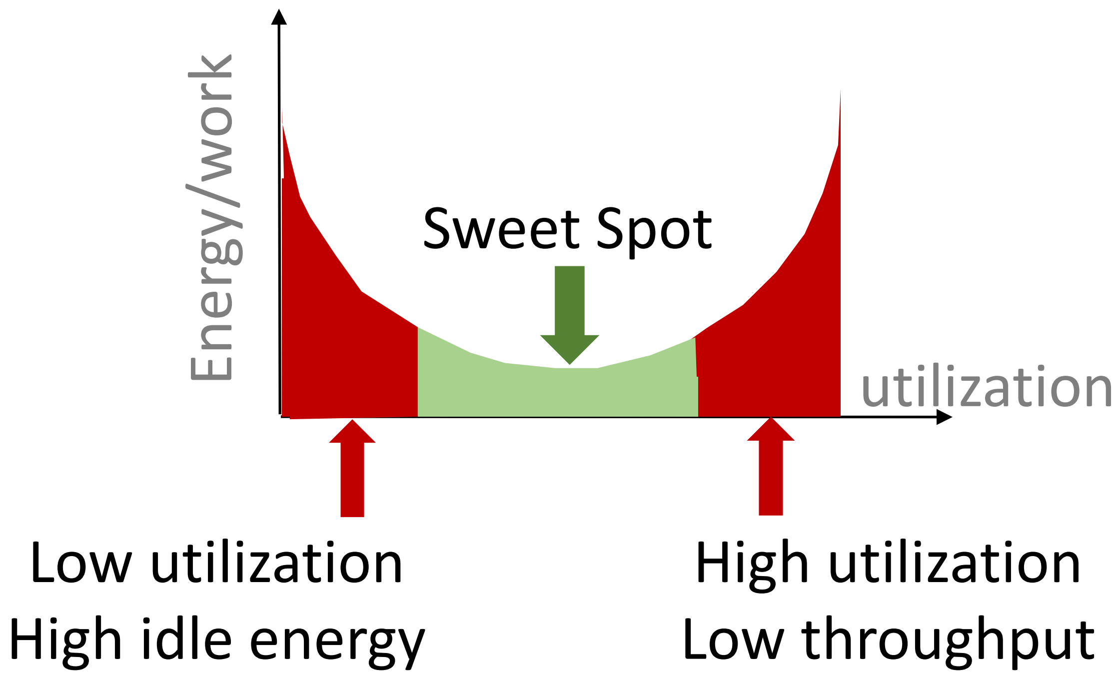
Pack hundreds, thousands of apps on shared infrastructure: keep utilization high



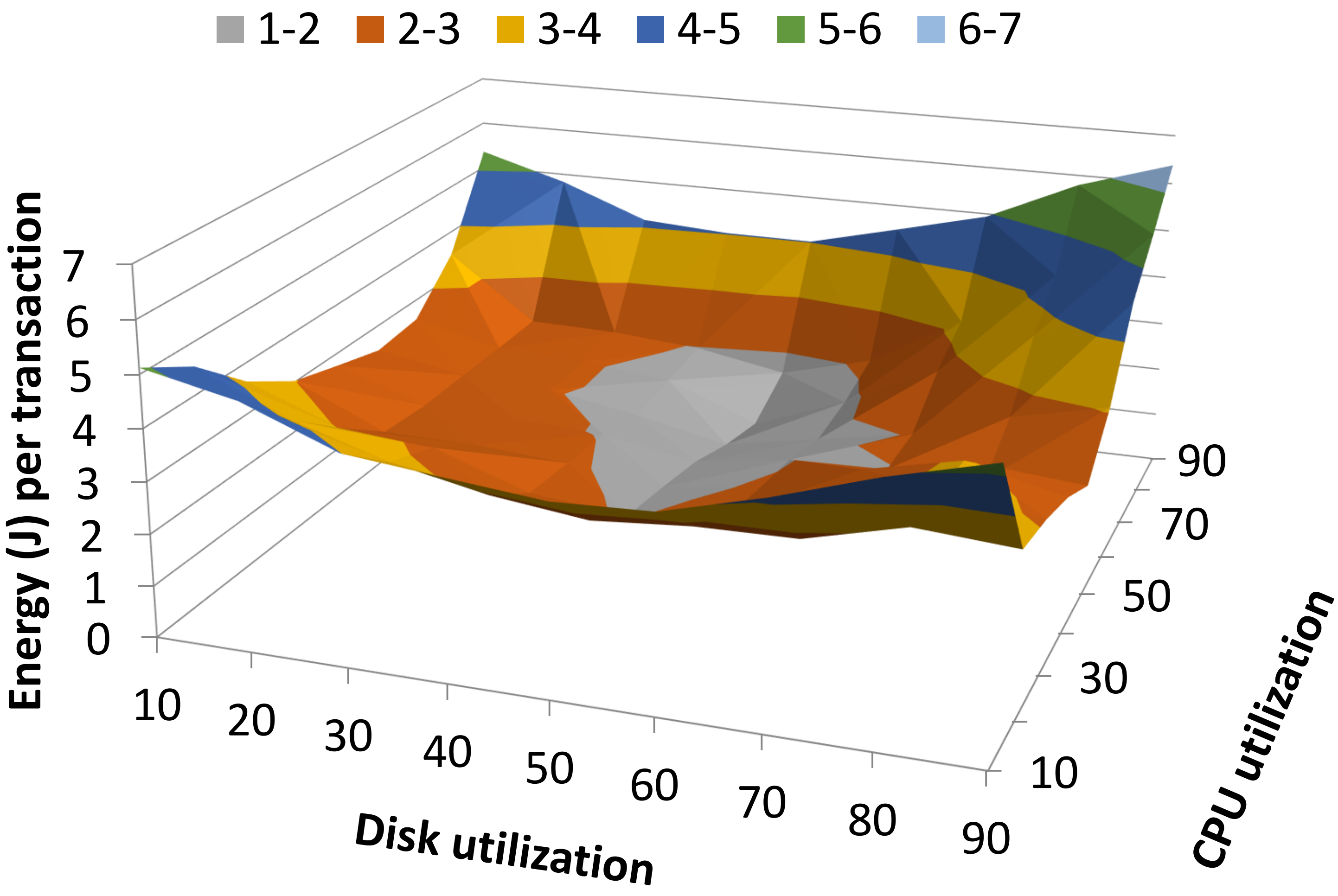
CPU + memory

storage

Consolidation Can Hurt Performance



Measurement



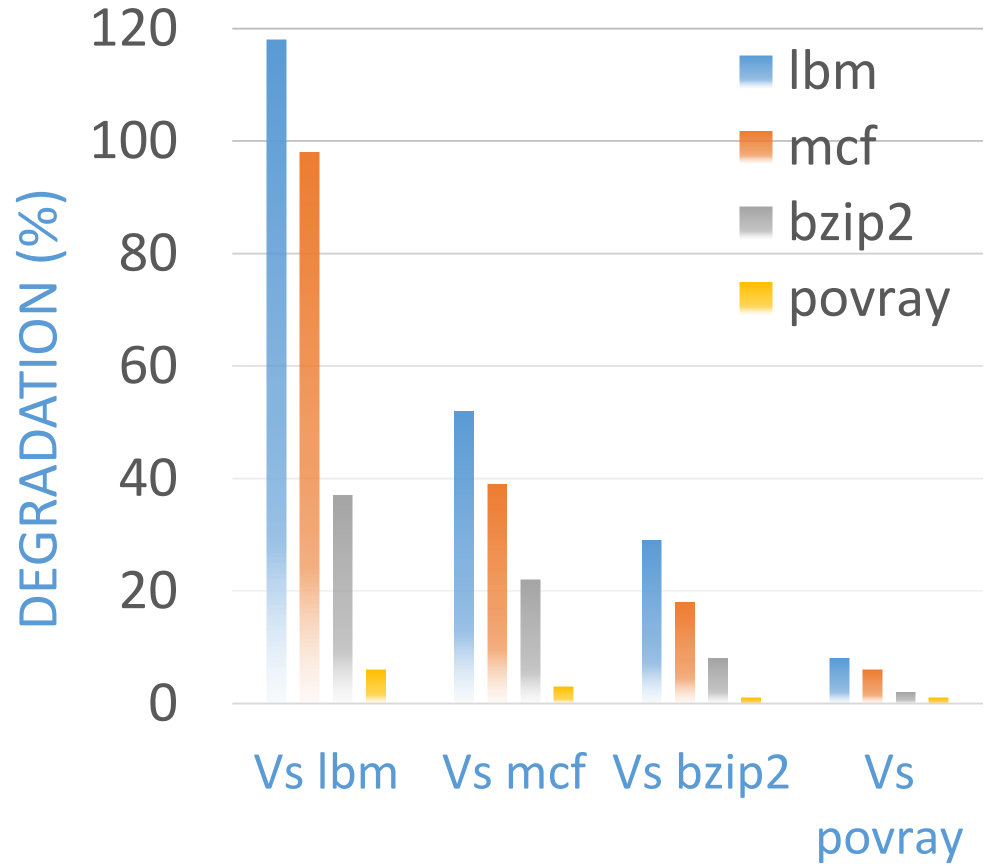
Power and performance for a toy web service with CPU and disk access

Virtualize to Isolate Resources

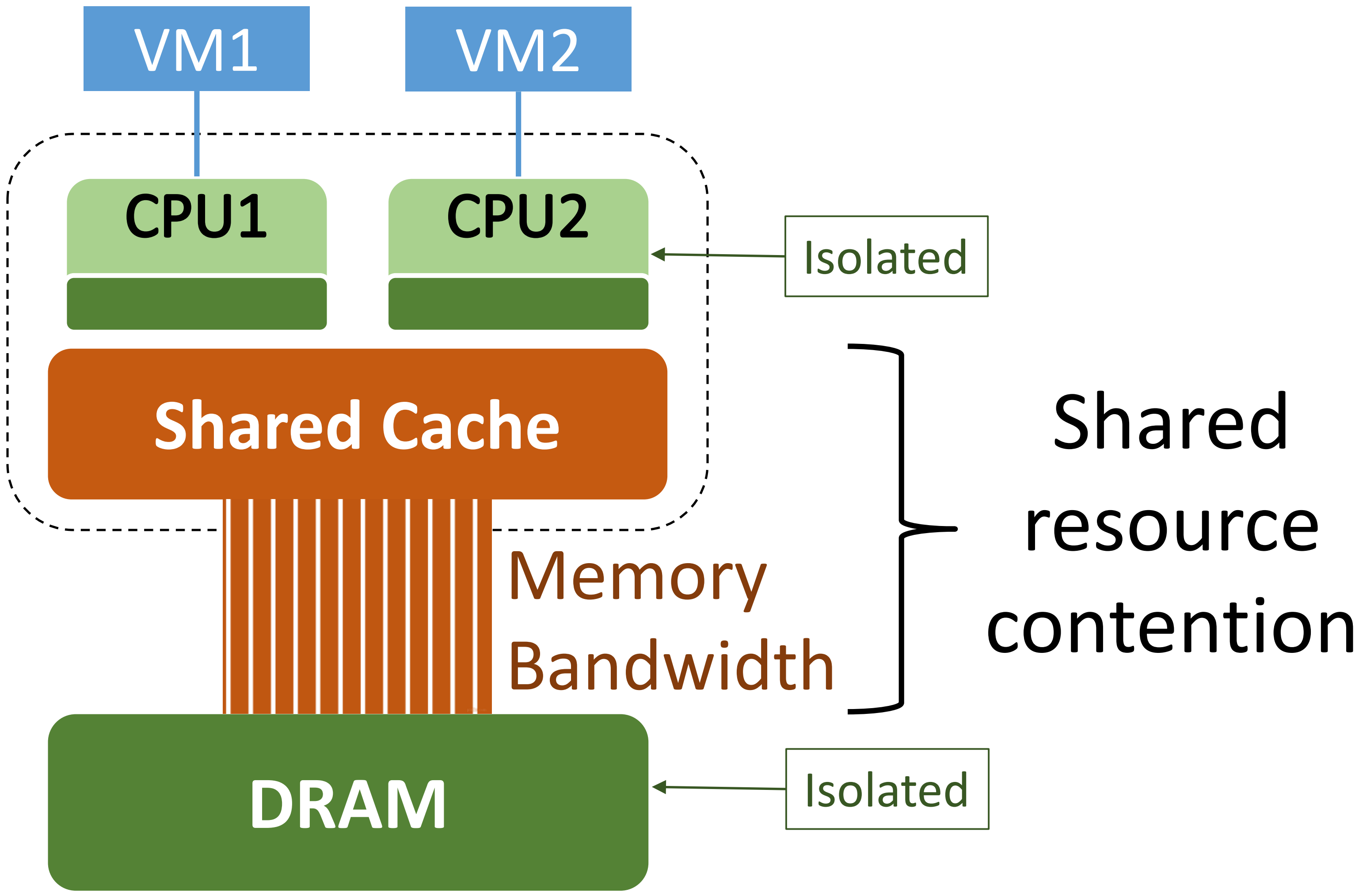
Not enough

Up to 125% degradation in Intel Core 2 Duo, Nehalem, AMD Opteron

Up to 40% measured on Google data center apps [Tang et al, ISCA'11]

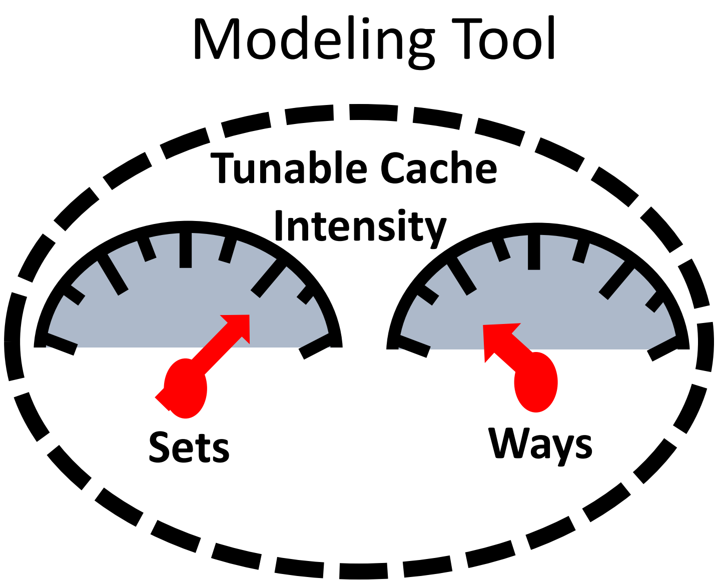


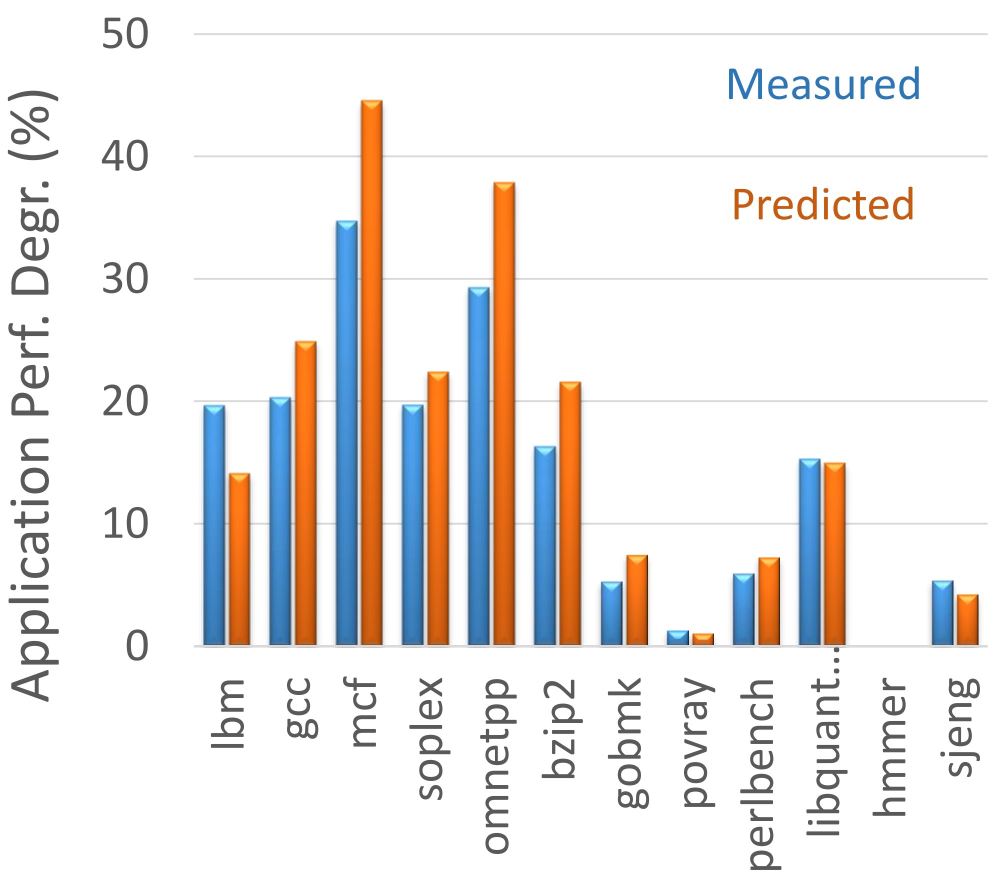
CPU: Isolation is not perfect



Interference Can Be Modeled

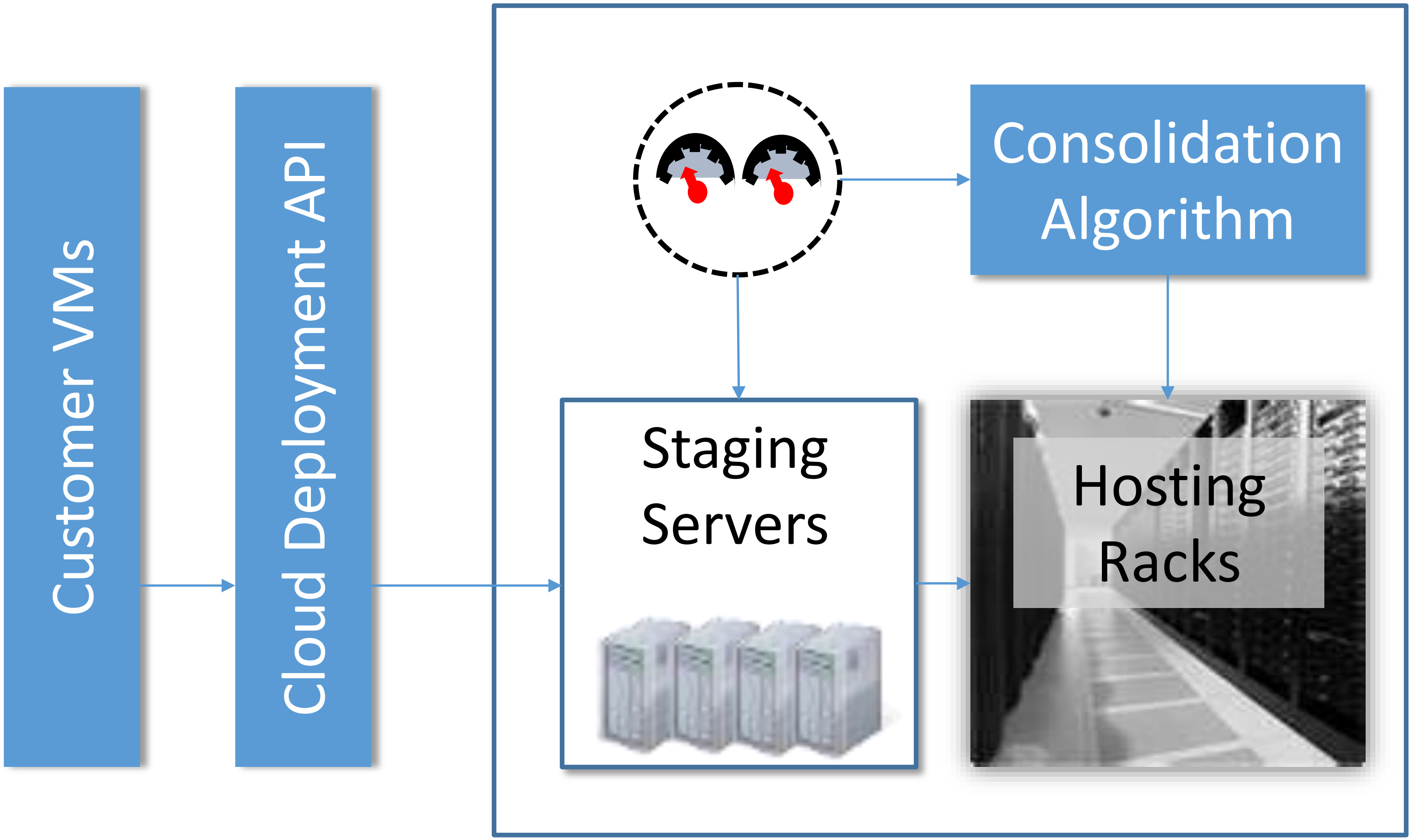
Individual modeling to predict all co-located sets





[Govindan et al, SOCC 2011]

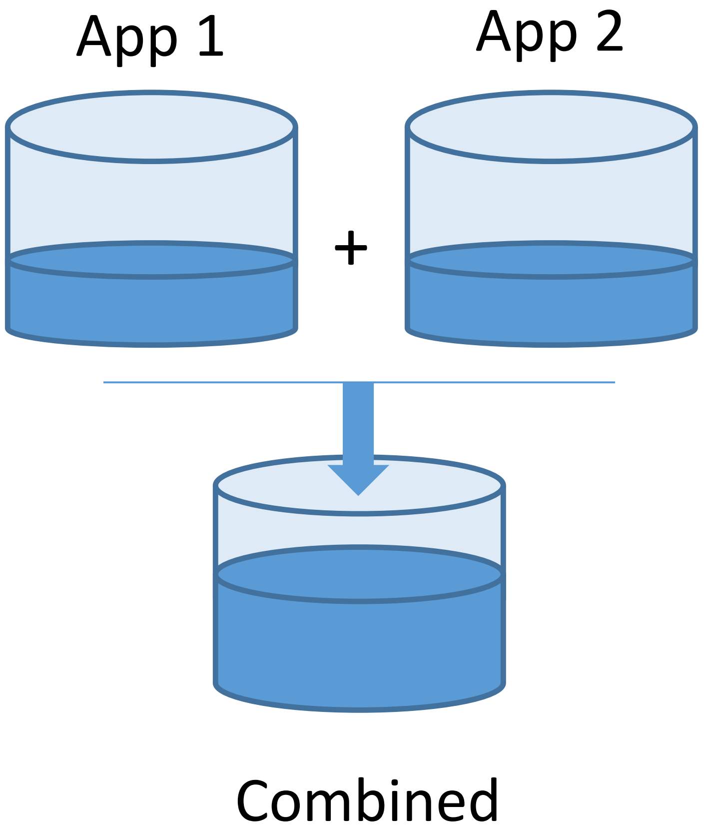
CPU: Performance Aware Consolidation



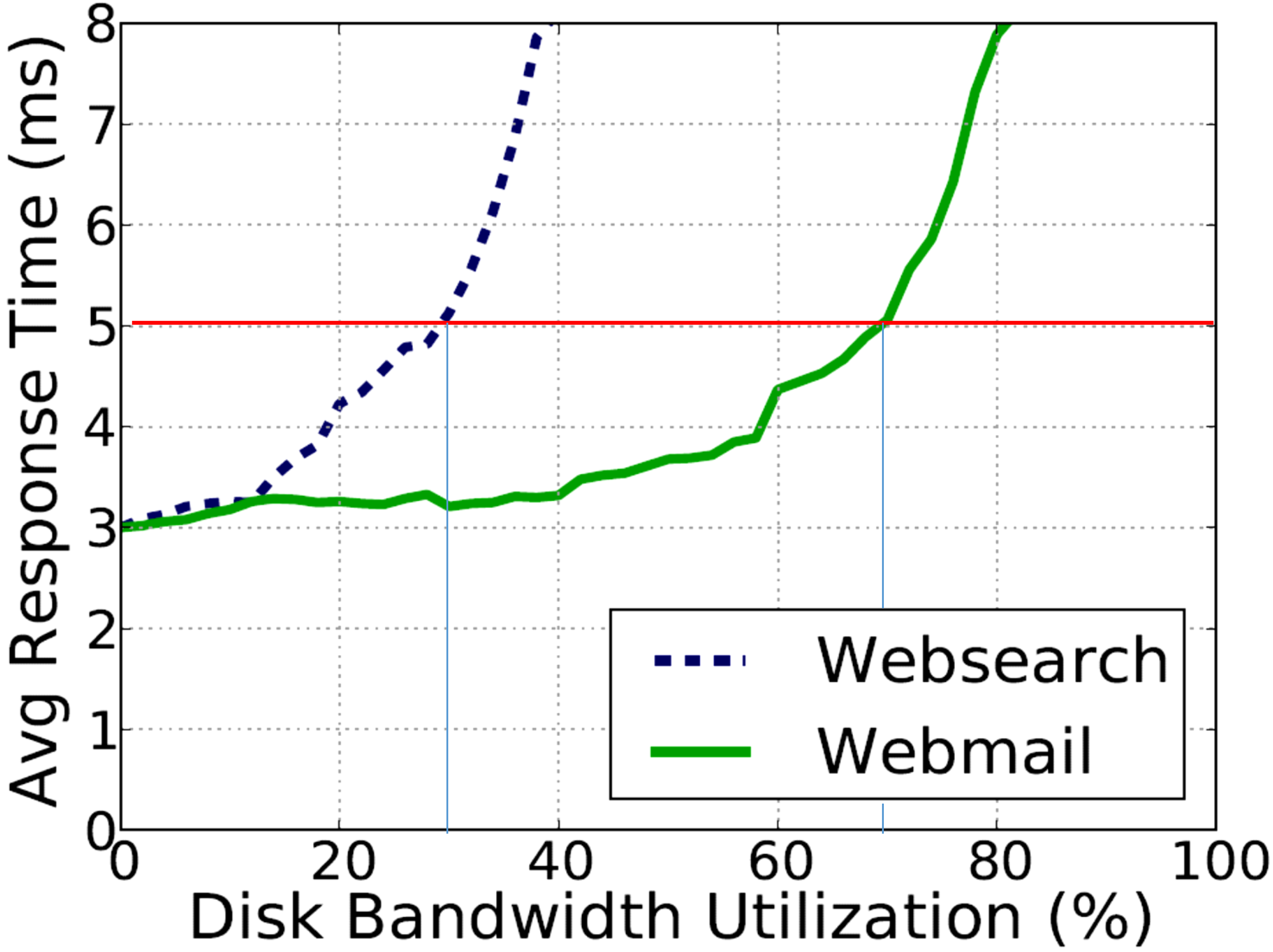
Consolidating Storage

Allocate required storage capacity

But performance depends on I/O bandwidth



Bandwidth is Not Additive



Sufficient Bandwidth

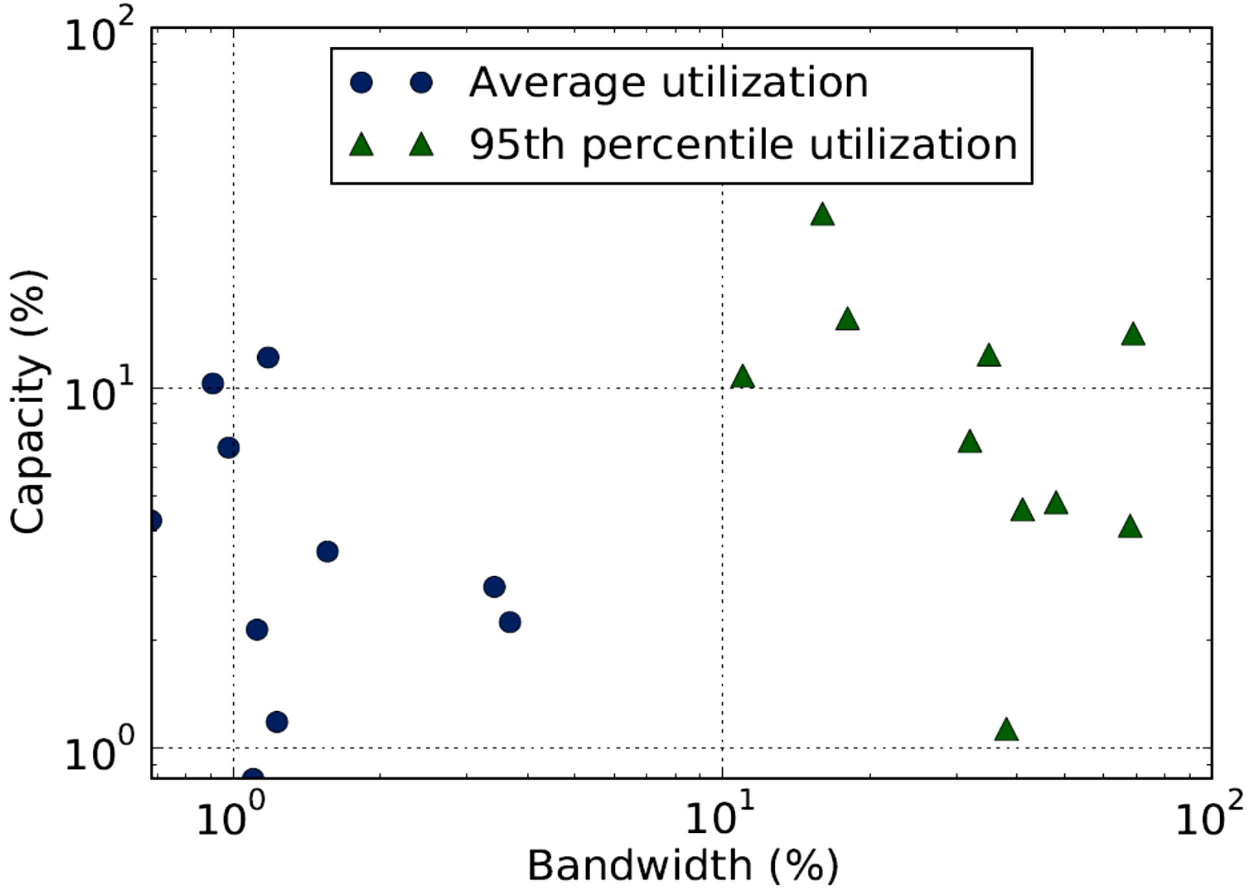
$B_{max}(A_i)$ = maximum bandwidth that app A can use within performance bound

$B(A_i)$ = current bandwidth usage of app A

$$\sum_{i=0}^{n} B(A_i) < \min_{i=1\ldots n} B_{max}(Ai)$$

Bandwidth Varies Over Time

More users active at certain times => more photos, emails



Storage Consolidation Savings

Strategy	Energy Savings	Performance
Capacity only	2.31	0.623
Bandwidth	1.35	0.970
Capacity, Bandwidth, Dynamics	3.18	0.982

Average savings across 10 Microsoft data center applications, relative to when hosted without consolidation (in research).

Summary: Don't forget the biggest slice

Look beyond energy use: infrastructure, IT

Use cheaper servers: tune for app needs

- CPU: fastest is not most efficient
- Storage: capacity is cheap, optimize for fast access (cache in RAM, stripe)
- Memory: larger RAM benefits interactive apps

Use fewer servers: do not waste idle capacity

- Consolidate: do more with less
- Bin packing is not enough, preserve performance

Acknowledgments

Sriram Govindan
Jie Liu
Suman Nath
Sriram Shankar
Kushagra Vaid
Feng Zhao
Christina Delimitrou
Christos Kozyrakis
Harold Lim
Alan Roytman
Shekhar Srikantaiah


Questions?
http://www.facebook.com/EfficientDataCenter