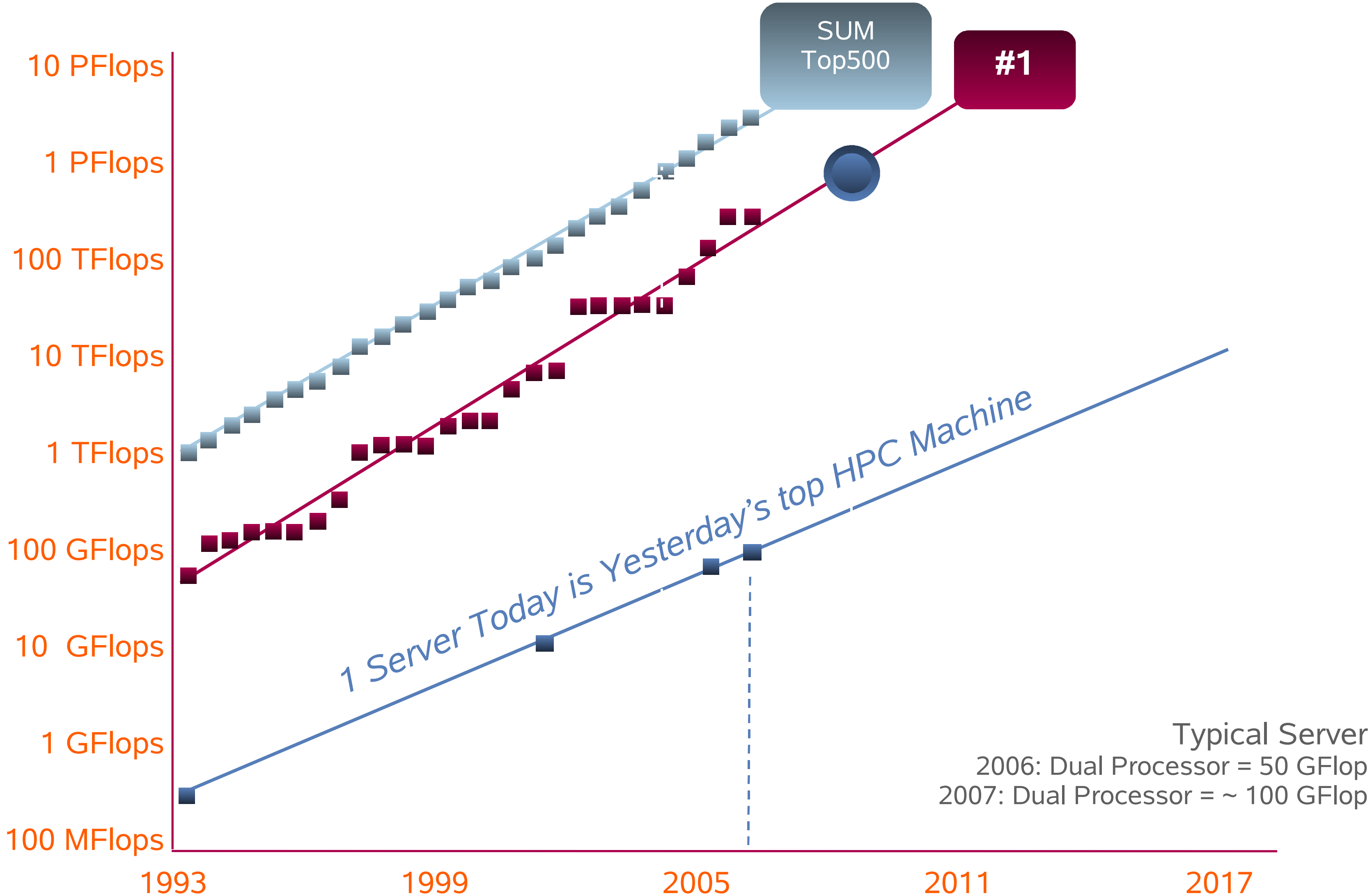
Is there a pathway to a Green Grid ??

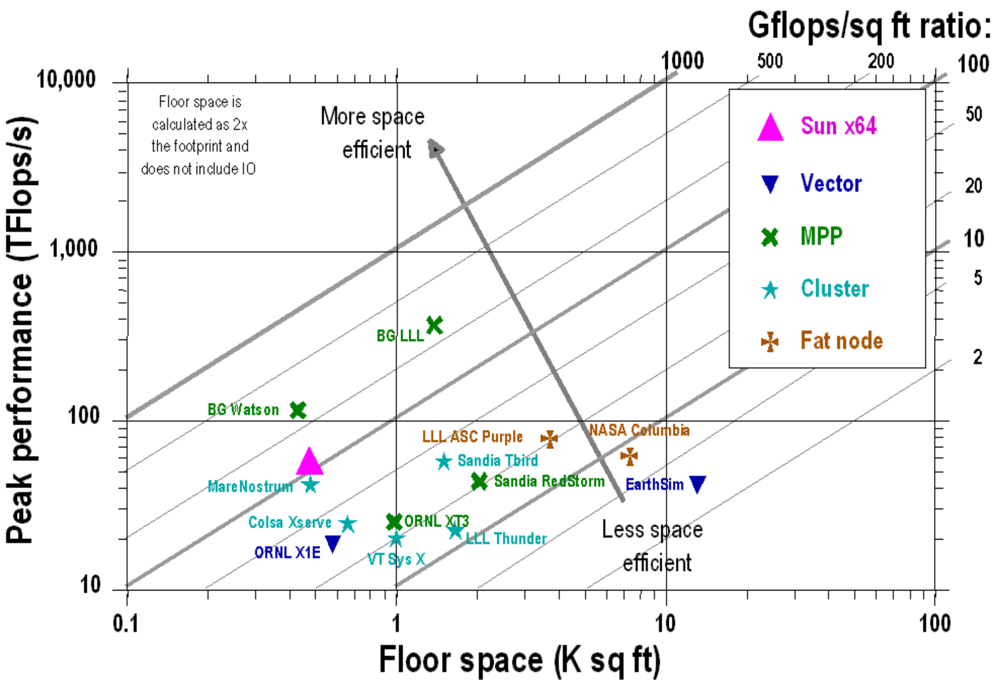
Simon See, Ph.D

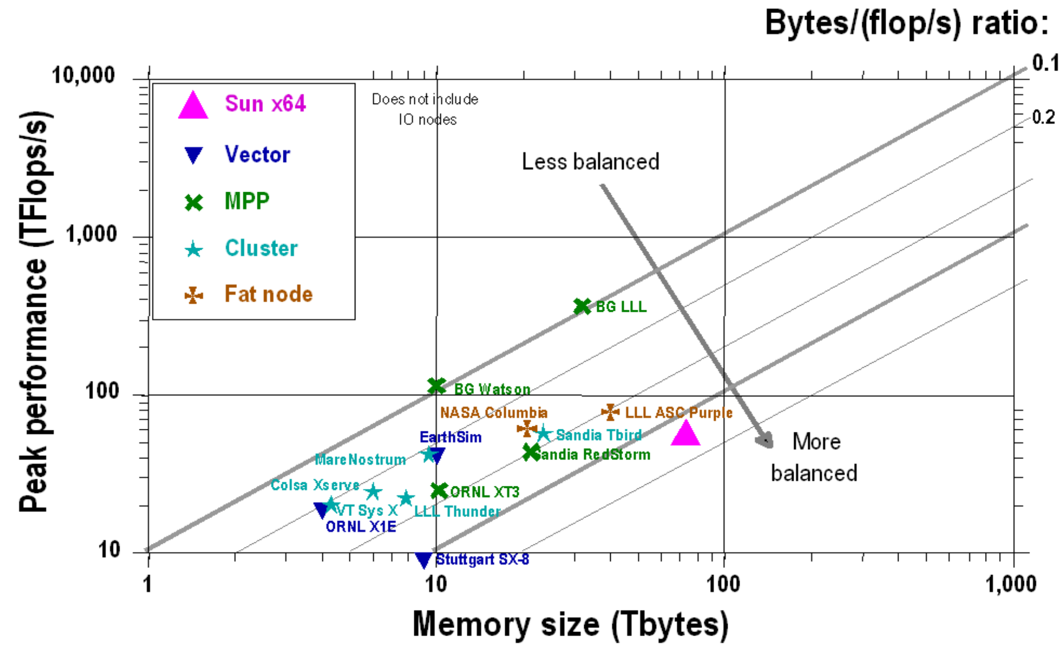
Director, Global HPC Solution
Sun Microsystesms Inc.

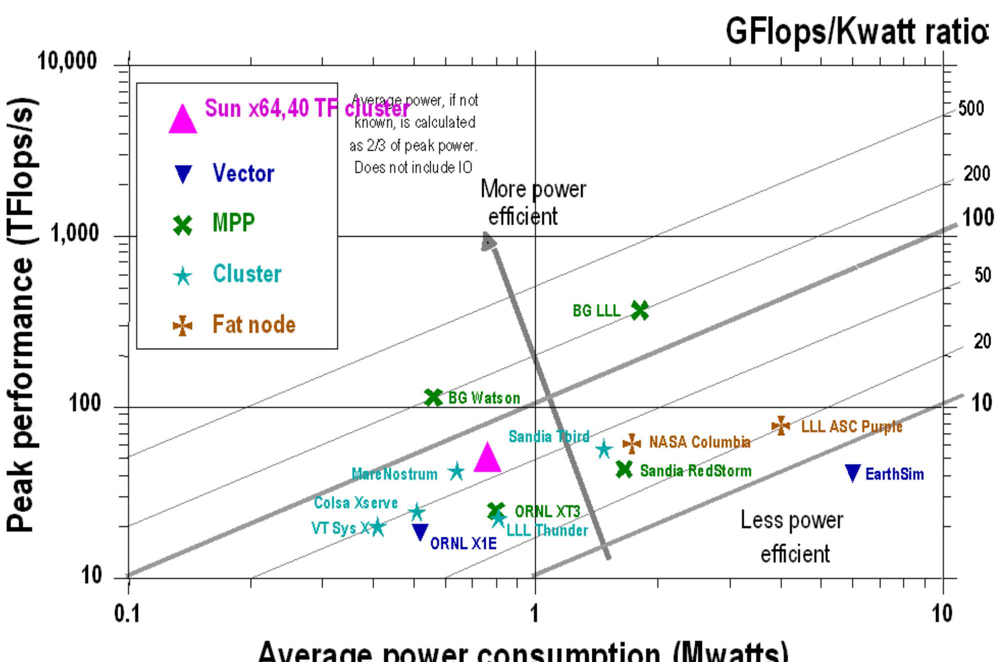
Associate Prof.
NTU and NUS

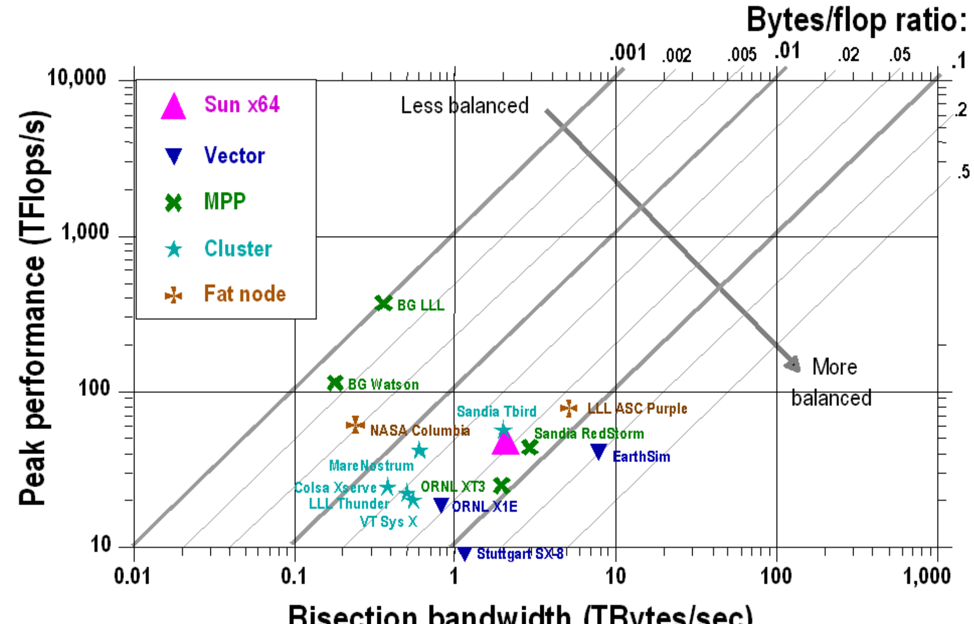
HPC Top500: An Example of Moore's Law




Gflops/sq ft ratio:
1000
500
200
100
50
20
10
5
2
10,000
1,000
100
10
Floor space is calculated as 2x the footprint and does not include IO
More space efficient
Less space efficient
Sun x64
Vector
MPP
Cluster
Fat node
BG LLL
BG Watson
LLL ASC Purple
NASA Columbia
Sandia Tbird
Sandia RedStorm
EarthSim
MareNostrum
Colsa Xserve
ORNL XT3
ORNL X1E
LLL Thunder
VT Sys X
0.1
1
10
100
Floor space (K sq ft)
Peak performance (TFlops/s)


Bytes/(flop/s) ratio:
0.1
0.2
10,000
1,000
100
10
Does not include IO nodes
Less balanced
More balanced
Sun x64
Vector
MPP
Cluster
Fat node
BG LLL
BG Watson
NASA Columbia
LLL ASC Purple
EarthSim
Sandia Tbird
Sandia RedStorm
MareNostrum
Colsa Xserve
ORNL XT3
VT Sys X
LLL Thunder
ORNL X1E
Stuttgart SX-8
1
10
100
1,000
Memory size (Tbytes)
Peak performance (TFlops/s)


GFlops/Kwatt ratio
500
200
100
50
20
10
10,000
1,000
100
10
Sun x64,40 TF cluster
Vector
MPP
Cluster
Fat node
Average power, if not known, is calculated as 2/3 of peak power. Does not include IO
More power efficient
Less power efficient
BG LLL
BG Watson
Sandia Tbird
NASA Columbia
LLL ASC Purple
MareNostrum
Sandia RedStorm
EarthSim
Colsa Xserve
ORNL XT3
VT Sys X
LLL Thunder
ORNL X1E
0.1
1
10
Average power consumption (Mwatts)
Peak performance (TFlops/s)


Bytes/flop ratio:
.001
.002
.005
.01
.02
.05
.1
.2
.5
10,000
1,000
100
10
Sun x64
Vector
MPP
Cluster
Fat node
Less balanced
More balanced
BG LLL
BG Watson
Sandia Tbird
LLL ASC Purple
NASA Columbia
Sandia RedStorm
MareNostrum
EarthSim
Colsa Xserve
ORNL XT3
LLL Thunder
ORNL X1E
VT Sys X
Stuttgart SX-8
0.01
0.1
1
10
100
1,000
Bisection bandwidth (TBytes/sec)
Peak performance (TFlops/s)

Number of Data Centers Growing

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Very Large	100	106	120	160	175	185	210	230	250	270
Large	900	870	880	900	920	990	1,040	1,100	1,170	1,250
Medium	1,405	1,385	1,395	1,420	1,490	1,585	1,665	1,765	1,870	1,975
Small	2,180	2,100	2,110	2,190	2,230	2,290	2,360	2,430	2,500	2,570
Total	**4,585**	**4,461**	**4,505**	**4,670**	**4,815**	**5,050**	**5,275**	**5,525**	**5,790**	**6,065**
Growth Rate		-2.70%	1.00%	3.70%	3.10%	4.90%	4.35%	4.74%	4.80%	4.75%

Small Data Center

- Between 350 – 500 Servers Installed
- 15,000 Sq Feet Of Raised Floor
- Predominately Volume Server Architecture, with 1 -3 High End Server Systems

Medium Data Center

- Between 1,500 – 1,700 Servers Installed
- 20,000 Sq Feet Of Raised Floor
- Four or Five High End Systems Form The Basis Of Enterprise Systems

Large Data Center

- Between 2,000 – 2,500 Servers Installed
- 35,000 Sq Feet Of Raised Floor
- House Up To 7 High End Systems

Very Large Data Center

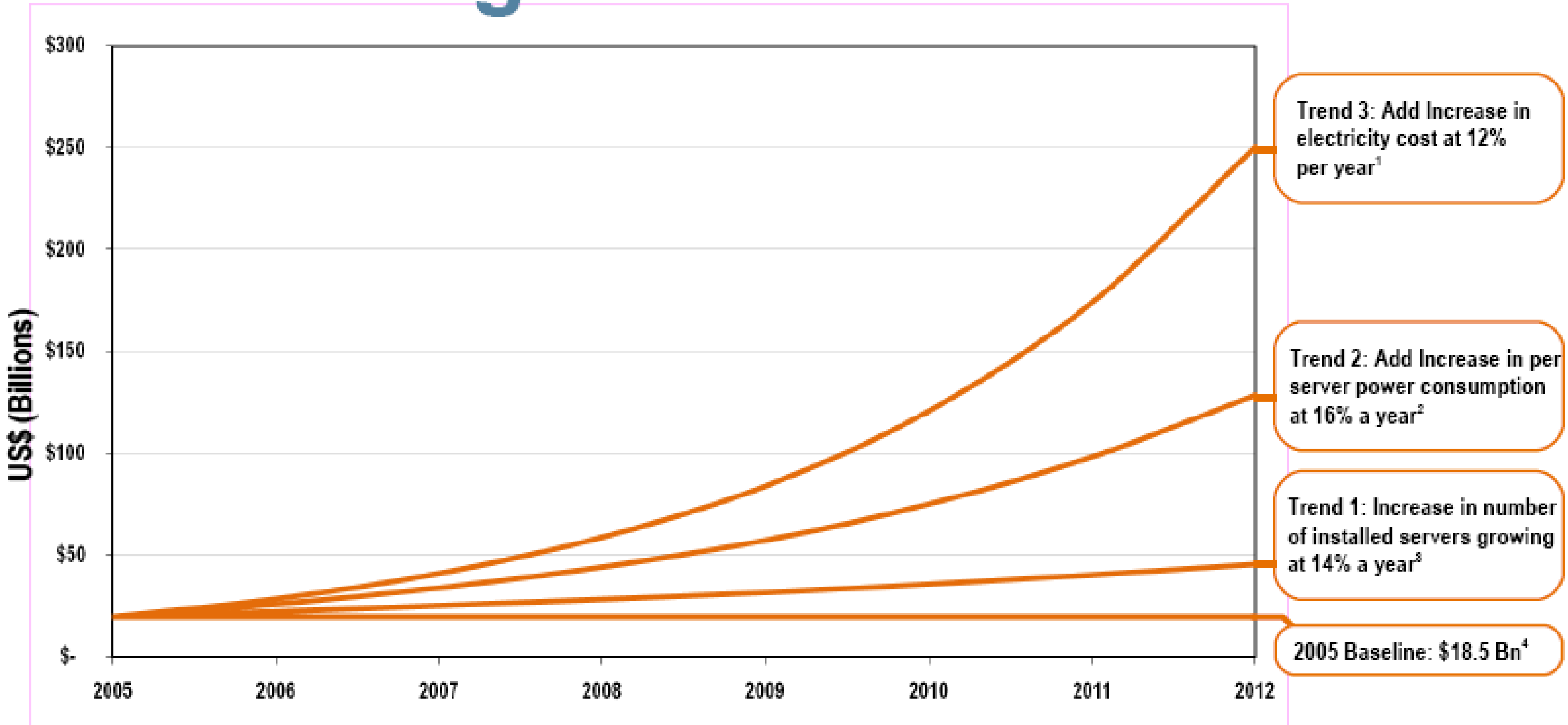
- Up to 25,000 Servers Installed
- 100,000+ Sq Feet Of Raised Floor
- Eight+ High End Systems

Source: Presentation by Vernon Turner from IDC at the Sun Summit on 21st Century Eco-Responsibility, November 14, 2005 for US Data Centers

Number of Servers in Data Centers Growing

	2008	2009	2010	2011	2012
Very Large	449	503	565	630	695
Greenfield	40	54	62	65	67
Brownfield	409	449	503	565	630
Servers in Very Large	11,227,357	12,572,577	14,115,316	15,754,862	17,509,290
Large	2,340	2,491	2,700	2,949	3,208
Greenfield	140	151	209	249	255
Brownfield	2,200	2,340	2,491	2,700	2,949
Servers in Large	5,265,040	5,603,777	6,075,723	6,635,948	7,295,537
Medium	3,746	3,987	4,333	4,714	5,095
Greenfield	220	241	346	381	385
Brownfield	3,526	3,746	3,987	4,333	4,714
Servers in Medium	5,994,227	6,379,685	6,932,461	7,542,168	8,196,942
Small	5,413	5,652	5,965	6,302	6,673
Greenfield	210	239	313	337	375
Brownfield	5,203	5,413	5,652	5,965	6,302
Servers in Small	2,300,426	2,401,961	2,535,233	2,678,327	2,833,263
Total Number of Servers	**24,787,050**	**26,958,000**	**29,658,733**	**32,611,304**	**35,835,032**

The Cubing Effect



By 2012 data center power consumption costs could grow to $250B worldwide – demanding proactive energy management solutions

1. U.S. Energy Information Administration (www.eia.doe.gov)
2. Sun primary research
3. IDC#34867 U.S. and Worldwide Server Installed Base 2006-2009 Forecast (February 2006)
4. IDC Worldwide Server Power and Cooling Expense 2006-2010 Forecast

Green Grid

Industry and user organization focused on Energy Efficient Data Centers and Enterprise IT

> Launched April 26th with 11 companies
> AMD, APC, Dell, HP, IBM, Intel, Microsoft, Rackable Systems, SprayCool, Sun Microsystems, and VMware
> Now at 40+ companies.

Mission Statement:

A global consortium dedicated to advancing energy efficiency in data centers and business computing ecosystems.

In furtherance of its mission, the Green Grid, in consultation with end-users, will:

- Define meaningful, end-user-centric models and metrics;
- Develop standards, measurement methods, processes and new technologies to improve performance against the defined metrics; and
- Promote the adoption of the energy efficient standards, processes, measurements and technologies.

SPECpower

1 Load varied in 10% decrements from 100% to 0%

2 SSJ_ops calculated at each load point

3 Power measured at wall socket with approved external meter at each load point

4 SSJ_ops divided by Watts at each load point

Benchmark Results Summary

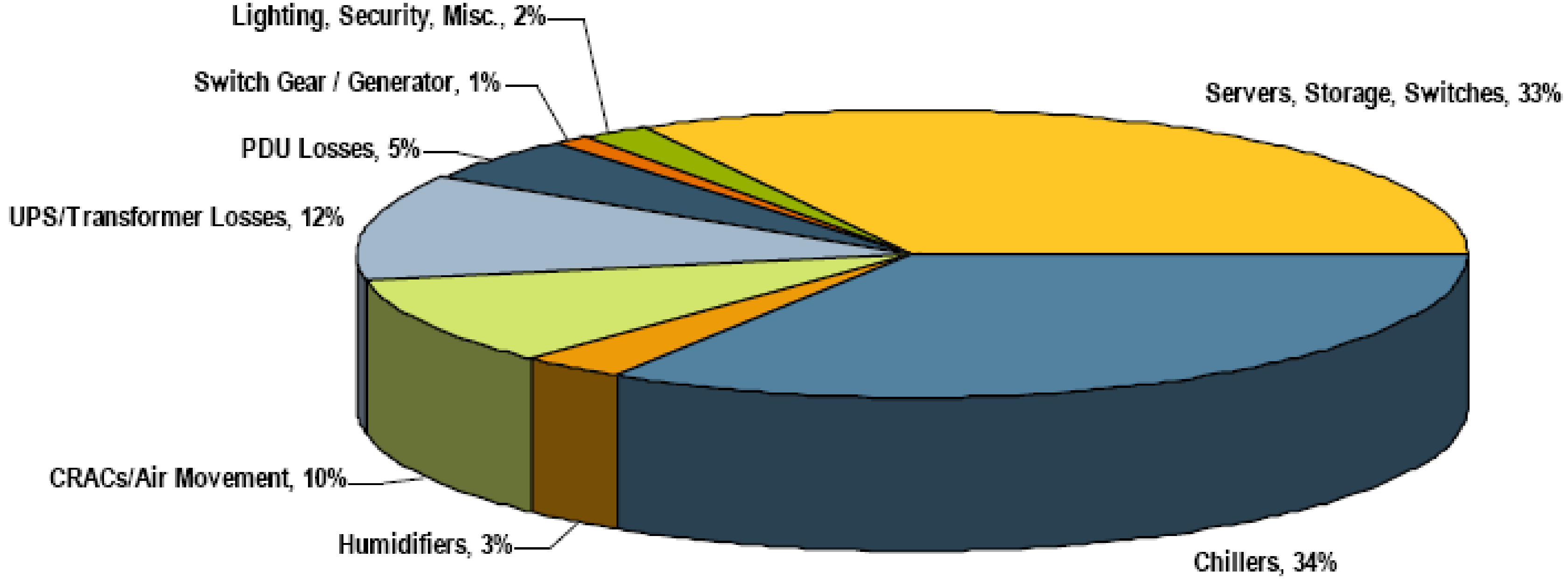
Performance			Power	Performance to Power Ratio
Target Load	Actual Load	ssj_ops	Average Power (W)	
100%	98.8%	84,913	288	295
90%	90.1%	77,489	280	277
80%	80.3%	69,012	268	257
70%	69.8%	59,971	258	232
60%	60.9%	52,386	251	209
50%	50.3%	43,226	243	178
40%	40.3%	34,638	236	146
30%	30.2%	25,952	230	113
20%	20.1%	17,275	224	77.1
10%	10.8%	9,326	220	42.4
	Active Idle	0	215	0
			∑ssj_ops / ∑power =	175

5 Total SSJ_ops at all load points (including 0%) divided by total Watts at all load points

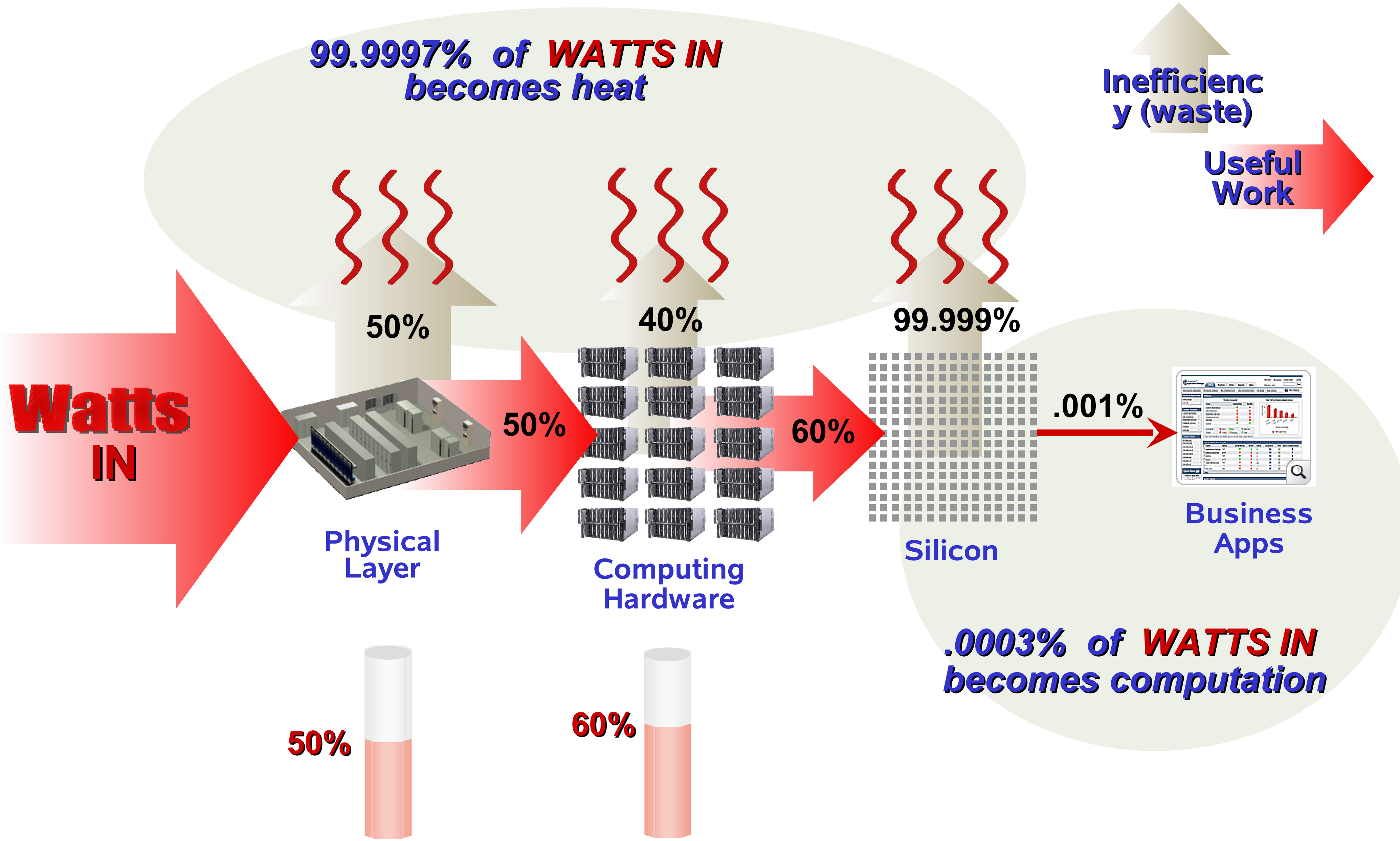
6 Final SPECpower number for system (higher is better)

Where has all the energy gone to?

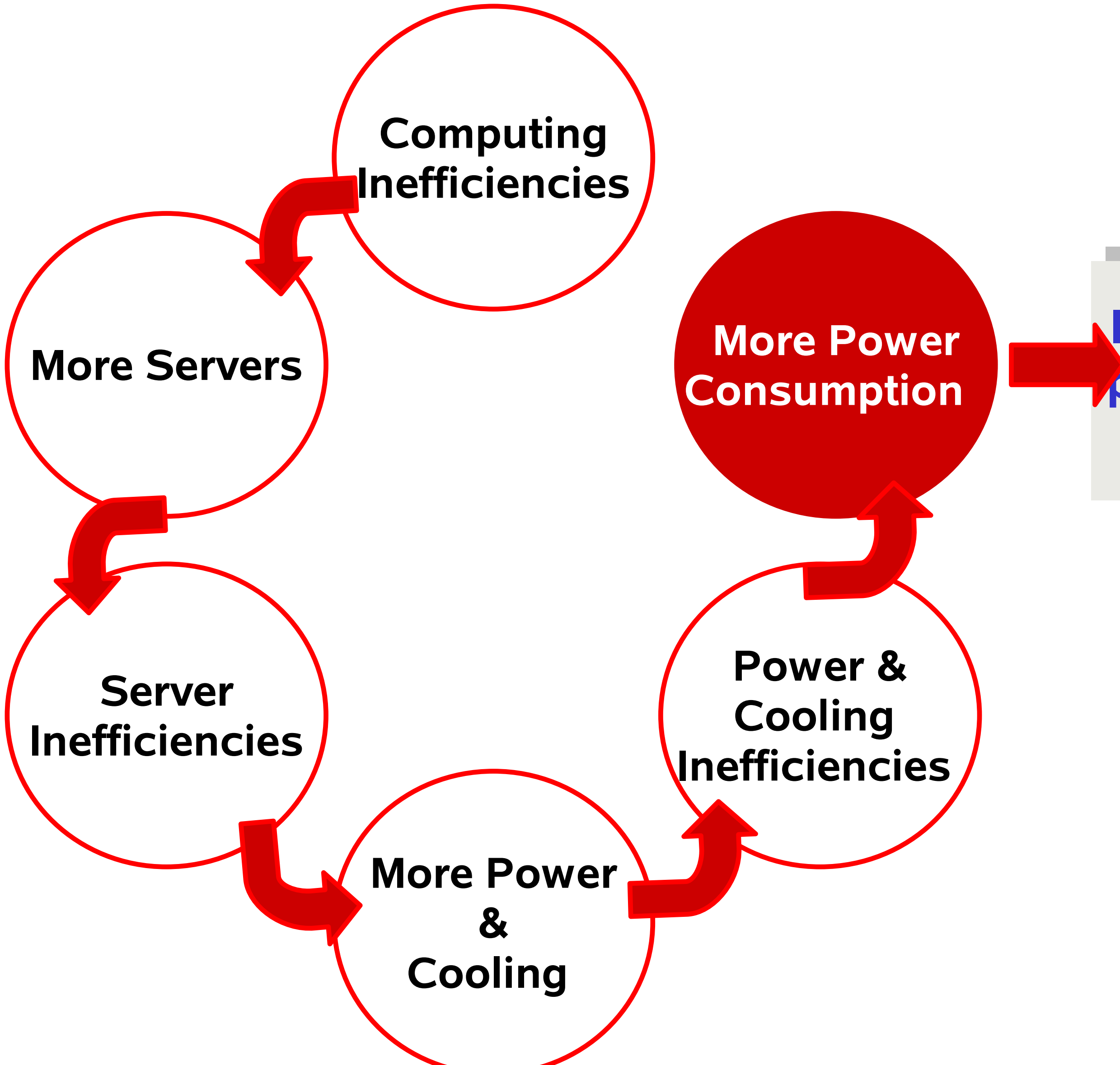
Where Does the Power in a Data Center Go?



Where Does Hardware Inefficiency Go?



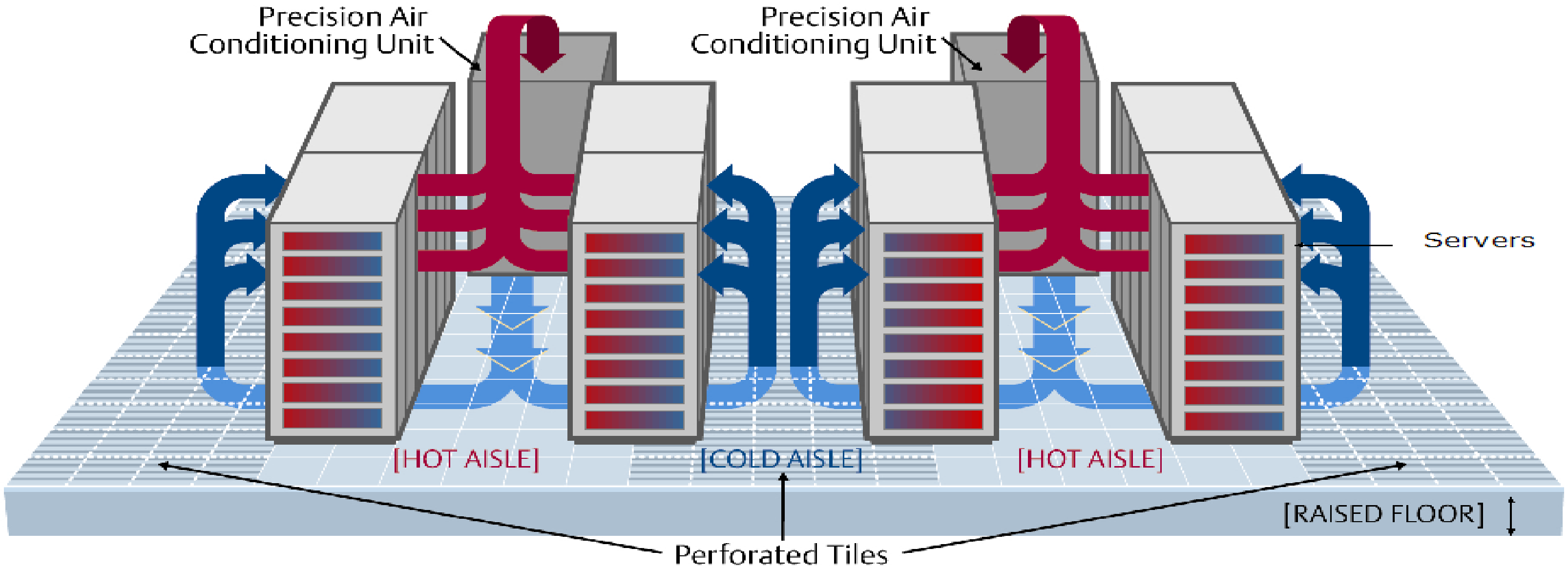
Inefficiencies Create Consumption



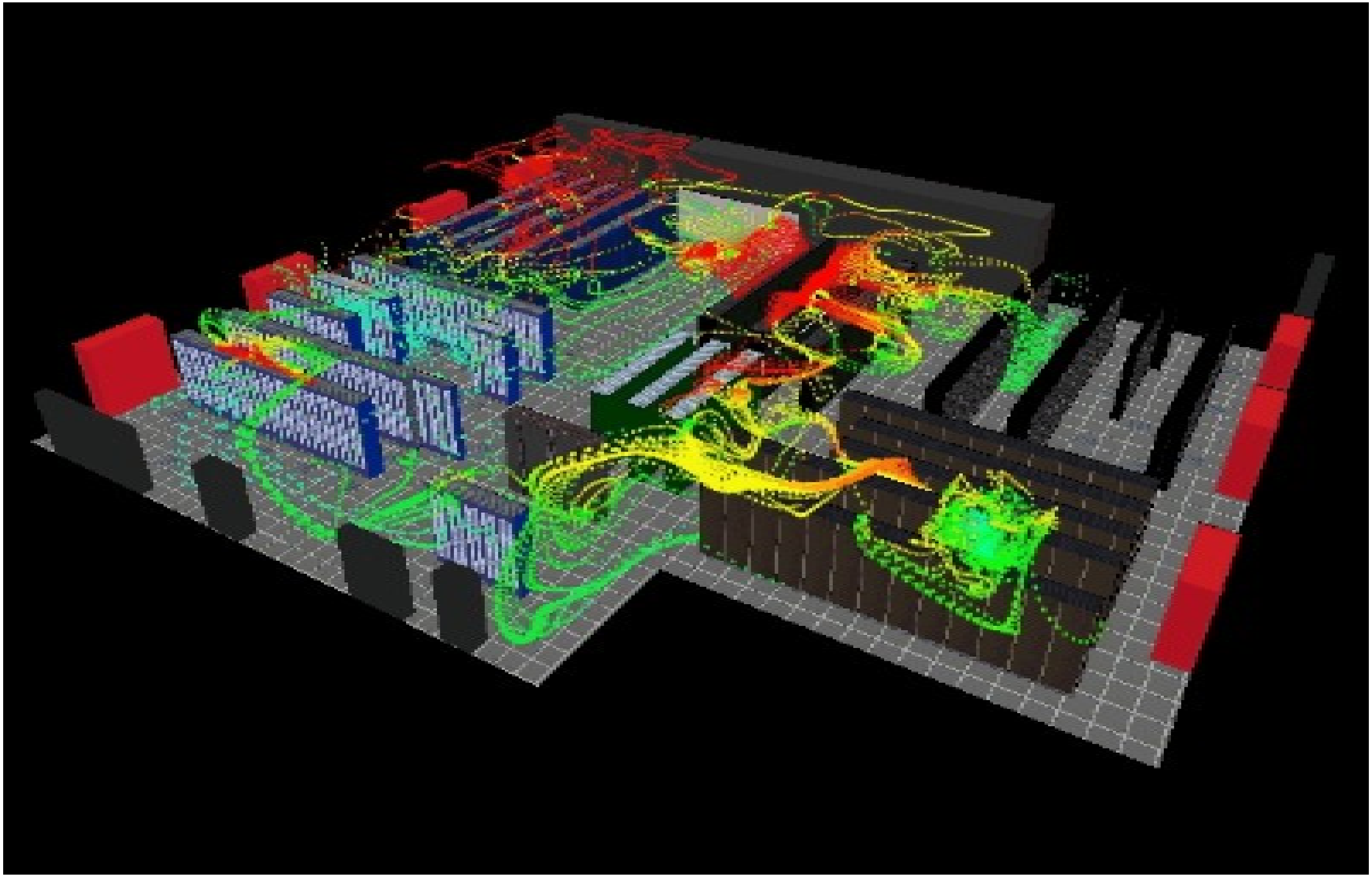
Inefficiencies drive both power consumption and material consumption

Cooling Basics

- All the heat generated by electronic equipment (server power) has to be removed out of the room.
- Traditional raised floor cooling can typically handle up to 5 kW per rack. This assumes:
  - > raised floor is high enough – higher than 24”
  - > no obstructions – cables, trays, etc…
  - > hot aisle / cold aisle equipment layout – servers front to front, back to back



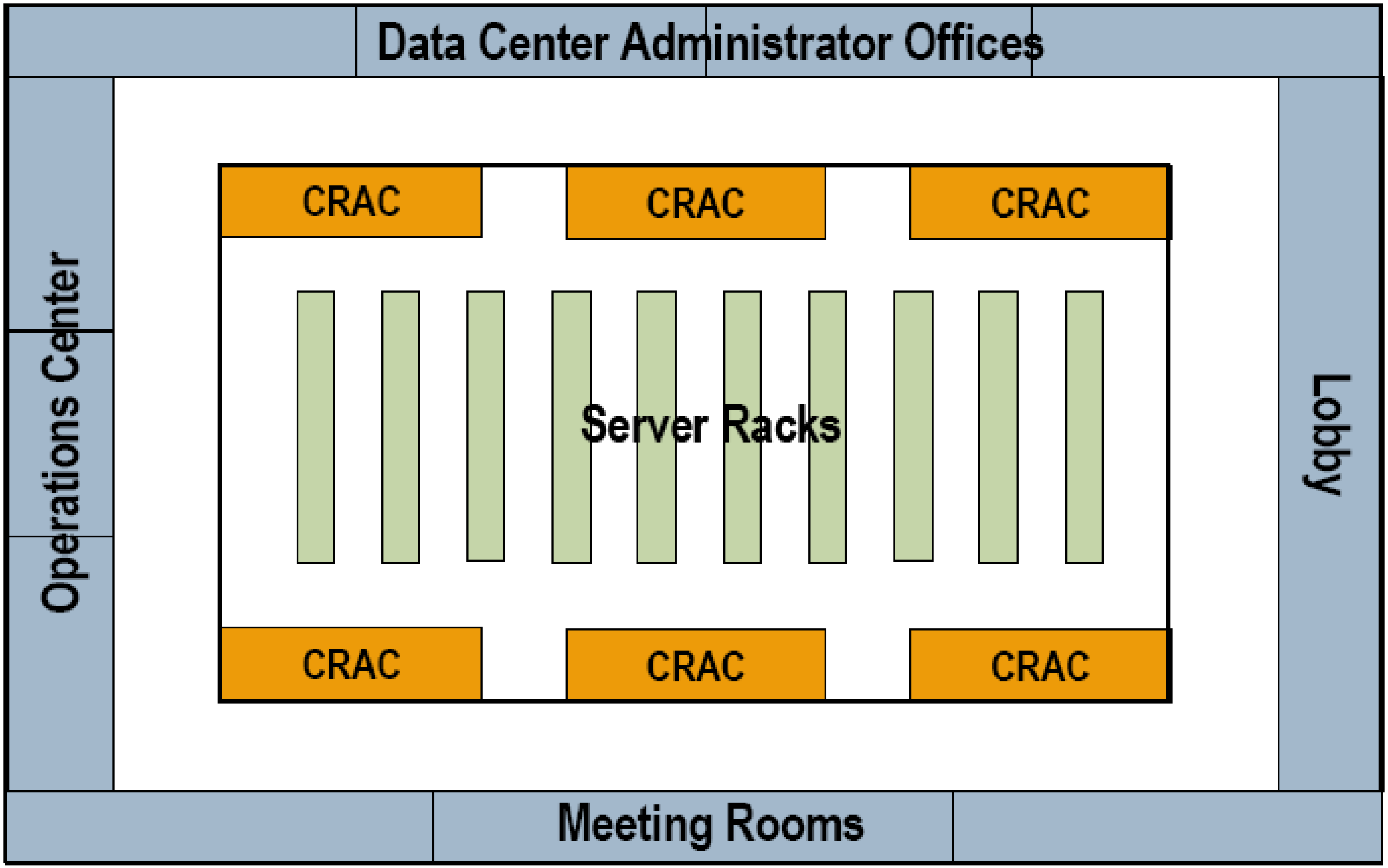
Data Center Mobile Hot Spots



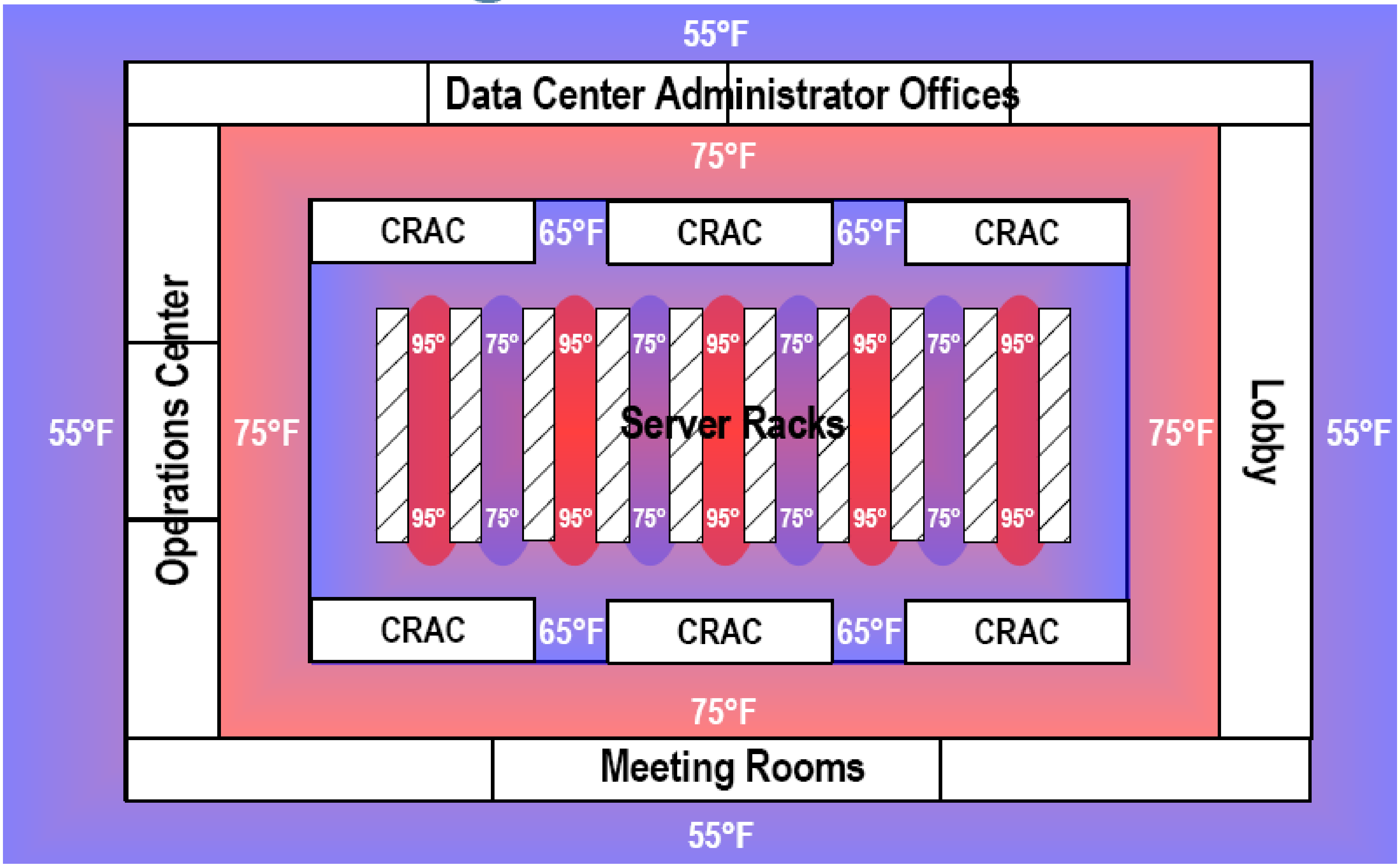
Power and Cooling Trends

- Raised Floor alone provide limited capabilities
- Rack Power Consumption > 10KW
- Blade Designs are increasing Density, Power, and Weight per Rack
- Close Coupling of Systems and Cooling will be required

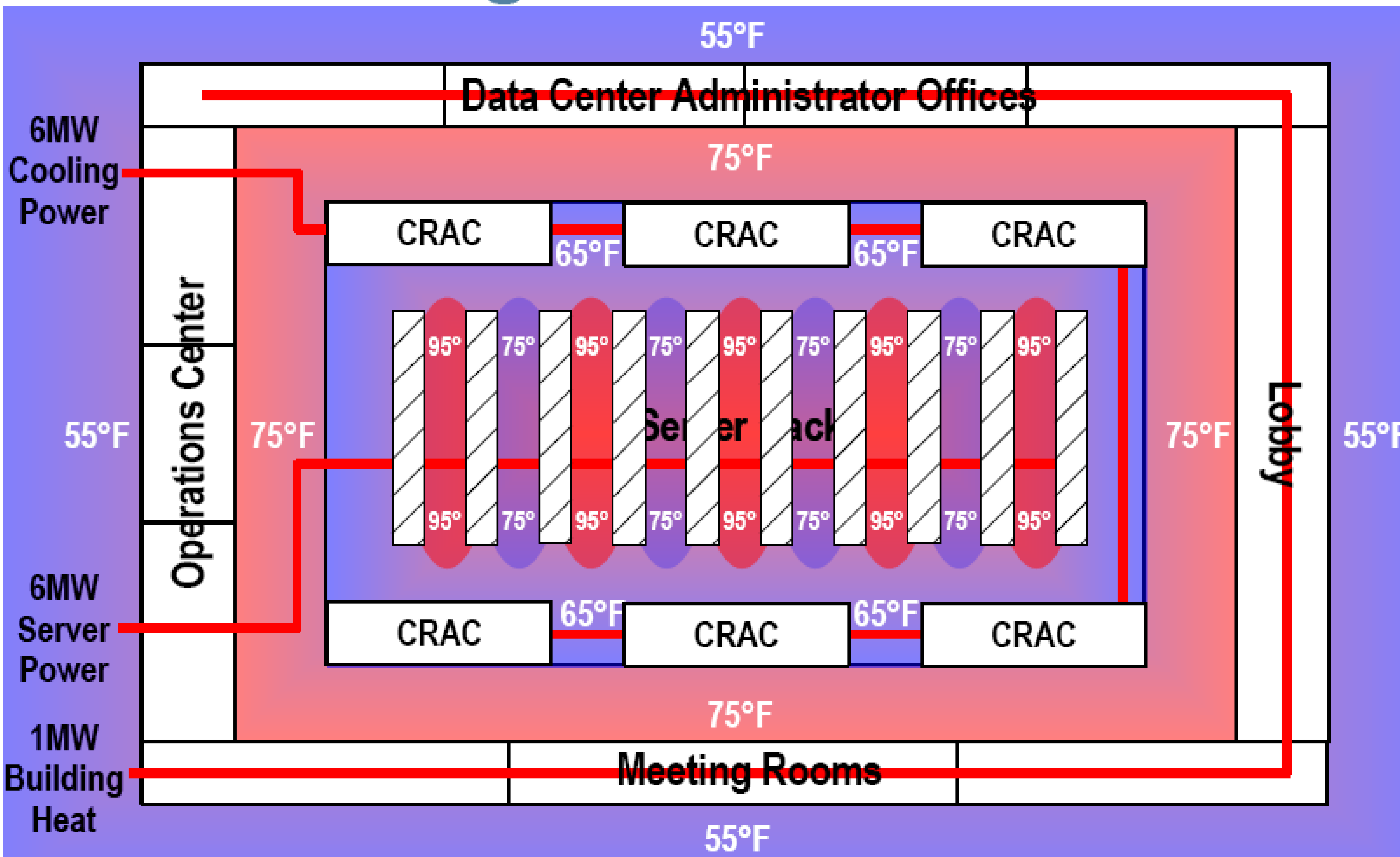
What's Wrong With This Data Center?



What's Wrong With This Data Center?



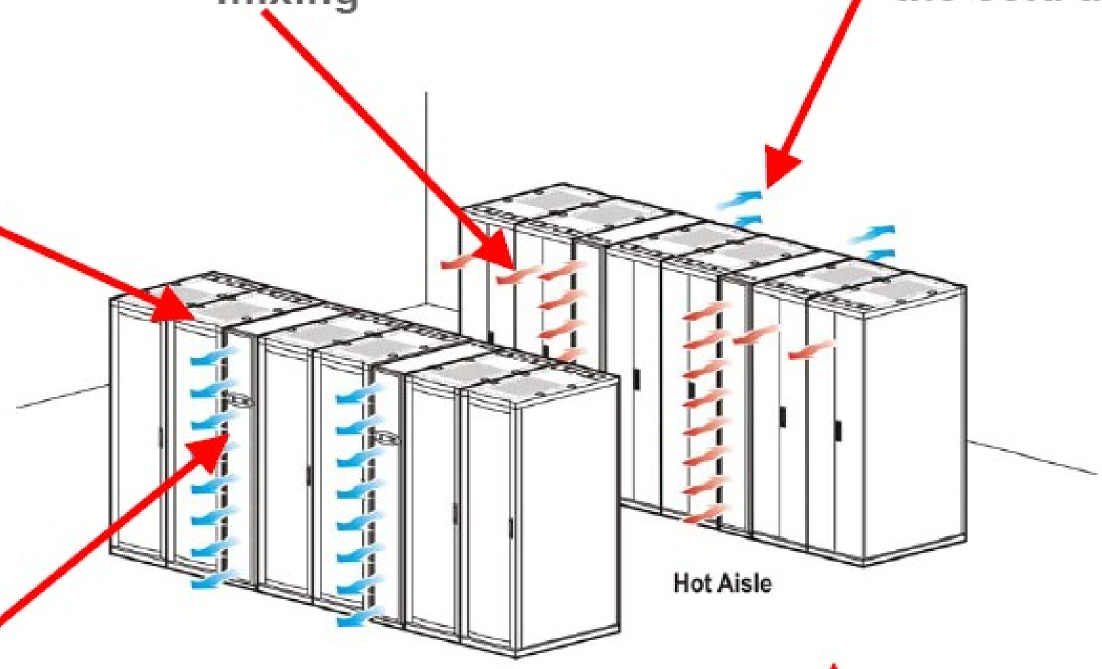
What's Wrong With This Data Center?



Sun Blade Cooling Option with APC

- Chilled Water or Refrigerant
- Variable Capacity Control
- kW Metering
- Front & Rear Serviceable
- Network Manageable

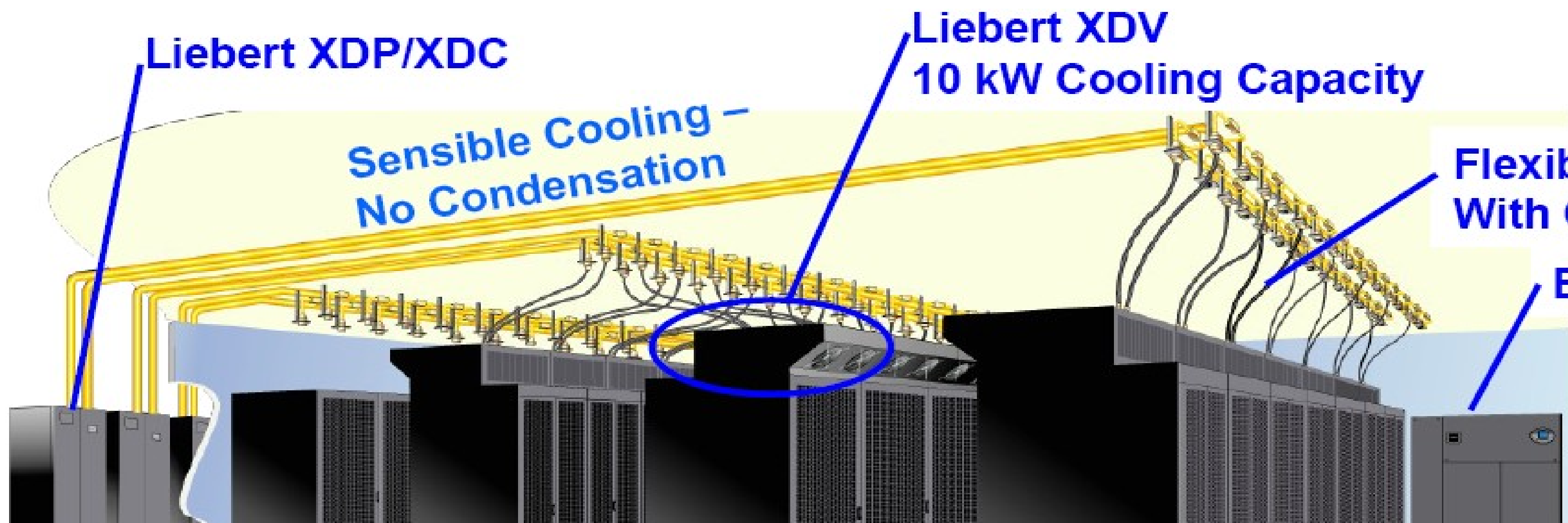
In-Row Chillers with Sun Blade 6000

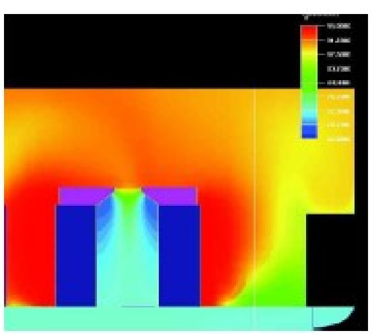


InfraStruXure InRow RC

Critical Power and Cooling Services

Liebert XDV





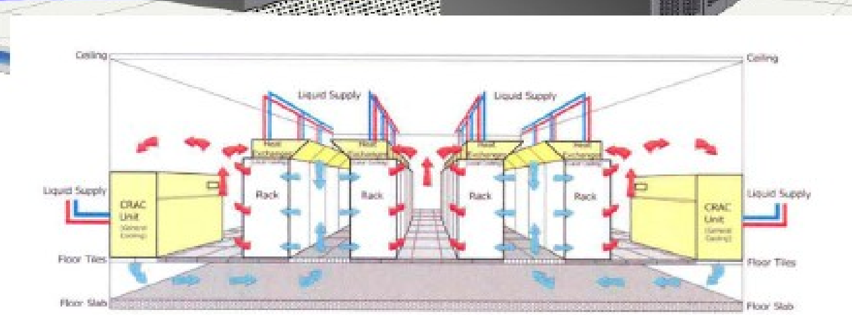


Figure 4-11 Local cooling distribution using overhead cooling units mounted to the ASHRAE, Datacom Equipment Power Trends and Cooling Applications, 2005. © American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., www.ashrae.org.

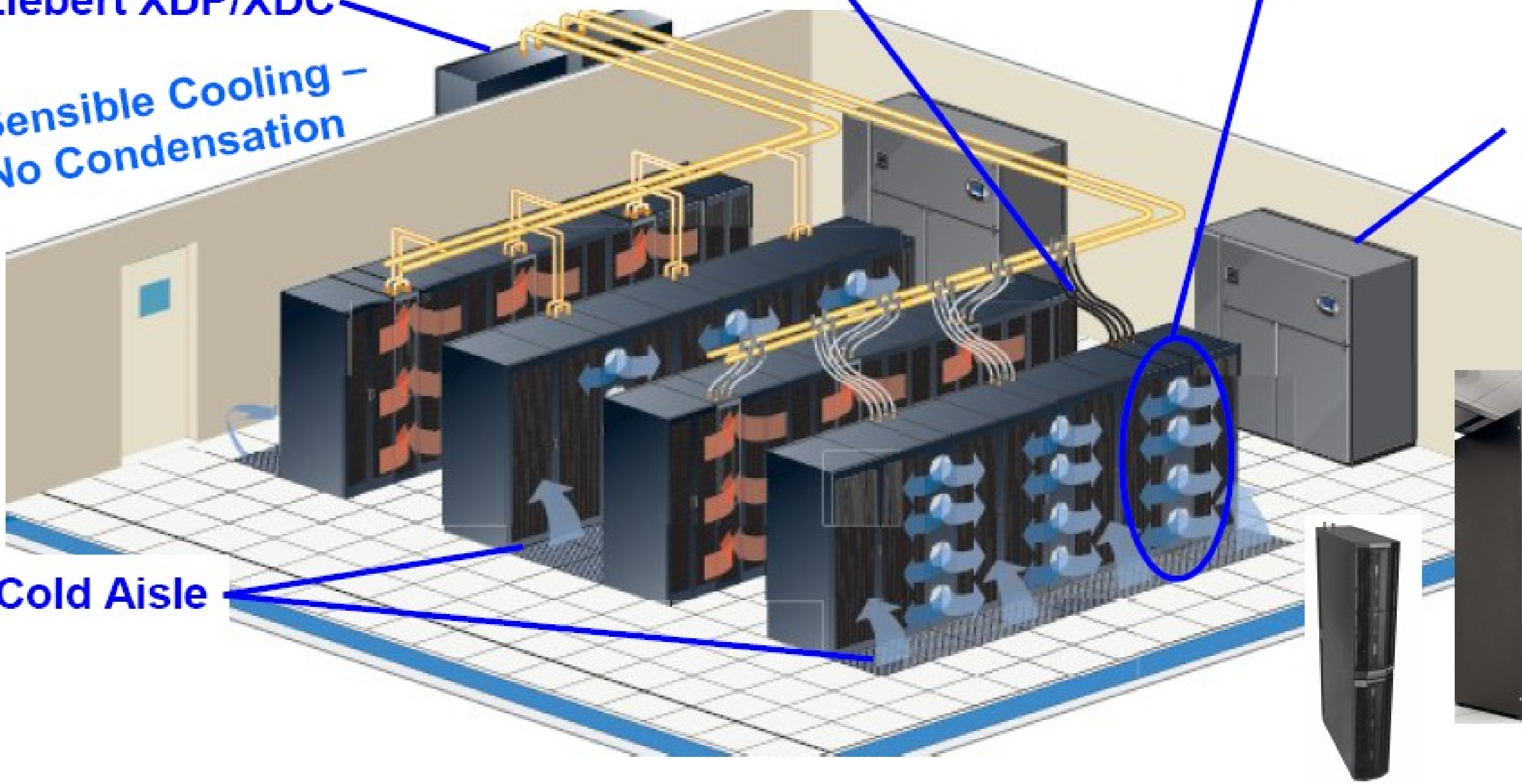
Double Stacked Liebert XDV

Liebert.

Sun Blade Applications

EMERSON. Network Power

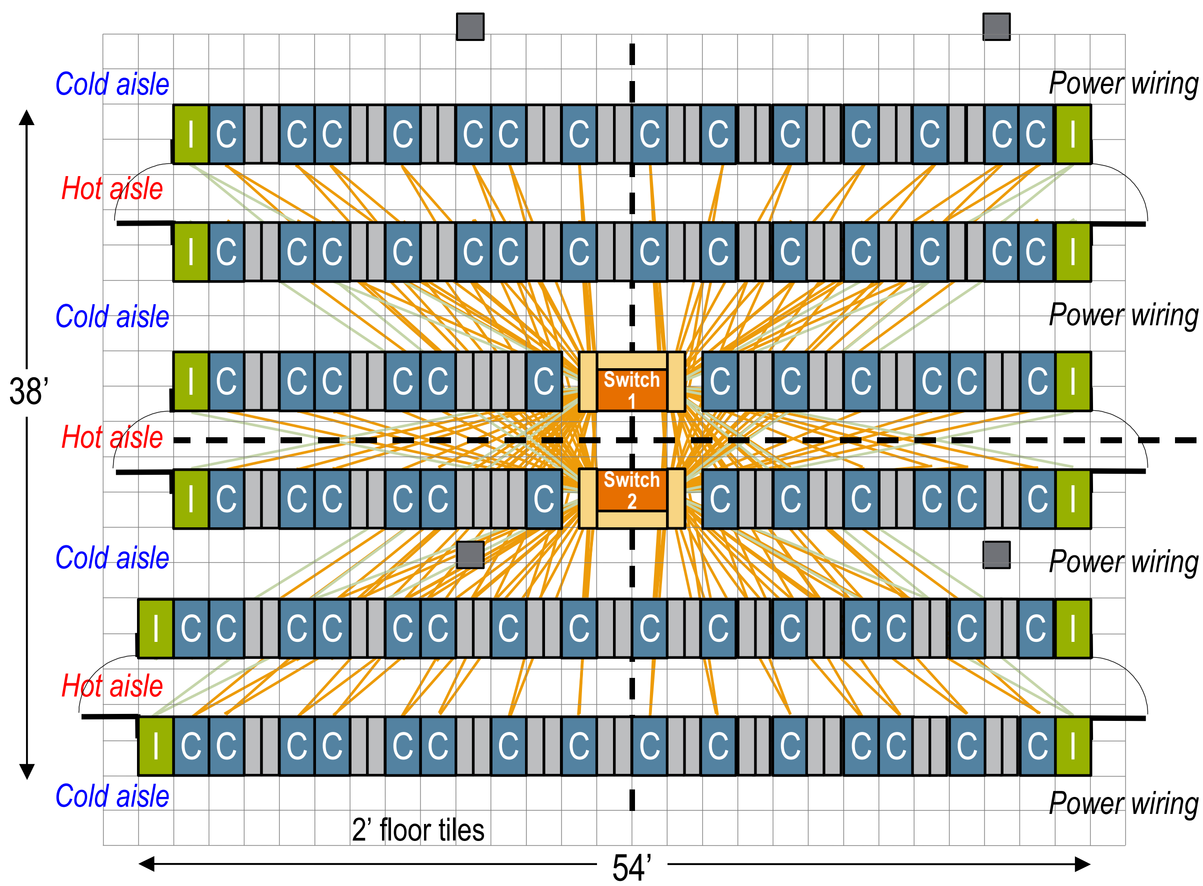
Liebert XDH



Liebert XDH

NSF TeriGrid - TACC Floorplan

Size: approximately half a basketball court



Switch — 2 Magnum switches, 16 line cards each (2,304 4x IB ports each)

C — 82 blade compute racks (3,936 4S blades)

I — 12 IO racks (25 X4600 4 RU, 72 X4500 4 RU)

112 APC Row coolers

1,312 12x cables (16 per rack) ***16 km total length***

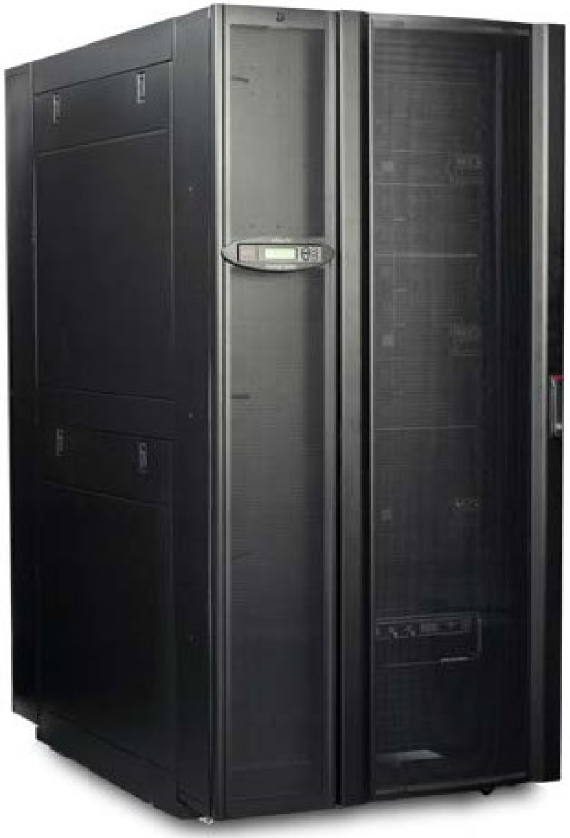
72 splitter cables, 6 per IO rack

12x cable lengths: 171 9m, 679 11m, 406 13m, 56 15m

Splitter cable lengths: 54 14m, 18 16m

Extreme Datacenter – Hot Air Containment

Rack Air Containment (RACS)



Hot Aisle Containment (HACS)

a company of Schneider Electric

BLADE READY

InfraStruXure
DATA CENTERS ON DEMAND

APC | MGE
Critical Power and Cooling Services

© a company of Schneider Electric

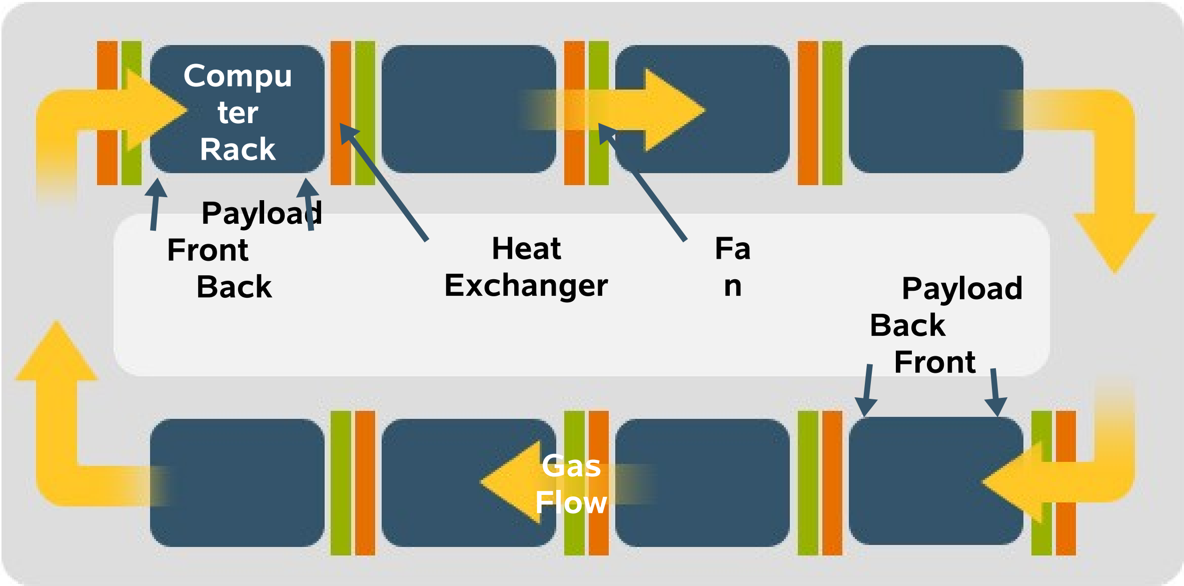
Sun microsystems

The World’s First Containment Datacenter

Module Data Center

Cooling

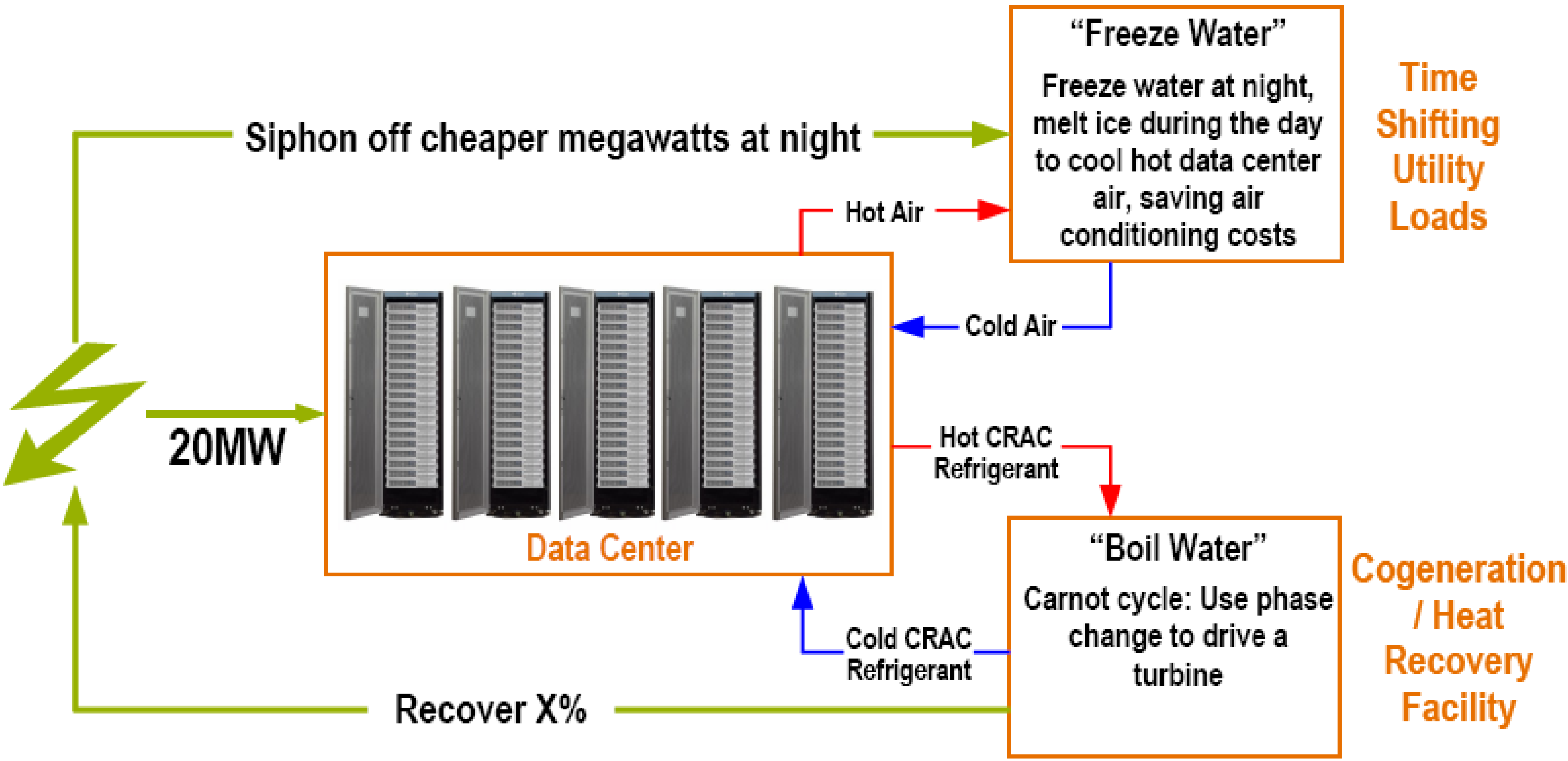
- Air flows in circular path with fans and heat exchanger per rack
  > Payload installed front to back
- Chiller size depends on the payload, 60-ton chiller for max 200kW load



Steam and Ice

"Today a data center just looks like a giant resistor in a multimegawatt circuit. It would be nice if it also was a capacitor."

– Mark Bramfitt, PG&E



Earth Pipes

Data Centers Underground

Disjointed Progress in the Data Center

"Facilities is from Mars, IT is from Venus"

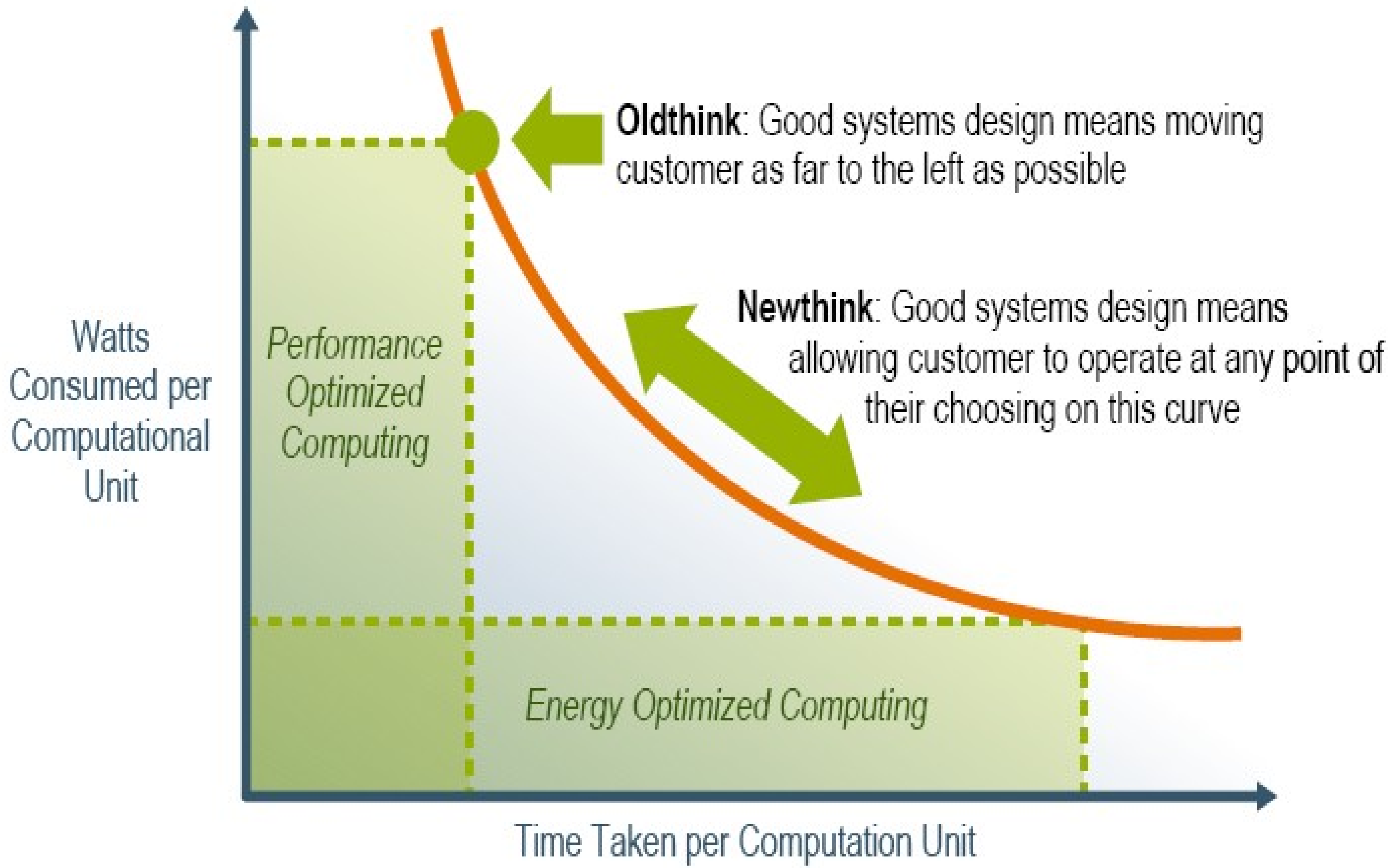
What Facilities is Doing

- Hot aisle containment
- Cold aisle containment
- Concrete slab floor
- Variable frequency drives
- Air side economizers

What IT is Doing

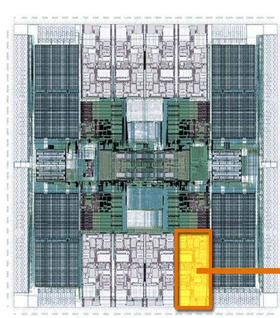
- Server refresh
- Consolidation
- Virtualization
- Utilization Management

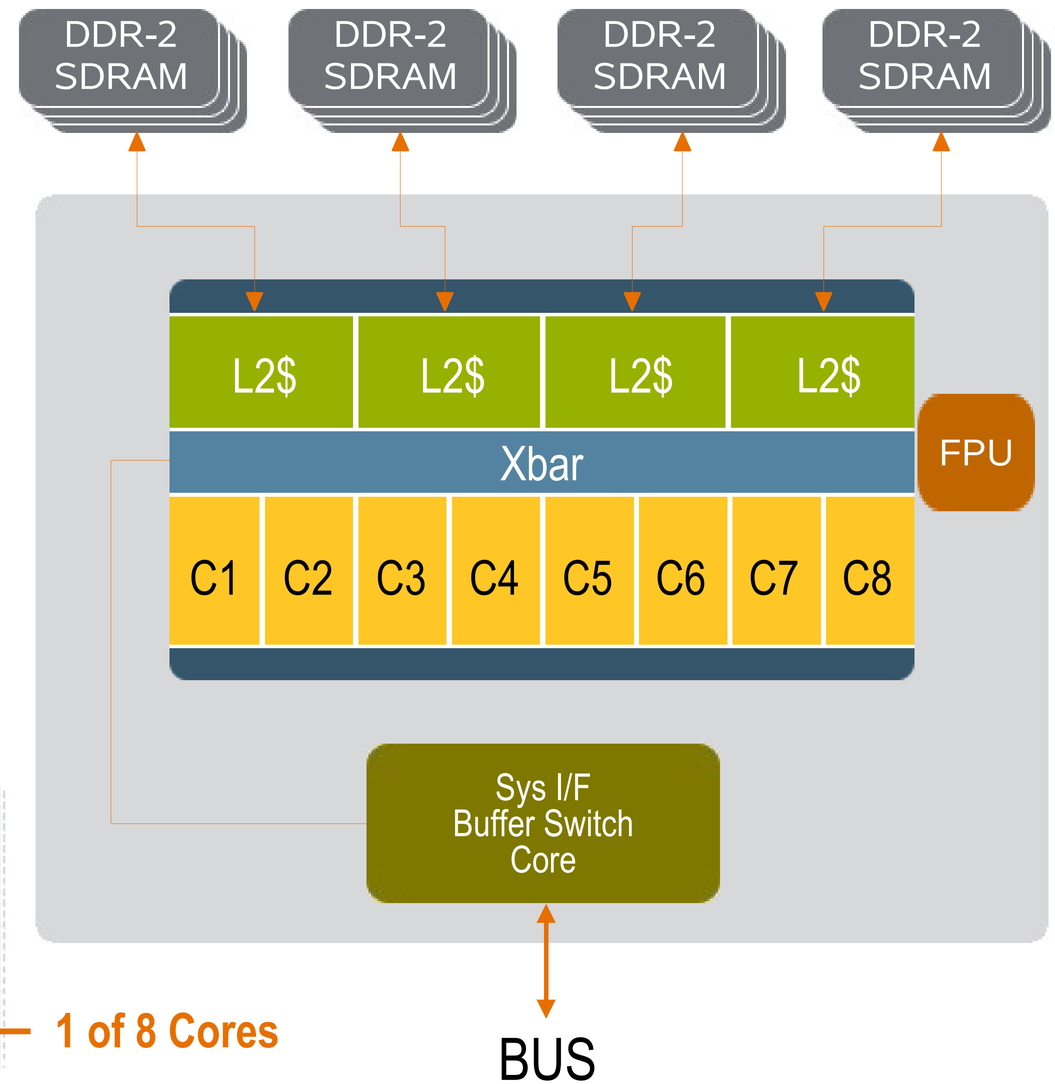
Newthinking in Systems/Grid Design



Introducing Niagara

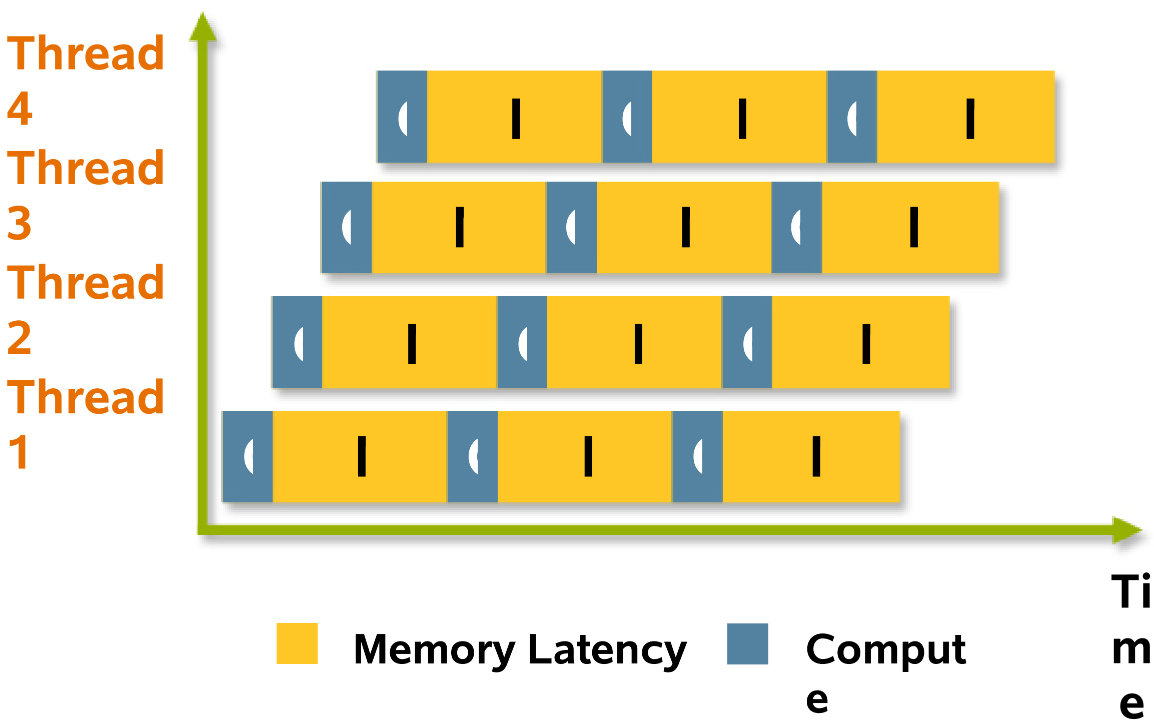
- SPARC V9 implementation
- Up to eight 4-way multi-threaded cores for up to 32 simultaneous threads
- All cores connected through a 90GB/sec crossbar switch
- High-bandwidth 12-way associative 3MB Level-2 cache on chip
- 4 DDR2 channels (23GB/s)
- Power : < 70W !
- ~300M transistors
- 378 sq. mm die


1 of 8 Cores

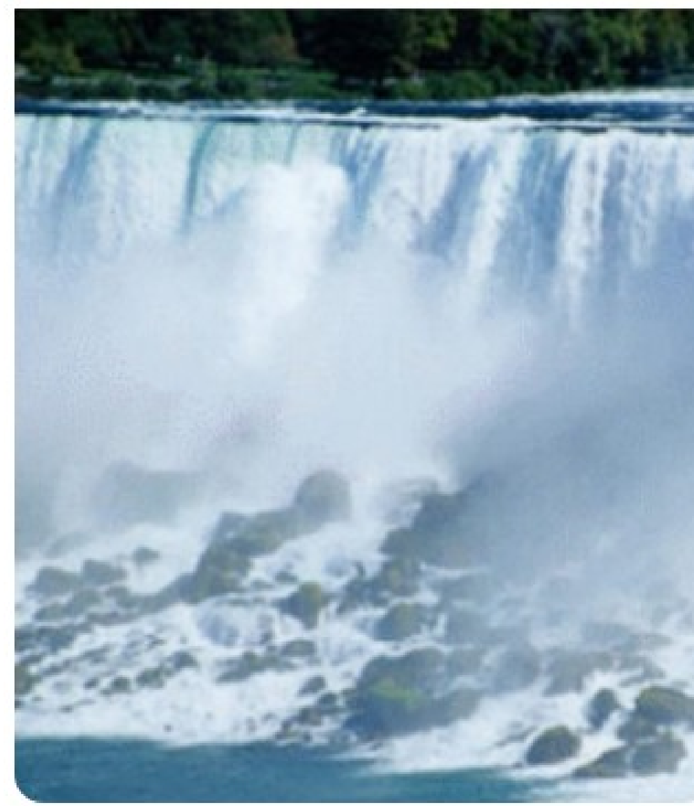


The Power of CMT

Niagara Processor Utilization: Up to 85%



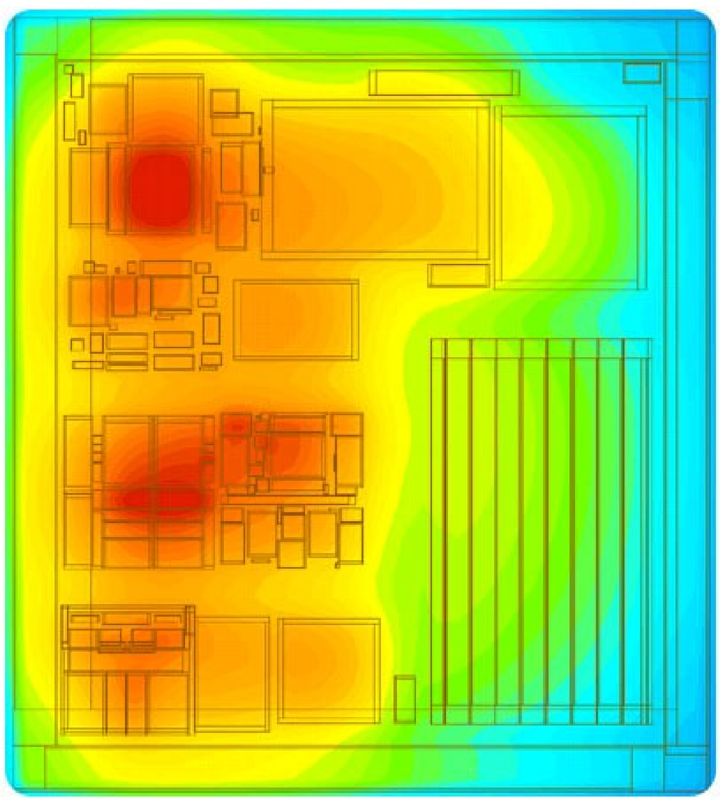
Chip Multi-threaded (CMT) Performance



Niagara's Power Advantage

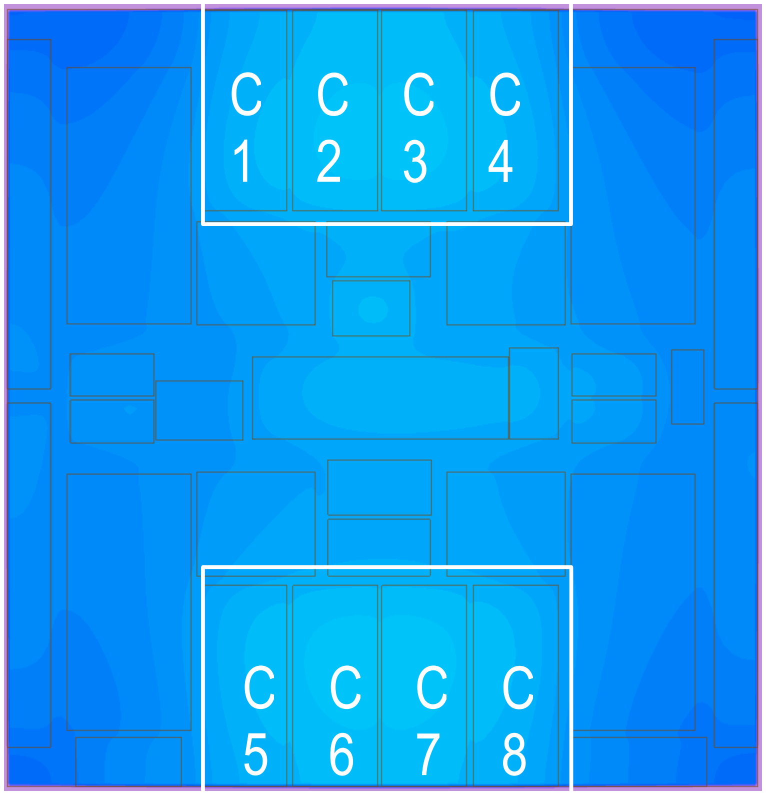
"Cool Threads" Dramatically Reduce Power Consumption

Uses a Fraction of the Power/Thread vs. Xeon



Single-core Processor

(Size Not to Scale)



CMT Processor

An integer linear programming (ILP) based static scheduling method that minimizes both thermal hot spots and temperature gradients to increase MPSoC reliability

TABLE III
ILP FORMULATION FOR MIN-TH&SP

Minimize $H + G$;

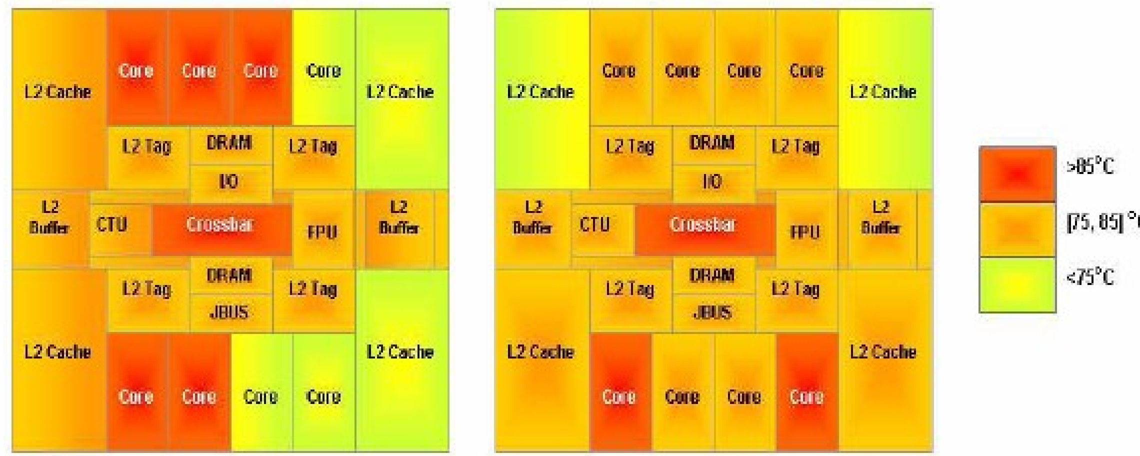
$H = max\{Q_p; p = 1...m$, for a system of m cores$\}$ where:

$$Q_p = \sum_{T_i \in T} \{x_{ip} \sum_{v_k} (y_{ik} q_{ik})\}$$

$$G = \sum_{p,r \in PU, p \neq r} \{n_{pr}\{ \sum_{i,j \in T, i \neq j} x_{ip} x_{jr} [p_{ij} d_{ij} (\tau_i - s_j) + p_{ji} d_{ji} (\tau_j - s_i)]\}\}$$

Subject to constraints:

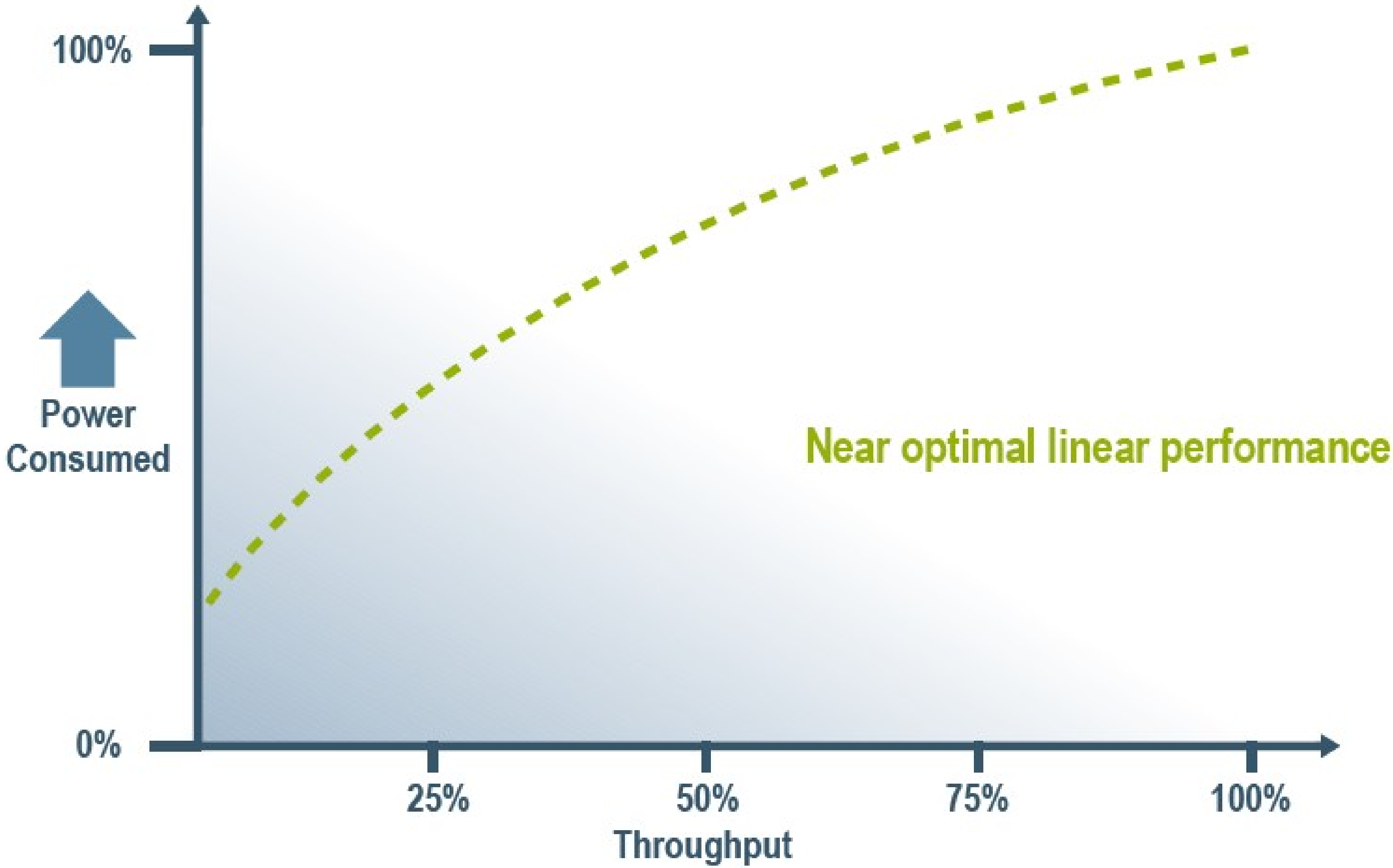
(a) $\forall T_i : \sum_p x_{ip} = 1$	Each task is assigned to only one PU
(b) $\forall T_i : \sum_k y_{ik} = 1$	Each task runs at only one V/f level
(c) $\tau_i = s_i + t_i$	Execution finish time for T_i
(d) $s_i \geq max_{E_{ji} \in E}\{\tau_j\}$	Task precedence
(e) $\tau_i \leq D_i$	Deadlines for all sink nodes
(f) $s_i \geq \tau_j;\ if\ p_{ji} = 1$	Precedence for tasks on the same core
(g) $p_{ij} + p_{ji} = 1$; if $x_{ip} = x_{jp} = 1$	If T_i and T_j are scheduled on the same core, either T_i precedes T_j, or vice versa



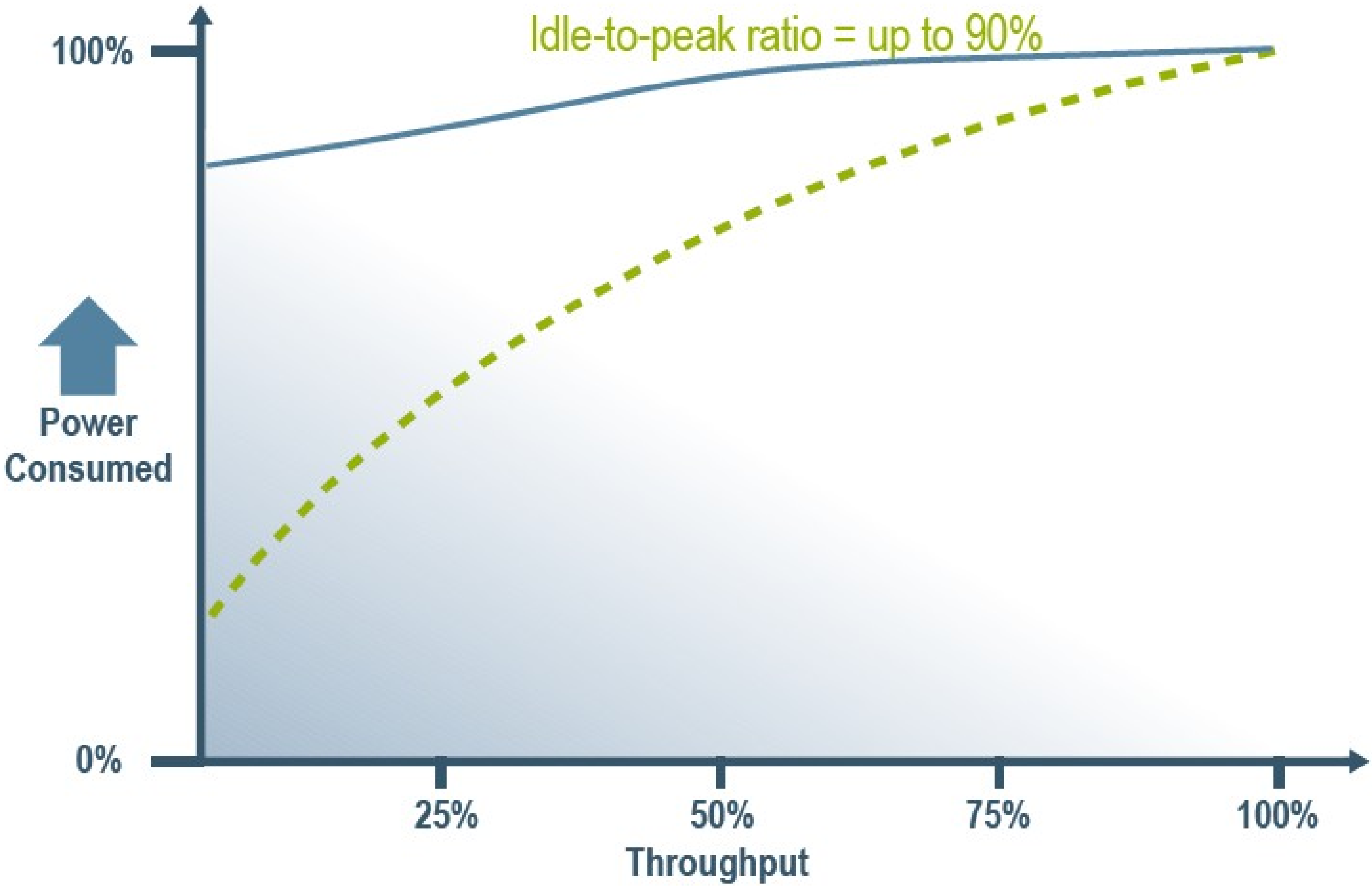
. Thermal maps: (a)DLB; (b)Adaptive-Random

Source : Ayse Coskun, Tajana Simuni ´c Rosing, Keith A. Whisnant, and Kenny C. Gross, Static and Dynamic Temperature-Aware Scheduling for Multiprocessor SoCs

Desired Power/Performance Curve

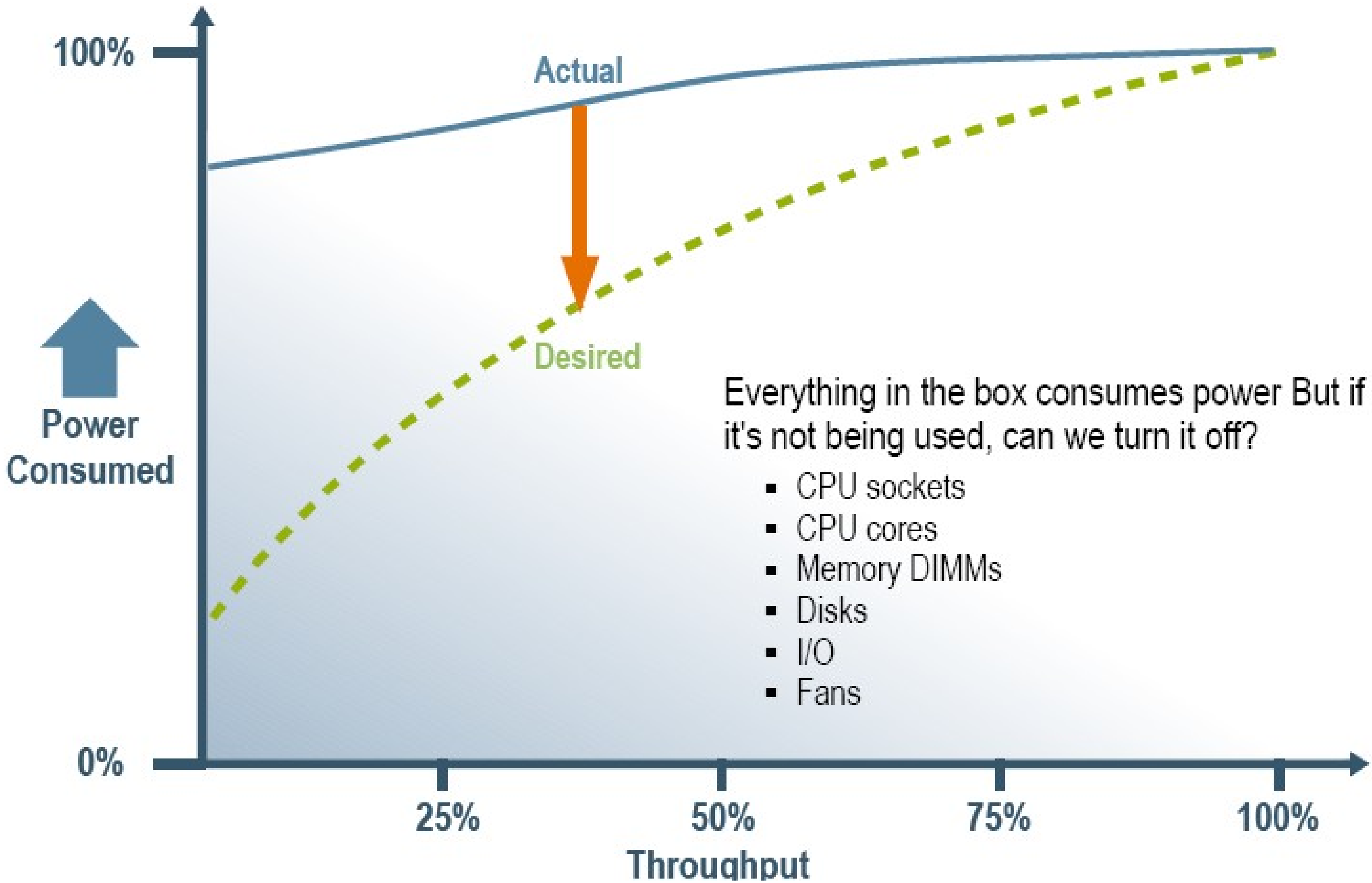


The Reality



Idle-to-peak ratio = up to 90%
100%
0%
Power
Consumed
25%
50%
75%
100%
Throughput

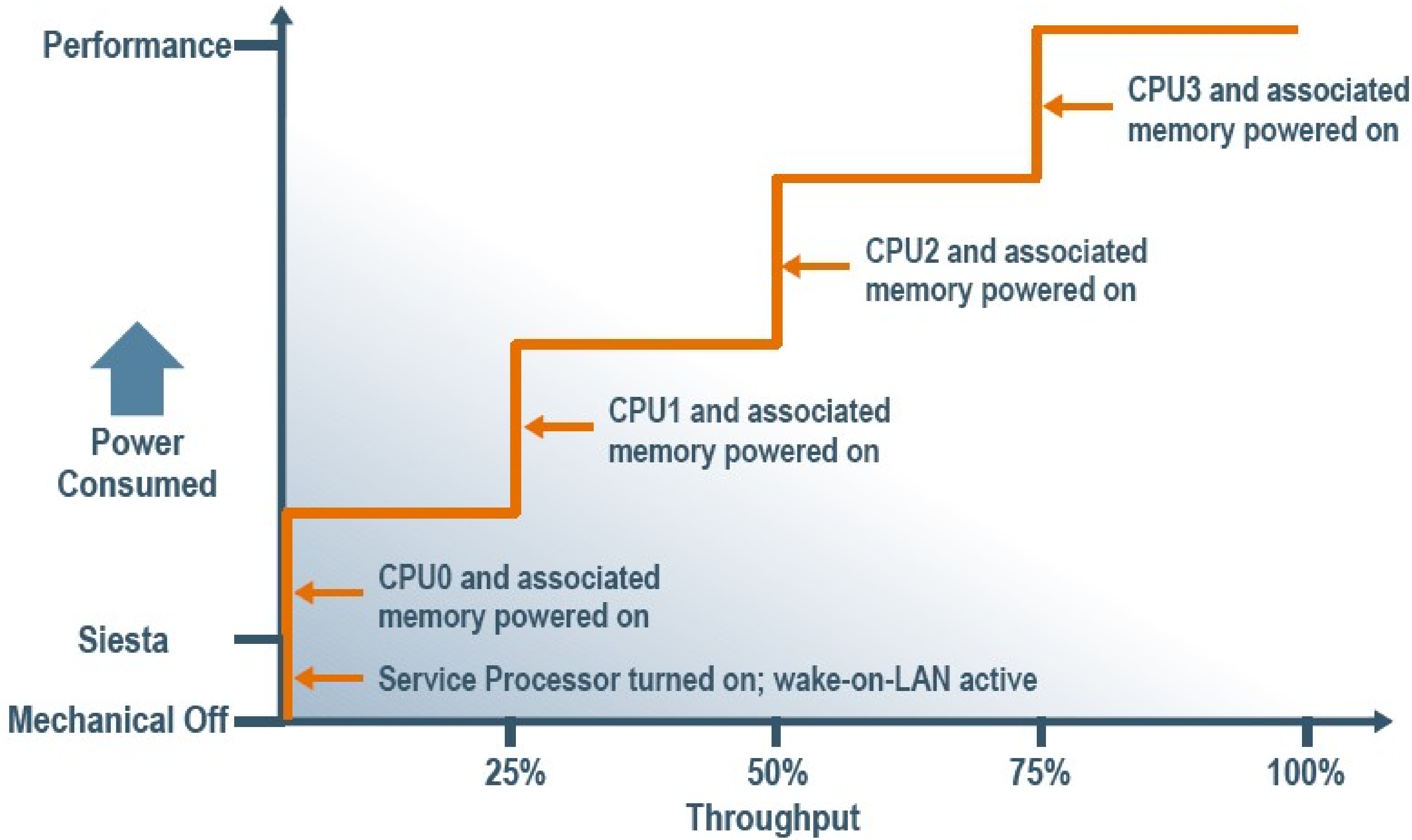
The Big Turn-Off



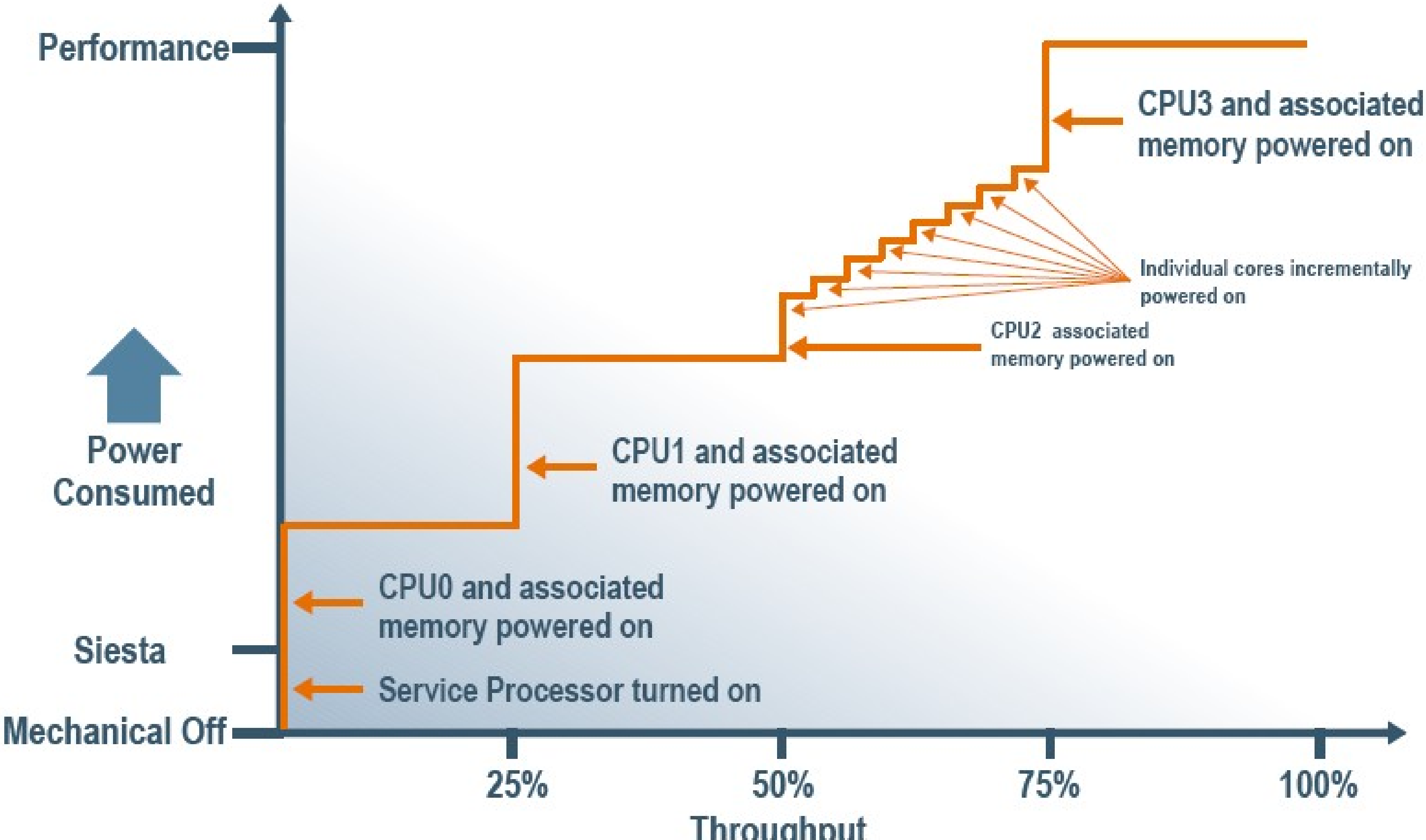
Everything in the box consumes power But if it's not being used, can we turn it off?

- CPU sockets
- CPU cores
- Memory DIMMs
- Disks
- I/O
- Fans

Desired Behavior of Future 4 Socket Servers

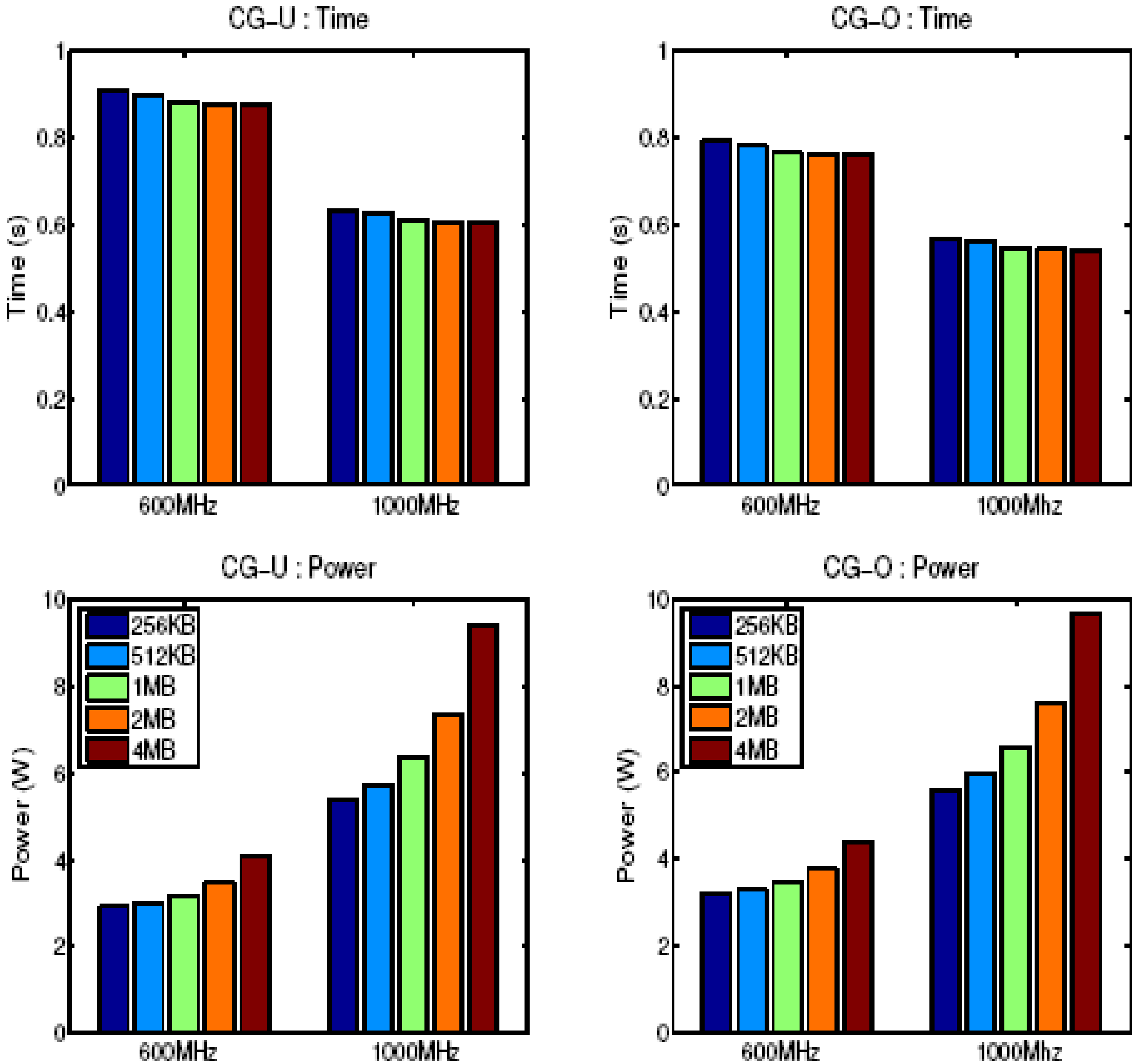


Getting to Greater Linearity

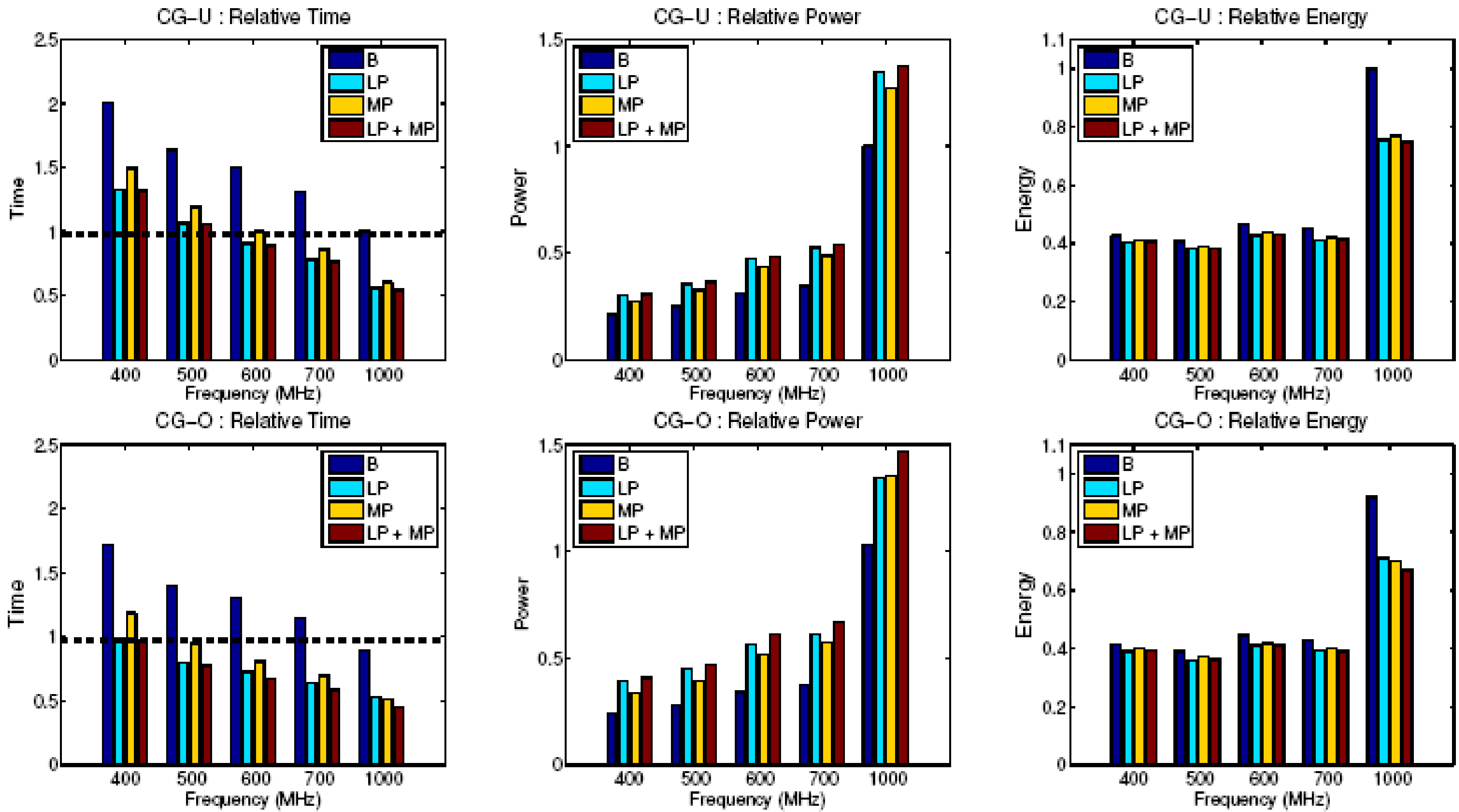


Algorithm Profile

Congugate Gradient Sparce Solver



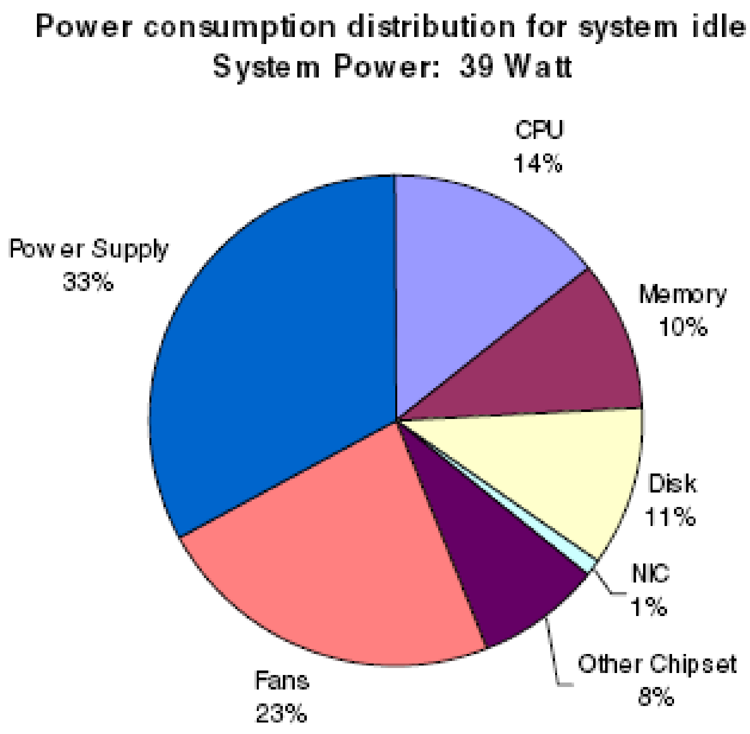
Source : Korad Malkowski, Ingyu Lee, Padma Raghavan, Mary Jane Irwin, Conjugate Gradient Sparse Solvers: Performance-Power Characteristics

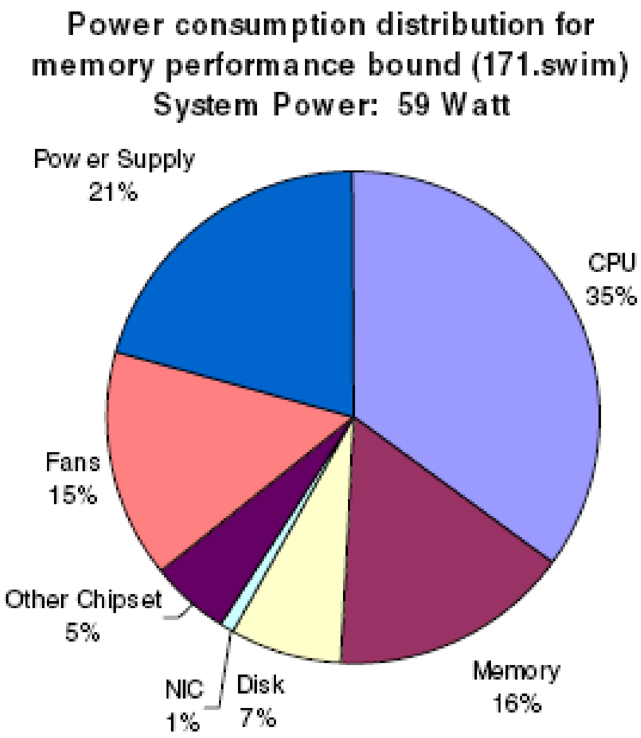


Source : Korad Malkowski, Ingyu Lee, Padma Raghavan, Mary Jane Irwin, Conjugate Gradient Sparse Solvers: Performance-Power Characteristics

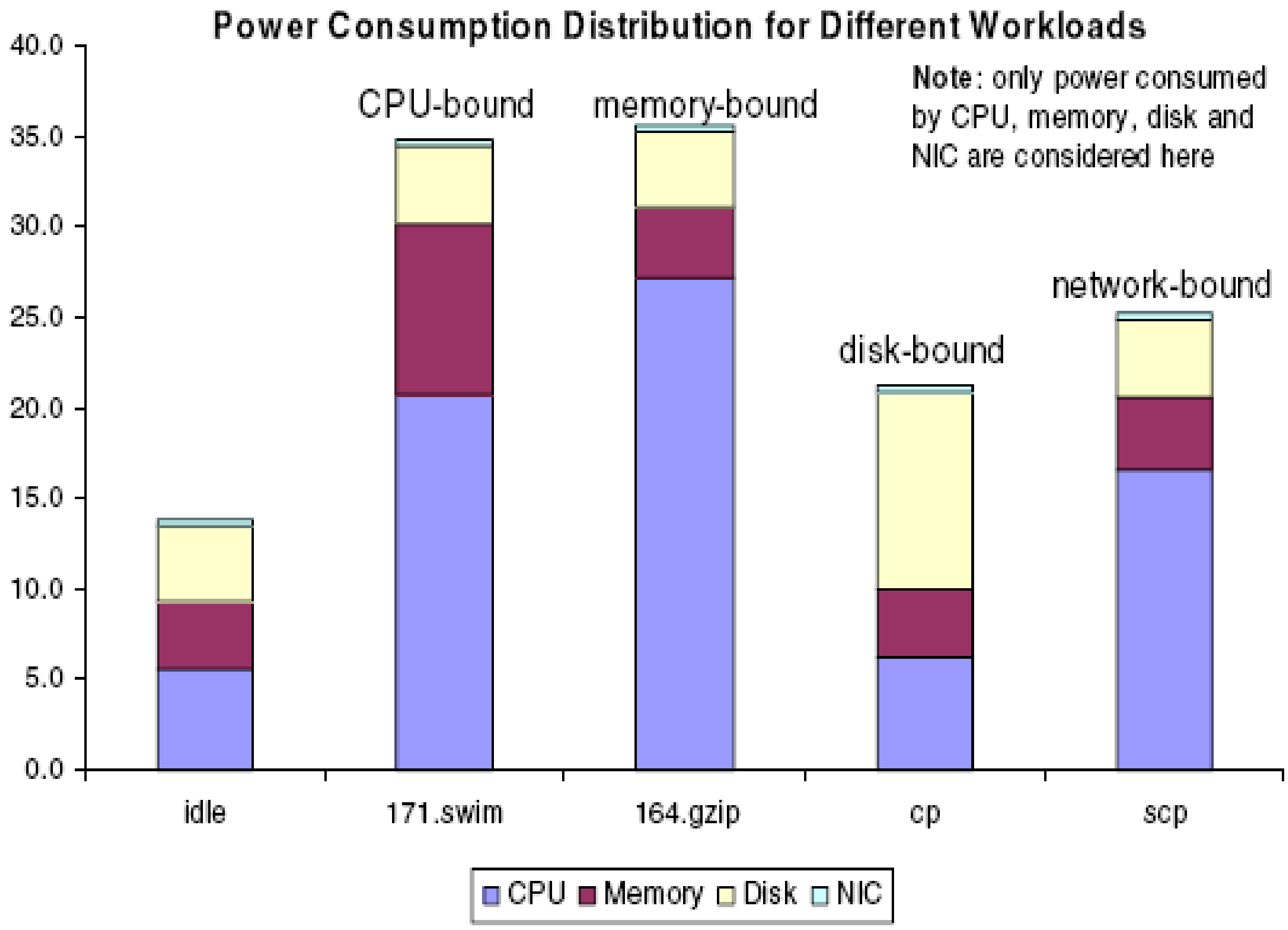
Application Power Profiling

Application Power Profile



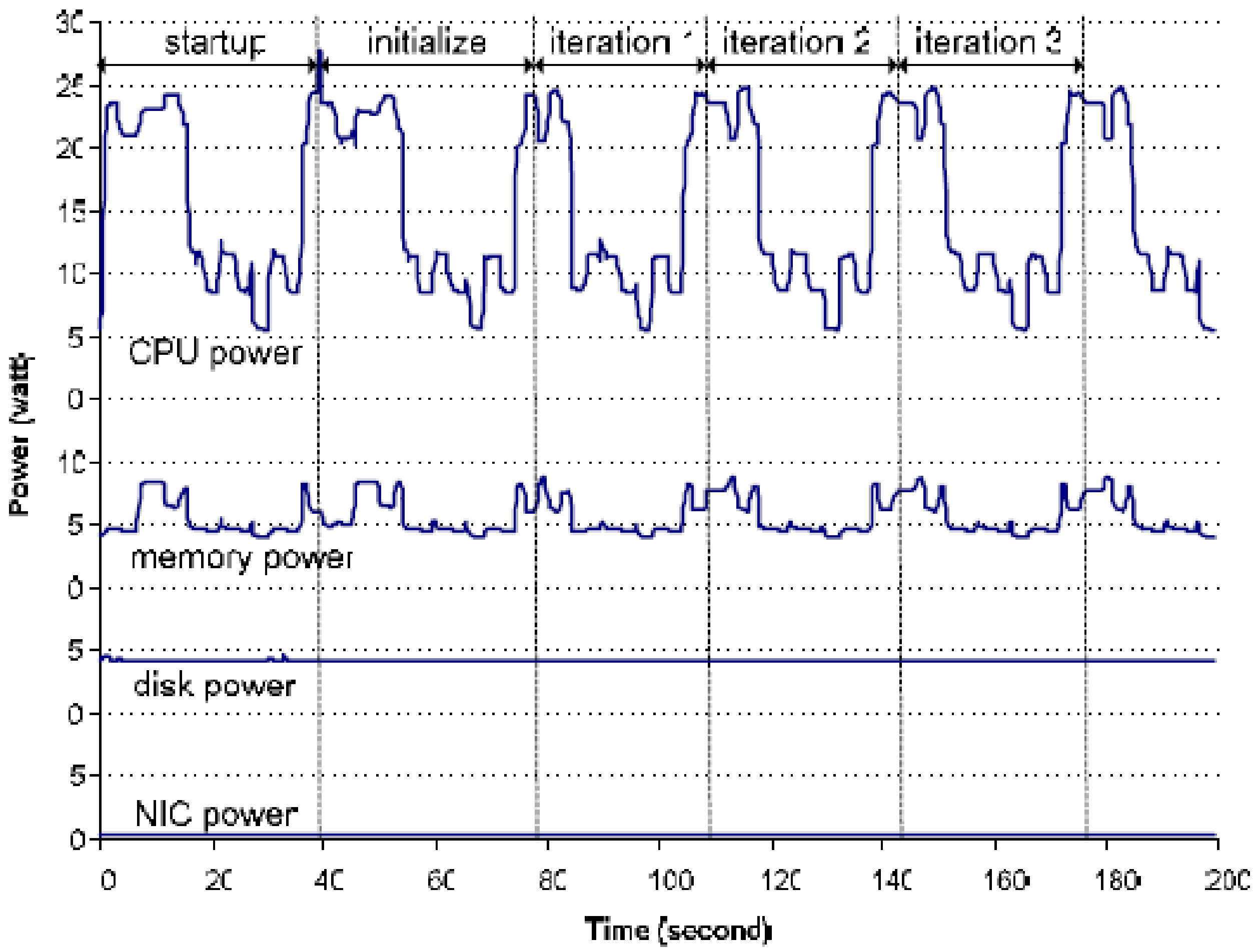


Source : Xizhou Feng, Rong Ge, Kirk W. Cameron, University of South Carolina, Columbia, SC 29208, Power and Energy Profiling of Scientific Applications on Distributed Systems

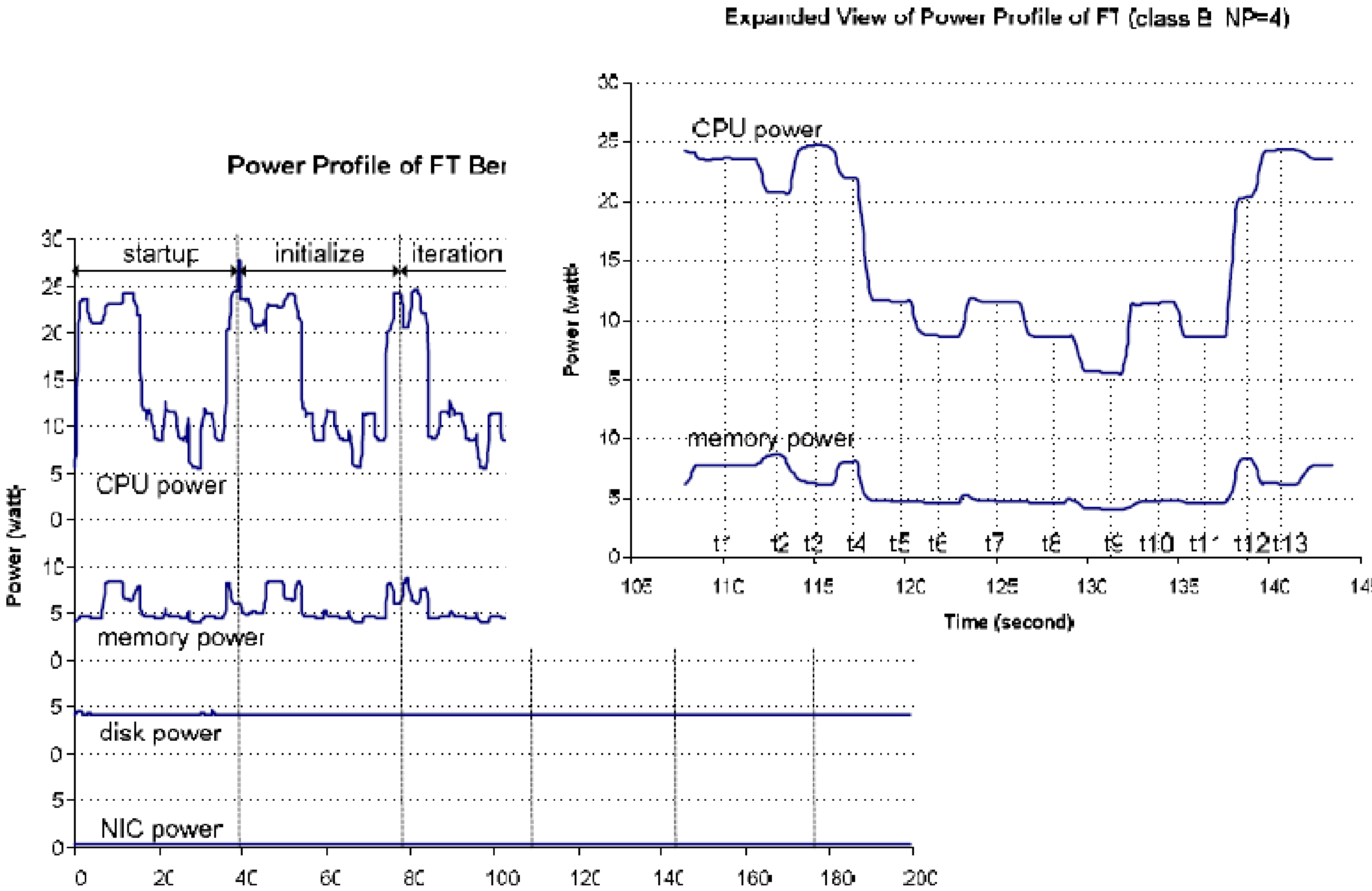


Source : Xizhou Feng, Rong Ge, Kirk W. Cameron, University of South Carolina, Columbia, SC 29208, Power and Energy Profiling of Scientific Applications on Distributed Systems

Power Profile of FT Benchmark (class B NP=4)

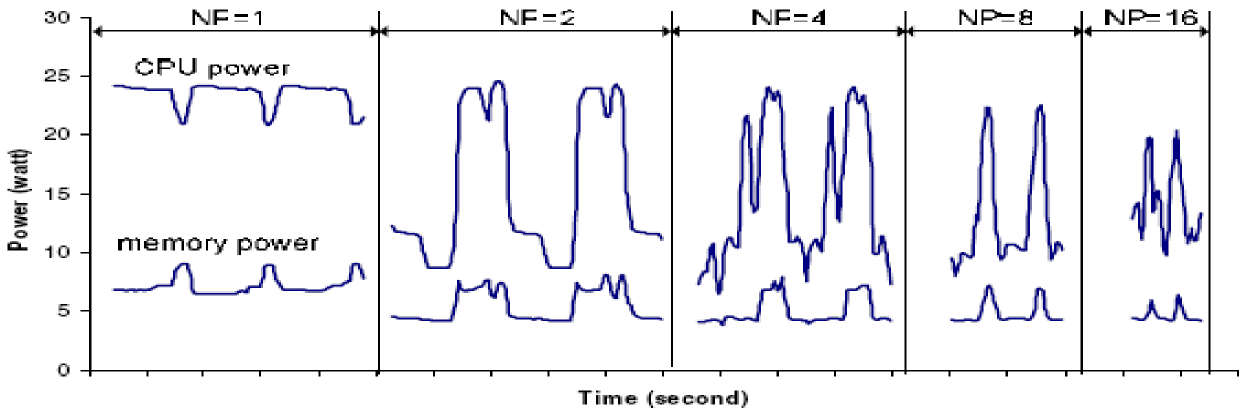


Source : Xizhou Feng, Rong Ge, Kirk W. Cameron, University of South Carolina, Columbia, SC 29208, Power and Energy Profiling of Scientific Applications on Distributed Systems

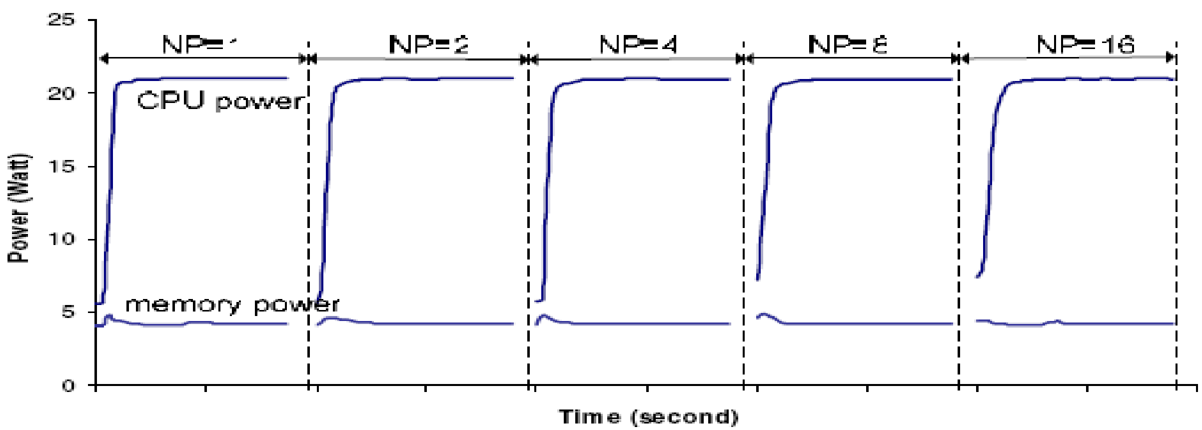


Source : Xizhou Feng, Rong Ge, Kirk W. Cameron, University of South Carolina, Columbia, SC 29208, Power and Energy Profiling of Scientific Applications on Distributed Systems

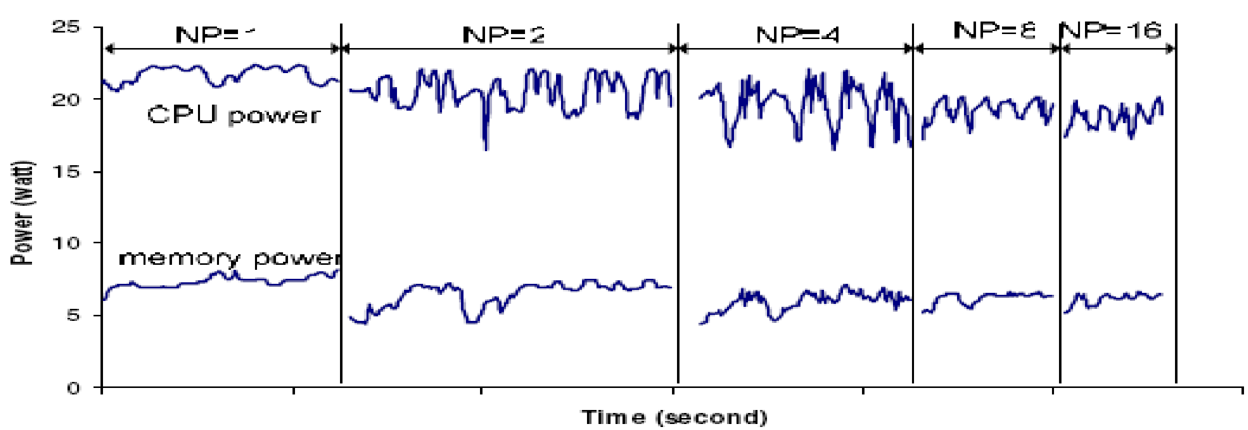
Power Profile of FT Benchmark (class A) with Different Number of Nodes



Power Profile of EP (class A) with Differentnumber of Nodes

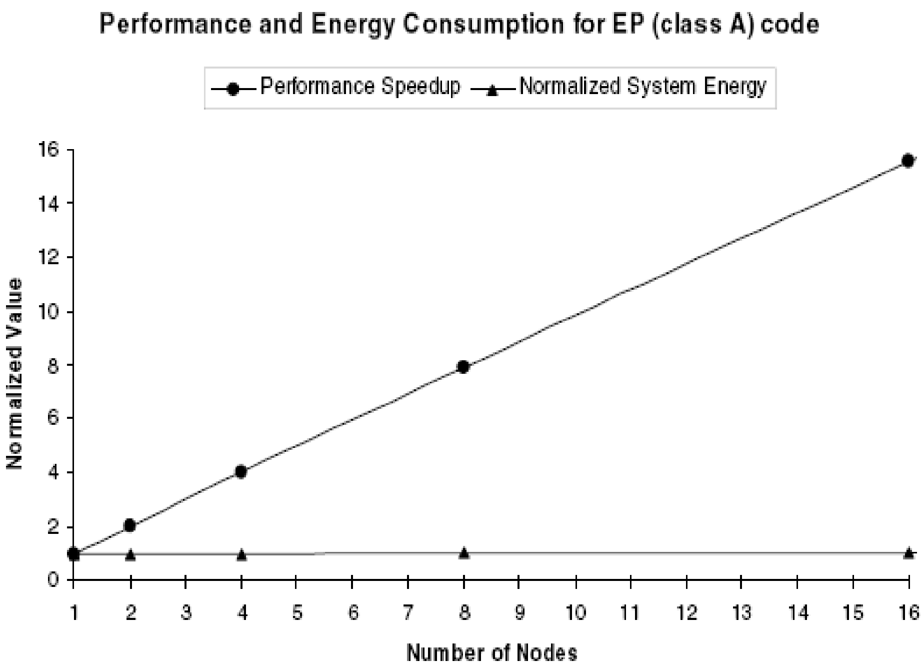


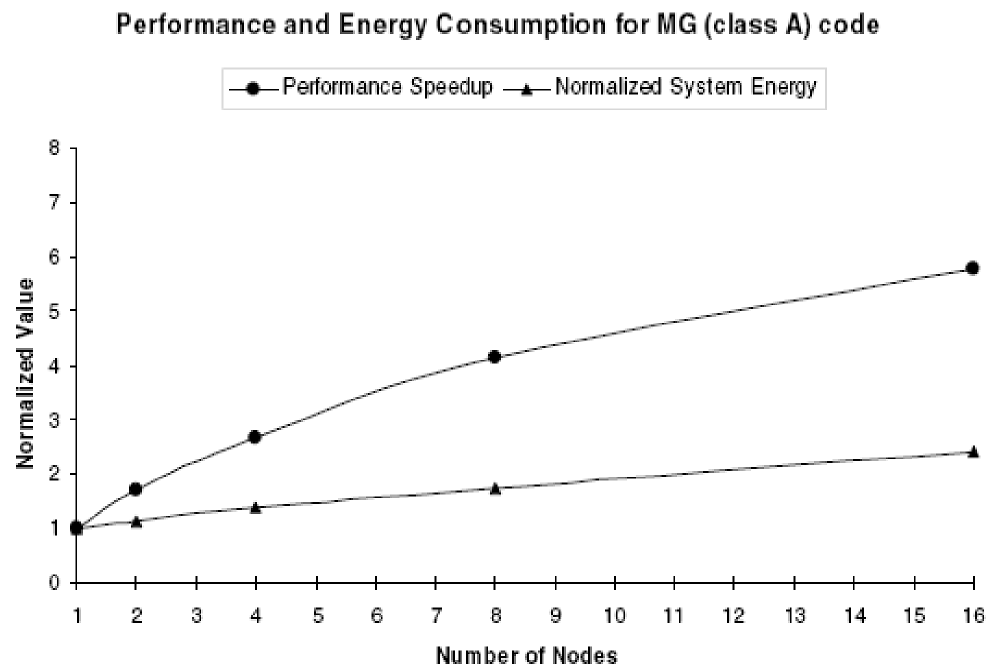
Power Profile of MG (Class A) with Different Number of Nodes

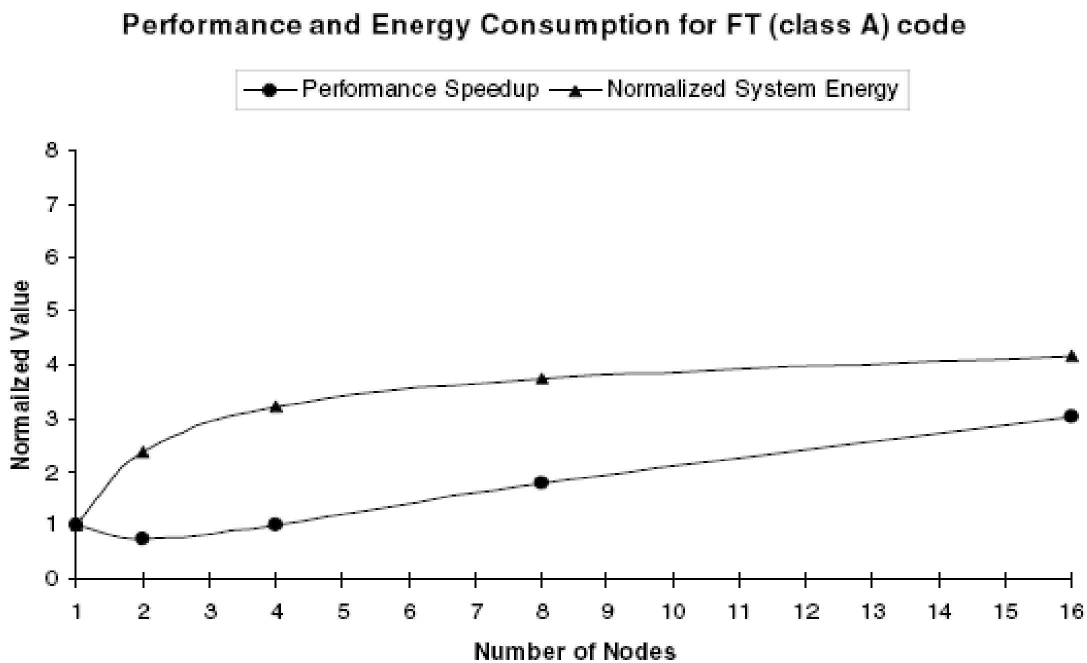


Observation of NPB

- CPU power consumption decreases as memory power consumption goes up;
- Both CPU power and memory power decrease with message communication among different nodes;
- For most parallel codes (except EP), the average power consumption goes down as the number of nodes increases;
- Communication distance and message size affects the power profile pattern (for example, LU has short and shallow power consumption in contrast with FT).







Source : Xizhou Feng, Rong Ge, Kirk W. Cameron, University of South Carolina, Columbia, SC 29208, Power and Energy Profiling of Scientific Applications on Distributed Systems

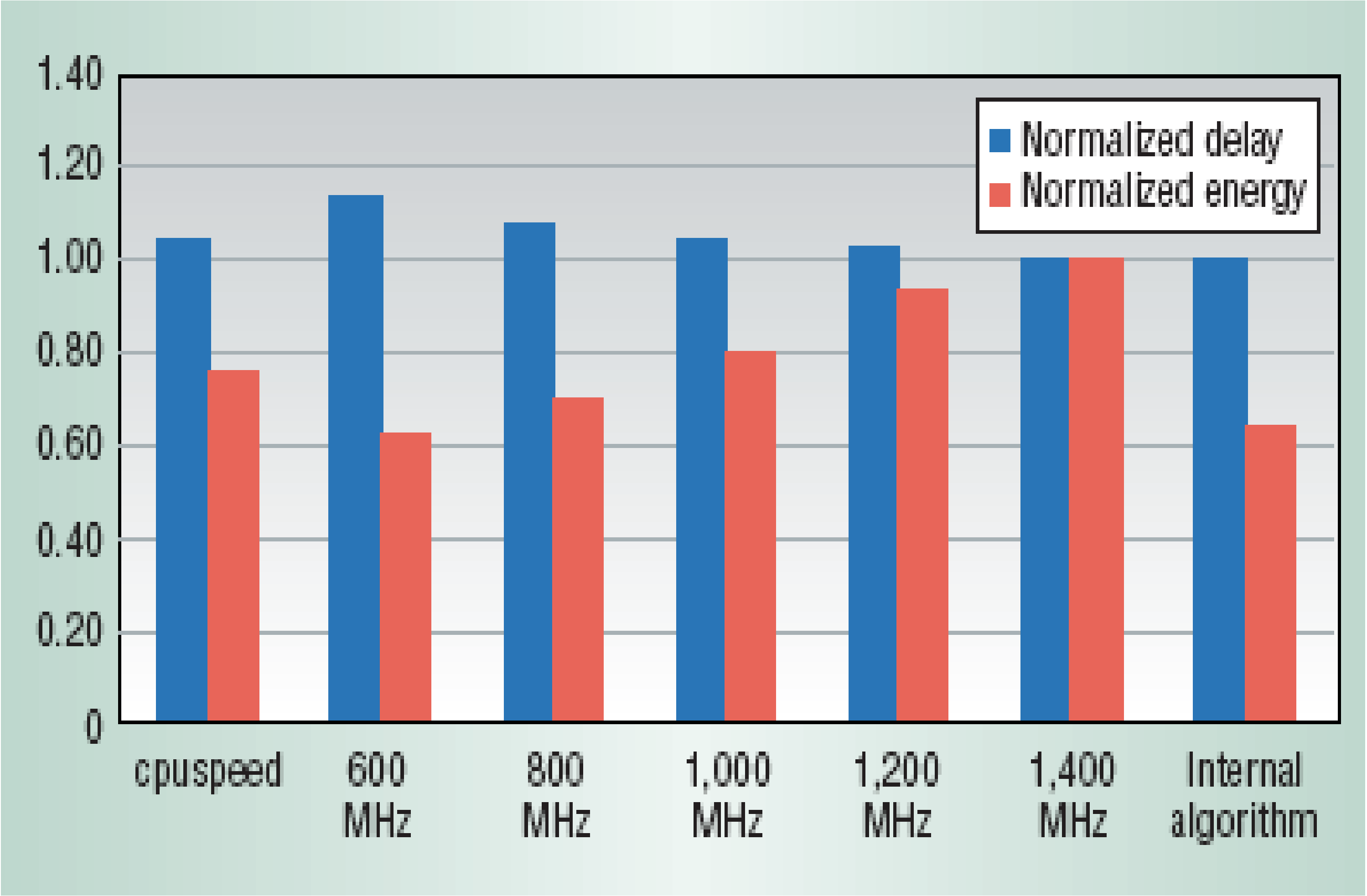
AMD Opteron 2218 : DVFS

Frequency (MHz)	Voltage (V)
1000	1.10
1800	1.15
2000	1.15
2200	1.20
2400	1.25
2600	1.30

Code	Frequency (MHz)					
	1000	1800	2000	2200	2400	2600
BT.C.16	1.66	1.17	1.08	1.07	1.05	1.00
	1.06	0.88	0.84	0.90	0.96	1.00
CG.C.16	1.47	1.15	1.11	1.07	1.03	1.00
	0.98	0.88	0.88	0.91	0.94	1.00
EP.C.16	2.57	1.45	1.30	1.18	1.08	1.00
	1.57	1.07	1.00	0.98	0.98	1.00
FT.C.16	1.40	1.10	1.06	1.04	1.02	1.00
	0.92	0.84	0.83	0.88	0.94	1.00
IS.C.16	1.52	1.07	0.99	1.01	1.01	1.00
	1.01	0.82	0.79	0.85	0.93	1.00
LU.C.16	1.62	1.13	1.05	1.02	1.06	1.00
	1.03	0.86	0.83	0.86	0.96	1.00
MG.C.16	1.41	1.11	1.03	1.05	0.99	1.00
	0.92	0.84	0.81	0.87	0.90	0.98
SP.C.16	1.53	1.08	1.03	1.02	1.05	1.00
	1.00	0.84	0.81	0.87	0.96	1.00

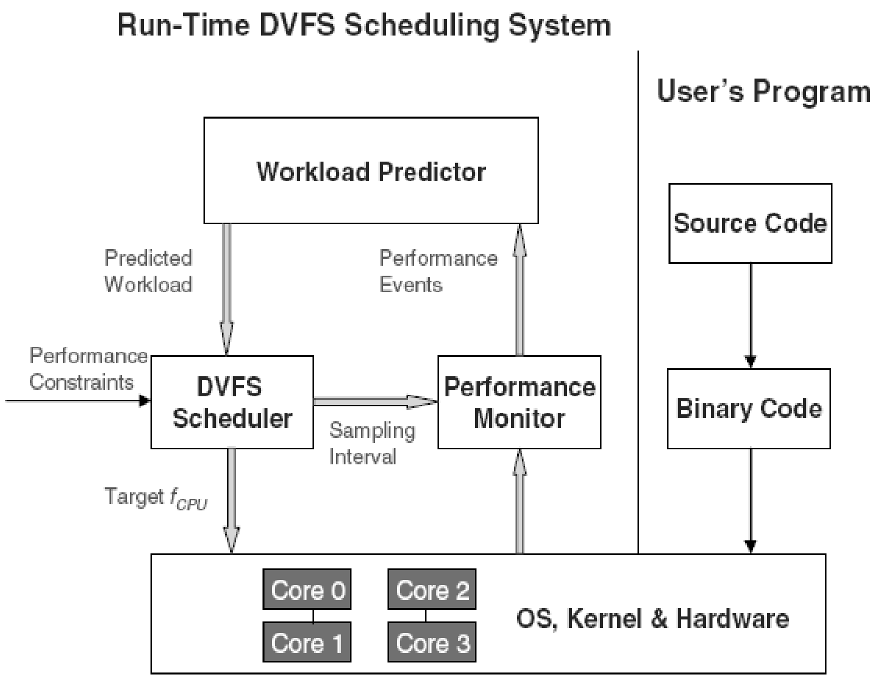
Source : Xizhou Feng, Rong Ge, Kirk W. Cameron, University of South Carolina, Columbia, SC 29208, Power and Energy Profiling of Scientific Applications on Distributed Systems

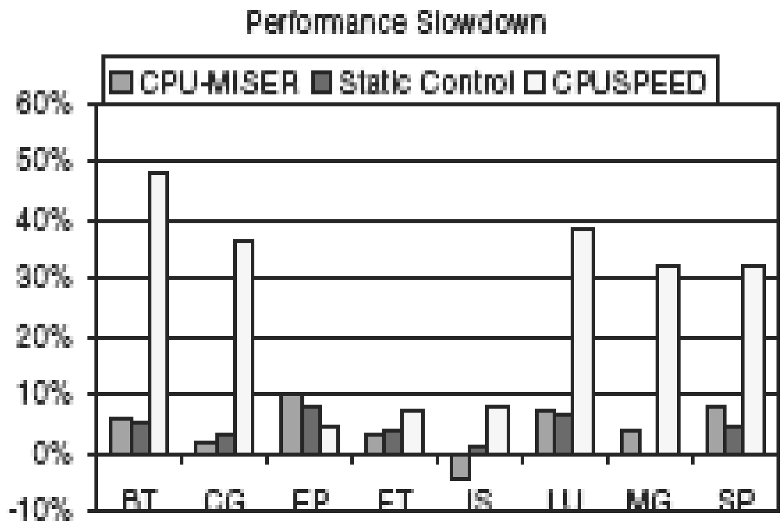
Intel 1.4 Ghz Pentium-M : FT Benchmark

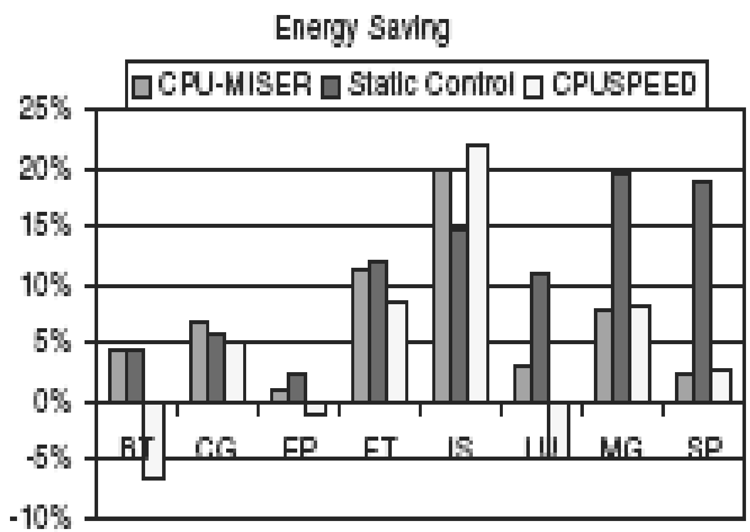


Source :K. Cameron, Rong Ge, Xizhou Feng High-Performance, Power-Aware Distributed Computing for Scientific Applications

CPU Miser


Run-Time DVFS Scheduling System
User's Program
Workload Predictor
Predicted Workload
Performance Events
Performance Constraints
DVFS Scheduler
Performance Monitor
Sampling Interval
Target f_{CPU}
Source Code
Binary Code
Core 0
Core 1
Core 2
Core 3
OS, Kernel & Hardware


Performance Slowdown
CPU-MISER
Static Control
CPUSPEED
60%
50%
40%
30%
20%
10%
0%
-10%
BT
CG
EP
FT
IS
LU
MG
SP


Energy Saving
CPU-MISER
Static Control
CPUSPEED
25%
20%
15%
10%
5%
0%
-5%
-10%
BT
CG
EP
FT
IS
LU
MG
SP

Green Destiny

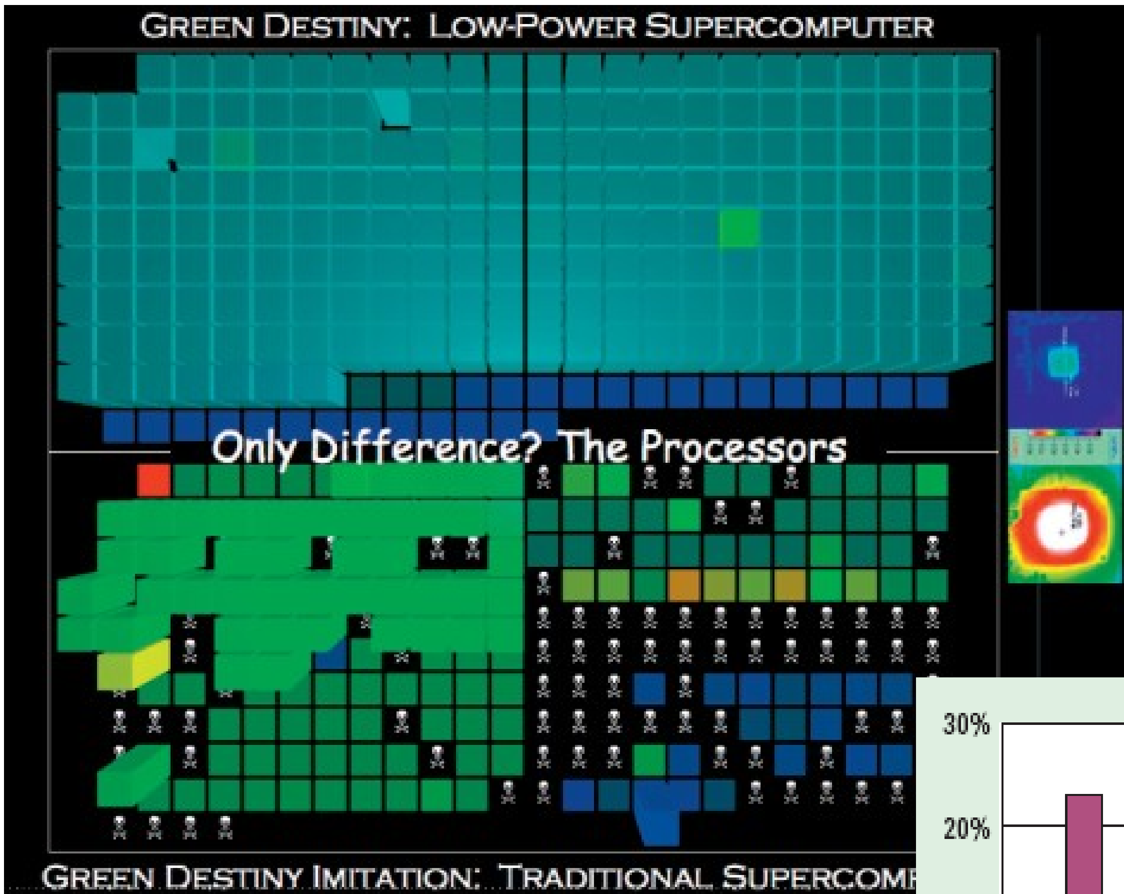
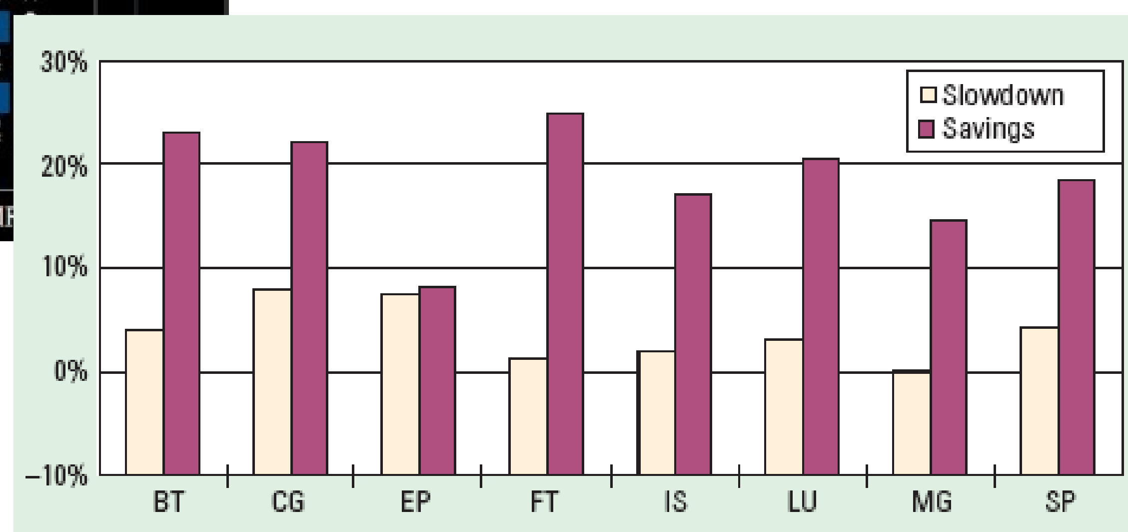


Table 1. Comparison of supercomputing systems on the Linpack benchmark. ASCI White's top performance is shown in italics; Green Destiny's appears in bold.

Performance Metric	ASCI White	Green Destiny
Year	2000	2002
Number of processors	8,192	240
Performance (Gflops)	*7,226*	101
Space (ft^2)	*9,920*	**5**
Power (kW)	*2,000*	**5**
DRAM (Gbytes)	*6,200*	150
Disk (Tbytes)	*160.0*	**4.8**
DRAM density (Mbytes/ft^2)	625	**30,000**
Disk density (Gbytes/ft^2)	16.1	**960.0**
Perf/space (Gflops/ft^2)	0.7	**20.2**
Perf/space (Gflops/kW)	4	**20**
Reliability (hours)	5.0 hours (2001), 40 hours (2003)	**No unscheduled downtime**

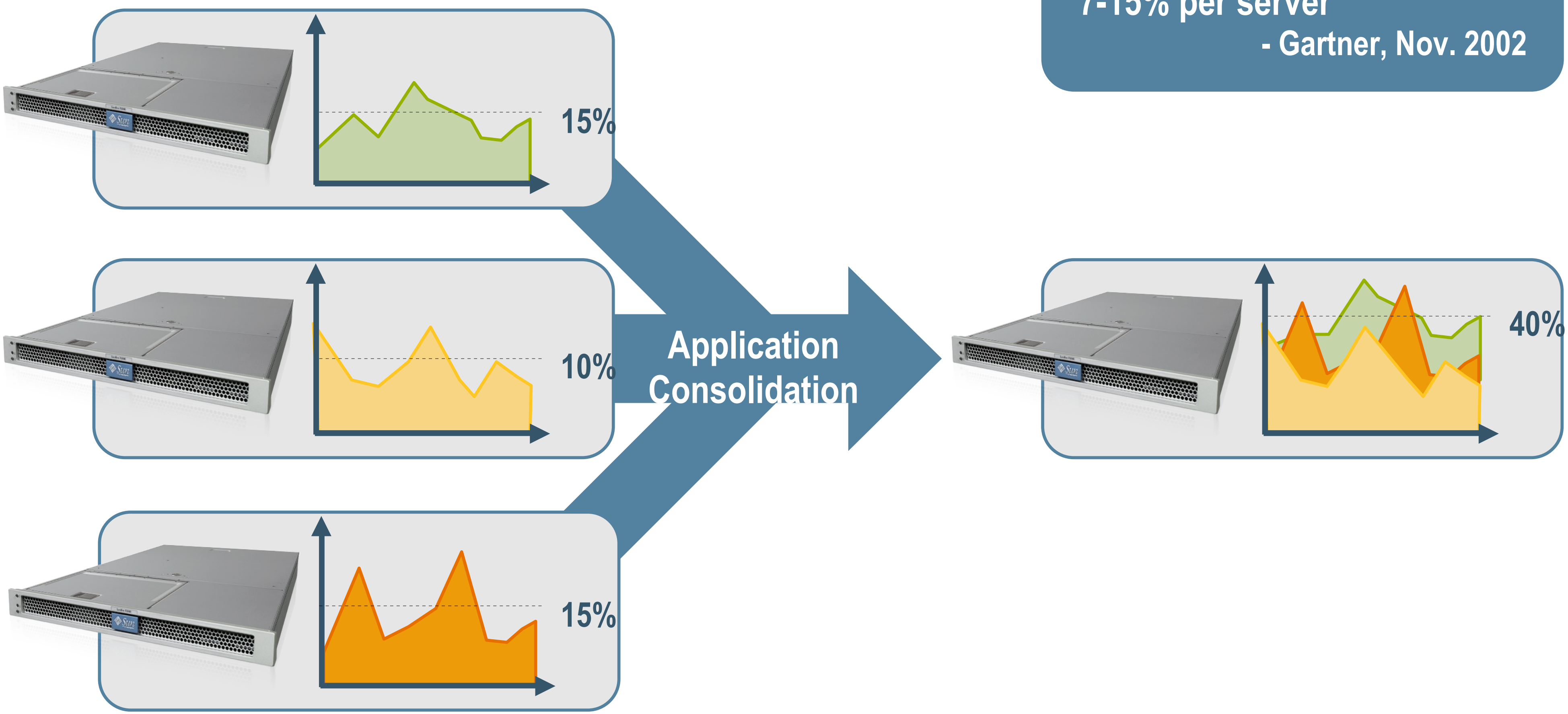


Source: Wu-chun Feng, Xizhou Feng, and Rong Ge, Green Computing Comes of Age

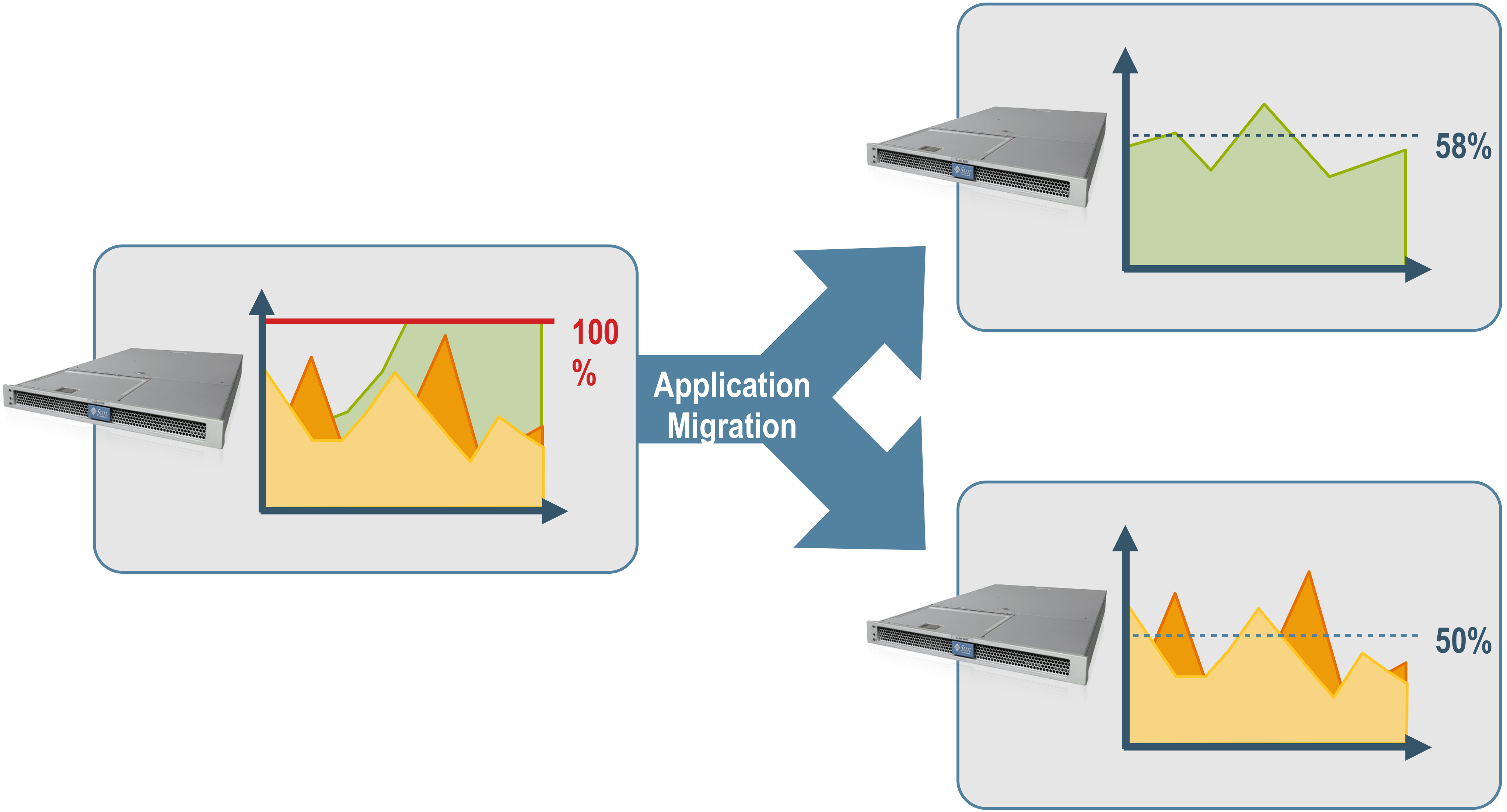
Virtualization & Grid

Efficient Resource Utilization: Consolidation

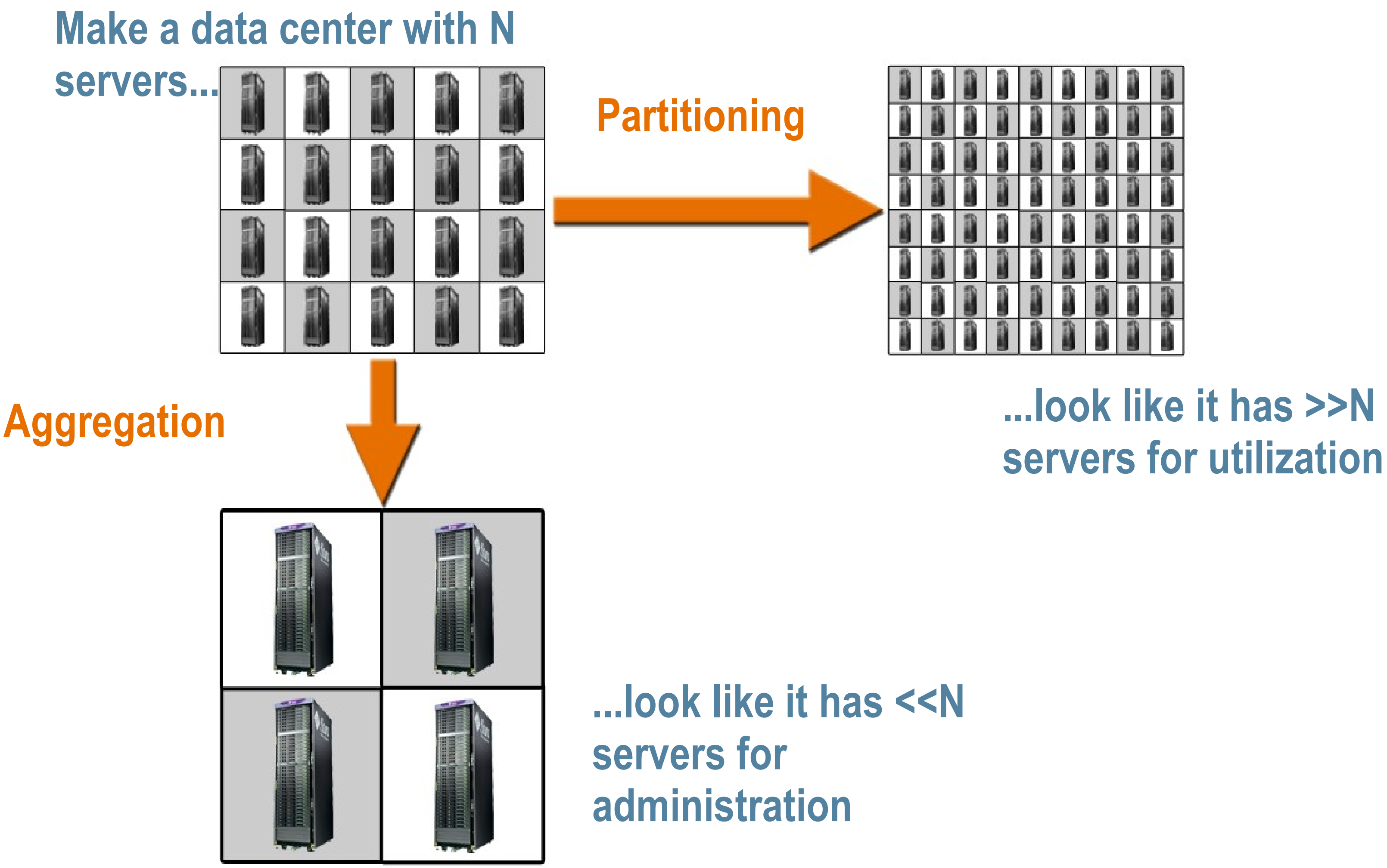
Today's average utilization:
7-15% per server
- Gartner, Nov. 2002



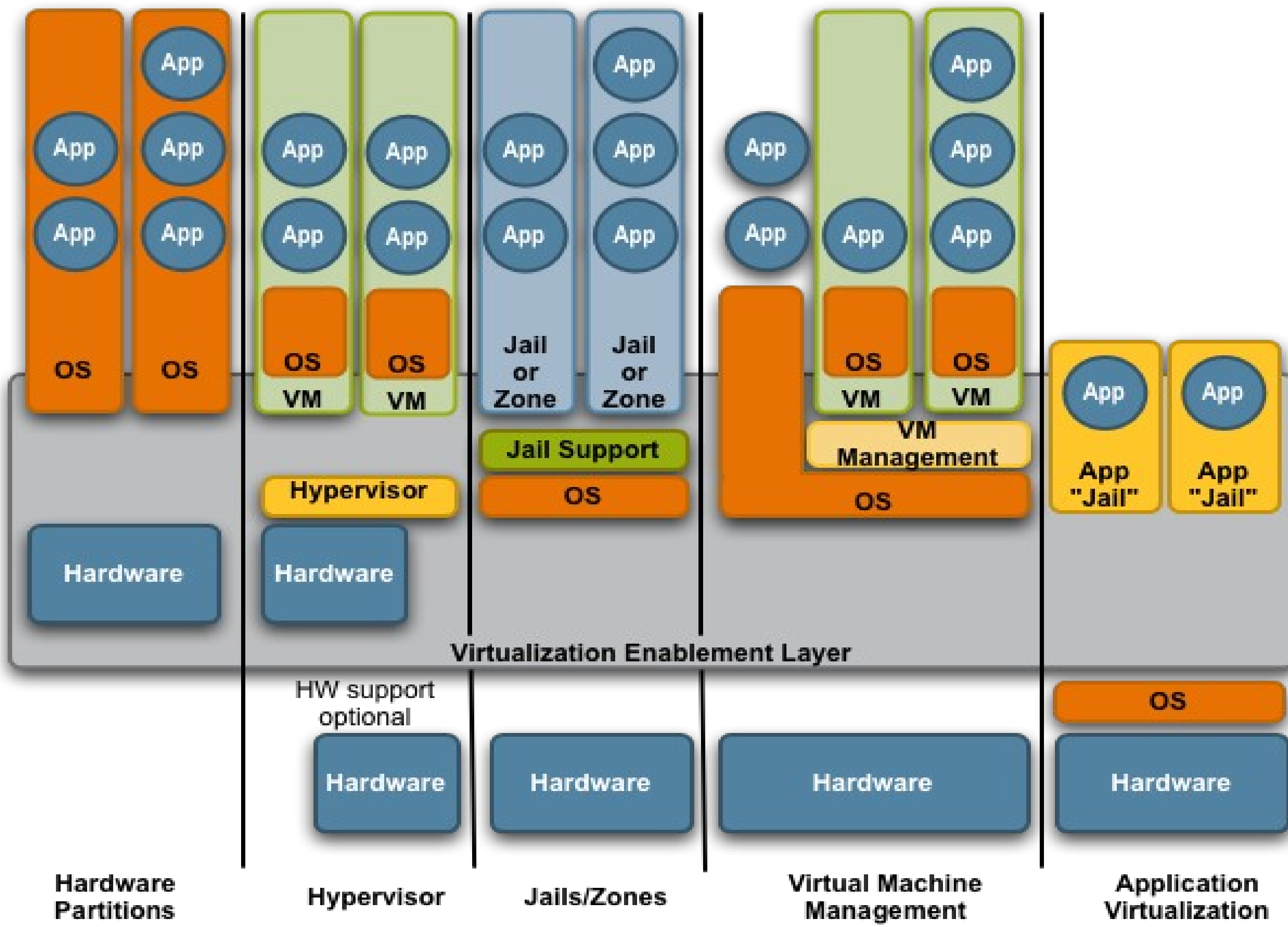
Efficient Resource Utilization: Migration



Two Dimensions of Virtualization



Virtualization



Virtualization Technology

Hardware Partitions

Technology	Vendor
Dynamic Systems Domain	Sun
LPAR	IBM
VPAR NPAR	HP

Hypervisor

Technology	Vendor
XVM Server	Sun
Virtual Infr 3 (ESX)	Vmware
Xen	Xensource & Sun
Viridian	Microsoft
Logical Domains	Sun
KVM (Linux	Community
VM	IBM

Virtualization Technology

OS Virtualization

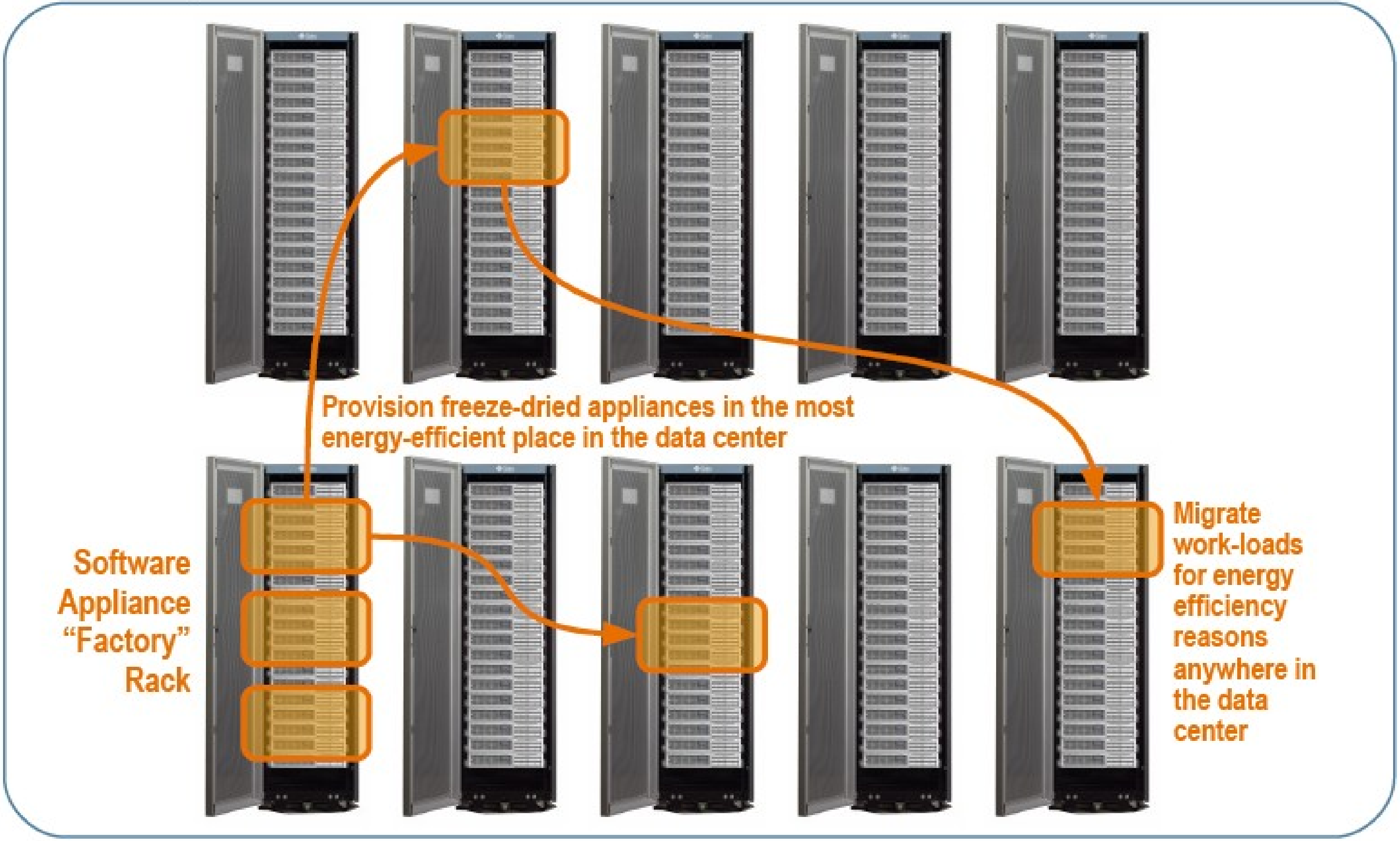
Technology	Vendor
Solaris Containers/Zones	Sun
IBM Wpars	IBM
BSD Jails	HP
Virtuozzo	Swsoft
OpenVZ	Community

Application Virtualization

Technology	Vendor
Etude	Sun
Trigence	Trigence
Softgrid	Softricity
SVS – Software Virt Soln	Alteris
Logical Domains	Sun
Project Tarpon	Citrix

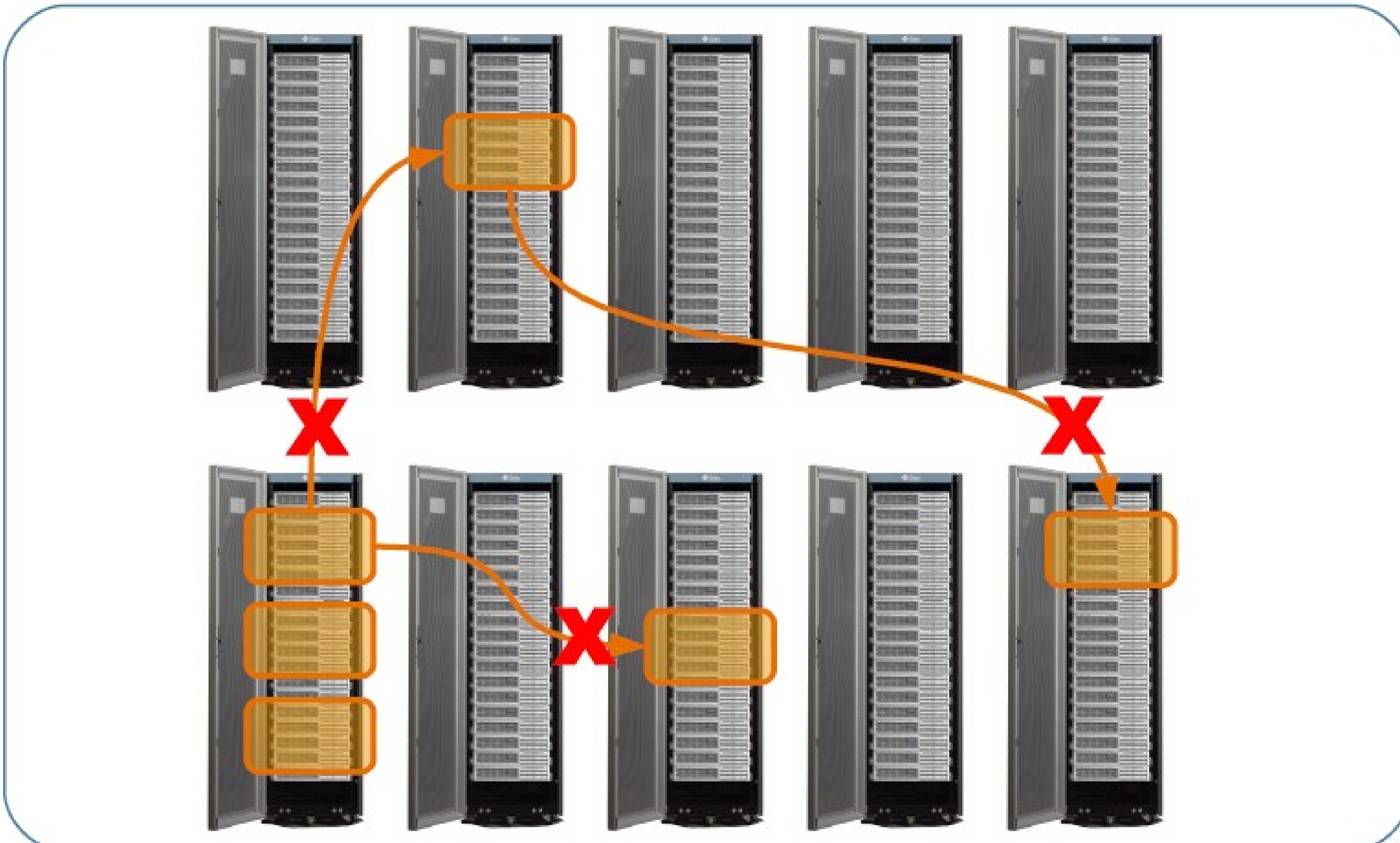
The Re-Entrant Grid

The Holy Grail



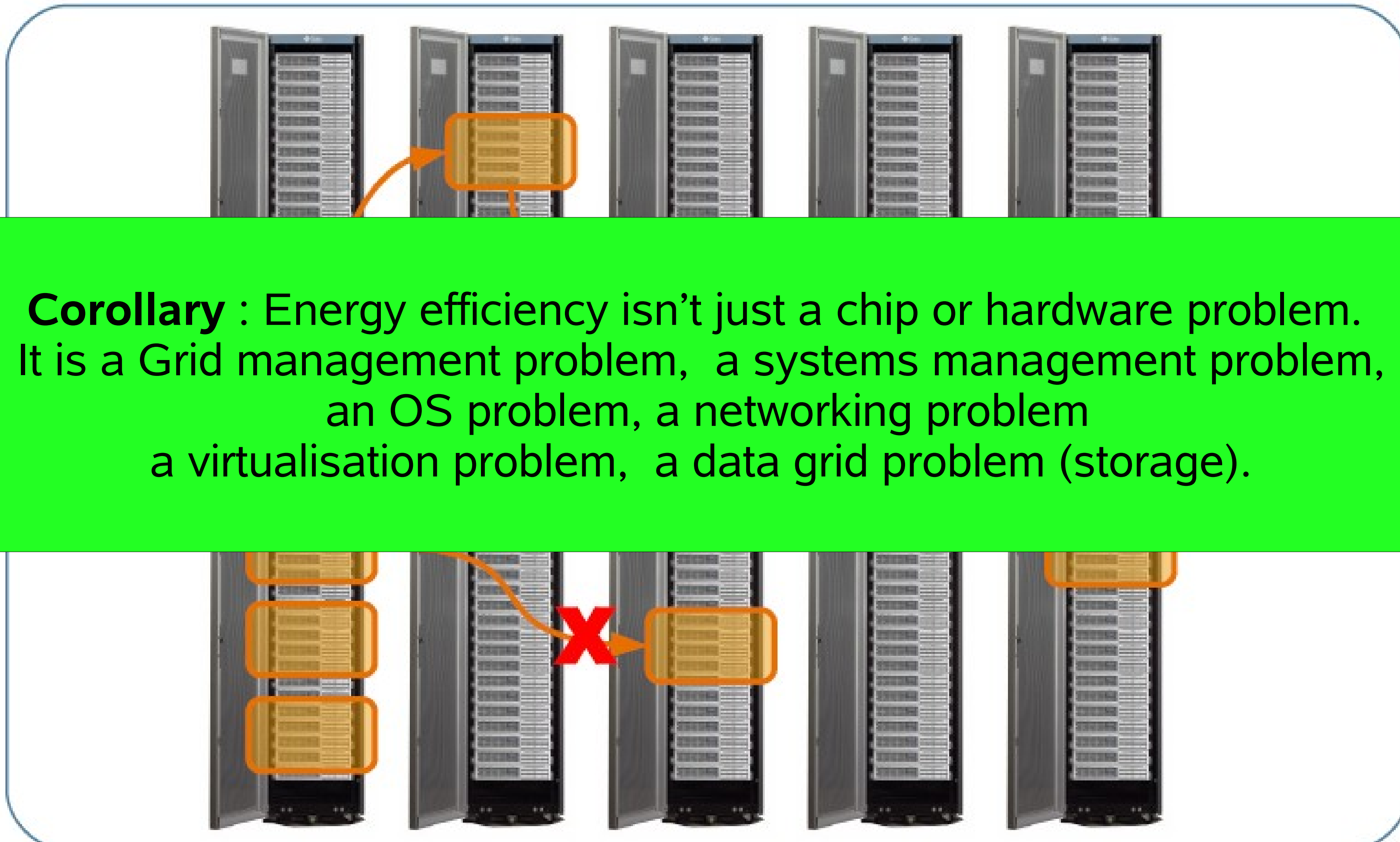
Why we cannot do this easily

Because The Network Needs To Get Involved



Why we cannot do this easily

Because The Network Needs To Get Involved



Corollary : Energy efficiency isn't just a chip or hardware problem.
It is a Grid management problem, a systems management problem,
an OS problem, a networking problem
a virtualisation problem, a data grid problem (storage).

What the community is up to?

- Spec power and performance
- Green Top 500
- Green Grid
- US Congress passed law 109-431
- EPA Report
- Others

Ackwoledgement

- Subodh Bapat, Sun Microsystems Inc.
- John Fragalla, Sun Microsystems Inc.
- Dave Douglas, Sun Microsystems Inc.
- Many others

Is there a pathway to a Green Grid ??

Dr Simon See
simon.see@sun.com
or
mcwsee@ntu.edu.sg