From Bits to Buildings: Energy Efficiency and the Path to Exaflops

Horst D. Simon
Lawrence Berkeley National Laboratory
and EECS Dept., UC Berkeley
hdsimon@lbl.gov

26e Forum ORAP, Paris, France
March 31, 2010

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

Acknowledgements

A large number of individuals have contributed to energy efficiency in computing at Berkeley Lab, UC Berkeley, and to this presentation: David Bailey (CRD), Michael Banda (CRD), Michael Bennett (ITD), ShoaibKamil (CRD), Jonathan Koomey (Stanford), Randy Katz (EECS), TsuJae King (EECS), Chuck McParland (CRD), Juan Meza (CRD), Bruce Nordman (EETD), Lenny Oliker (CRD), EkowOtoo (CRD), Vern Paxson (UCB/ICSI/CRD), DoronRotem (CRD), Dale Sartor (EETD), John Shalf (NERSC), Erich Strohmaier (CRD), Bill Tschudi (EETD), Howard Walter (NERSC), Michael Wehner (CRD), Kathy Yelick (NERSC/CRD) … and many others

Check the LBNL website for more details:

http://www.lbl.gov/cs/html/energy_efficient_computing.html

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Outline

Trends in Power Consumption and Energy Efficiency for HPC

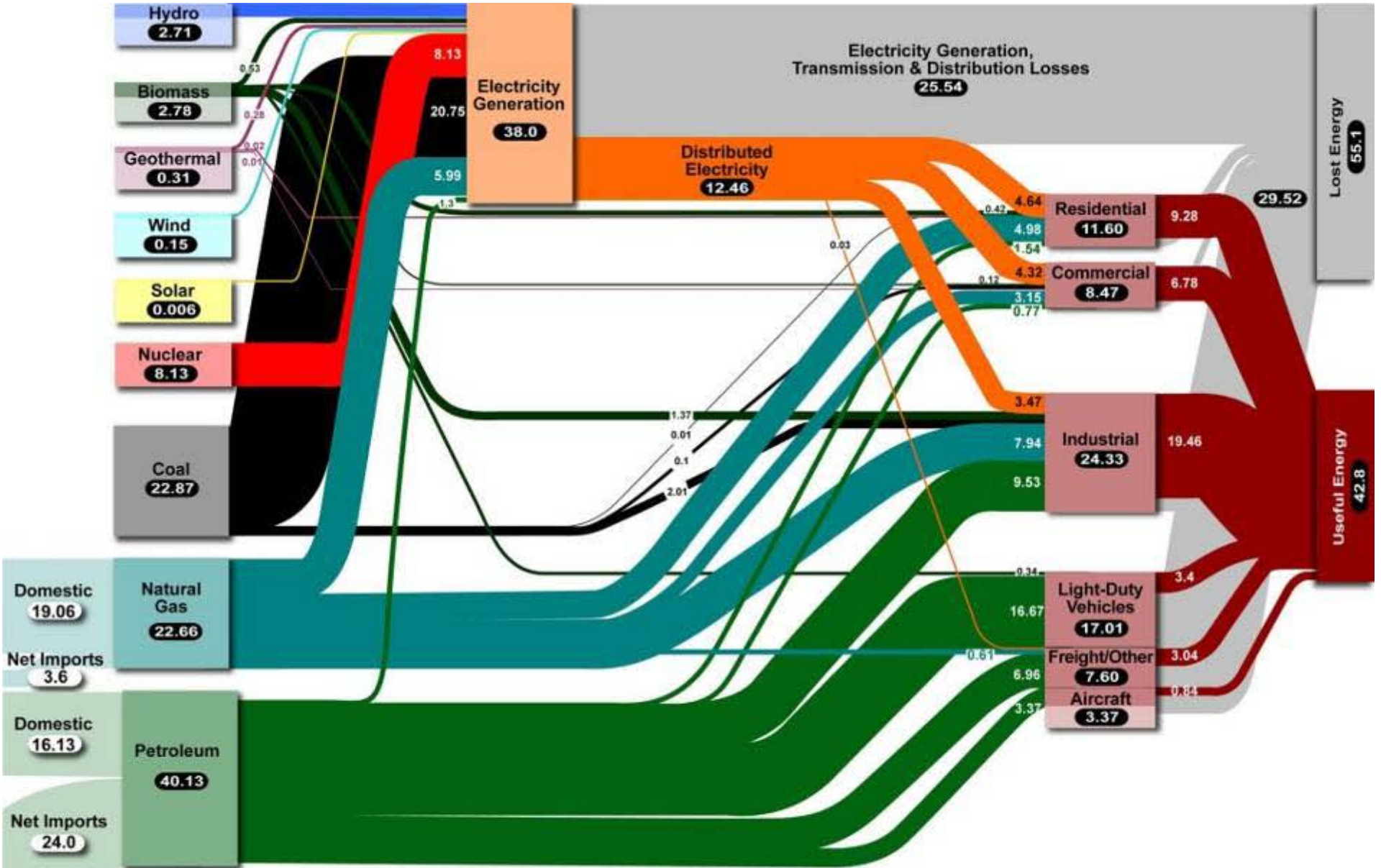
Building and infrastructure problem -- continued increase in demand for computing ("buildings")

Computer technology problem -- no more power density scaling ("bits")

Summary of the Talk

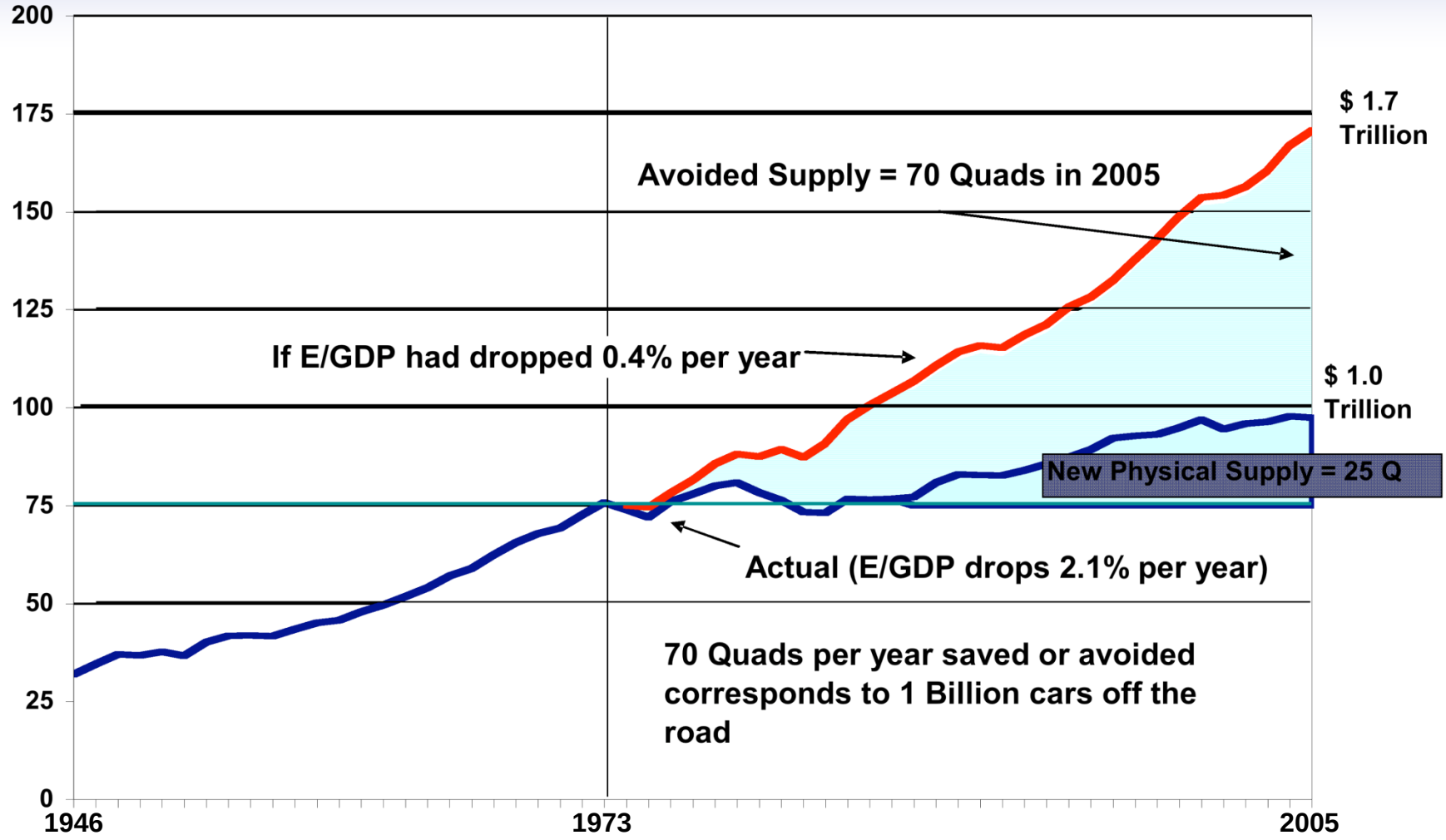
Power has become THE dominant problem in computing

EnergyProduction and Use in the U.S.



Why does saving energy matter?

Energy Consumption in the United States 1949 - 2005



Source: Art Rosenfeld, California Energy Commission, http://www.energy.ca.gov/commission/commissioners/rosenfeld_docs/index.html

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

An Honest Question?

Does the HPC community really care about reducing the carbon footprint?

NO!

HPC Interests

- **Energy efficiency in computer rooms**
  - **Spend more resources on computing than on infrastructure**
- **Energy efficient technology**
  - **Maintain performance growth and get things done that could not be done before**

Khazzoom-Brookes Postulate

- **Energy efficiency at the micro-level leads to higher energy consumption at the macro-level**
  - **cheaper energy increases use**
  - **increased energy efficiency leads to economic growth**
  - **increased efficiency in one bottleneck resource increases use of companion technologies**
- **HPC follows Khazzoom-Brookes**

Energy and IT

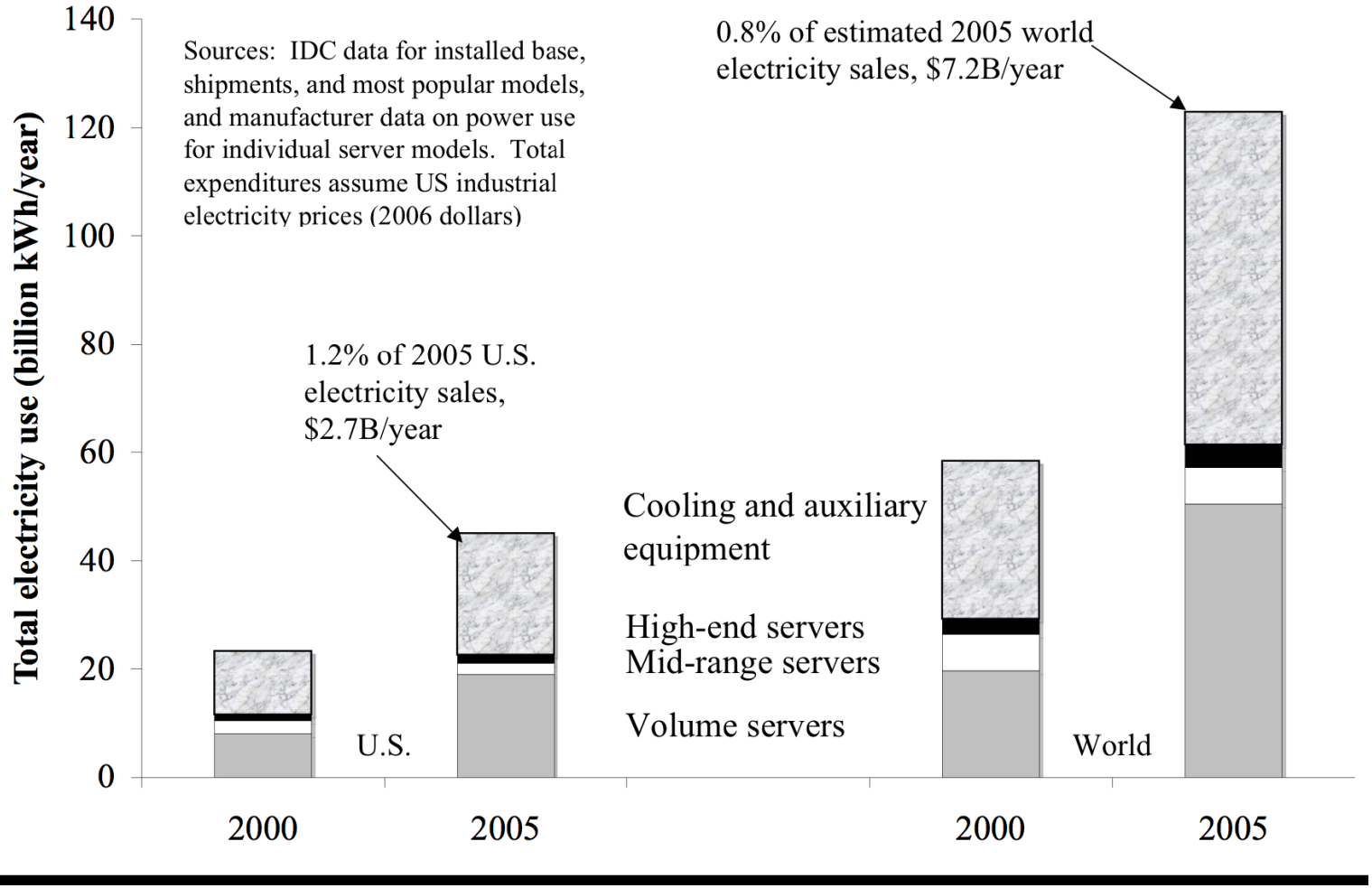
Numbers represent U.S. only

- **"Big IT" – all electronics**
  - **PCs / etc., consumer electronics, telephony**
    - Residential, commercial, industrial
  - **More than 200 TWh/year**
  - **$16 billion/year**
    - Based on .08$/KWh
  - **Nearly 150 million tons of CO_2 per year**
    - Roughly equivalent to 30 million cars!

One central baseload power plant (about 7 TWh/yr)

… and IT electricity use is increasing

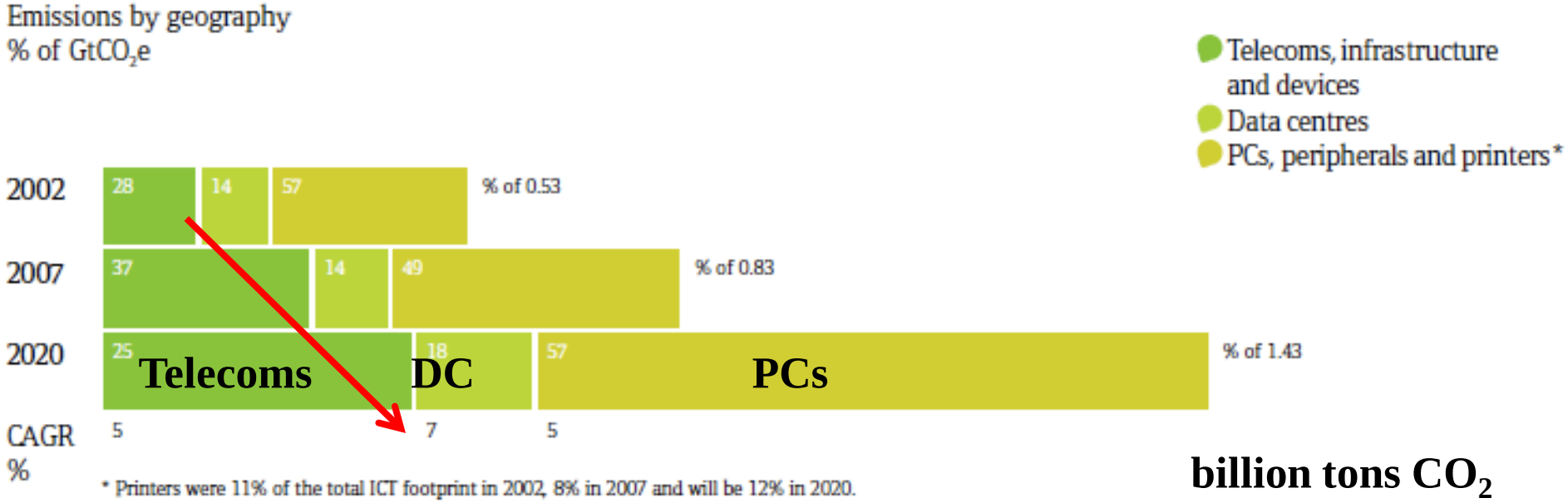
data taken from: Jonathan Koomey, "Estimating Total Power Consumption by Servers in the U.S. and the World"
Available at: http://www.koomey.com/publications.html



2020 IT Carbon Footprint

"SMART 2020: Enabling the Low Carbon Economy in the Information Age", The Climate Group

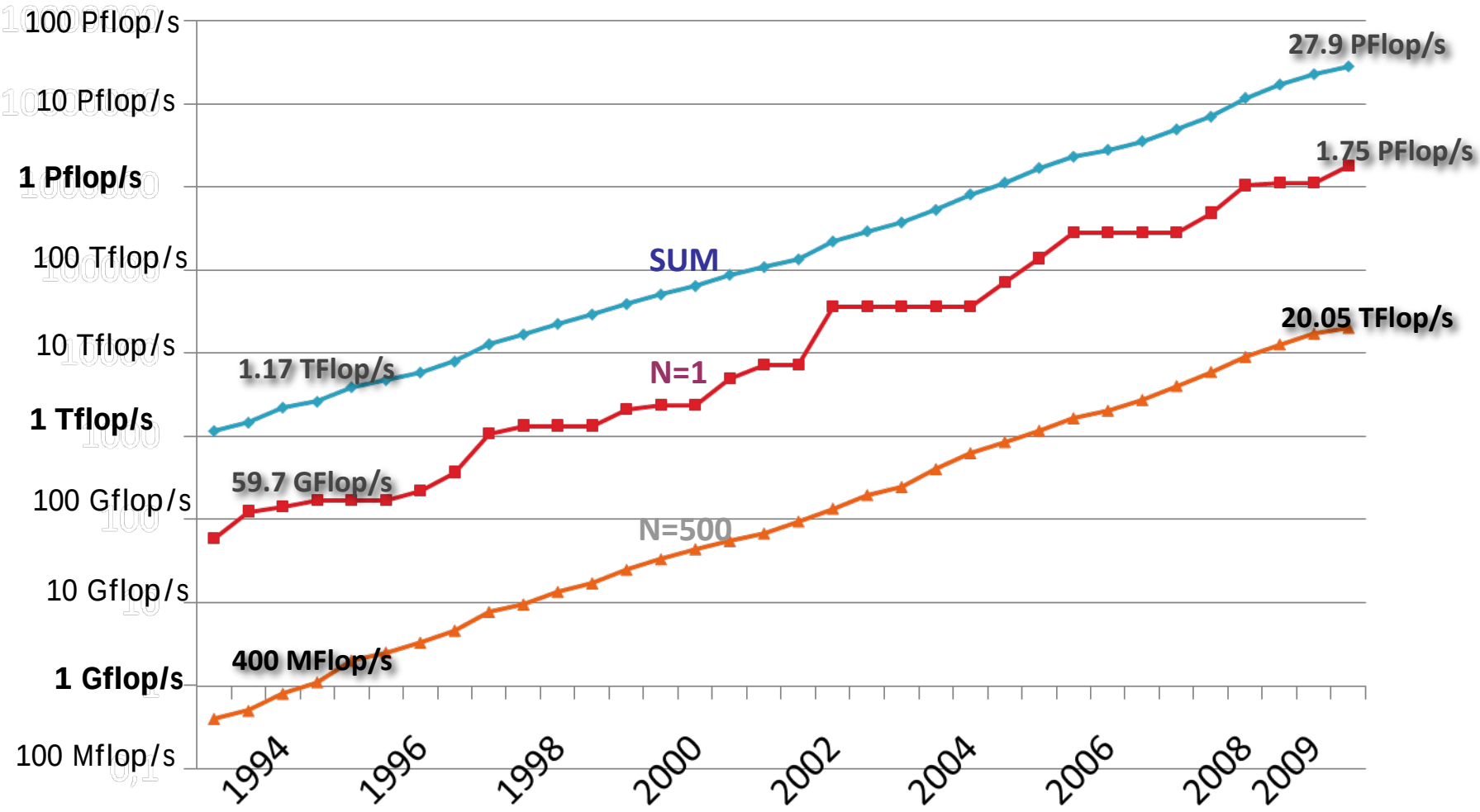
Fig. 2.3 The global footprint by subsector



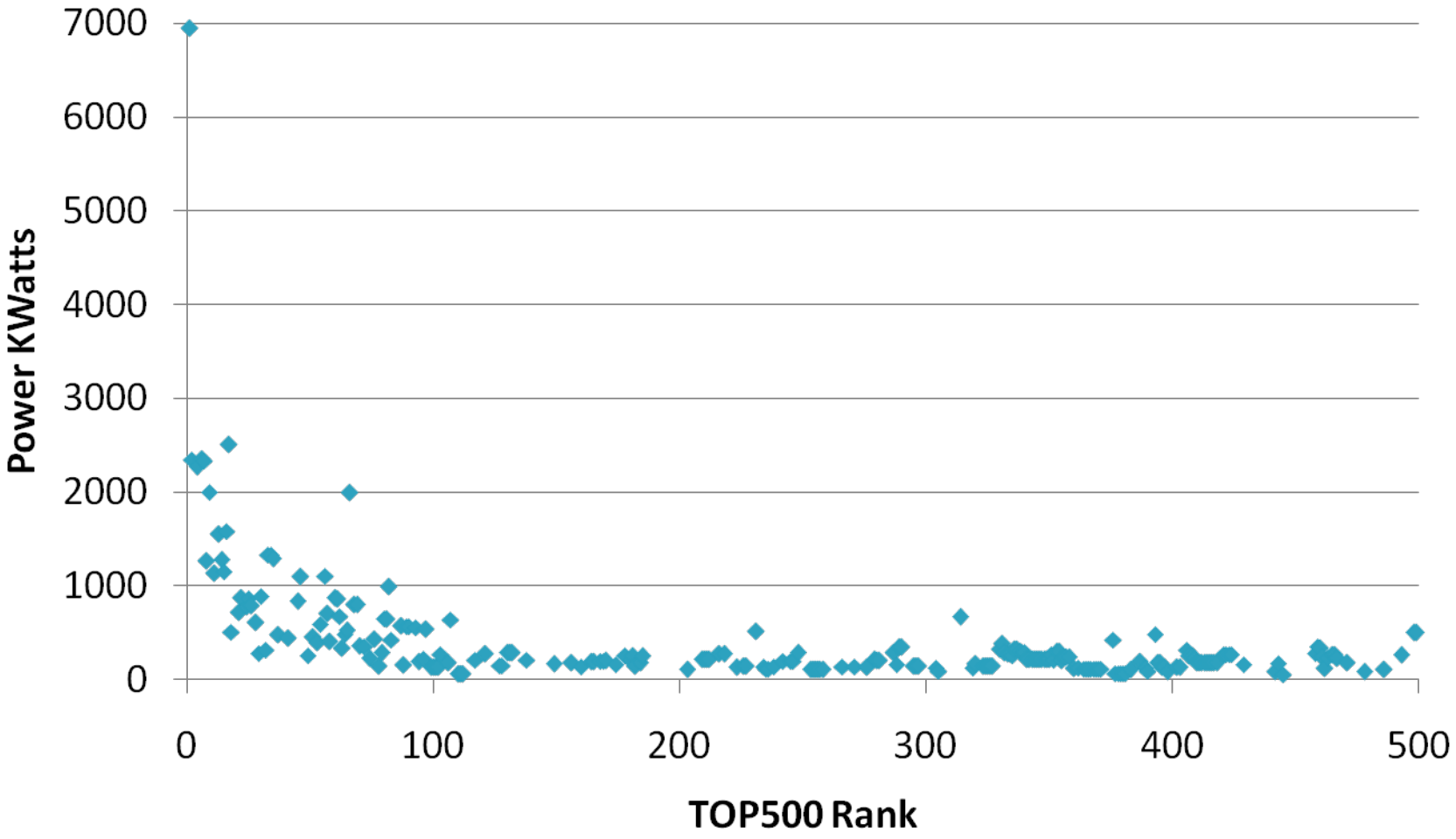
billion tons CO_2

Datacenters: Owned by single entity interested in reducing opex

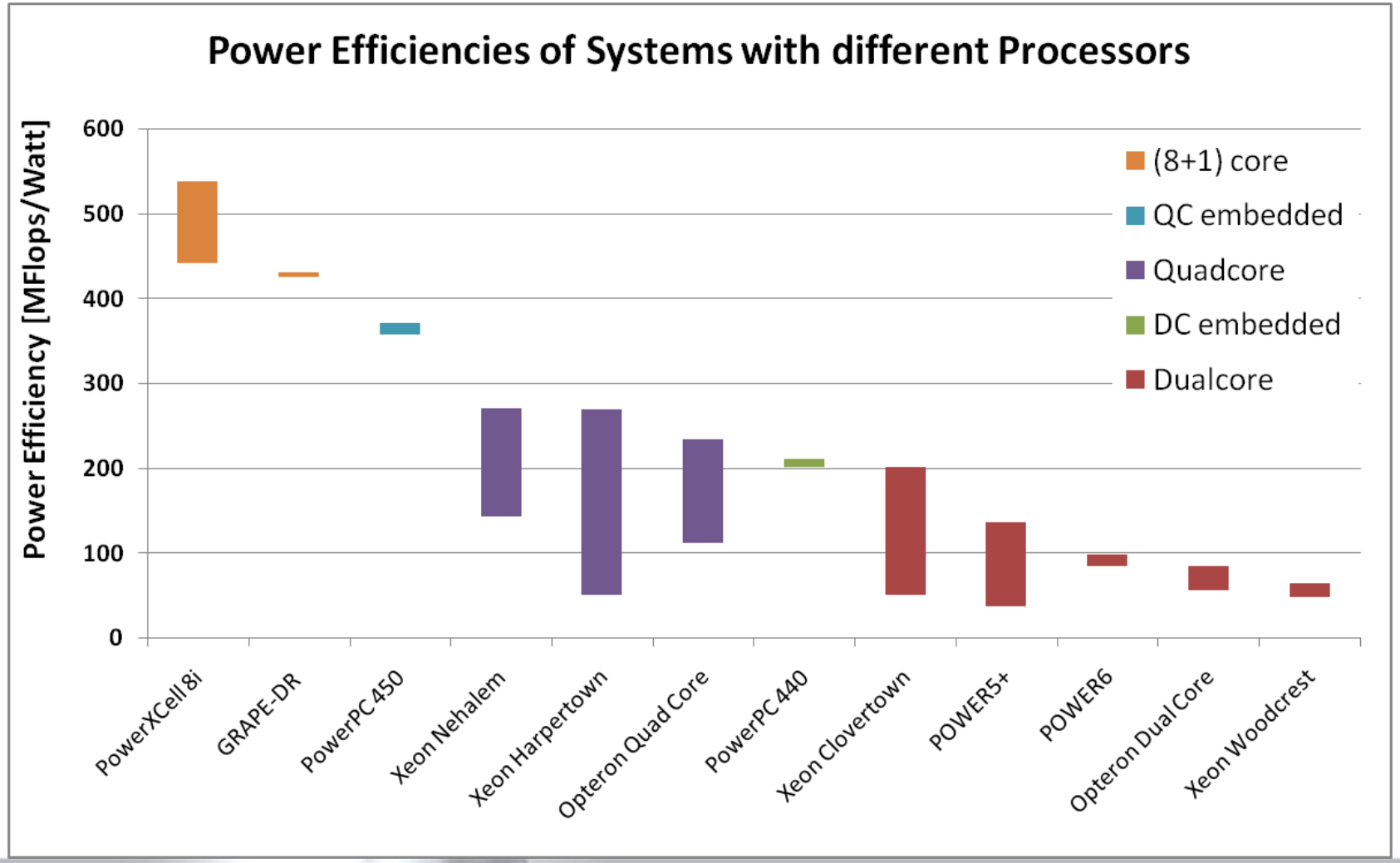
Performance Development



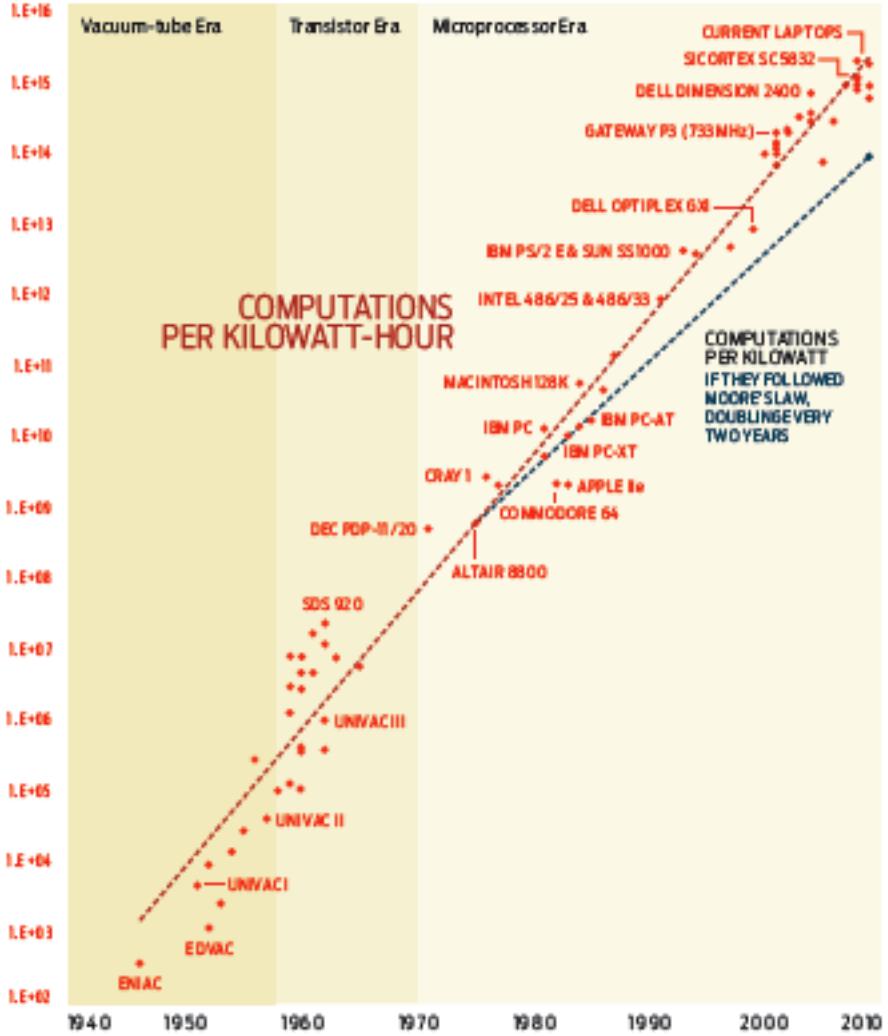
Absolute Power Levels



Power Efficiency related to Processors



Koomey's Law



The complete study is available from the Intel Web site at http://download.intel.com/pressroom/pdf/computestrendsrelease.pdf

- **Computations per kWh have improved by a factor about 1.5 per year**
- **"Assessing Trends in Electrical Efficiency over Time", see IEEE Spectrum, March 2010**

BERKELEY LAB

Trend Analysis

- **Processors and Systems have become more energy efficient over time**
  - **Koomey's Law shows factor of 1.5 improvement in kWh/computations**
- **Supercomputers have become more powerful over time**
  - **TOP500 data show factor of 1.86 increase of computations/sec**
- **Consequently power/system increases by about 1.24 per year**

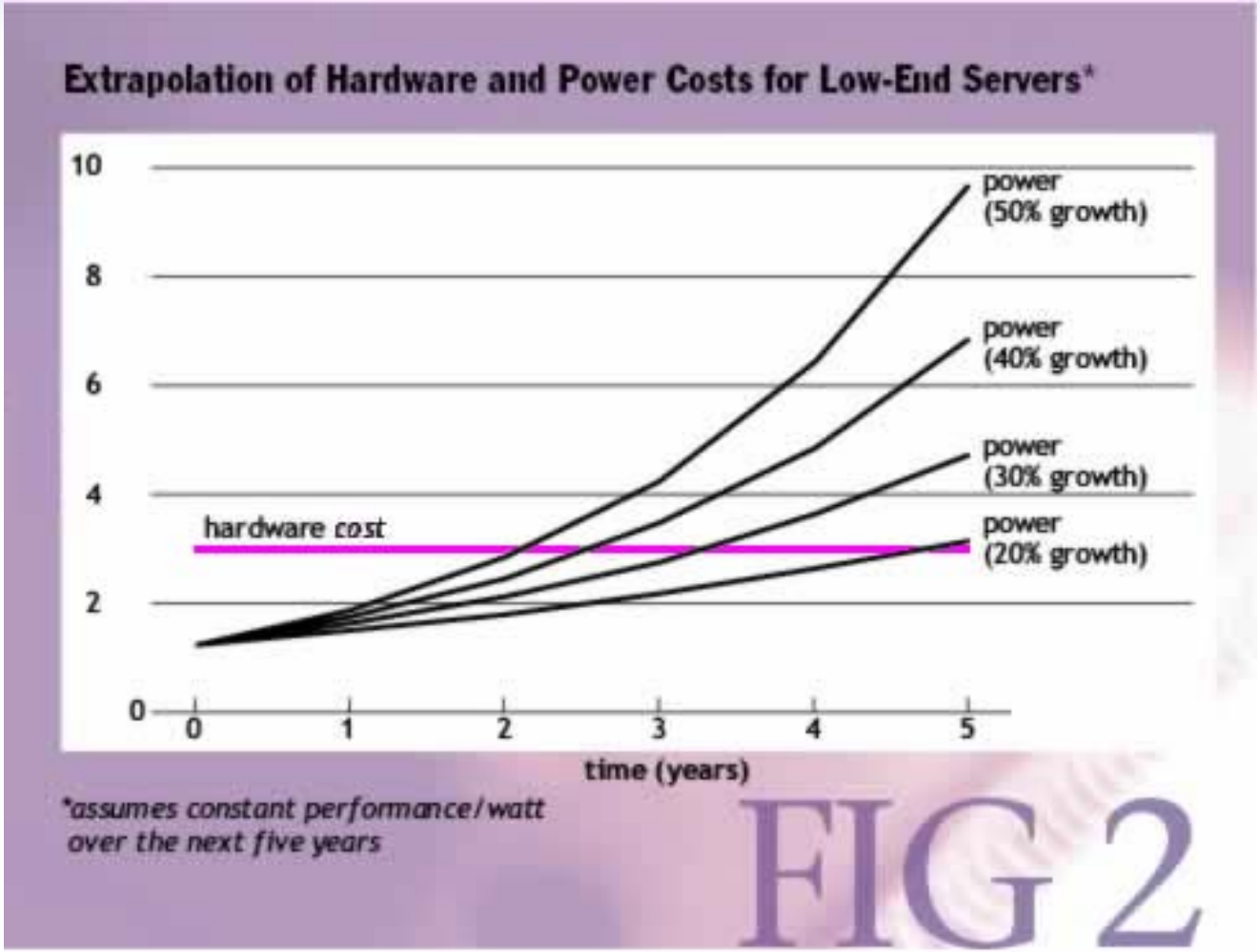
Outline

Trends in Power Consumption and Energy Efficiency for HPC

Building and infrastructure problem -- continued increase in demand for computing ("buildings")

Computer technology problem -- no more power density scaling ("bits")

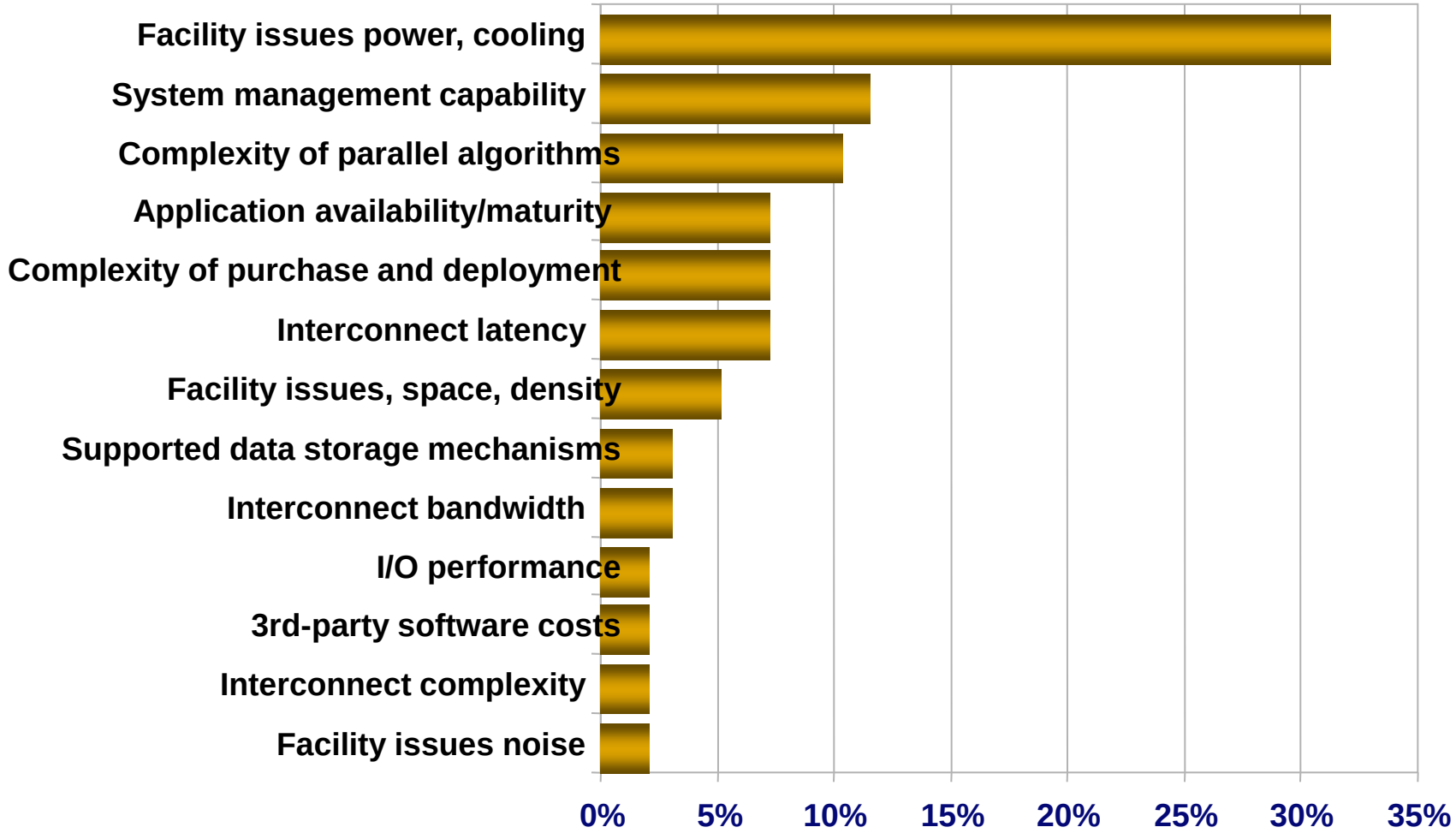
The Problem



Unrestrained IT power consumption could eclipse hardware costs and put great pressure on affordability, data center infrastructure, and the environment.

Source: Luiz André Barroso (Google), "The Price of Performance," *ACM Queue*, Vol. 2, No. 7, pp. 48-53, September 2005
(Modified with permission)

Top Challenges to Clusters

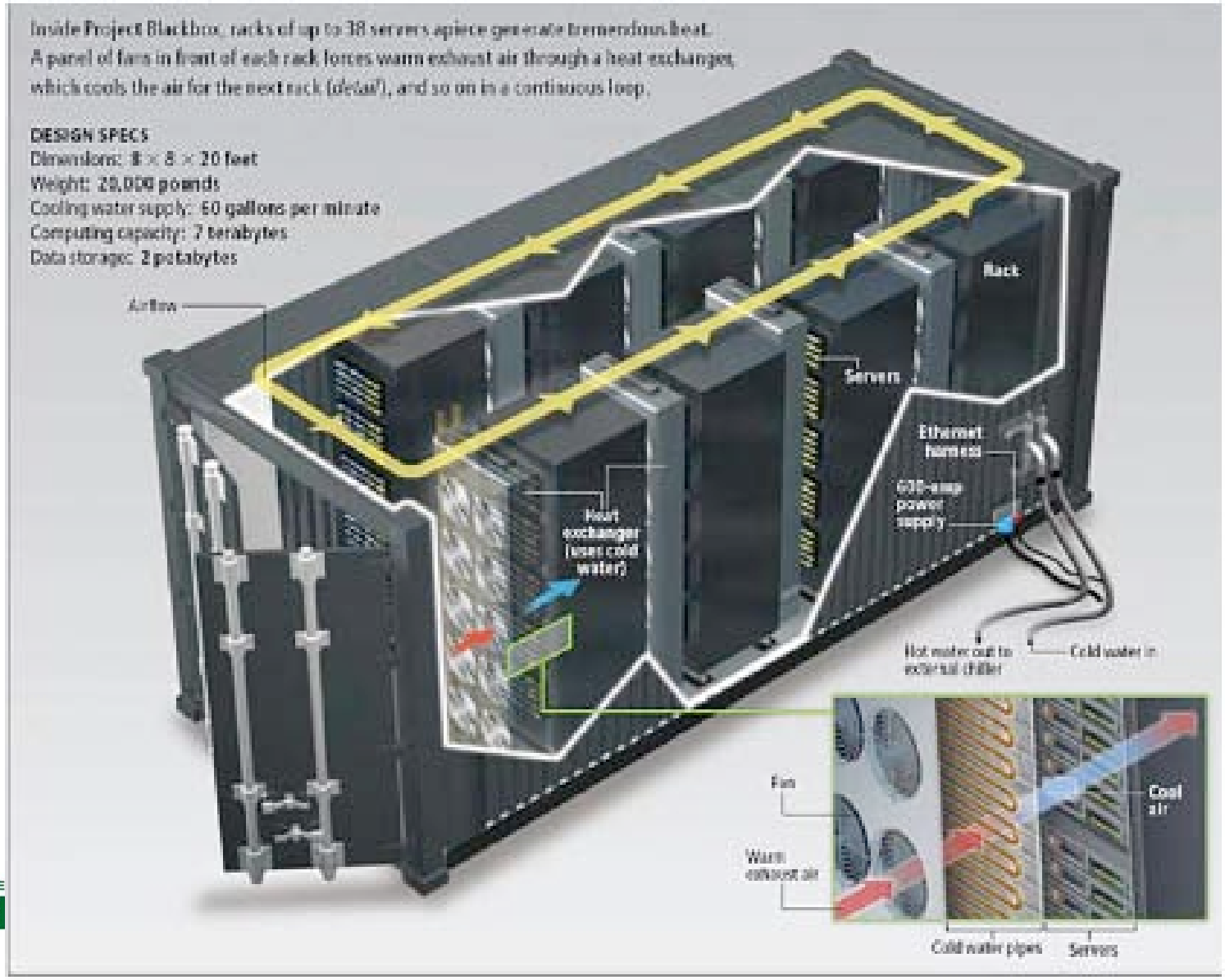


n = 96

Responses

- Cloud
- Containerized data centers
- Large scale data “factories”
- Increased emphasis on computer room and building efficiency

Containerized Datacenter Mechanical-Electrical Design



Data Center Economic Reality (2006)

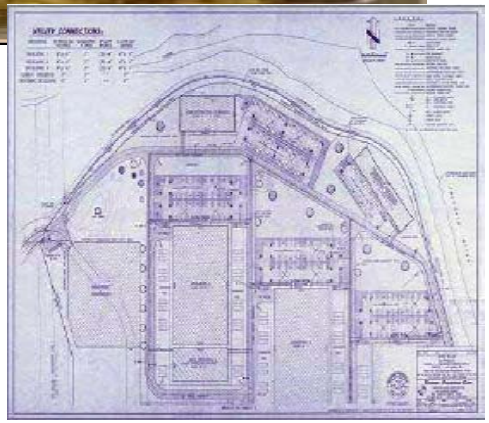
- **June 2006 - Google begins building a new data center near the Columbia River on the border between Washington and Oregon**
  - Because the location is “at the intersection of cheap electricity and readily accessible data networking”

    “Hiding in Plain Sight, Google Seeks More Power”
    by John Markoff, NYT, June 14, 2006

- **Microsoft and Yahoo are building big data centers upstream in Wenatchee and Quincy, Wash.**
  - To keep up with Google, which means they need cheap electricity and readily accessible data networking

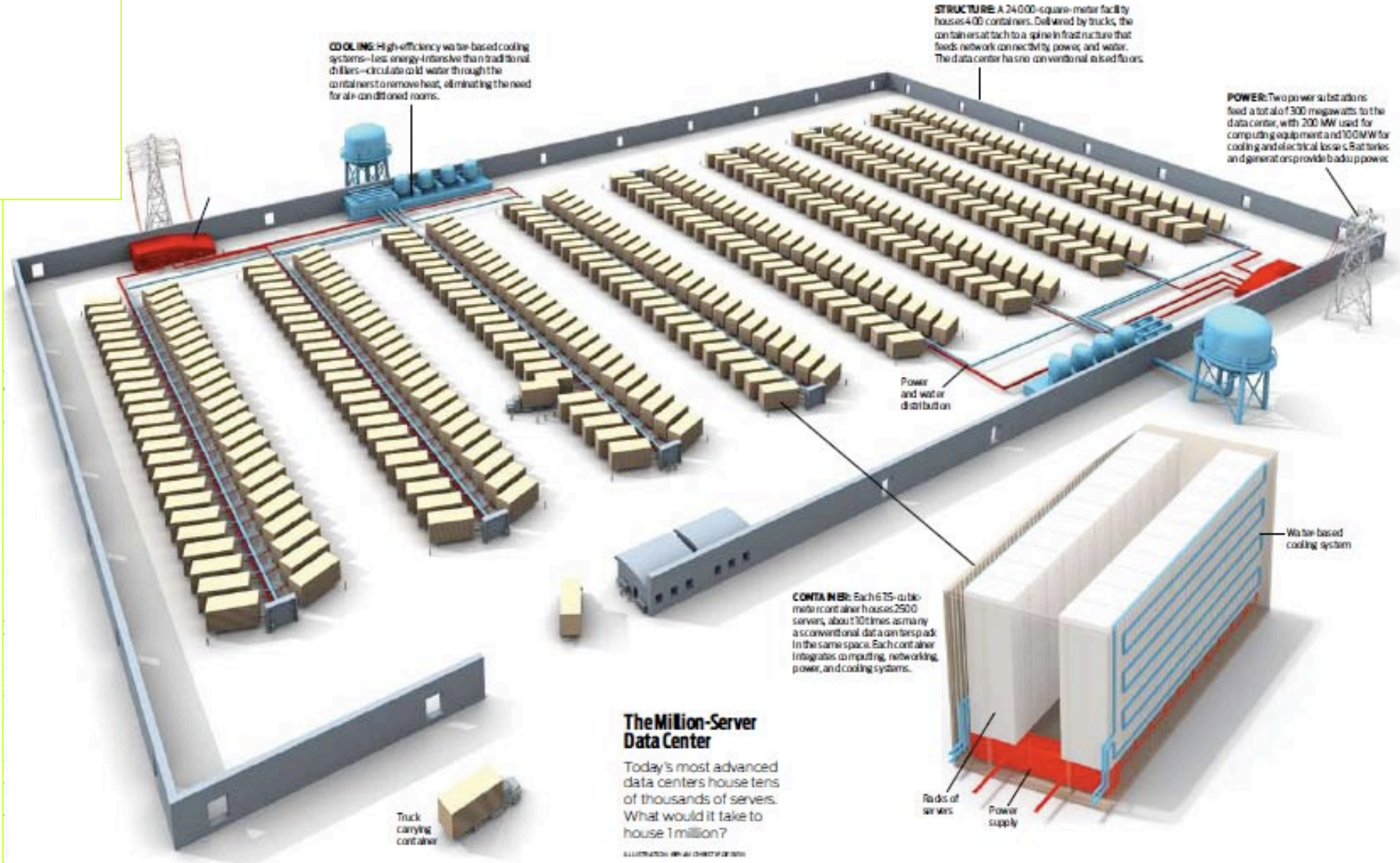
Source: New York Times, June 14, 2006

Google Dalles Oregon Facility
68,680 Sq Ft Per Pod



Source: Levy and Snowhorn, Data Center Power Trends, February 18, 2008

Microsoft's Chicago Modular Datacenter



The Million Server Datacenter

- 24000 sq. m housing 400 containers
  - Each container contains 2500 servers
  - Integrated computing, networking, power, cooling systems
- 300 MW supplied from two power substations situated on opposite sides of the datacenter
- Dual water-based cooling systems circulate cold water to containers, eliminating need for air conditioned rooms

Potential Benefits of Improved Data Center Energy Efficiency:

- 20-40% savings typically possible
- Aggressive strategies can yield better than 50% savings
- Extend life and capacity of existing data center infrastructures
- But is my center good or bad?

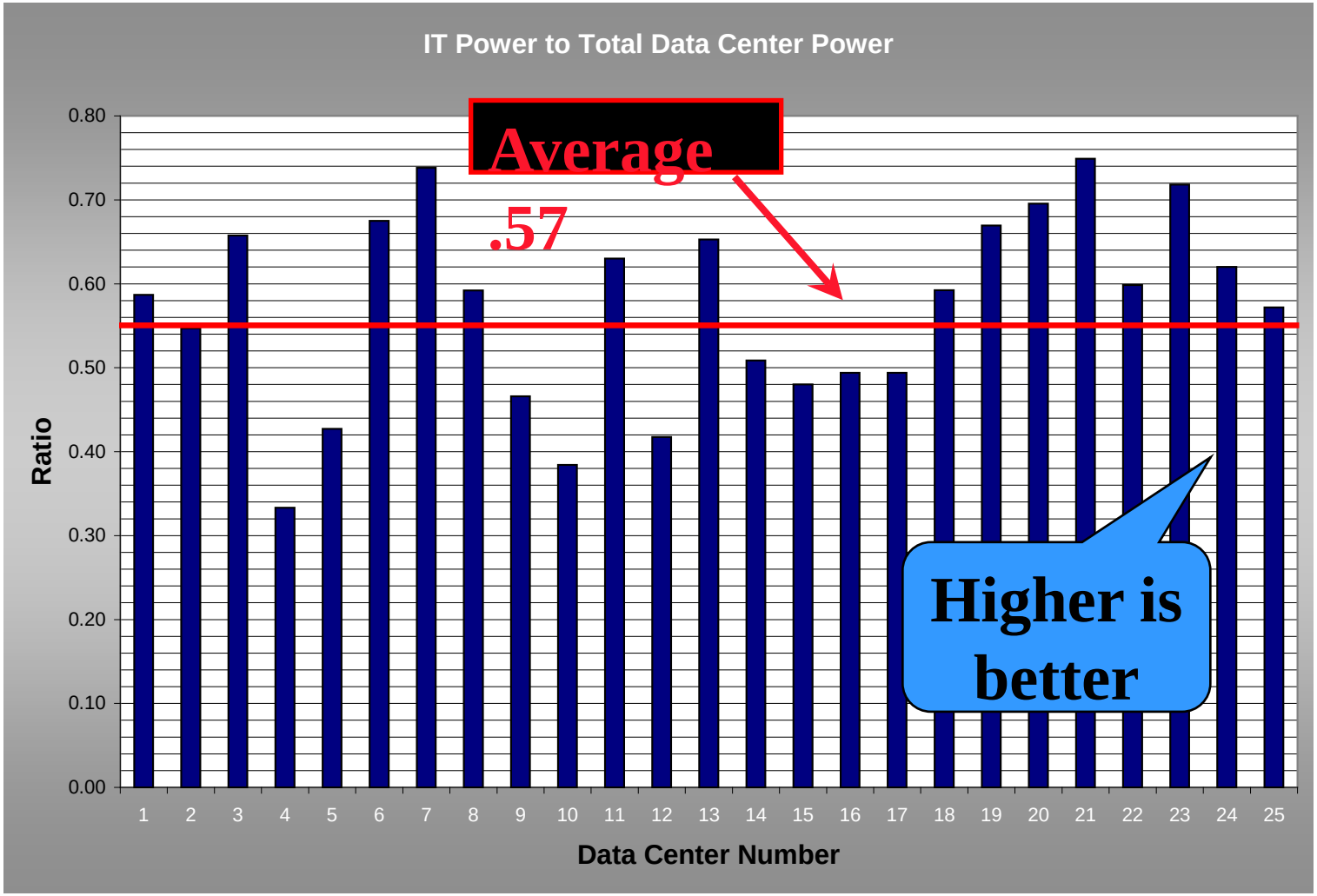
Benchmarking for Energy Performance Improvement:

Energy benchmarking can allow comparison to peers and help identify best practices

LBNL conducted studies of over 30 data centers:

- **Found wide variation in performance**
- **Identified best practices**

High Level Metric—
Data Center Infrastructure Efficiency (DCiE)
Ratio of Electricity Delivered to IT Equipment to Total



U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

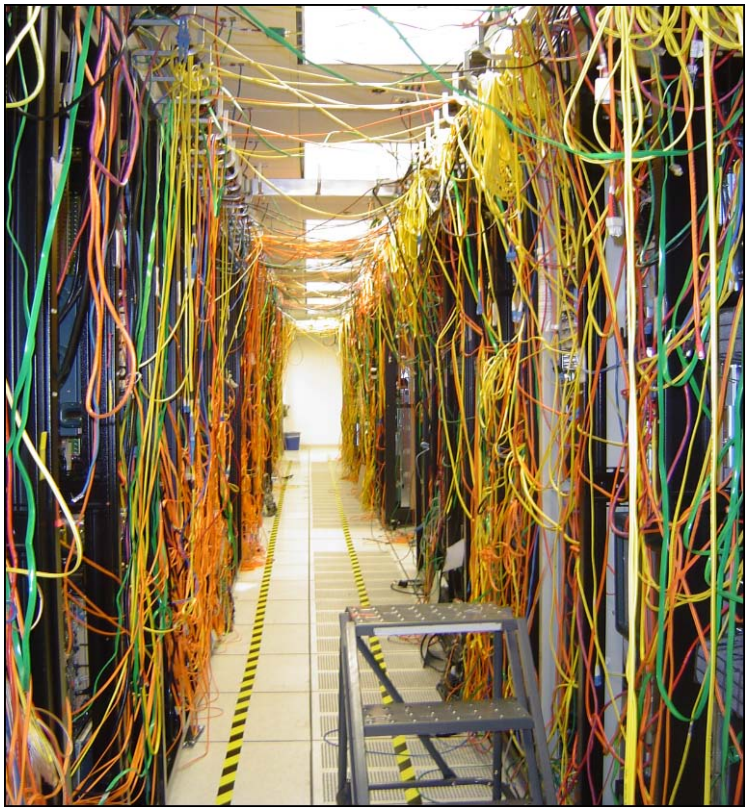
Focus on PUE

- **PUE = “power usage effectiveness” metric promoted by “Green Grid”**
- **PUE = total facility power/ computer equipment power = 1/DCiE**
- **Reduce PUE by consistent application of facilities improvements**
- **Take all PUE claims with a grain of salt**

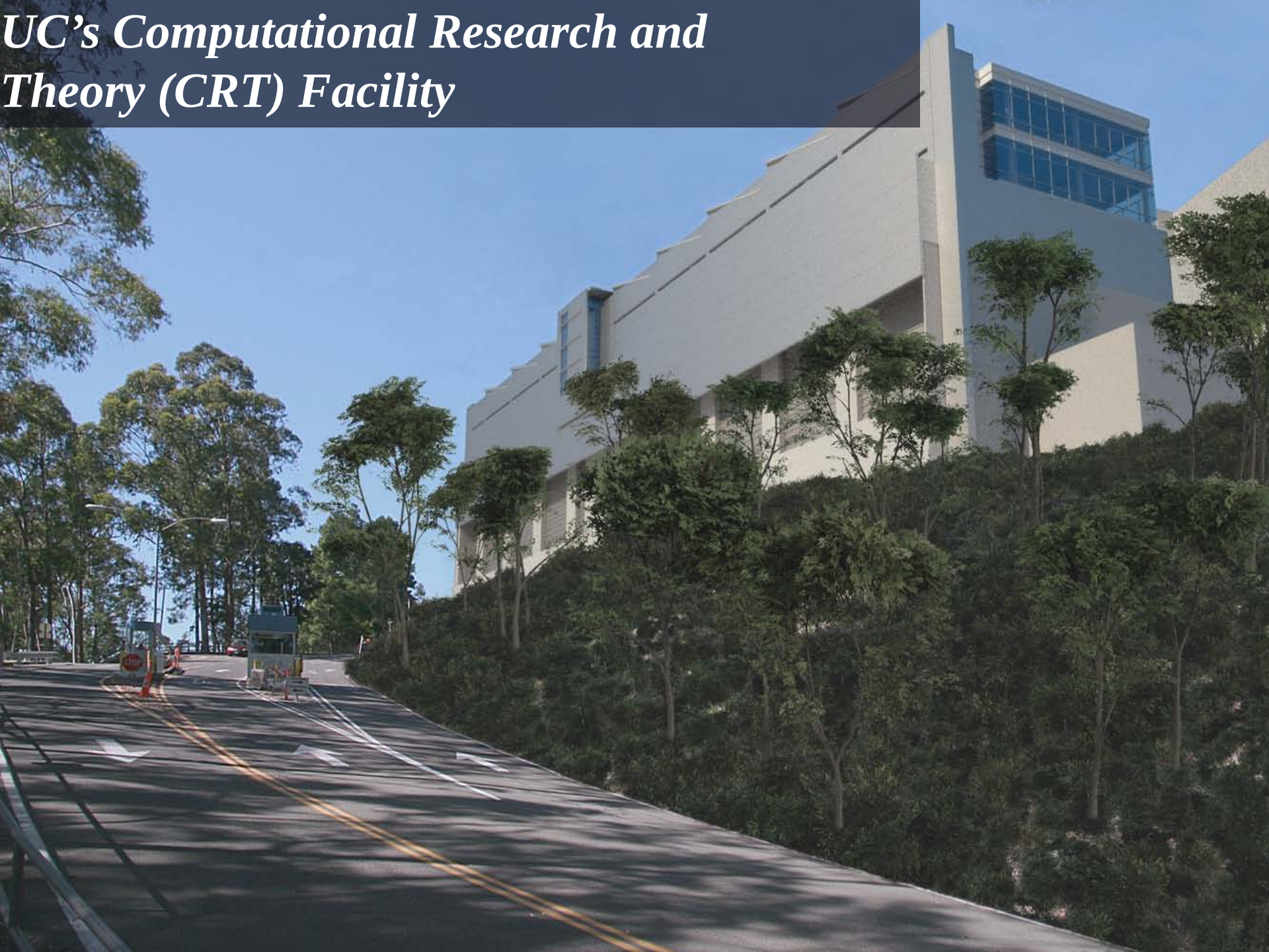
	PUE
Current Trends	1.9
Improved Operations	1.7
Best Practices	1.3
State-of-the-Art	1.2

Using benchmark results to find best practices:

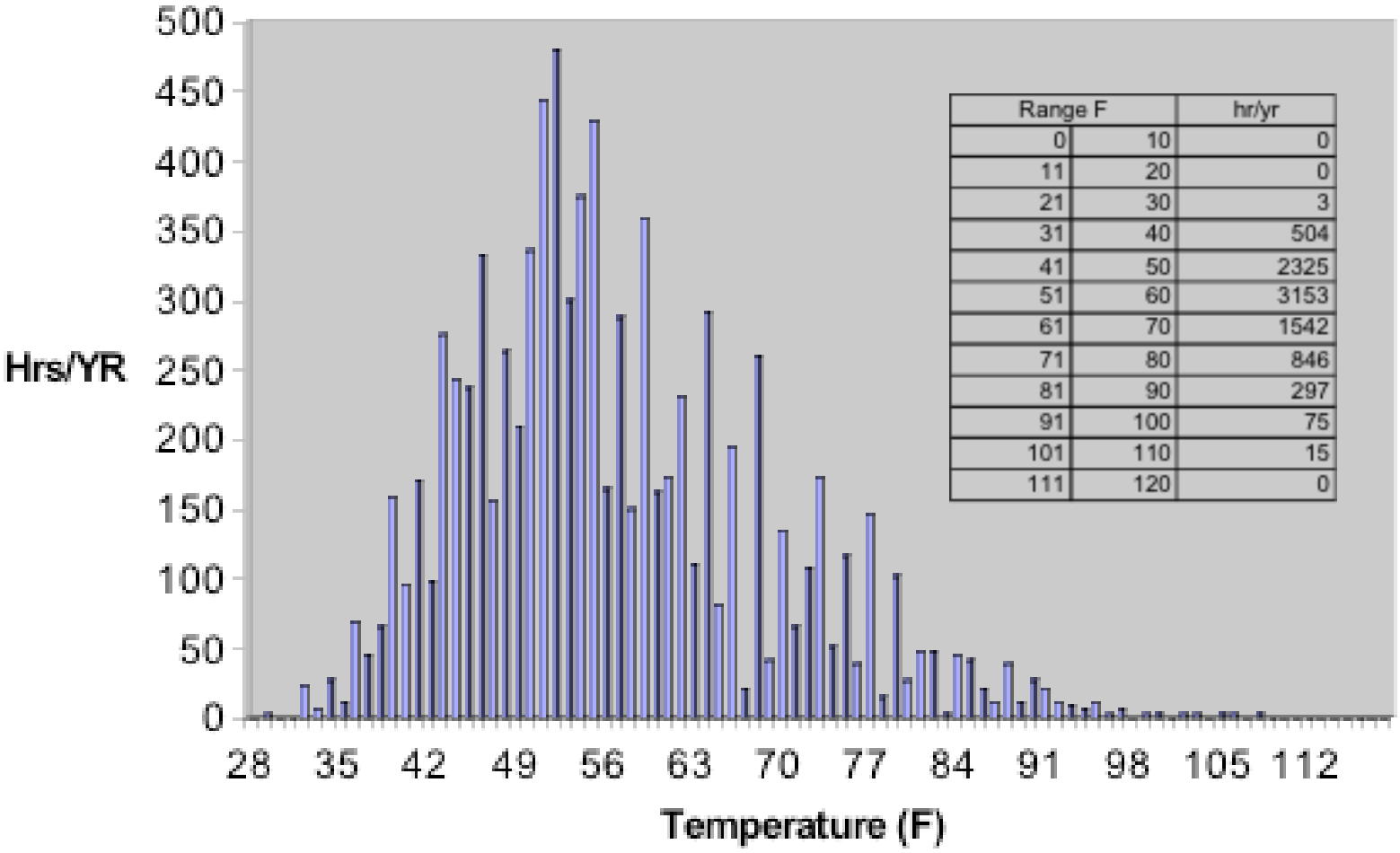
- **Air management**
- Right-sizing
- Central plant optimization
- Efficient air handling
- **Liquid cooling**
- **Free cooling**
- **Humidity control**
- **Improve power chain**
- On-site generation
- **Design and M&O processes**



UC's Computational Research and Theory (CRT) Facility



Berkeley Weather



Range F		hr/yr
0	10	0
11	20	0
21	30	3
31	40	504
41	50	2325
51	60	3153
61	70	1542
71	80	846
81	90	297
91	100	75
101	110	15
111	120	0

Use Free Cooling:

- **Water-side Economizers**
  - **No contamination question**
  - **Can be in series with chiller**
- **Outside-Air Economizers**
  - **Can be very effective (24/7 load)**
  - **Must consider humidity**

Water Cooling: Four-pipe System

- Allows multiple temperature feeds at server locations through mixing of CHW & TRW
- Closed-loop treated cooling water from cooling towers (via heat exchanger)
- Chilled water from chillers
- Headers, valves and caps for modularity and future flexibility

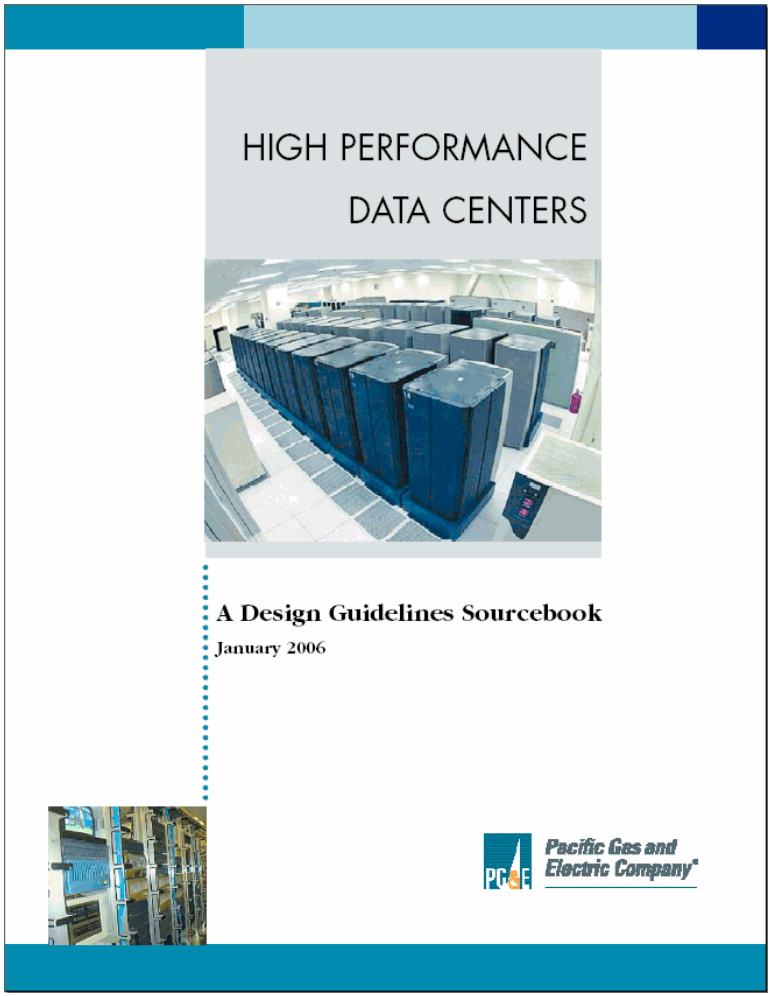
Predicted CRT Performance

- DCIE of 0.95 based on annual energy
- DCIE of 0.88 based on peak power

Design Guidelines Are Available

- Design Guides were developed based upon the observed best practices
- Guides are available through PG&E and LBNL websites
- Self benchmarking protocol also available

http://hightech.lbl.gov/datacenters.html



U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Links to Get Started

DOE Website: Sign up to stay up to date on new developments
www.eere.energy.gov/datacenters

Lawrence Berkeley National Laboratory (LBNL)
http://hightech.lbl.gov/datacenters.html

LBNL Best Practices Guidelines (cooling, power, IT systems)
http://hightech.lbl.gov/datacenters-bpg.html

ASHRAE Data Center technical guidebooks
http://tc99.ashraetcs.org/

The Green Grid Association – White papers on metrics
http://www.thegreengrid.org/gg_content/

Energy Star® Program
http://www.energystar.gov/index.cfm?c=prod_development.server_efficiency

Uptime Institute white papers
www.uptimeinstitute.org

Outline

Trends in Power Consumption and Energy Efficiency for HPC

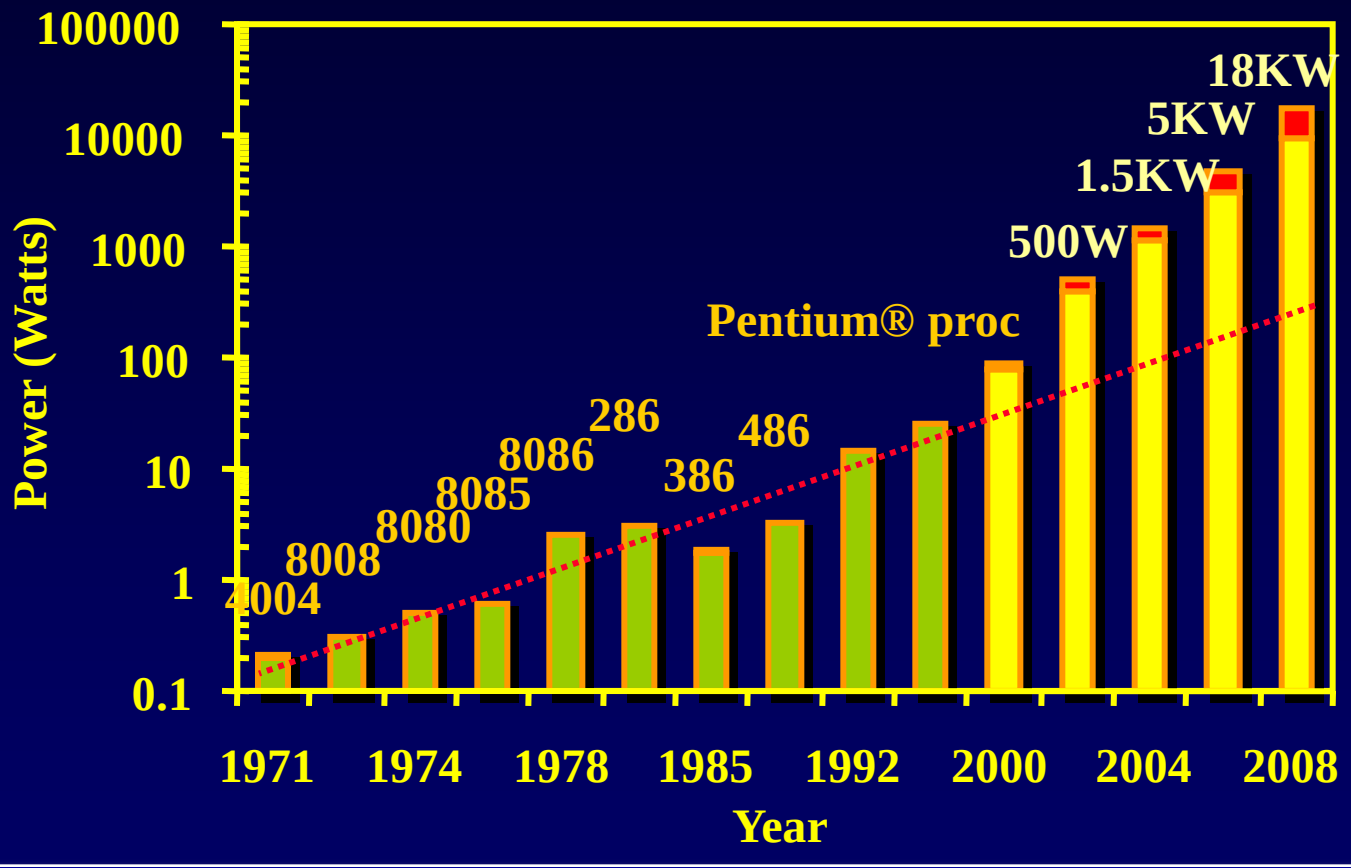
Building and infrastructure problem -- continued increase in demand for computing ("buildings")

Computer technology problem -- no more power density scaling ("bits")

An Early Warning

- **Presented by ShekharBorkar in Berkeley in November 2000**

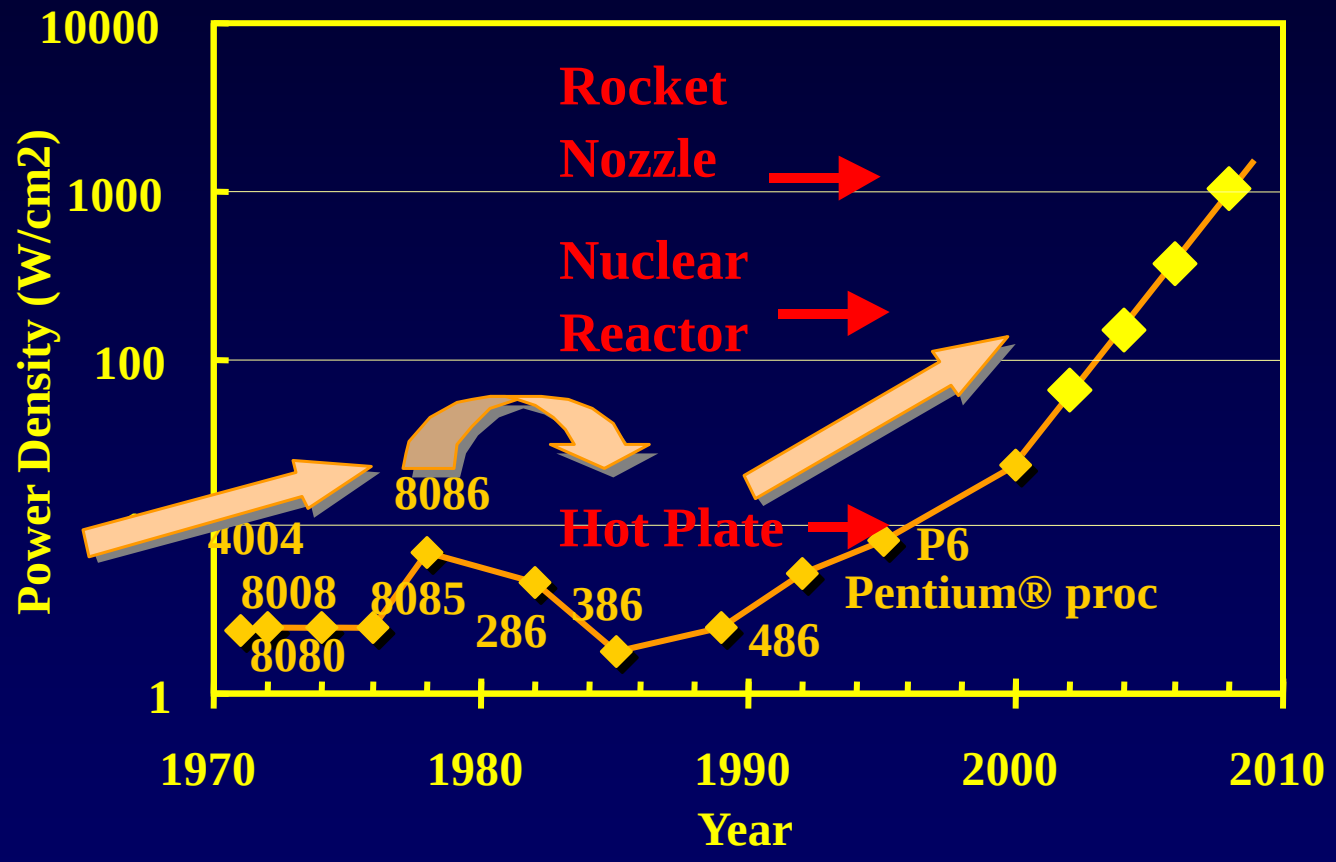
Power will be a problem



Power delivery and dissipation will be prohibitive

intel®

Power density will increase



Power density too high to keep junctions at low temp

intel®

Traditional Sources of Performance Improvement are Flat-Lining (2004)

- **New Constraints**
  - **15 years of *exponential* clock rate growth has ended**
- **Moore's Law reinterpreted:**
  - **How do we use all of those transistors to keep performance increasing at historical rates?**
  - **Industry Response: #cores per chip doubles every 18 months *instead* of clock frequency!**
  - **multicore**

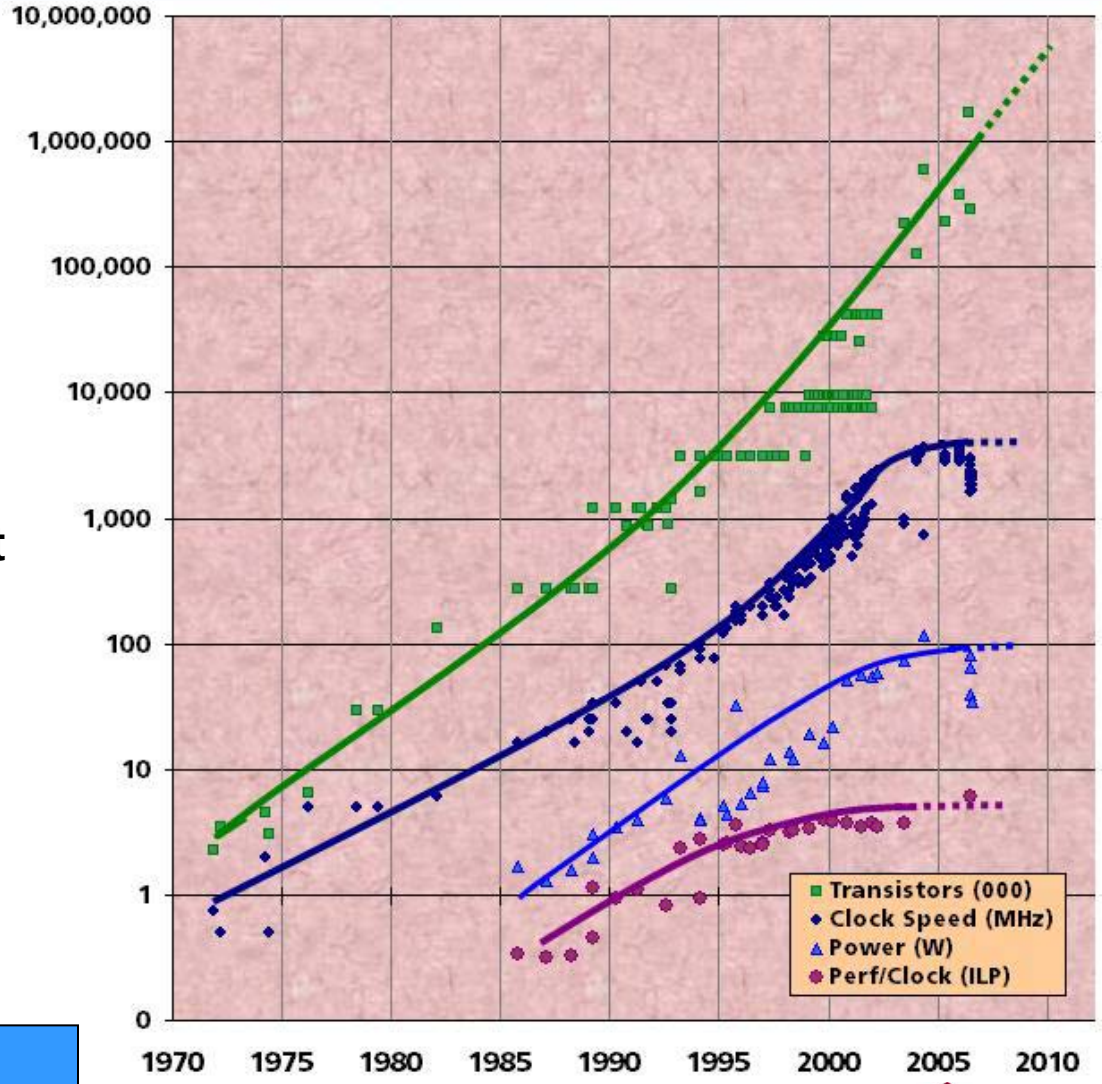
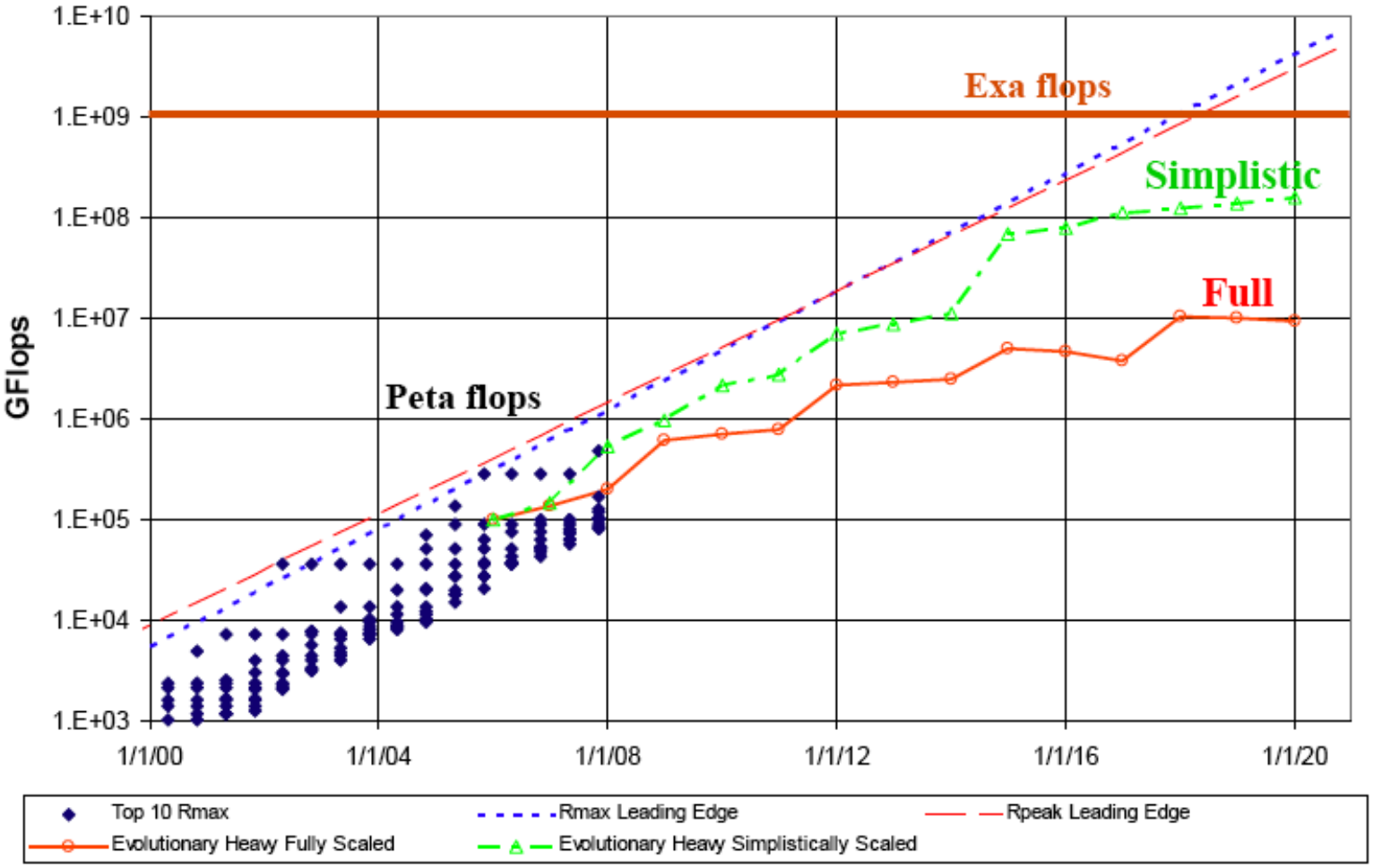


Figure courtesy of KunleOlukotun, Lance Hammond, Herb Sutter, and Burton Smith

DARPA Exascale Study

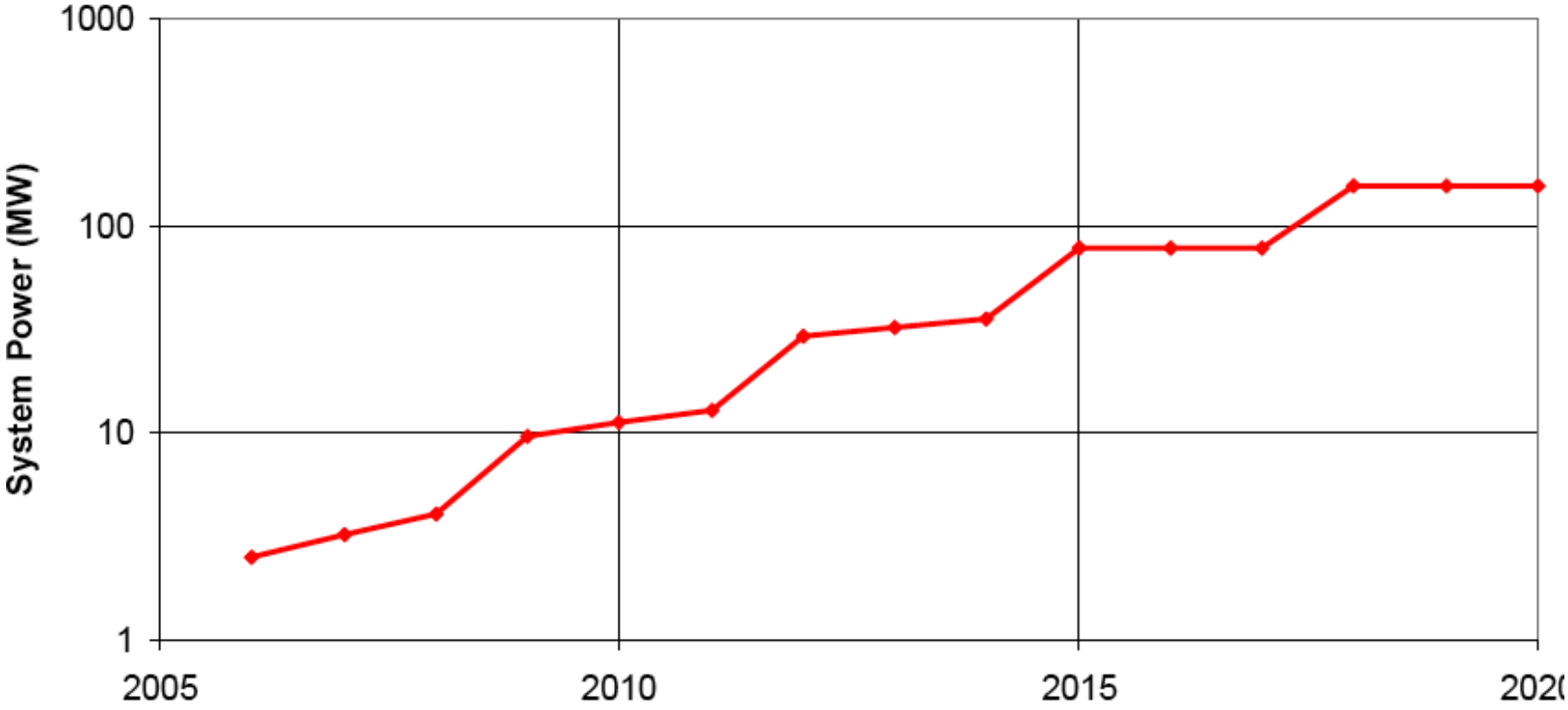
- **Commissioned by DARPA to explore the challenges for Exaflop computing**
- **Two model for future performance growth**
  - **Simplistic: ITRS roadmap; power for memory grows linear with #of chips; power for interconnect stays constant**
  - **Fully scaled: same as simplistic, but memory and router power grow with peak flops per chip**

We won't reach Exaflops with this approach



From Peter Kogge, DARPA Exascale Study

… and the power costs will still be staggering



From Peter Kogge,
DARPA Exascale Study

U.S. DEPARTMENT OF ENERGY

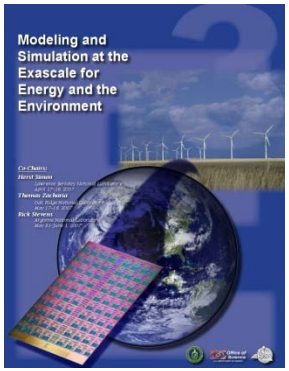
A decadal DOE plan for providing exascale applications and technologies for DOE mission needs

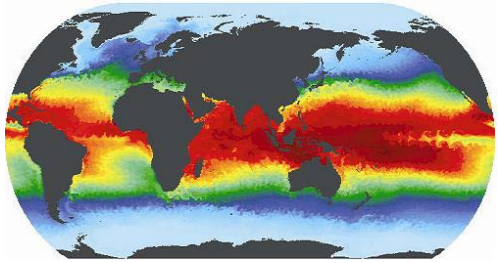
Rick Stevens and Andy White, co-chairs

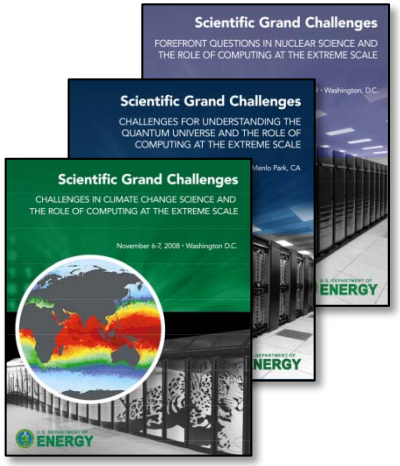
Pete Beckman, Ray Bair-ANL; Jim Hack, Jeff Nichols, Al Geist-ORNL; Horst Simon, Kathy Yelick, John Shalf-LBNL; Steve Ashby, Moe Khaleel-PNNL; Michel McCoy, Mark Seager, Brent Gorda-LLNL; John Morrison, Cheryl Wampler-LANL; James Peery, Sudip Dosanjh, Jim Ang-SNL; Jim Davenport, Tom Schlagel, BNL; Fred Johnson, Paul Messina, ex officio

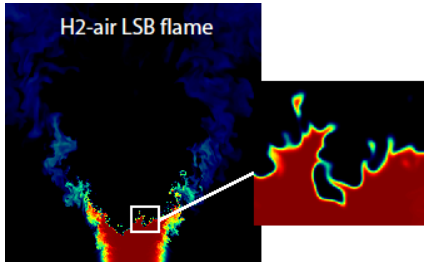
Process for identifying exascale applications and technology for DOE missions ensures broad community input

- Town Hall Meetings April-June 2007
- Scientific Grand Challenges Workshops Nov, 2008 – Oct, 2009
  - Climate Science (11/08),
  - High Energy Physics (12/08),
  - Nuclear Physics (1/09),
  - Fusion Energy (3/09),
  - Nuclear Energy (5/09),
  - Biology (8/09),
  - Material Science and Chemistry (8/09),
  - National Security (10/09)
  - Cross-cutting technologies (2/10)
- Exascale Steering Committee
  - "Denver" vendor NDA visits 8/2009
  - SC09 vendor feedback meetings
  - Extreme Architecture and Technology Workshop 12/2009
- International Exascale Software Project
  - Santa Fe, NM 4/2009; Paris, France 6/2009; Tsukuba, Japan 10/2009

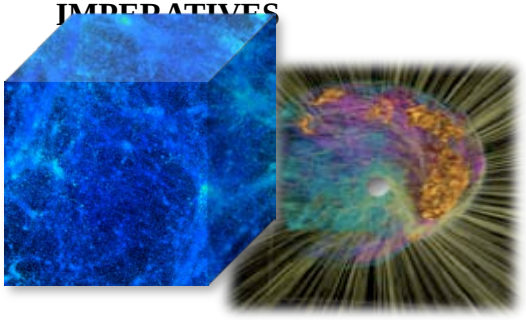








MISSION IMPERATIVES



FUNDAMENTAL SCIENCE

U.S. DEPARTMENT OF ENERGY

Potential System Architecture Targets

System attributes	2010	"2015"		"2018"	
System peak	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	6 MW	15 MW		20 MW	
System memory	0.3 PB	5 PB		32-64 PB	
Node performance	125 GF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW	25 GB/s	0.1TB/sec	1 TB/sec	0.4TB/sec	4 TB/sec
Node concurrency	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	18,700	50,000	5,000	1,000,000	100,000
Total Node Interconnect BW	1.5 GB/s	20 GB/sec		200GB/sec	
MTTI	days	O(1day)		O(1 day)	

U.S. DEPARTMENT OF ENERGY

What are critical exascale technology investments?

- **System power** is a first class constraint on exascale system performance and effectiveness.
- M**emory** is an important component of meeting exascale power and applications goals.
- **Programming model**. Early investment in several efforts to decide in 2013 on exascale programming model, allowing exemplar applications effective access to 2015 system for both mission and science.
- **Investment in exascale processor design** to achieve an exascale-like system in 2015.
- **Operating System strategy for exascale** is critical for node performance at scale and for efficient support of new programming models and run time systems.
- **Reliability and resiliency are critical at this** scale and require applications neutral movement of the file system (for check pointing, in particular) closer to the running apps.
- ***HPC co-design strategy and implementation*** *requires a set of a hierarchical performance models and simulators as well as commitment from apps, software and architecture communities.*

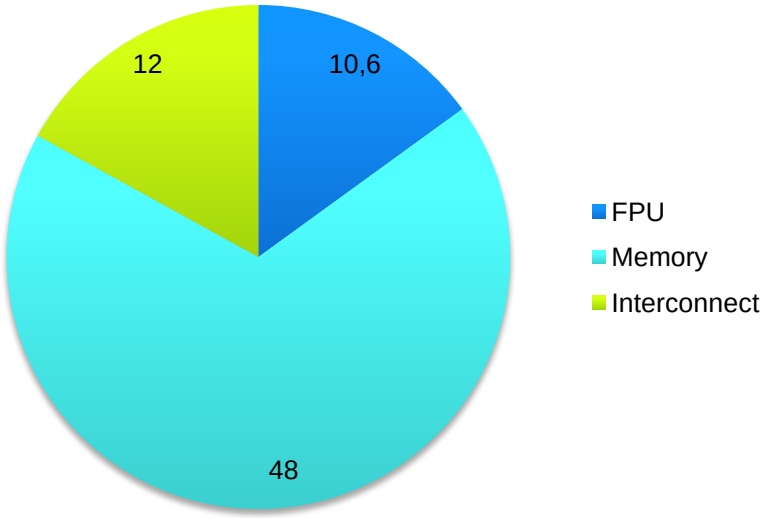
DOE Exascale Technology Roadmap

Key Observations from DOE Exascale Architecture and Technology Workshop, San Diego, Dec. 2009

- **supercomputers are power limited**
- **the biggest energy delta is off-chip data movement**

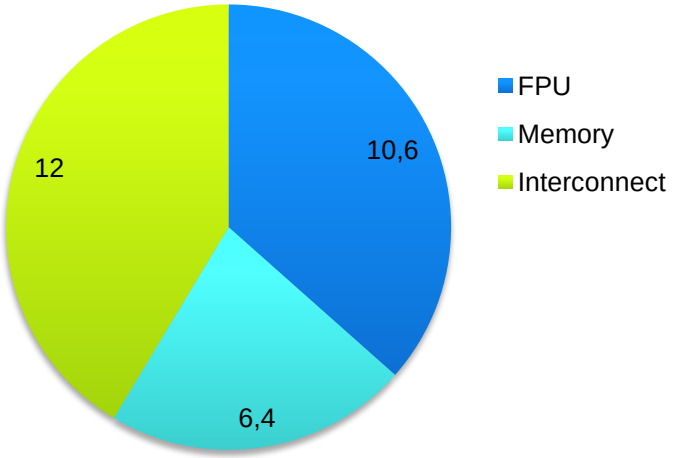
Memory Power Consumption

- Power Consumption with standard Technology Roadmap



70 Megawatts total

- Power Consumption with Investment in Advanced Memory Technology

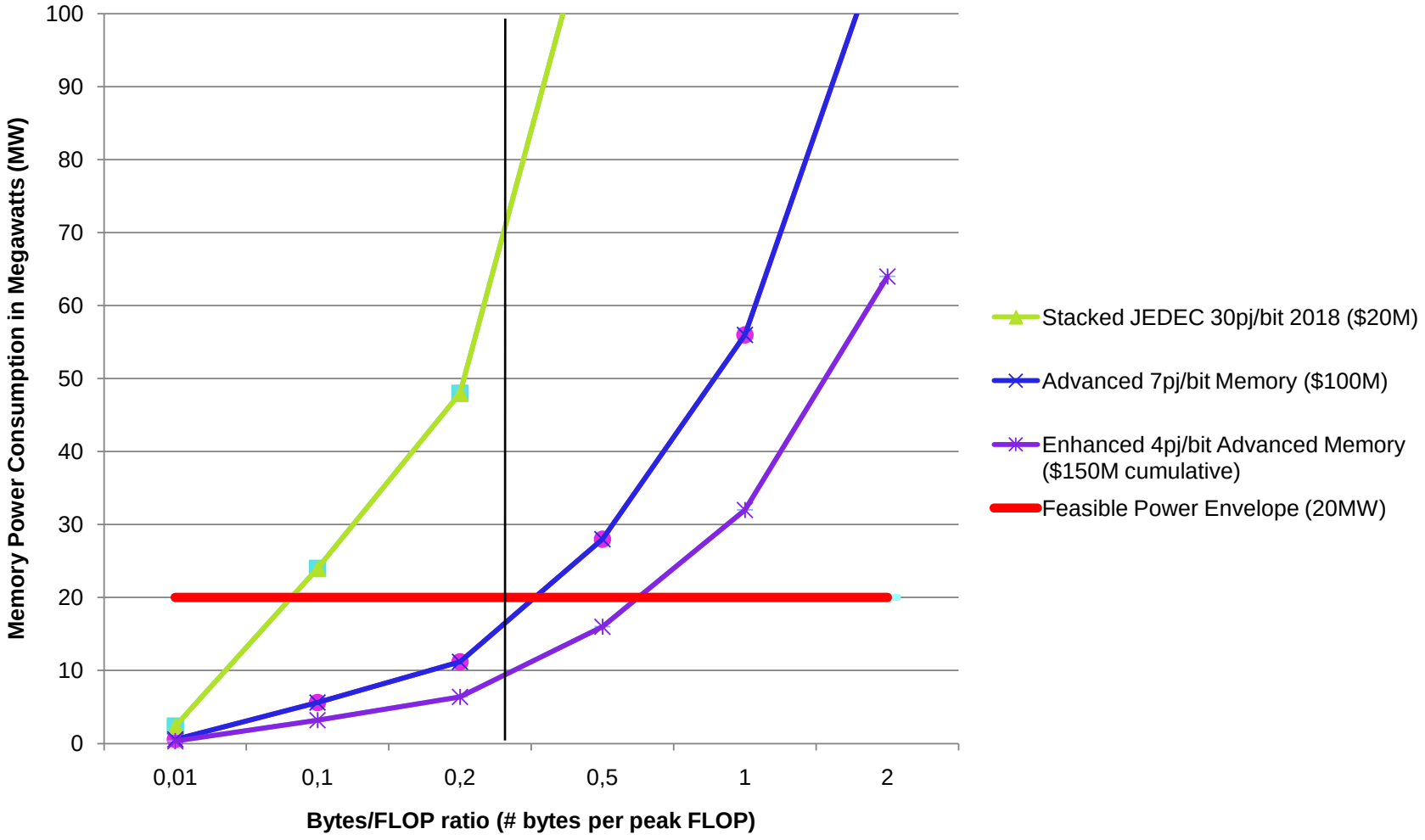


20 Megawatts total

	Phase I	Phase II	Phase III
Capacity/Cube:	512MB →	2GB	2GB
Bandwidth/Cube:	128 GB/s	128 GB/s →	256GB/s
Energy/bit:	10.0 pJ →	7.0 pJ →	5.0 pJ

U.S. DEPARTMENT OF ENERGY | Office of Science

Memory Technology
Bandwidth costs power



Green Flash: Ultra-Efficient Climate Modeling

- **Project by Shalf, Oliker, Wehner and others at LBNL**
- **A route to exascale computing**
  - **Target specific machine designs to answer a scientific question**
  - **Use of new technologies driven by the consumer market.**

Ultra-Efficient "Green Flash" Computing at NERSC: 100x over Business as Usual

Radically change HPC system development via application-driven hardware/software co-design

- ***Achieve 100x power efficiency and 100x capability of mainstream HPC approach for targeted high-impact applications***
- Accelerate development cycle for exascale HPC systems
- Approach is applicable to numerous scientific applications
- Proposed pilot application: Ultra-high resolution climate change simulation

Path to Power Efficiency

Reducing Waste in Computing

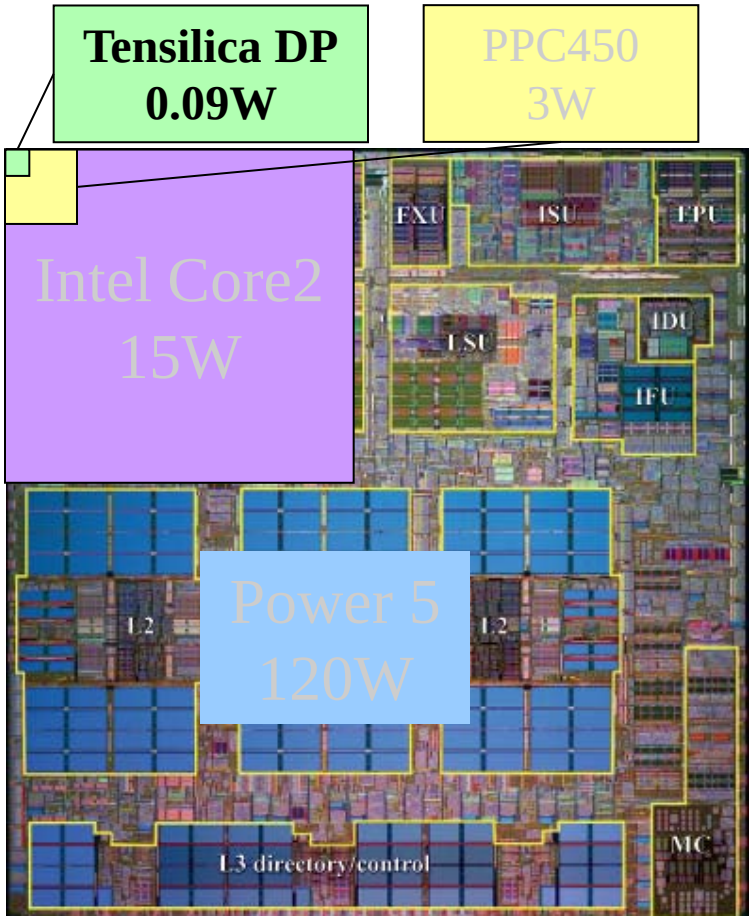
- **Examine methodology of low-power embedded computing market**
  - **optimized for low power, low cost and high computational efficiency**

"Years of research in low-power embedded computing have shown only one design technique to reduce power:reducewaste."

—Mark Horowitz, Stanford University &Rambus Inc.

- **Sources of waste**
  - **Wasted transistors (surface area)**
  - **Wasted computation (useless work/speculation/stalls)**
  - **Wasted bandwidth (data movement)**
  - **Designing for serial performance**

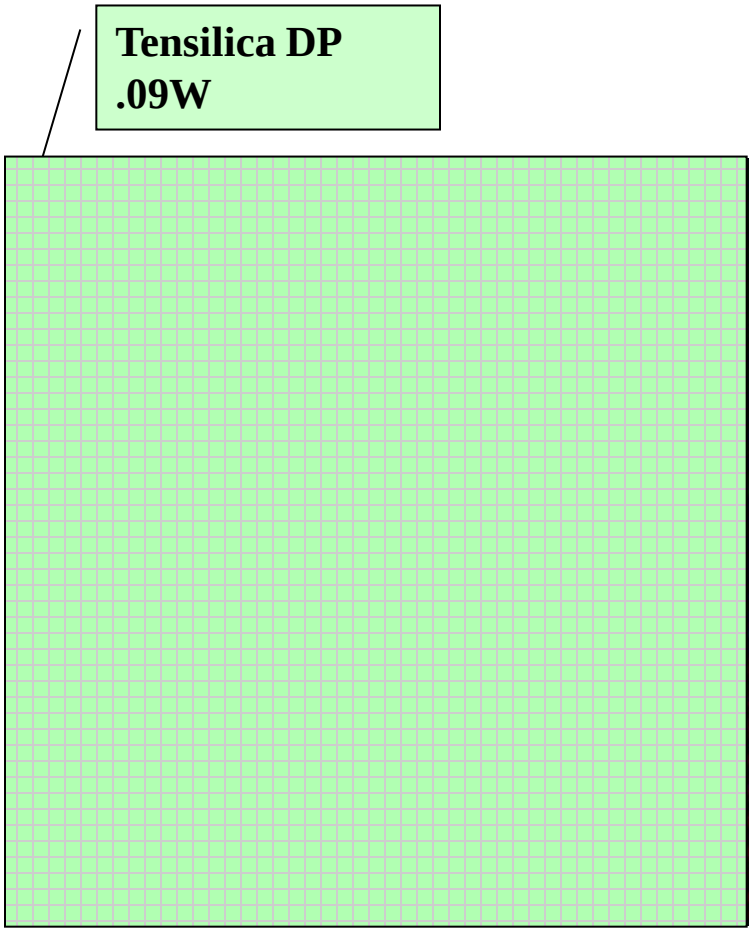
Design for Low Power: More Concurrency



- Cubic power improvement with lower clock rate due to V^2F
- Slower clock rates enable use of simpler cores
- Simpler cores use less area (lower leakage) and reduce cost
- Tailor design to application to reduce waste

This is how iPhones and MP3 players are designed to maximize battery life and minimize cost

Low Power Design Principles



- IBM Power5 (server)
  - 120W@1900MHz
  - Baseline
- Intel Core2 sc (laptop) :
  - 15W@1000MHz
  - *4x more FLOPs/watt than baseline*
- IBM PPC 450 (BG/P - low power)
  - 0.625W@800MHz
  - 90x more
- Tensilica XTensa (Moto Razor) :
  - 0.09W@600MHz
  - 400x more

Even if each core operates at 1/3 to 1/10th efficiency of largest chip, you can pack 100s more cores onto a chip and consume 1/20 the power

Green Flash Strawman System Design

We examined three different approaches (in 2008 technology)

Computation .015ºX.02ºX100L: 10 PFlops sustained, ~200 PFlops peak

- AMD Opteron: **Commodity approach, lower efficiency for scientific applications offset by cost efficiencies of mass market**
- BlueGene: **Generic embedded processor core and customize system-on-chip (SoC) to improve power efficiency for scientific applications**
- Tensilica XTensa: **Customized embedded CPU w/SoC provides further power efficiency benefits but maintains programmability**

Processor	Clock	Peak/ Core (Gflops)	Cores/ Socket	Sockets	Cores	Power	Cost 2008
AMD Opteron	2.8GHz	5.6	2	890K	1.7M	179 MW	$1B+
IBM BG/P	850MHz	3.4	4	740K	3.0M	20 MW	$1B+
Green Flash / Tensilica XTensa	650MHz	2.7	32	120K	4.0M	3 MW	$75M

Climate System Design Concept

Strawman Design Study

32 boards per rack

100 racks @ ~25KW

power + comms

32 chip + memory clusters per board (2.7 TFLOPS @ 700W

RAM RAM

Proc Array

RAM RAM

8 DRAM per processor chip: ~50 GB/s

VLIW CPU:

- 128b load-store + 2 DP MUL/ADD + integer op/ DMA per cycle:
- Synthesizable at 650MHz in commodity 65nm
- 1mm² core, 1.8-2.8mm² with inst cache, data cache data RAM, DMA interface, 0.25mW/MHz
- Double precision SIMD FP : 4 ops/cycle (2.7GFLOPs)
- Vectorizing compiler, cycle-accurate simulator, debugger GUI (Existing part of Tensilica Tool Set)
- 8 channel DMA for streaming from on/off chip DRAM
- Nearest neighbor 2D communications grid

32K I | 8 chan DMA

CPU

64-128K D 2x128b

External DRAM interface

Master Processor

External DRAM interface

Opt. 8MB embedded DRAM

External DRAM interface

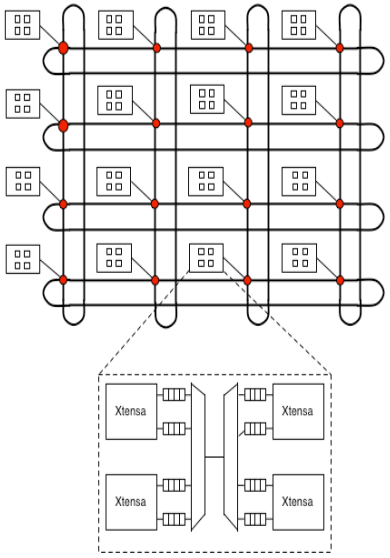
External DRAM interface

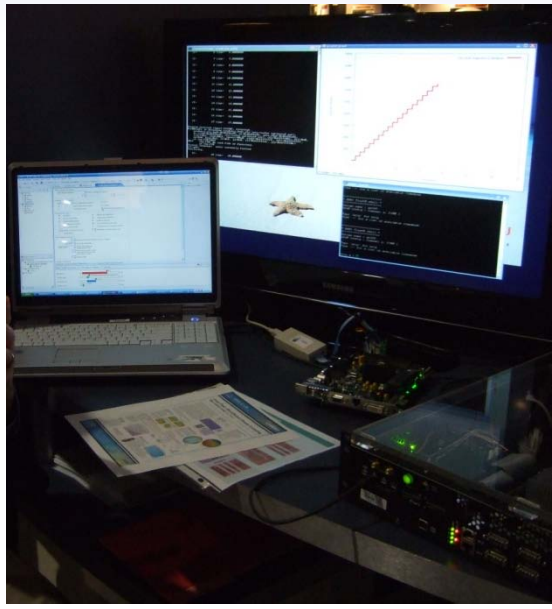
Comm Link Control

32 processors per 65nm chip 83 GFLOPS @ 7W

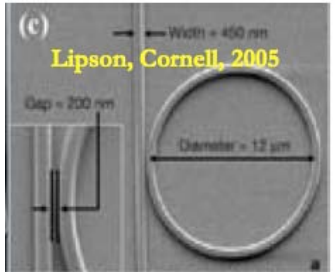
Green Flash Hardware Demo at SC08 and SC09

- **Demonstrated during SC ’08 and ‘09**
- **Proof of concept**
  - CSU atmospheric model ported to Tensilica Architecture
  - Single Tensilica processor running atmospheric model at 50MHz
- **Emulation performance advantage**
  - Processor running at 50MHz vs. Functional model at 100 kHz
  - 500x Speedup
- **Actual code running - not representative benchmark**

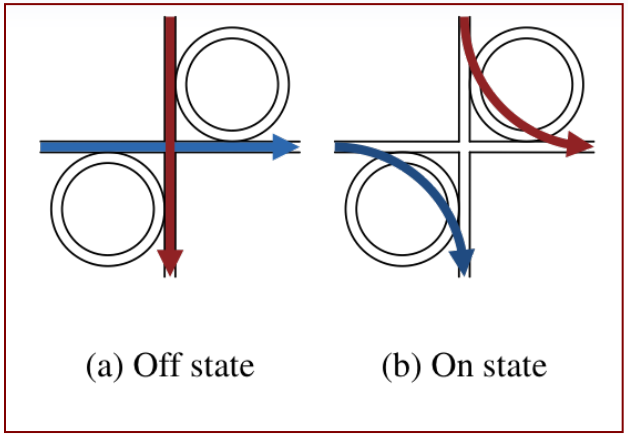




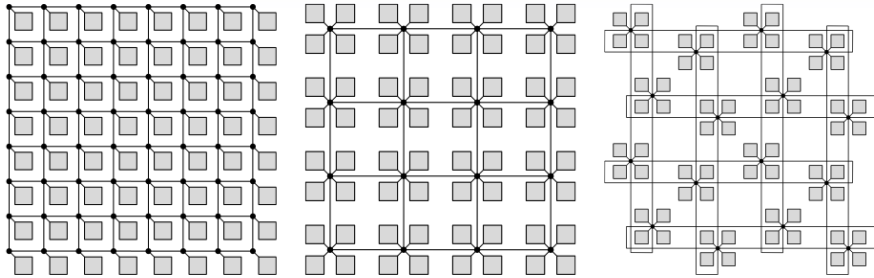
Silicon Photonics for Energy-Efficient Communication



Silicon Photonic Ring Resonator

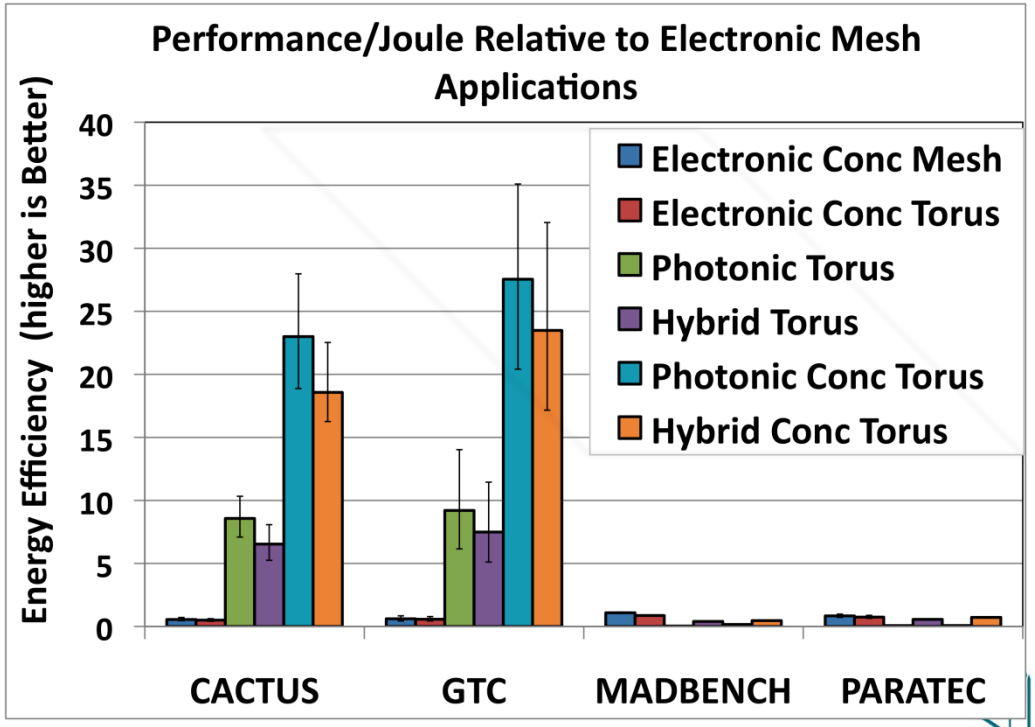


(a) Off state (b) On state



(a) Mesh (b) Concentrated Mesh (c) Concentrated Torus

- **Silicon photonics enables optics to be integrated with conventional CMOS**
- **Enables up to 27x improvement in communication energy efficiency!**



U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Summary

- Power consumption is a huge problem in HPC
  - "Bits": we may not be able to scale to Exaflops without new technologies
    - In particular need technology investment in memory
    - Be prepared for low byte/Flop ratios
  - "Buildings": we may have to spend more $$ on infrastructure and less on computing
    - Get ready for cloud computing as power consumption is going to change economics of computing

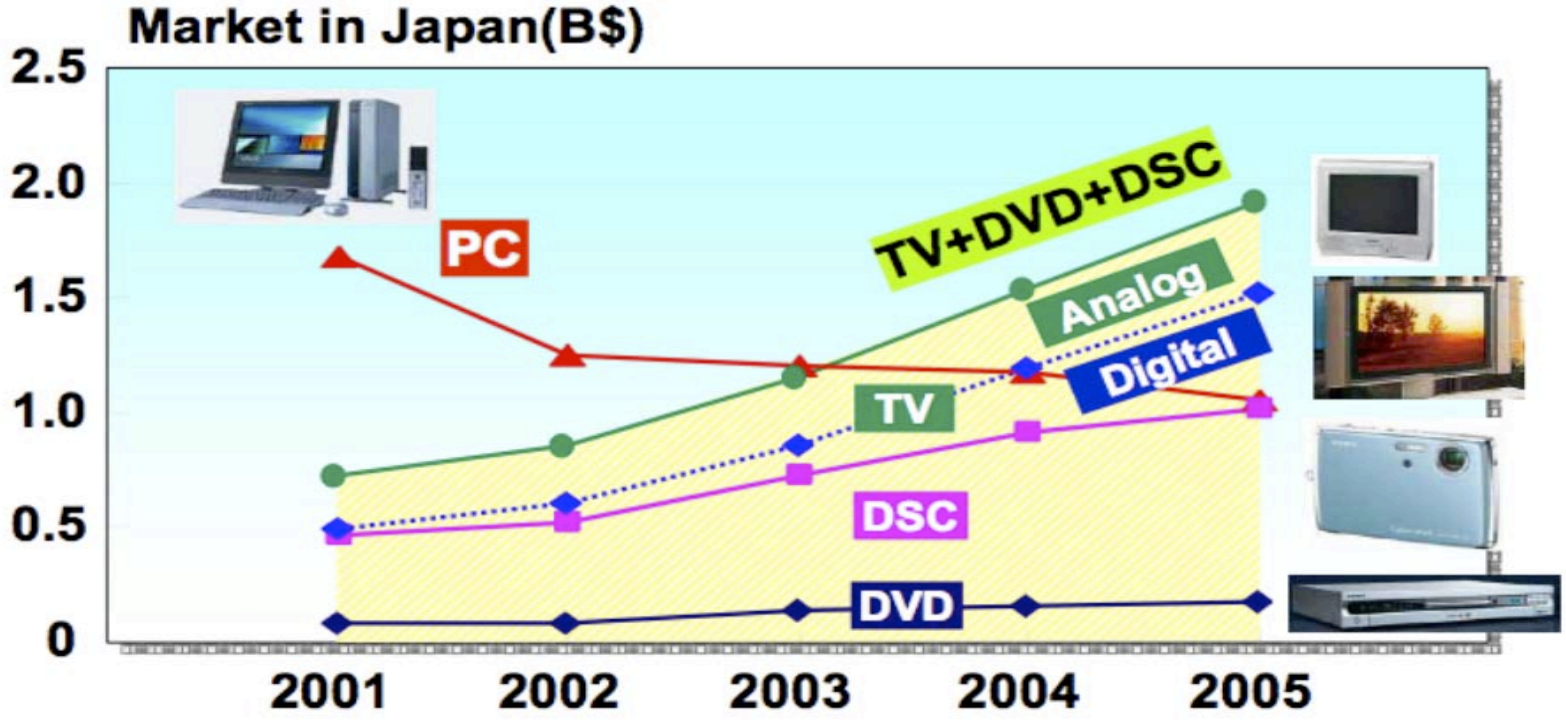
Extra Slides

Outline

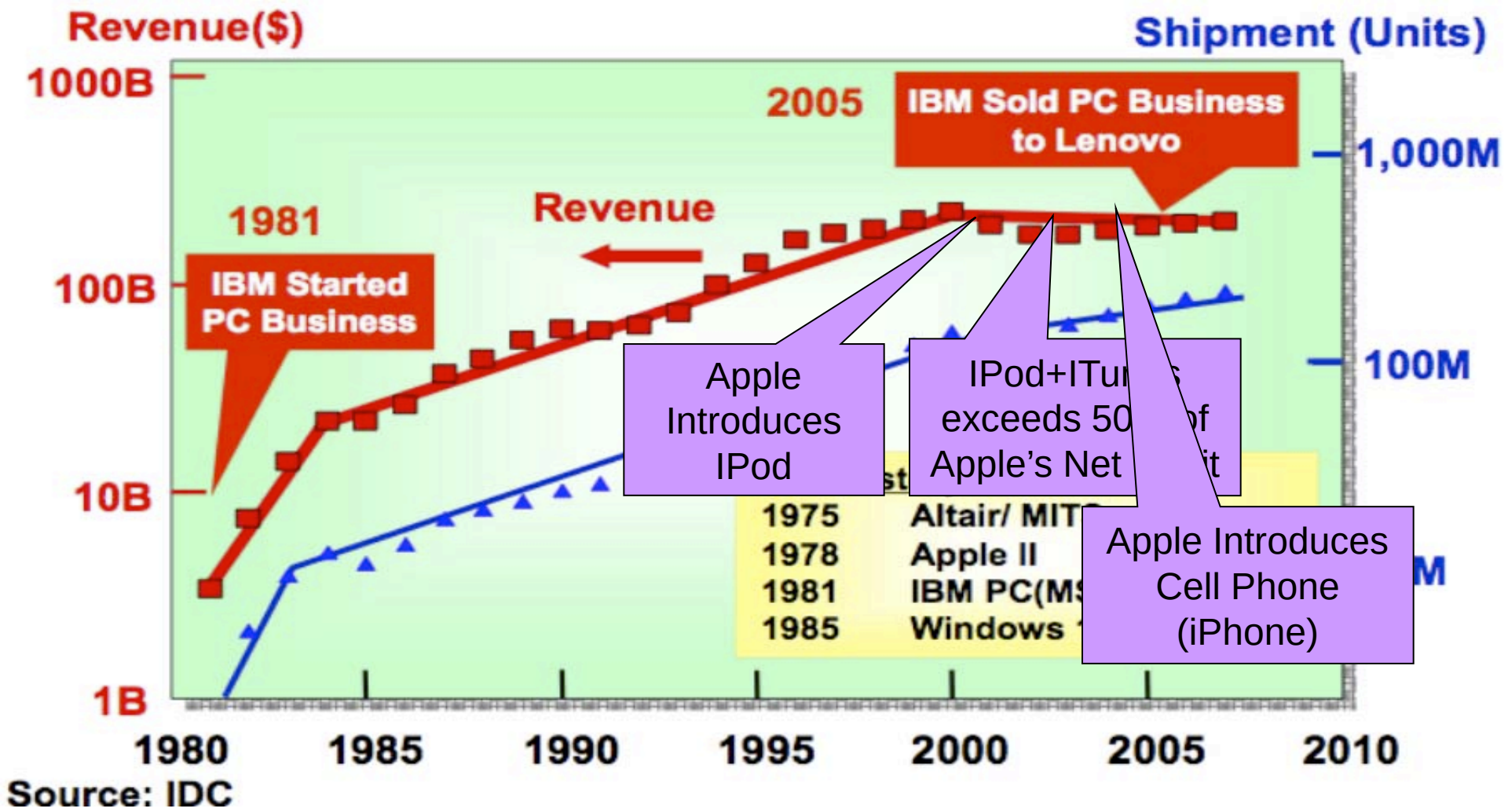
1. Power consumption has become an industry-wide issue for computing
2. Building and computer room energy efficiency
3. Computer architecture for energy efficiency- the Green Flash project
4. Future

Processor Technology Trend

- 1990s - R&D computing hardware dominated by desktop/COTS
  - **Had to learn how to use COTS technology for HPC**
- 2010 - R&D investments moving rapidly to consumer electronics/ embedded processing
  - **Must learn how to leverage embedded processor technology for future HPC systems**



Consumer Electronics has Replaced PCs as the Dominant Market Force in CPU Design!!



Source: IDC

Power fundamentals 2018--2020

Presented at STF Workshop, Sept. 2008 by Bill Camp, Intel

Processor budget: **15 MW** for a sustained HPL Exaflops (10pJ/op) {250}

Memory budget: **25– 50 MW** (25 pJ/op) {300}
[1/2 Byte/sec/Flops]

Interconnect budget: **50 MW** (5 pJ/op) [0.1 B/F] {30}

I/O Budget: **5 MW** (5 pJ/byte) 1 petabyte/sec

Power and Cooling Budget @30%: **30 MW**

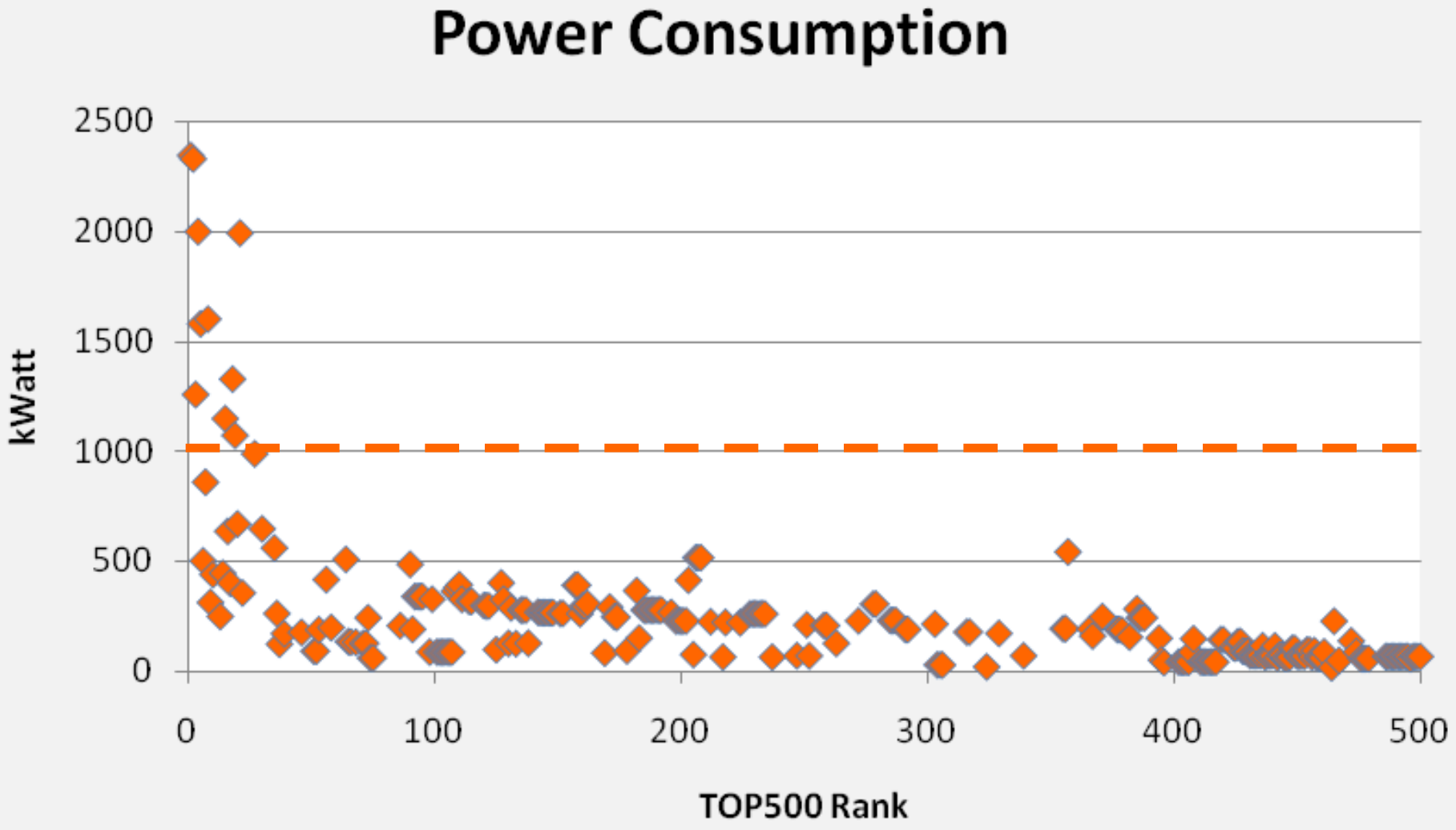
Total Power required 125 MW!

The Transition to Low-Power Technology is Inevitable

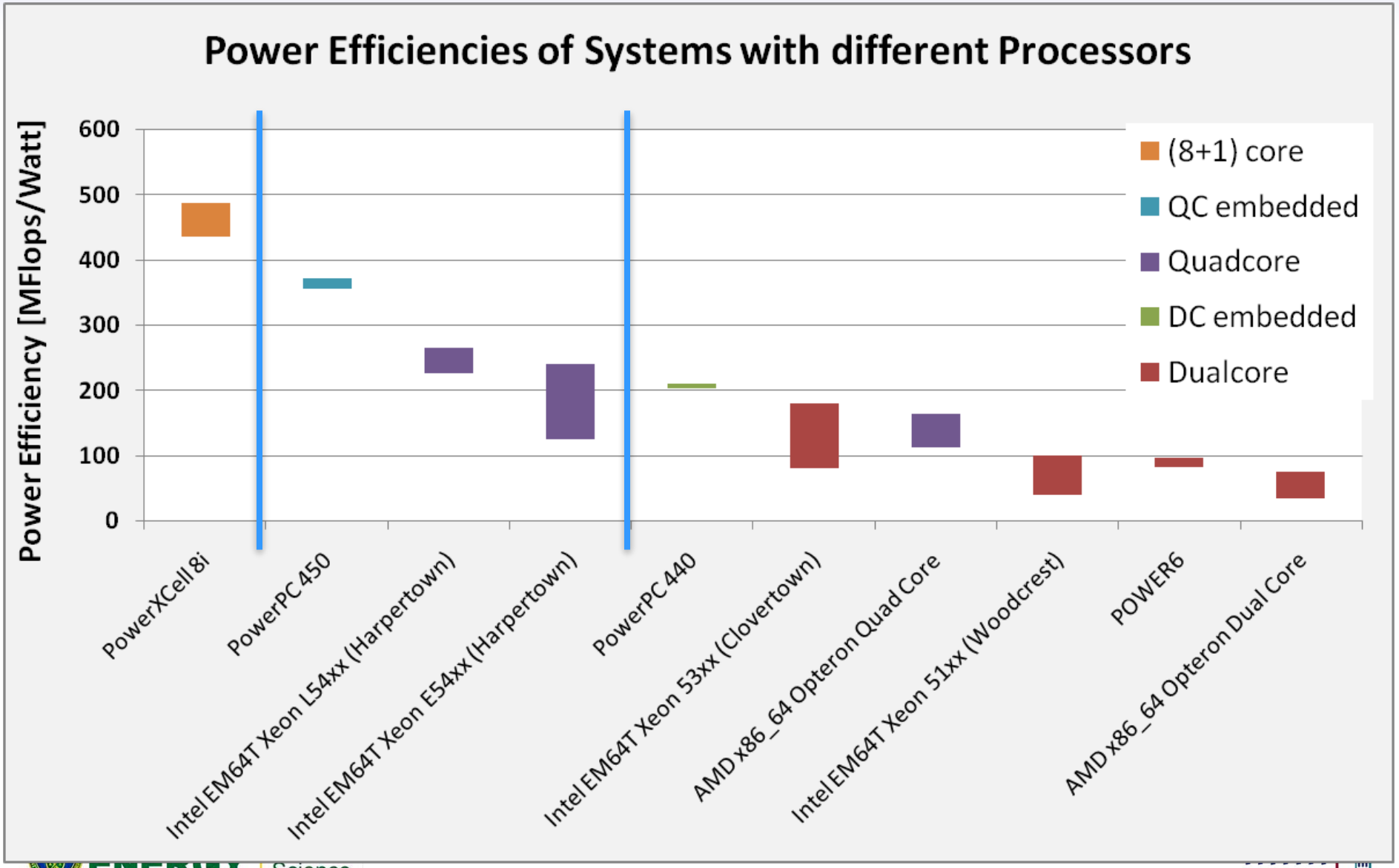
Does it make sense to build systems that require the electric power equivalent of an aluminum smelter?

- Information “factories” are only affordable for a few government labs and large commercial companies (Google, MSN, Yahoo …)
  - Midrange installations will soon hit the 1 - 2 MW wall, requiring costly new installations
  - Economics will change if operating expenses of a server exceed acquisition cost
- The industry will switch to low-power technology within 2 - 3 years
- Embedded processors or game processors will be the next step (BG, Cell, Nvidia, SiCortex, Tensilica)
  - Example RR, first Petaflops system

Absolute Power Levels

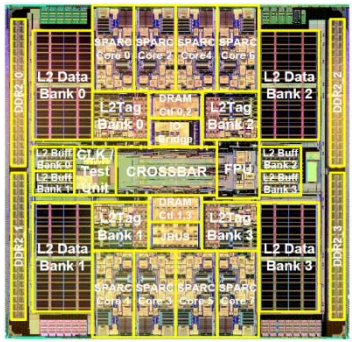


Power Efficiency related to Processors



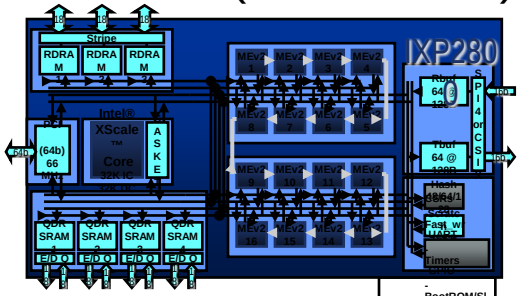
Convergence of Platforms

- Multiple parallel general-purpose processors (GPPs)
- Multiple application-specific processors (ASPs)

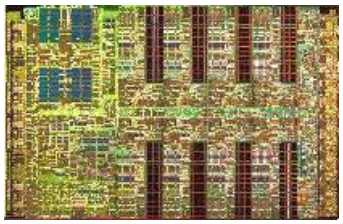


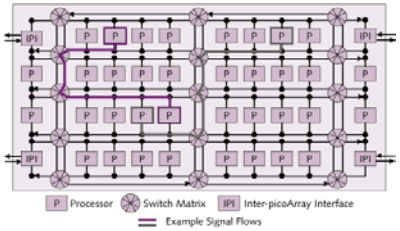
Sun Niagara
8 GPP cores (32 threads)

Intel Network Processor
1 GPP Core
16 ASPs (128 threads)



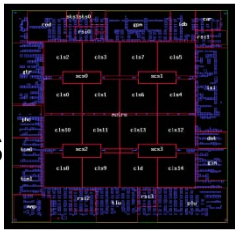
IBM Cell
1 GPP (2 threads)
8 ASPs

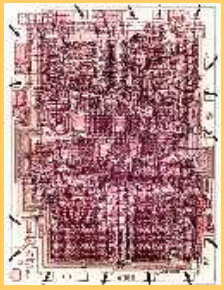




Picochip DSP
1 GPP core
248 ASPs

Cisco CRS-1
188 Tensilica GPPs





Intel 4004 (1971):
4-bit processor,
2312 transistors,
~100 KIPS,
10 micron PMOS,
11 mm^2 chip

1000s of processor cores per die

"The Processor is the new Transistor" [Rowen]

Summary (1)

- **LBNL has taken a comprehensive approach to the power in computing problem**
  - **Component level (investigate use of low-power components and build new system)**
  - **System level (measuring and understanding energy consumption of system**
  - **Computer Room level (understand airflow and cooling technology)**
  - **Building Level (enforce rigorous energy standards in new computer building and use of innovative energy savings technology)**

Summary (2)

- **Economic factors are driving us already to more energy efficient solutions in computing**
- **Incremental improvements are well on track, but we may ultimately need revolutionary new technology to reach the Exaflop/s level and beyond**

Outline

1. **Power consumption has become an industry-wide issue for computing**
2. **Building and computer room energy efficiency**
3. **Computer architecture for energy efficiency- the Green Flash project**
4. **Towards a better understanding of "green computing"**