The Greening of HPC - Will Power Consumption Become the Limiting Factor for Future Growth in HPC?

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

August 21, 2008

Acknowledgements

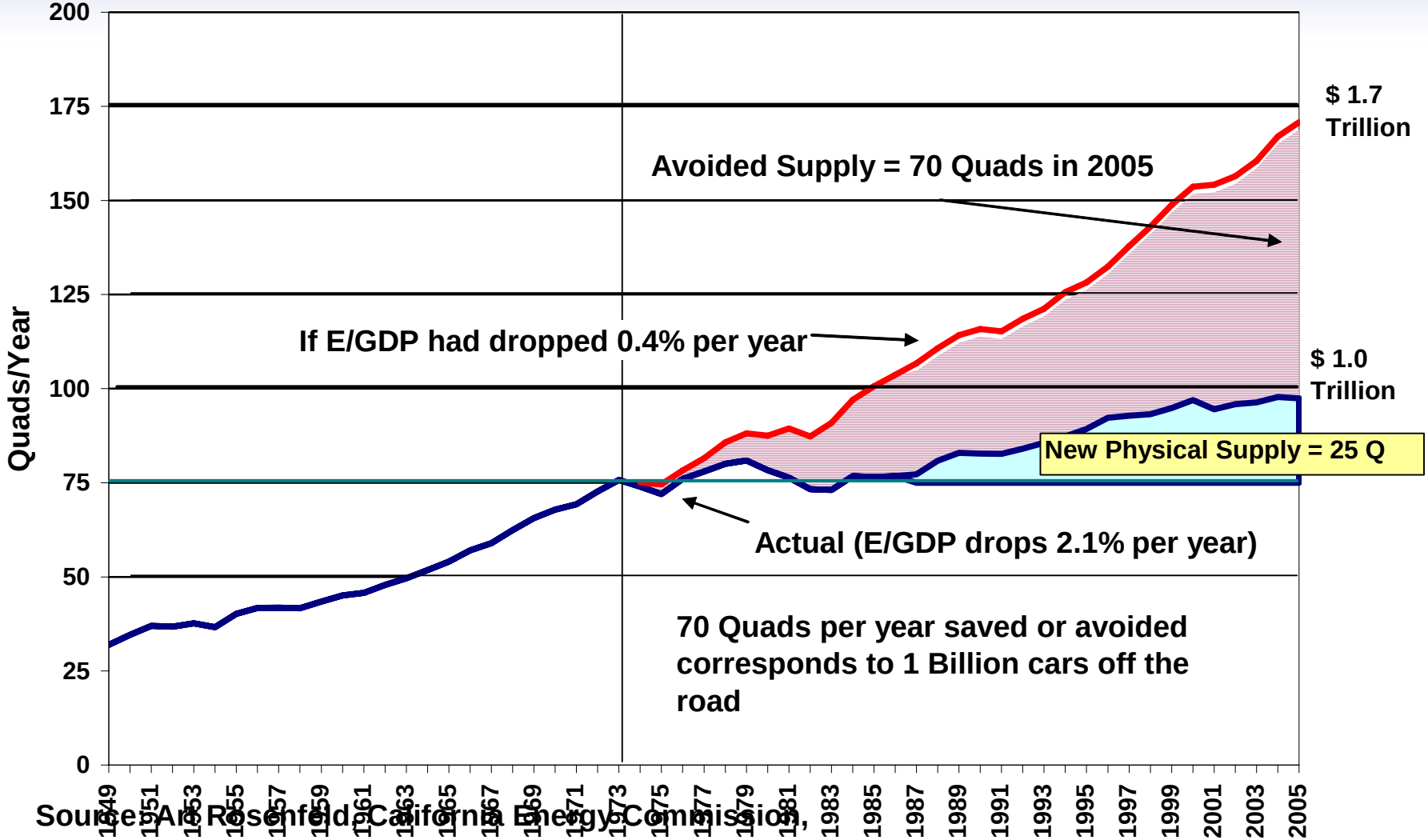
A large number of individuals have contributed to energy efficiency in computing at the Lab and to this presentation:

David Bailey (CRD), Michael Banda (CRD), Michael Bennett (ITD), Shoaib Kamil (CRD), Jonathan Koomey (Stanford), Chuck McParland (CRD), Bruce Nordman (EETD), Lenny Oliker (CRD), Ekow Otoo (CRD), Vern Paxson (UCB/ICSI/CRD), Doron Rotem (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

Almost all Berkeley resources about eneregy efficiency are available at http://www.lbl.gov/CS/html/energy%20efficient%20computing.html

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

Outline

1. Power consumption has become an industry-wide issue for computing
2. Understanding the problem
3. Building and computer room energy efficiency
4. Computer architecture for energy efficiency
5. Future plans

The Problem

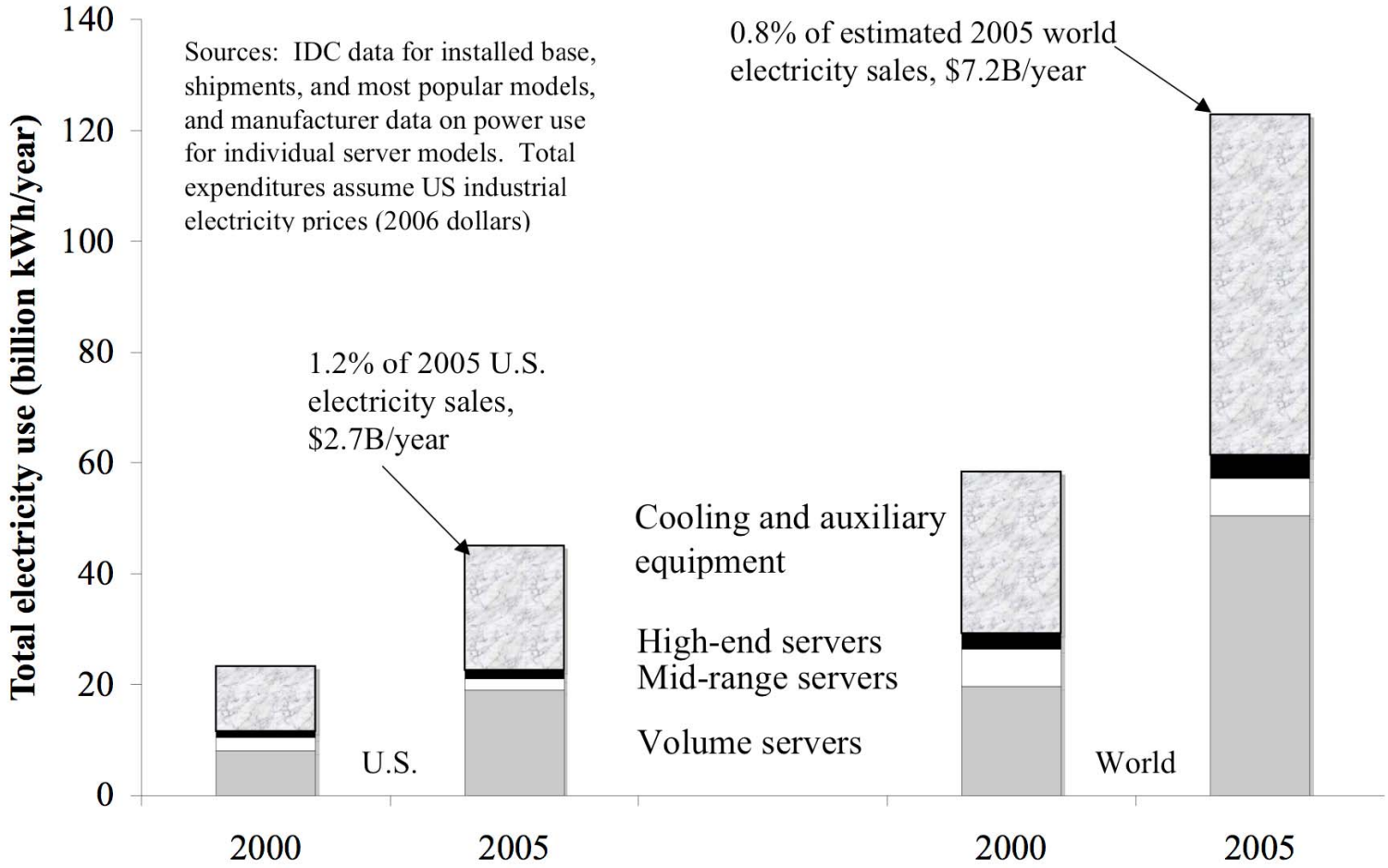
- **"Big IT" – all electronics**
  - **PCs / etc., consumer electronics, telephony**
    - Residential, commercial, industrial
  - **More than 200 TWh/year**
  - **$16 billion/year**
    - Based on .08$/KWh
  - **Nearly 150 million tons of CO_2 per year**
    - Roughly equivalent to 30 million cars!

Numbers represent U.S. only

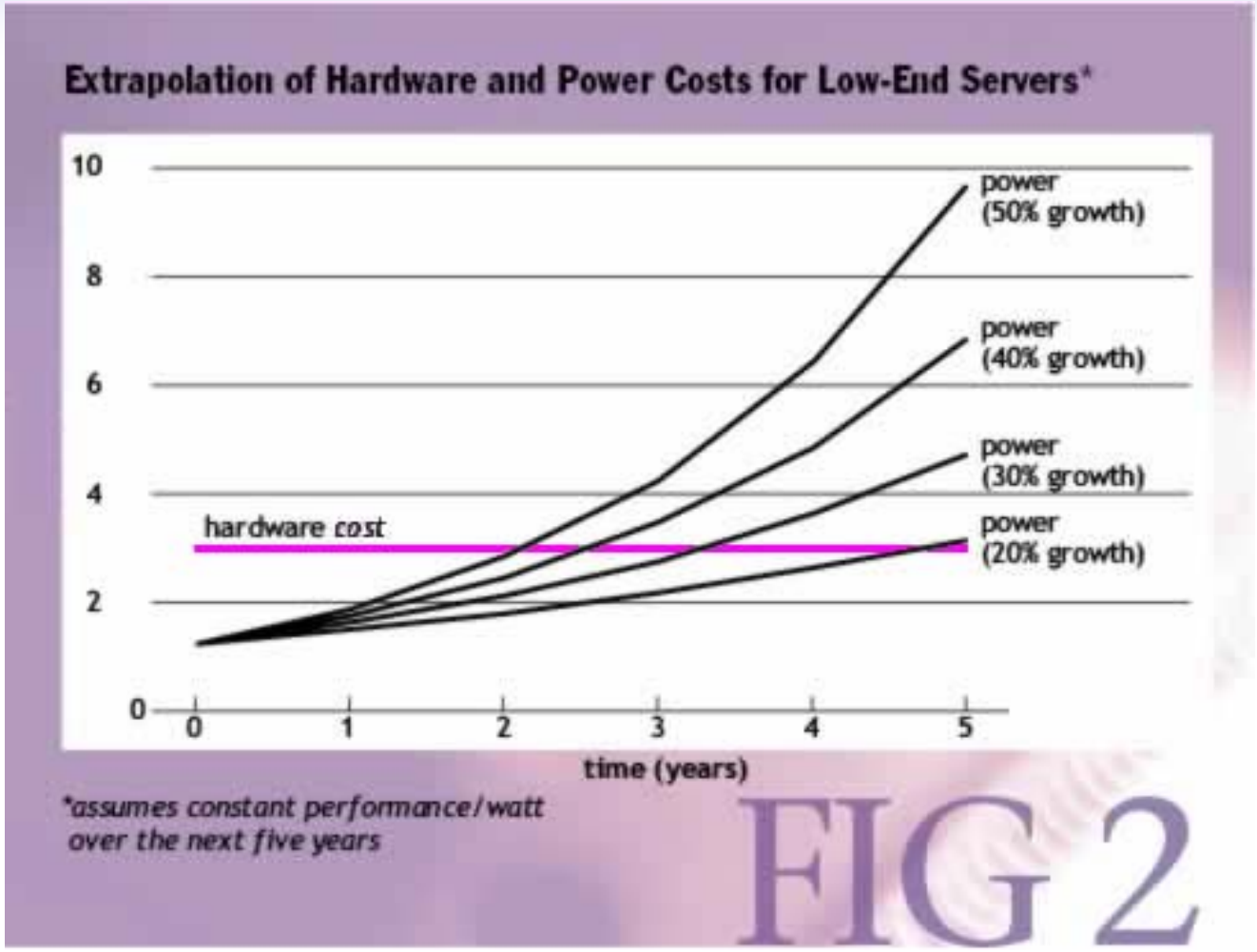
One central baseload power plant (about 7 TWh/yr)

… 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



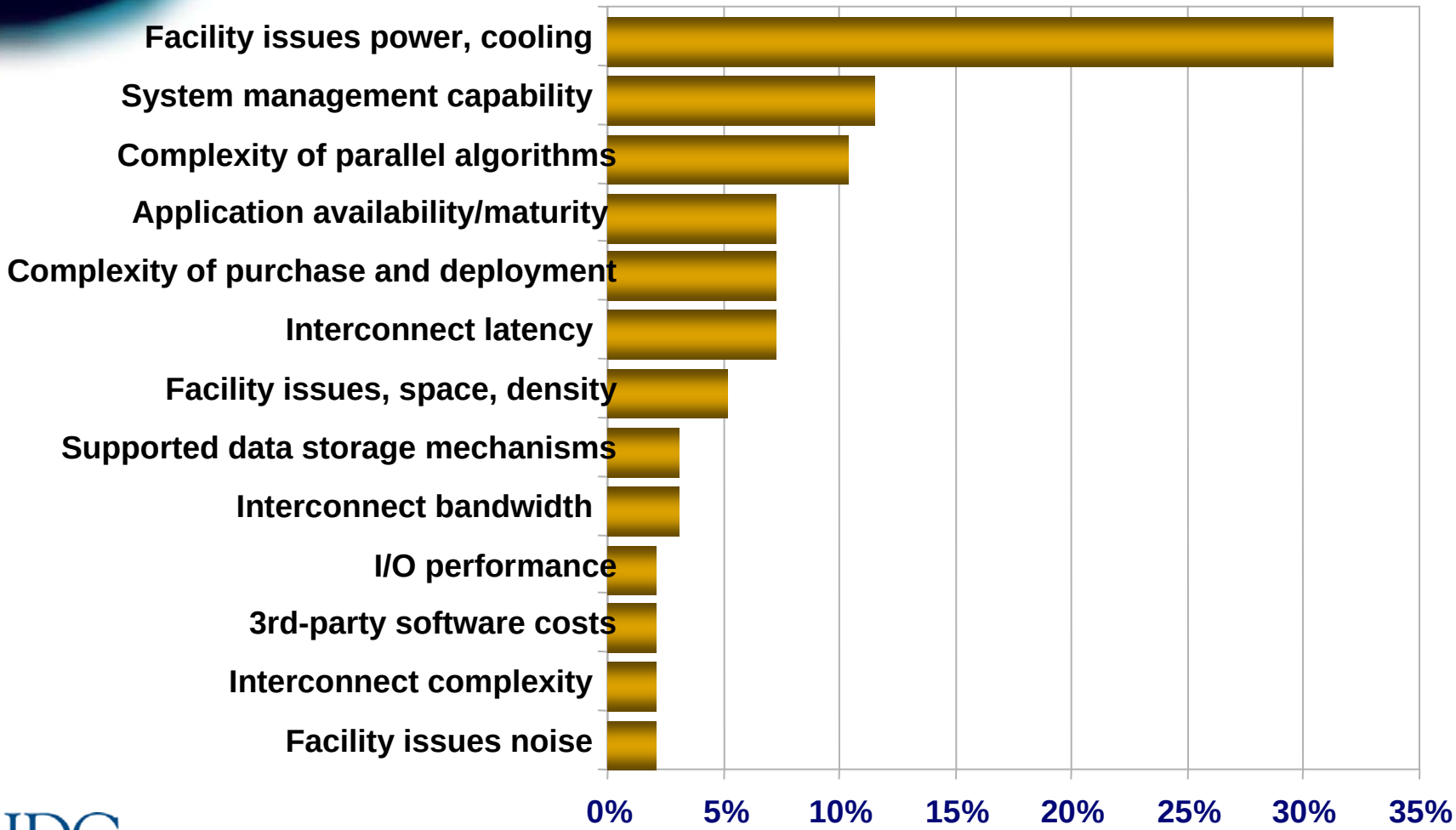
The Problem



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

Combined facilities score — 38.6

©2007 IDC

Even Consumers See the HPC Heat Issue

76-inch
HDTV:
200 watts

Video game
with
IBM Cell BE
processor:
380 watts,
twice what the
TV uses!

Source: John Gustafson, ClearSpeed

NERSC

BERKELEY LAB

Office of Science
U.S. DEPARTMENT OF ENERGY

Data Center Economic Reality

- **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"

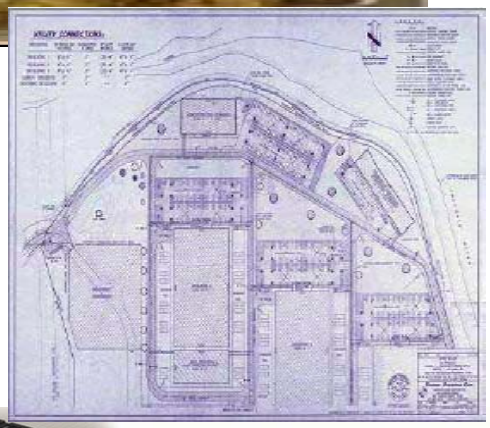
    "Hiding in Plain Sight, Google Seeks More Power"
    by John Markoff, NYT, June 14, 2006

- **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 Quincy, Wash.
470,000 Sq Ft, 47MW!

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

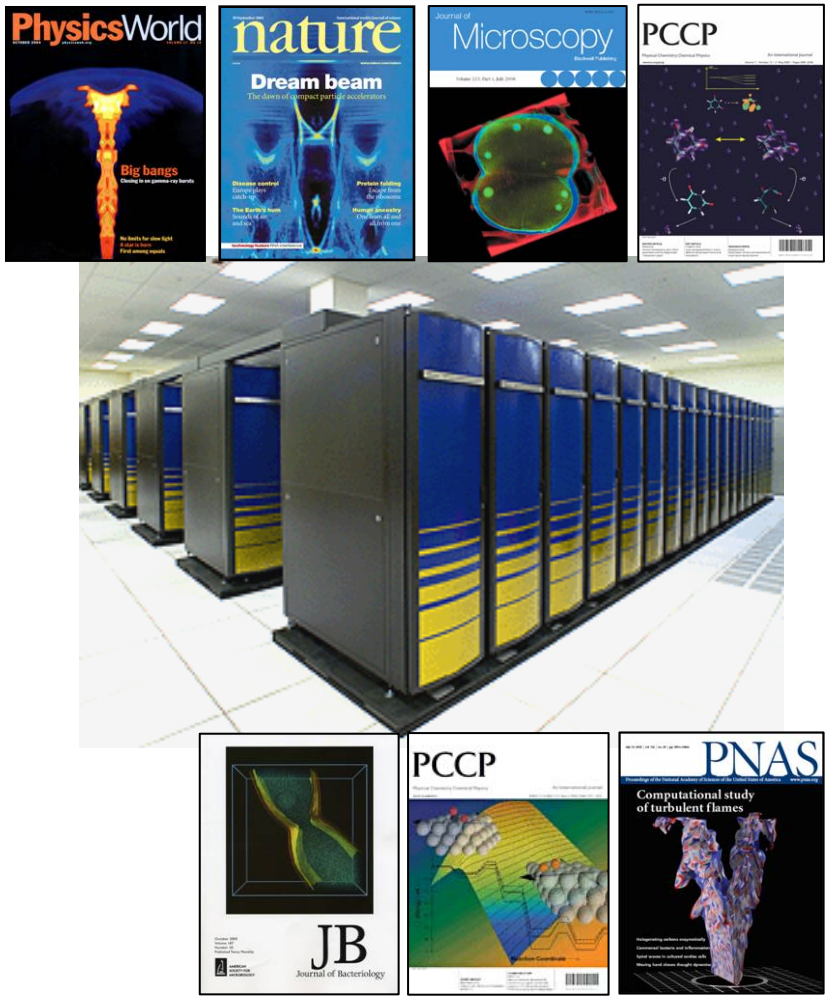
Outline

1. Power consumption has become an industry-wide issue for computing
2. Understanding the problem for HPC
3. Building and computer room energy efficiency
4. Computer architecture for energy efficiency
5. Future plans

Understanding Power Consumption in HPC Systems

- Until about 2 - 3 years ago there has been a complete lack of interest in power-related issues
- There are virtually no data and no methodology to address power issues in computer architecture
- Project at NERSC/LBNL to develop measurement standards and better quantitative understanding

New 100 Tflops Cray XT-4 at NERSC



Cray XT-4 "Franklin"

19,344 compute cores

102 Tflop/sec peak

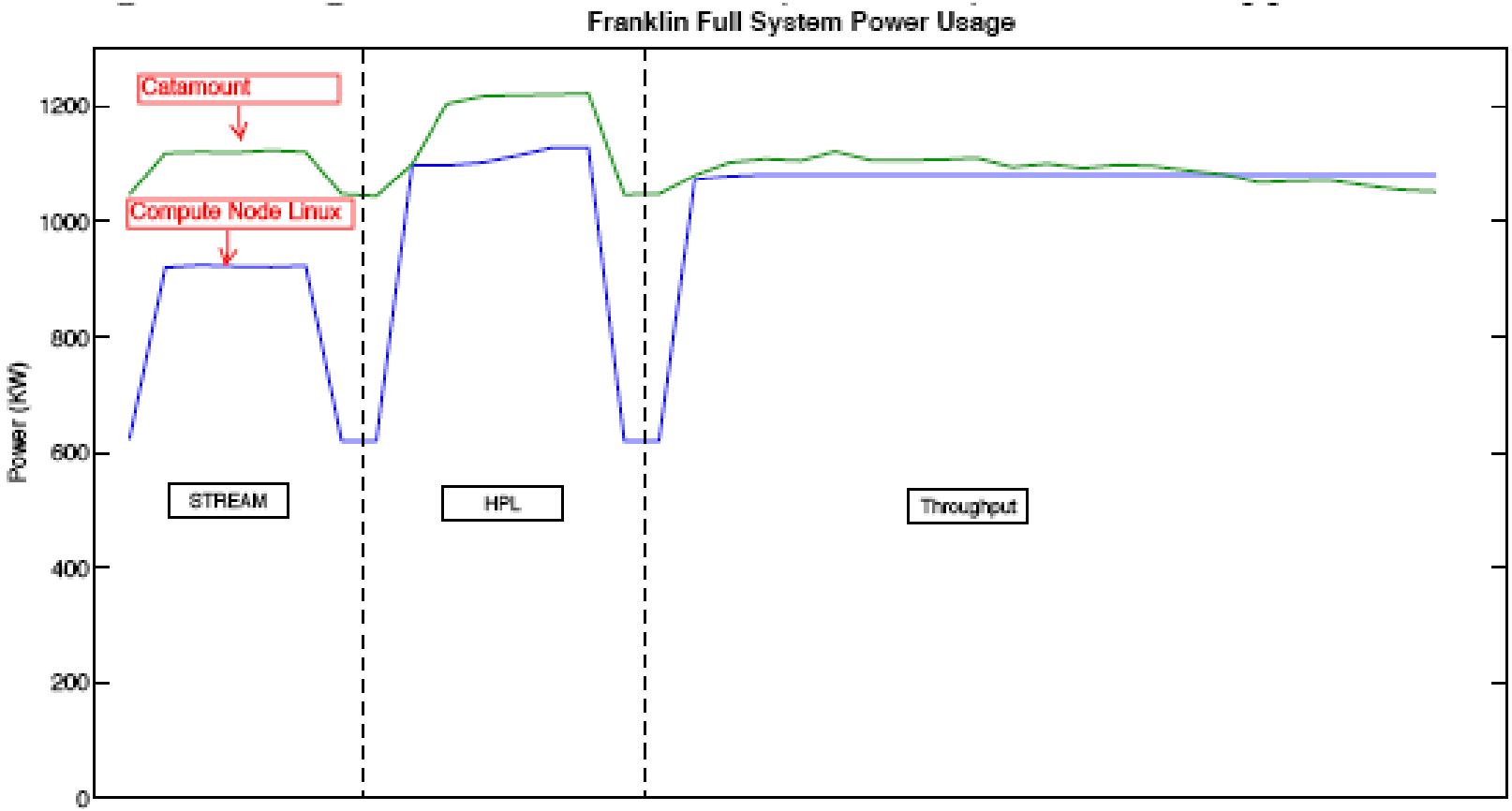
39 TB memory

350 TB usable disk space

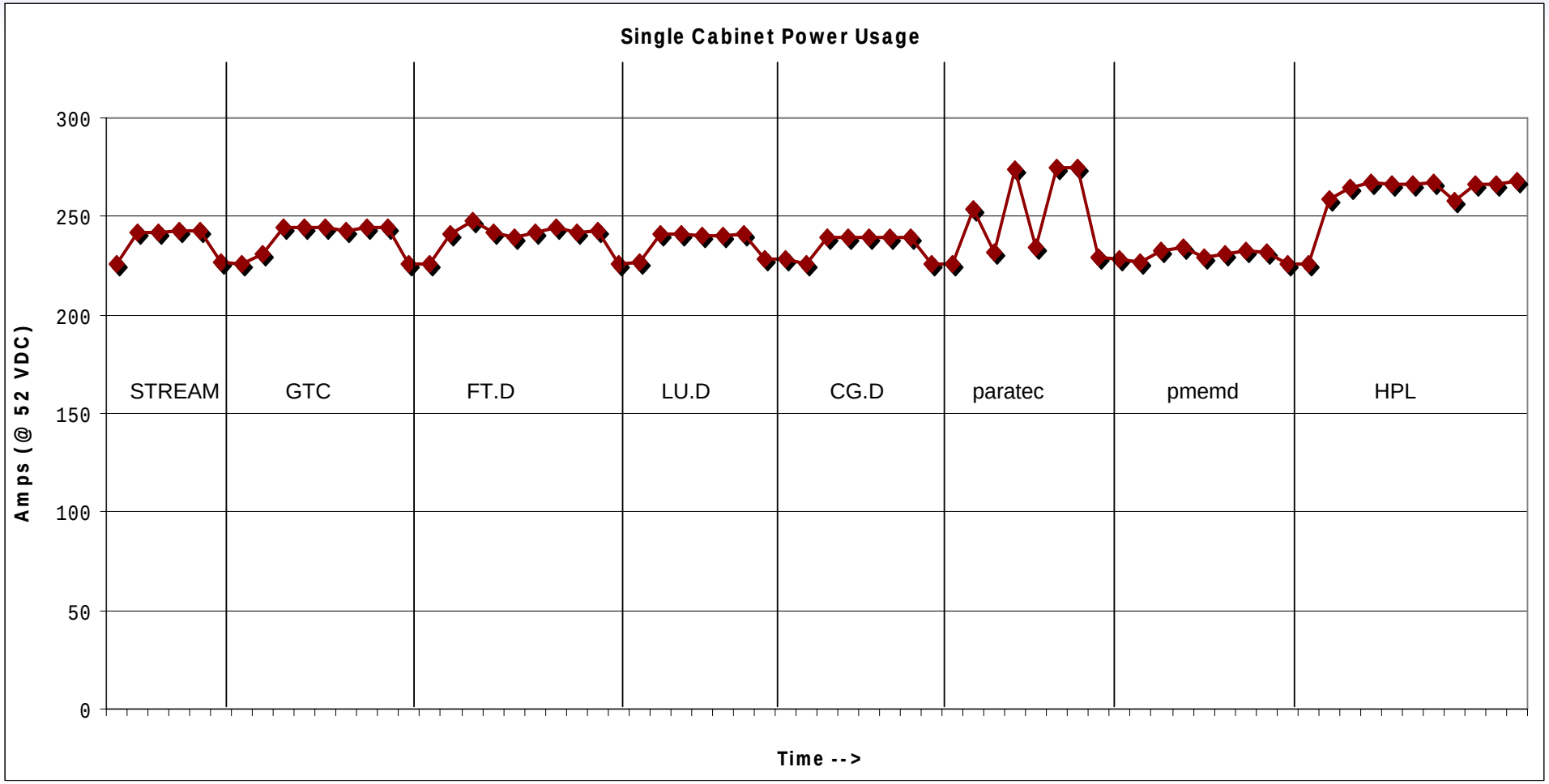
50 PB storage archive

NERSC is enabling new science

Franklin Full System Power Usage



Single Rack Tests



- Administrative utility gives rack DC amps & voltage
- HPL & Paratec are highest power usage

Conclusions

- **Power utilization under an HPL/Linpack load is a good estimator for power usage under mixed workloads for single nodes, cabinets / clusters, and large scale systems**
  - Idle power is not
  - Nameplate and CPU power are not

- **LINPACK running on one node or rack consumes approximately same power as the node would consume if it were part of full-system parallel LINPACK job**

- **We can estimate overall power usage using a subset of the entire HPC system and extrapolating to total number of nodes using a variety of power measurement techniques**
  - And the estimates mostly agree with one another!
  - Details see paper by Kamil, Shalf, and Strohmaier, IPDPS, 2008

Outline

1. Power consumption has become an industry-wide issue for computing
2. Understanding the problem
3. Building and computer room energy efficiency
4. Computer architecture for energy efficiency
5. Future plans

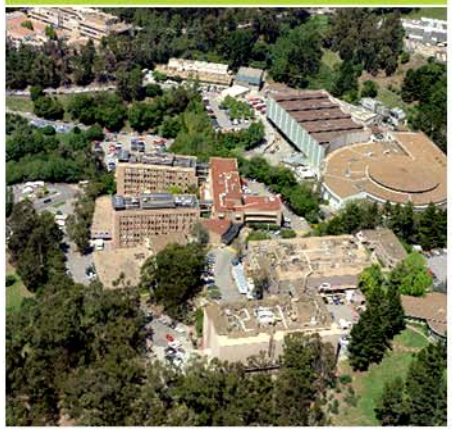
Understanding Power Consumption in HPC Computer Room Environment

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

- EETD has long-term experience in computer room energy efficiency for data centers (power distribution, air flow, cooling technology)
- EETD is the nationwide expert in this technology
- However, usage patterns are significantly different in HPC centers
- EETD partnered with NERSC and ITD to understand and improve computer room issues for HPC centers (e.g., instrumentation of LBNL computer room with wireless sensors)

PERKINS
+ WILL

Ideas + buildings
that honor the broader
goals of society



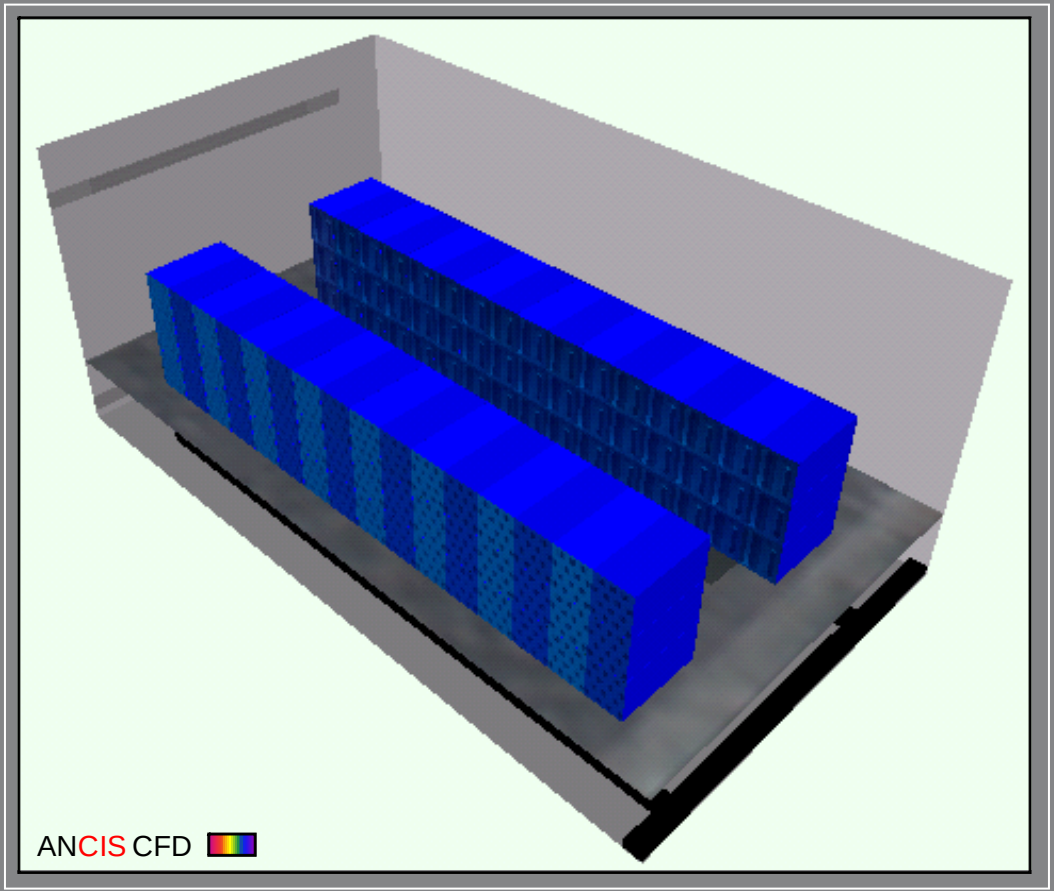
Computational Research & Theory Facility - LBNL

50% Schematic Design

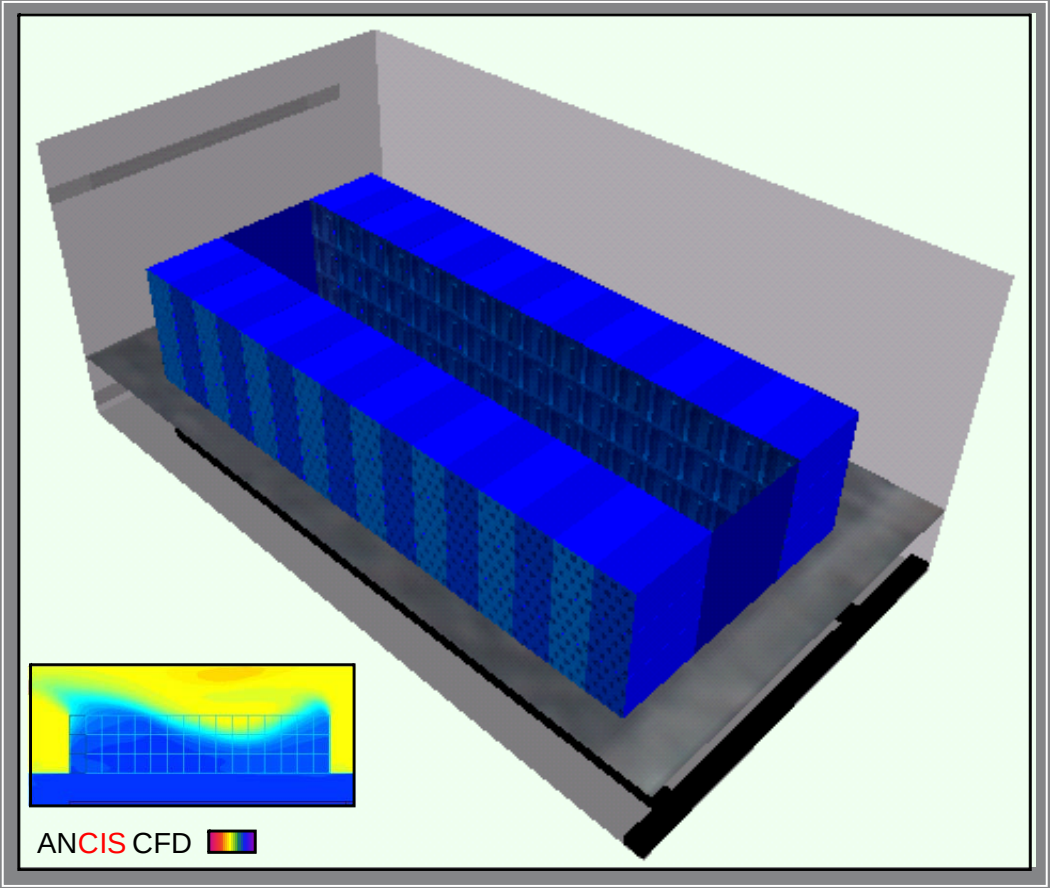
09.14.07

eyp mission critical facilities

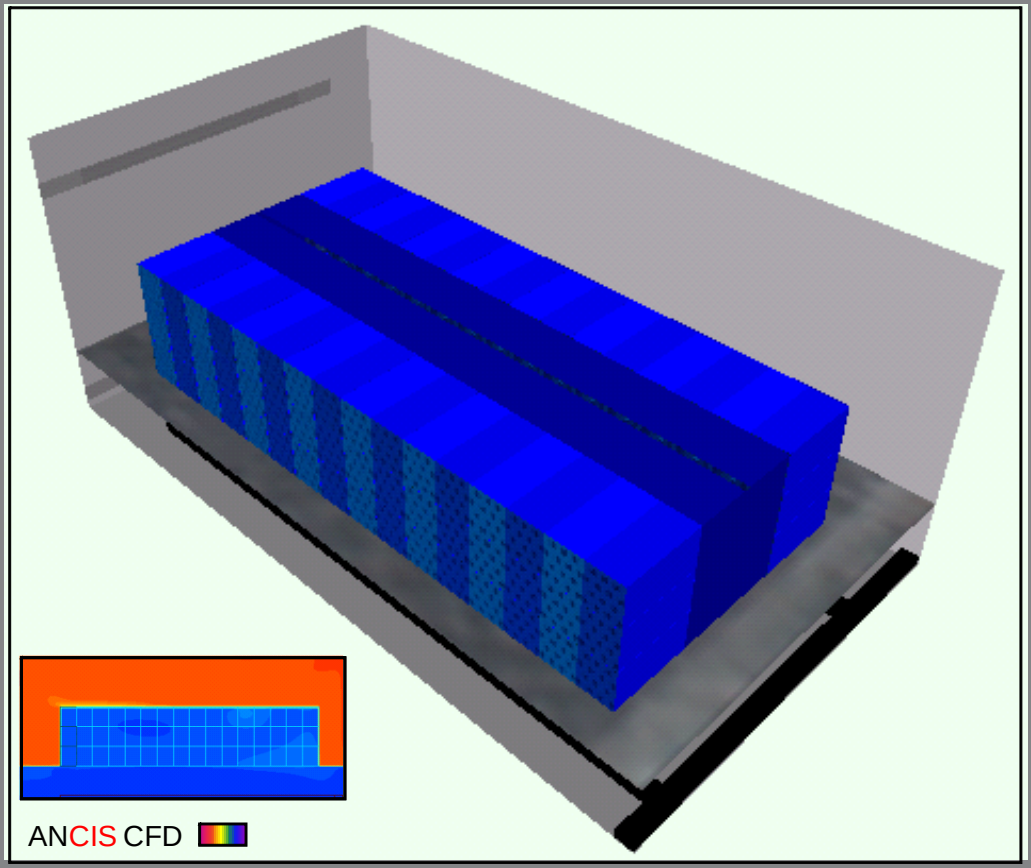
BERKELEY LAB



Proof of Concept Simulations



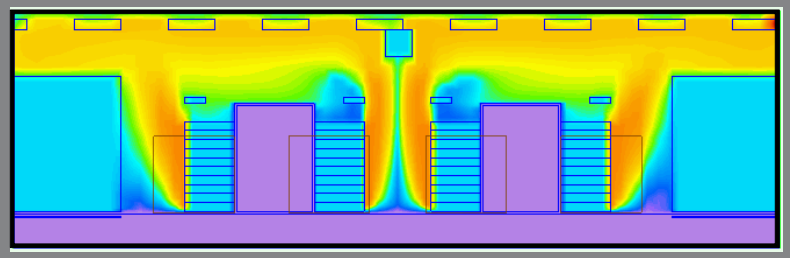
Cold-Aisle Doors



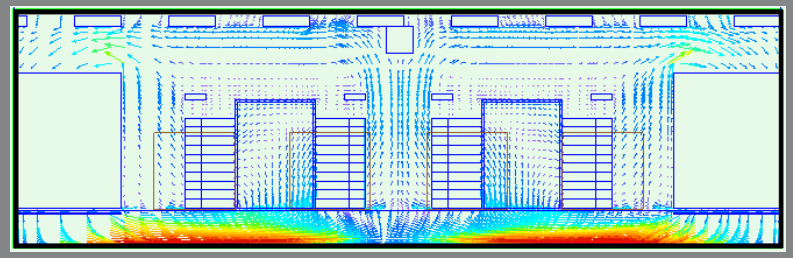
Cold-Aisle Enclosure

CFD Modeling of Alternatives

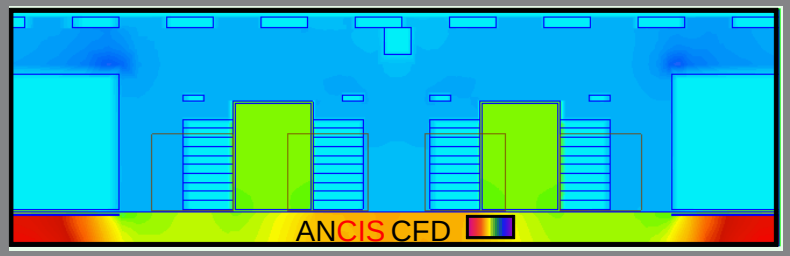
Example:
Enclosed cold equipment aisles



Temperature

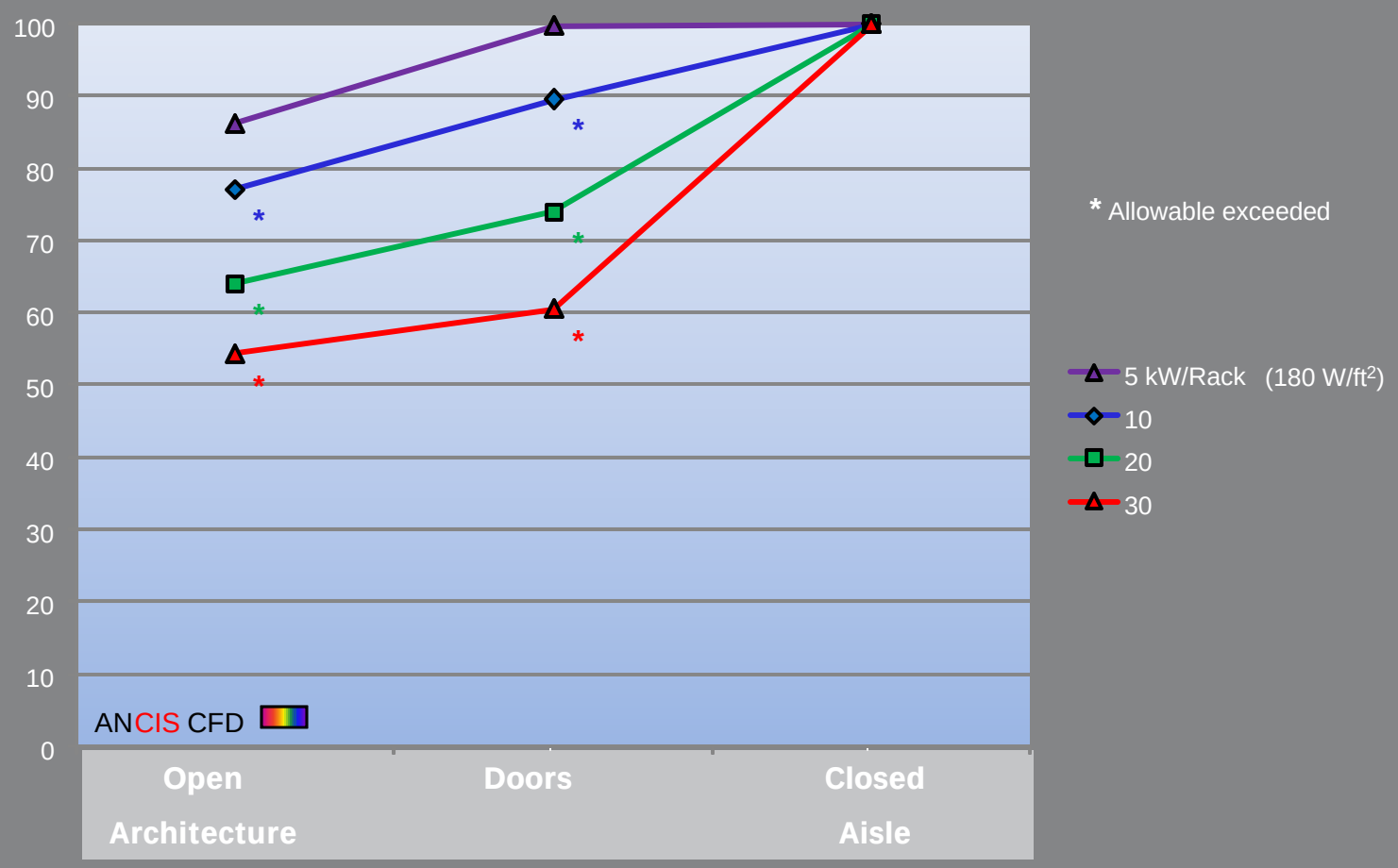


Velocity



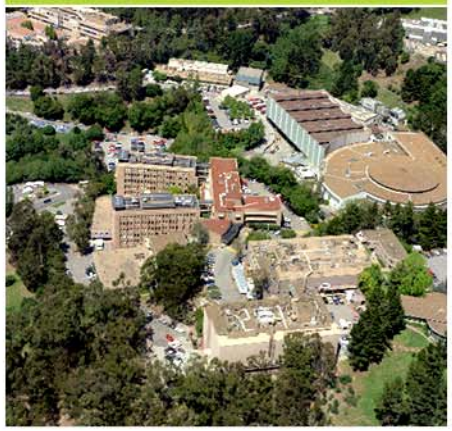
Pressure

RCI vs. Architecture



PERKINS
+ WILL

Ideas + buildings
that honor the broader
goals of society

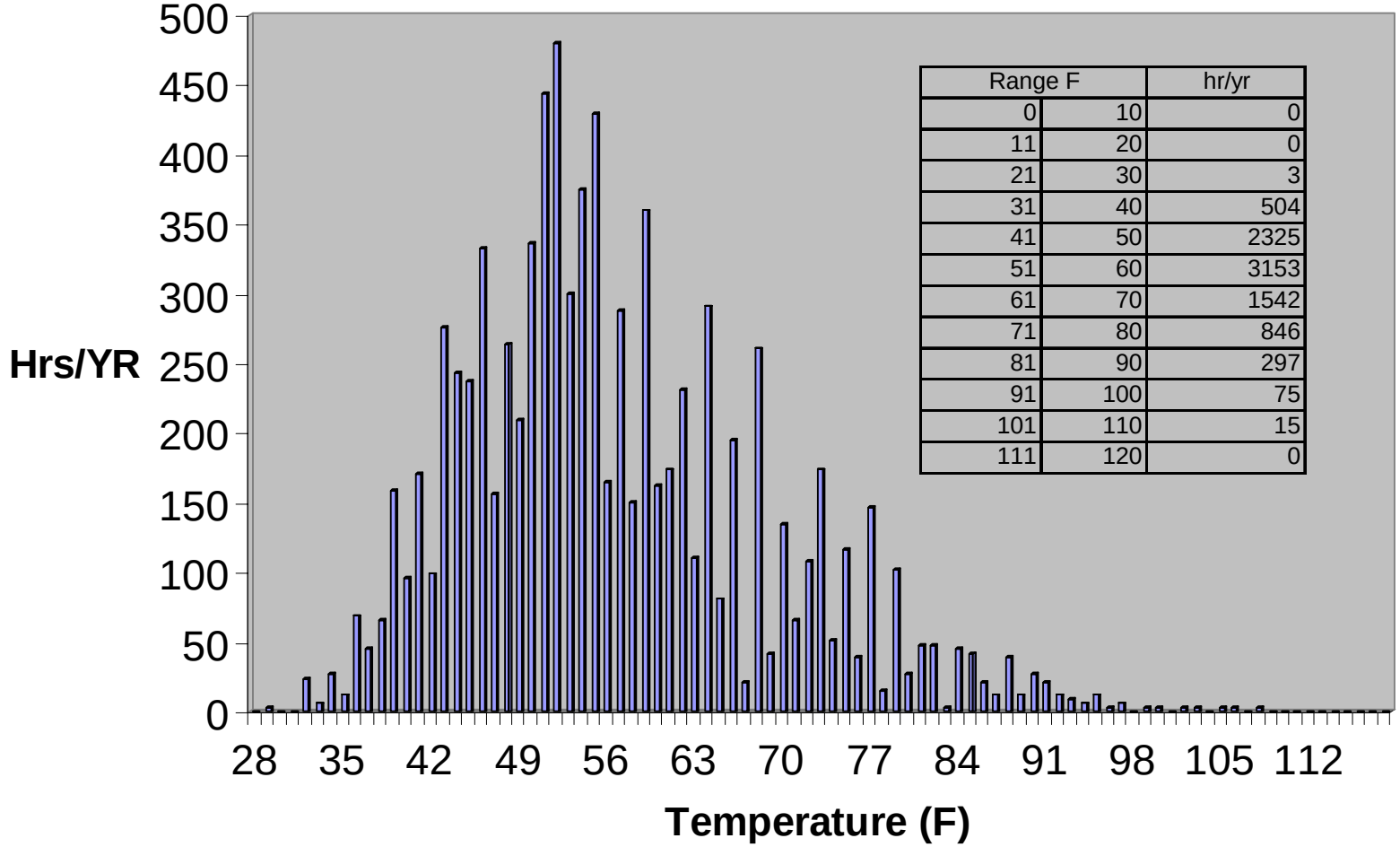


Computational Research & Theory Facility - LBNL

50% Schematic Design

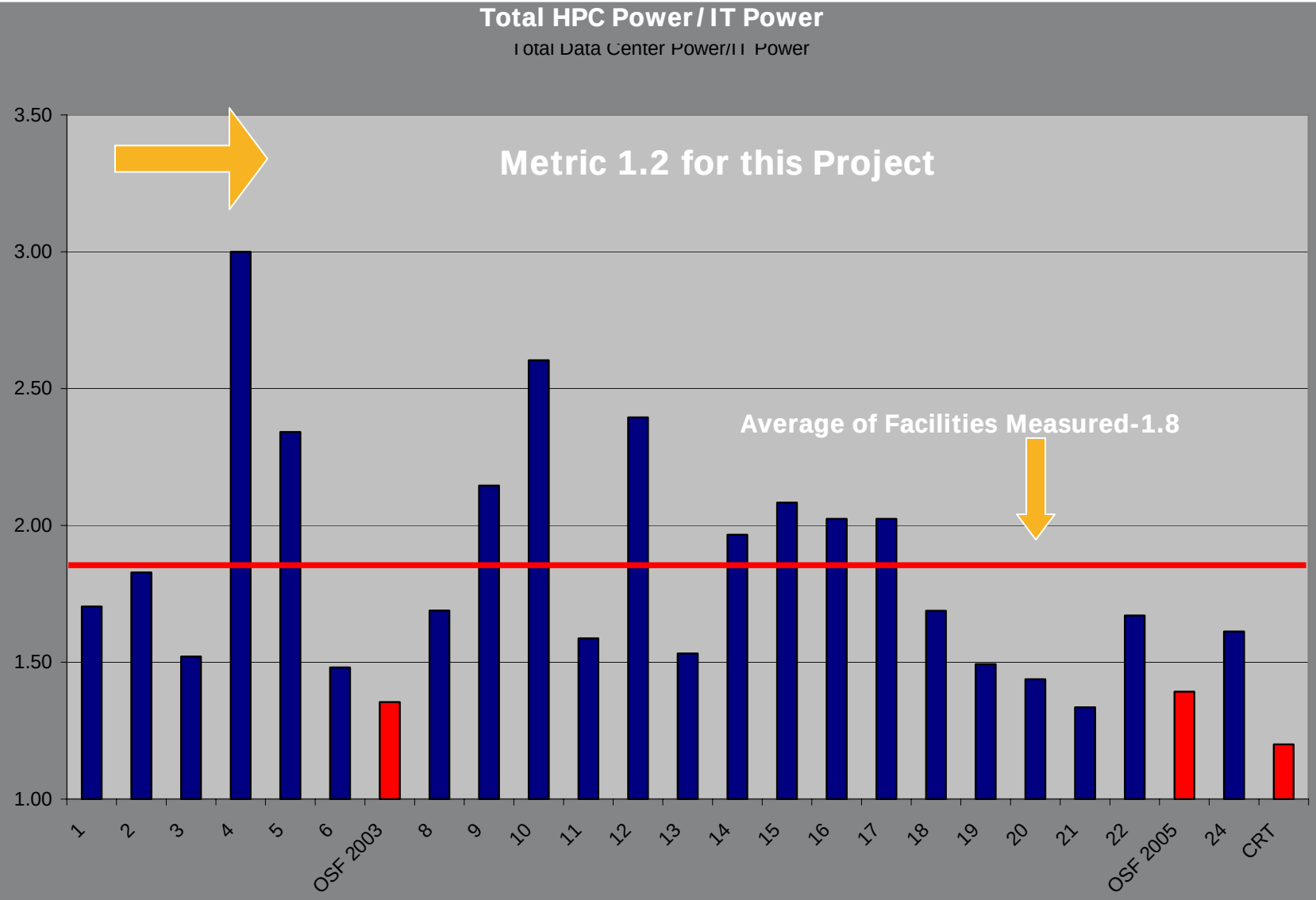
09.14.07

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

Objectives Synopsis
Project Energy Systems



Outline

1. Power consumption has become an industry-wide issue for computing
2. Understanding the problem
3. Building and computer room energy efficiency
4. Computer architecture for energy efficiency
5. Future plans

Ultra-Efficient 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 areas in the DOE Office of Science
- Proposed pilot application: Ultra-high resolution climate change simulation

Estimated Exascale Power Requirements

- LBNL IJHPCA Study for ~1/5 Exaflop for Climate Science
  - Extrapolation of Blue Gene and AMD design trends
  - Estimate: **20 MW** for BG and **179 MW** for AMD
- DOE E3 Report
  - Extrapolation of existing design trends to exascale in 2016
  - Estimate: **130 MW**
- DARPA Study
  - More detailed assessment of component technologies for exascale system
  - Estimate: **20 MW** just for memory alone, **60 MW** aggregate extrapolated from current design trends
- **The current approach is not sustainable!**

Path to Power Efficiency

Reducing Waste in Computing

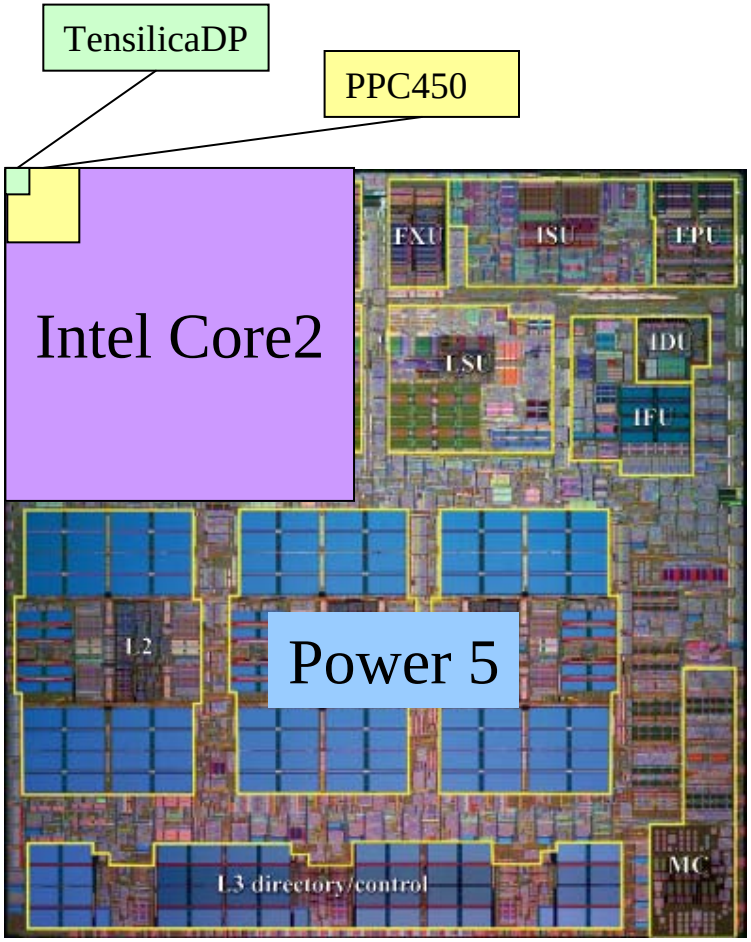
- Examine methodology of low-power embedded computing market
  - optimized for low power, low cost and high computational efficiency

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

—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

How Small Is "Small"?

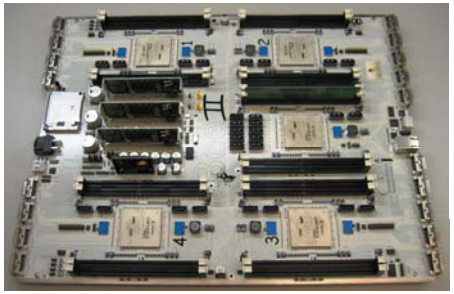


- Power5 (Server)
  - 389 mm^2
  - 120 W @ 1900 MHz
- Intel Core2 sc (Laptop)
  - 130 mm^2
  - 15 W @ 1000 MHz
- PowerPC450 (BlueGene/P)
  - 8 mm^2
  - 3 W @ 850 MHz
- Tensilica DP (cell phones)
  - 0.8 mm^2
  - 0.09 W @ 650 MHz

Each core operates at 1/3 to 1/10th efficiency of largest chip, but you can pack 100x more cores onto a chip and consume 1/20 the power!

Partnerships for Power-Efficient Computing

- Identify high-impact exascale Office of Science projects!
- Embark on targeted program of tightly coupled hardware/software co-design
  - Impossible using the typical two-year hardware lead times
  - Break slow feedback loop for system designs via RAMP hardware emulation platform and auto-tuned code generation
  - Technology partners:
    - UC Berkeley: K. Yelick, D. Patterson, K. Asanovic, J. Kubiatowicz
    - Stanford University / Rambus Inc.: M. Horowitz
    - Tensilica Inc.: C. Rowen



- Pilot application: kilometer-scale climate model
  - Provides important answers to question with multi-trillion-dollar ramifications
  - Climate community partners: Michael Wehner, Bill Collins, David Randall, et al.

NERSC

BERKELEY LAB

Office of Science
U.S. DEPARTMENT OF ENERGY

Climate Strawman System Design in 2008

- Design system around the requirements of the massively parallel application
- Example: kilometer-scale climate model application

We examined three different approaches:

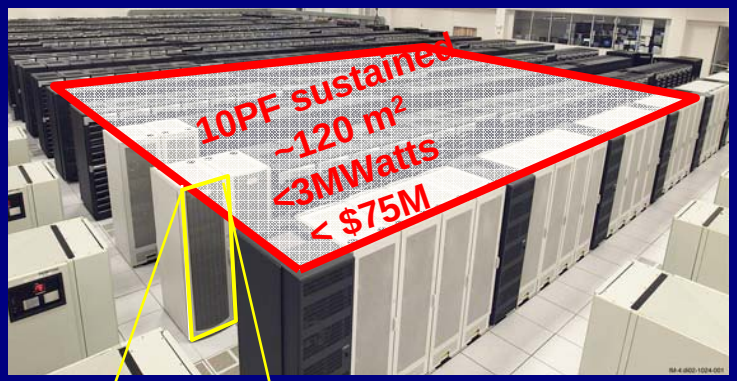
- AMD Opteron: Commodity approach, lower efficiency for scientific applications offset by cost efficiencies of mass market
- BlueGene: Generic embedded processor core and customize system-on-chip (SoC) services to improve power efficiency for scientific applications
- Tensilica: Customized embedded CPU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

Processor	Clock	Peak/ Core (Gflops)	Cores/ Socket	Mem/ BW (GB/s)	Network BW (GB/s)	Sockets	Power	Cost 2008
AMD Opteron	2.8GHz	5.6	2	6.4	4.5	890K	179 MW	
IBM BG/P	850MHz	3.4	4	5.5	2.2	740K	20 MW	
Climate computer	650MHz	2.7	32	51.2	34.5	120K	3 MW	$75M

Solve an exascale problem without building an exaflop/s machine!

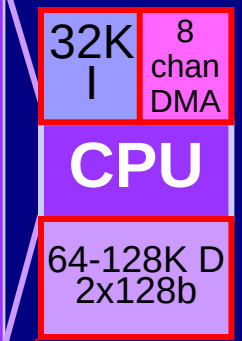
Climate System Design Concept
Strawman Design Study

From Chris Rowen, Tensilica & Mark Horowitz, Stanford

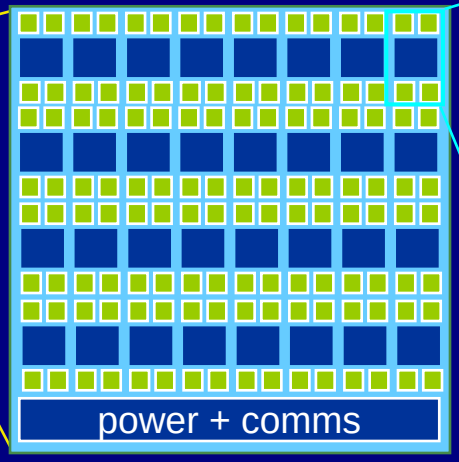


VLIW CPU:

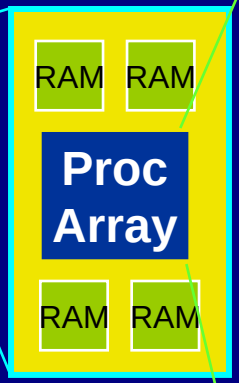
- 128b load-store + 2 DP MUL/ADD + integer op/ DMA per cycle:
- Synthesizable at 650 MHz in commodity 65 nm
- 1mm² core, 1.8-2.8 mm² with inst cache, data cache data RAM, DMA interface, 0.25 mW/MHz
- Double precision SIMD FP : 4 ops/cycle (2.7 Gflops)
- Vectorizing compiler, cycle-accurate simulator,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



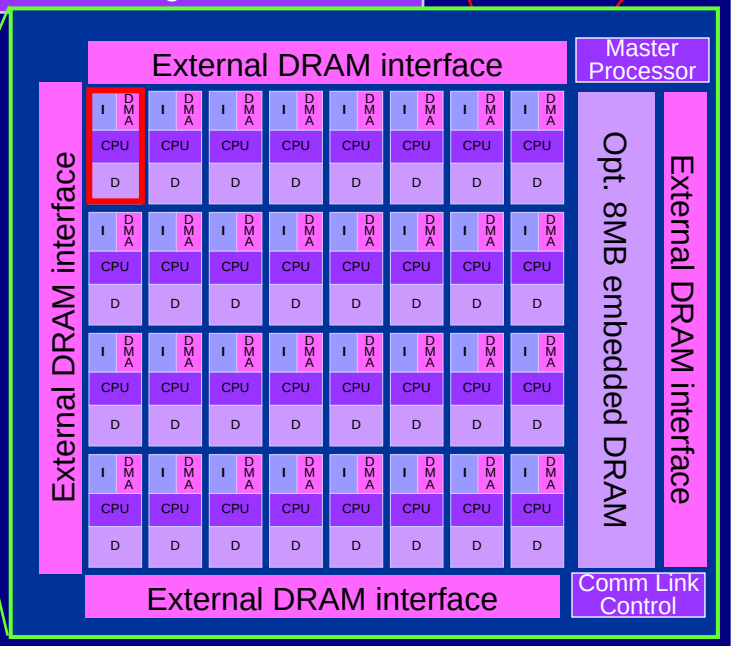
100 racks @ ~25KW



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



8 DRAM per processor chip: 50 GB/s



32 processors per 65nm chip
83 GFLOPS @ 7W

Outline

1. Power consumption has become an industry-wide issue for computing

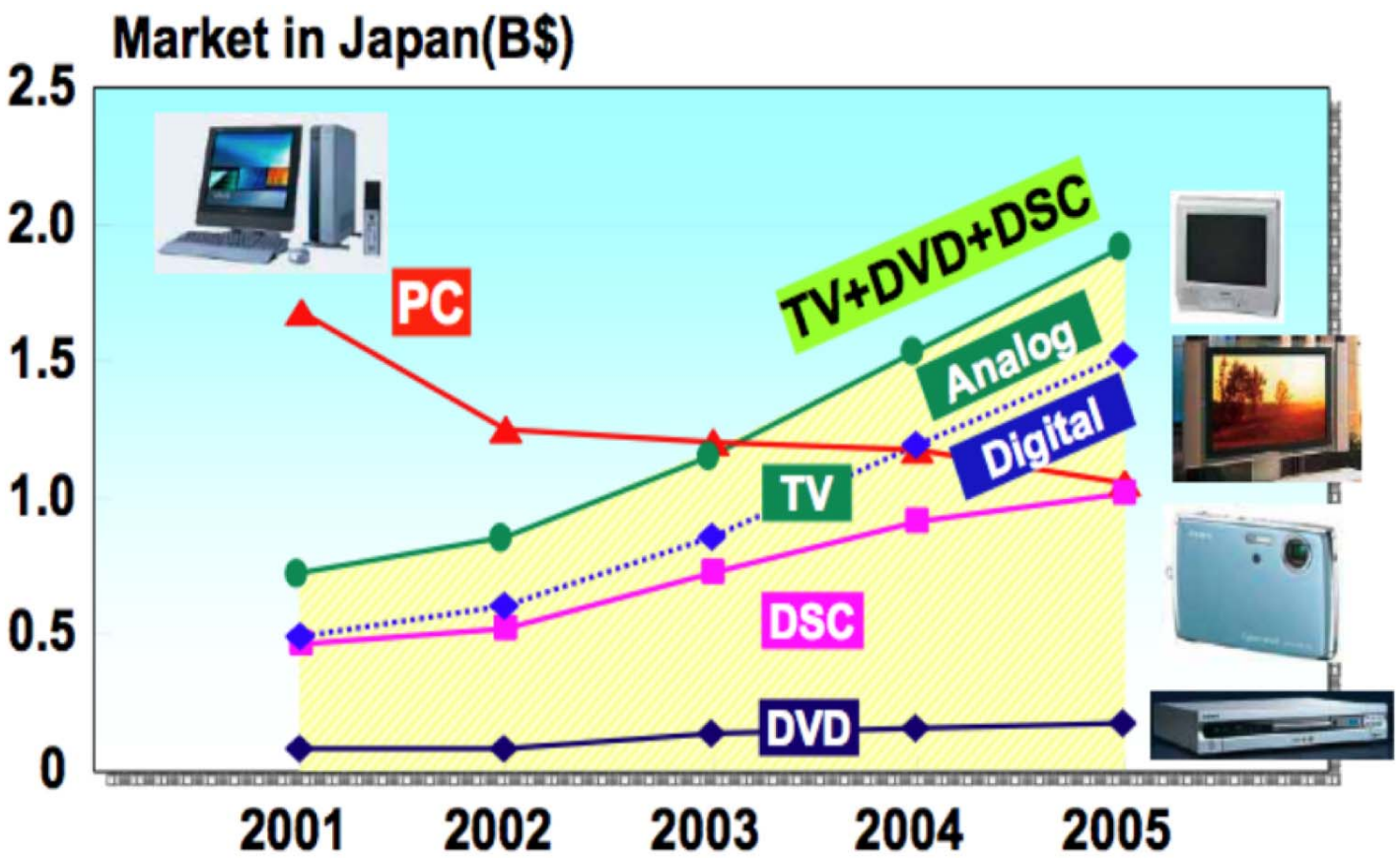
2. Understanding the problem

3. Building and computer room energy efficiency

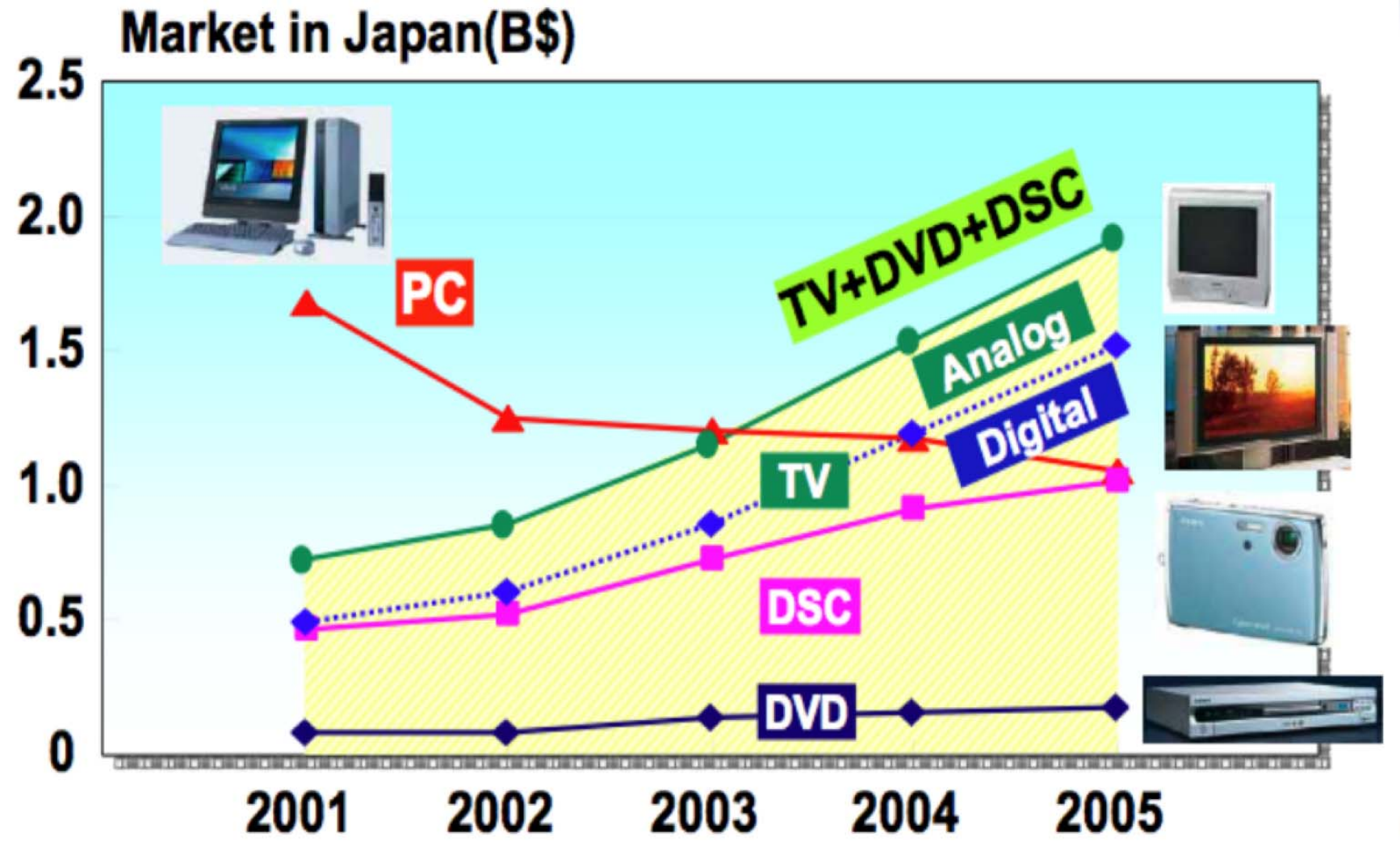
4. Computer architecture for energy efficiency

5. Future

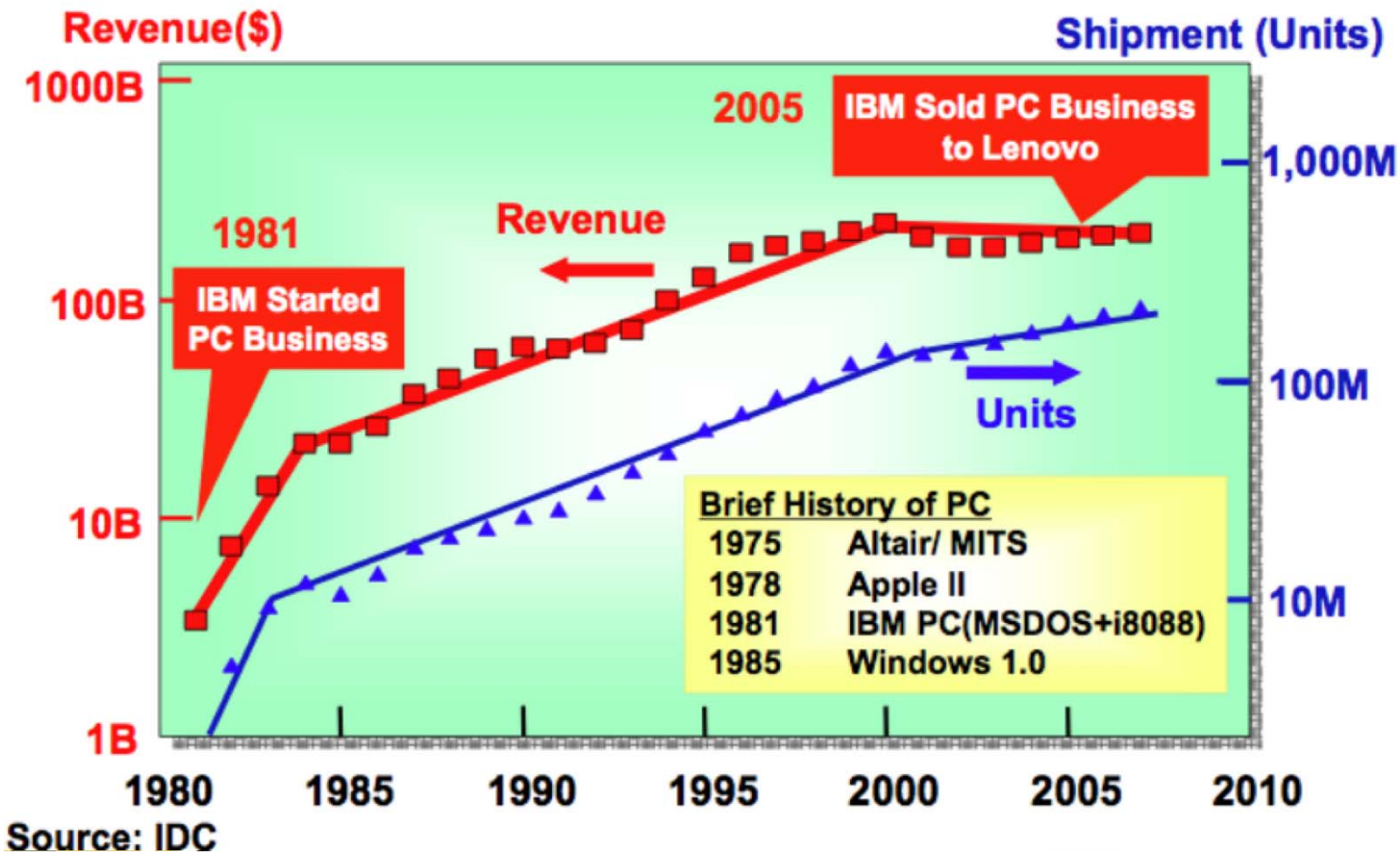
Consumer Electronics Convergence



Consumer Electronics Convergence



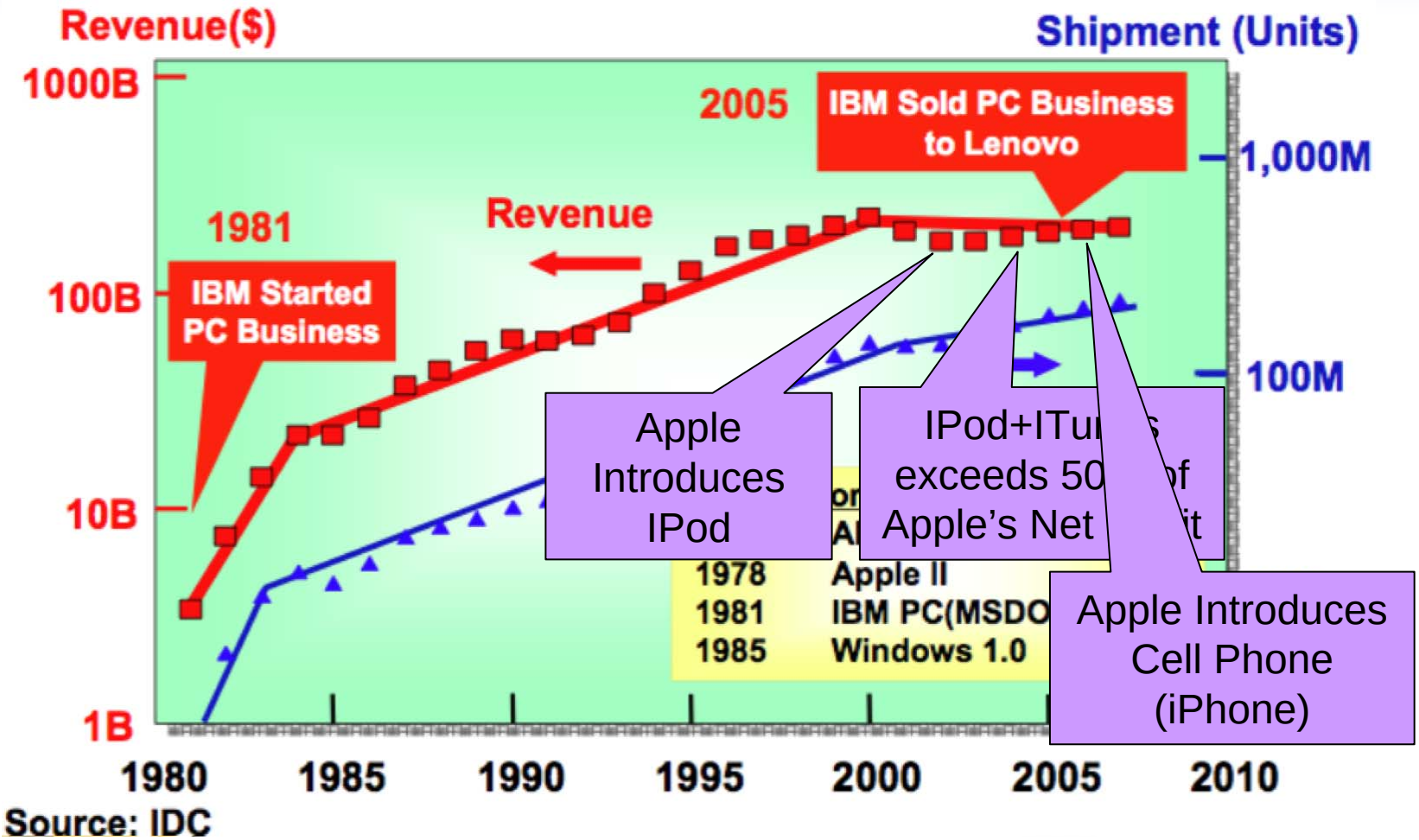
Consumer Electronics has Replaced PCs as the Dominant Market Force in CPU Design!!



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

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

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



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

are

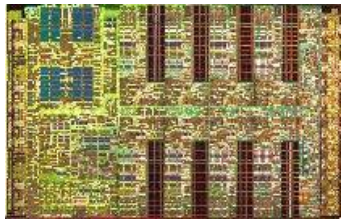
Apple Introduces IPod

IPod+ITu S exceeds 50 of Apple's Net it

Apple Introduces Cell Phone (iPhone)

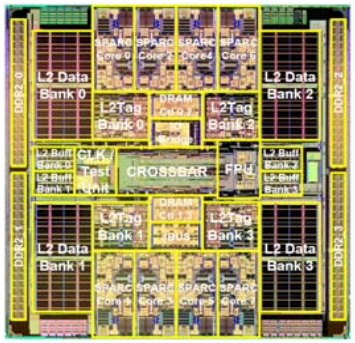
Convergence of Platforms

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

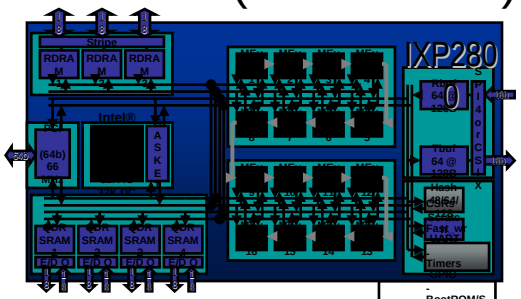


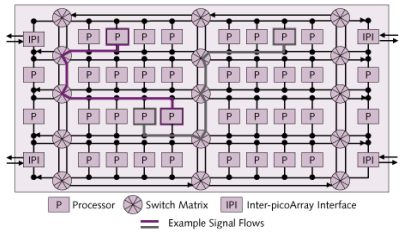
IBM Cell
1 GPP (2 threads)
8 ASPs

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

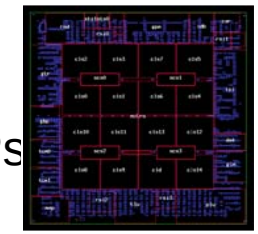


Sun Niagara
8 GPP cores (32 threads)

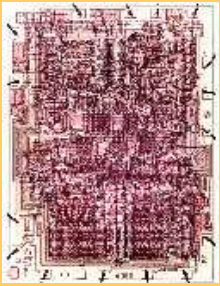




Picochip DSP
1 GPP core
248 ASPs



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]

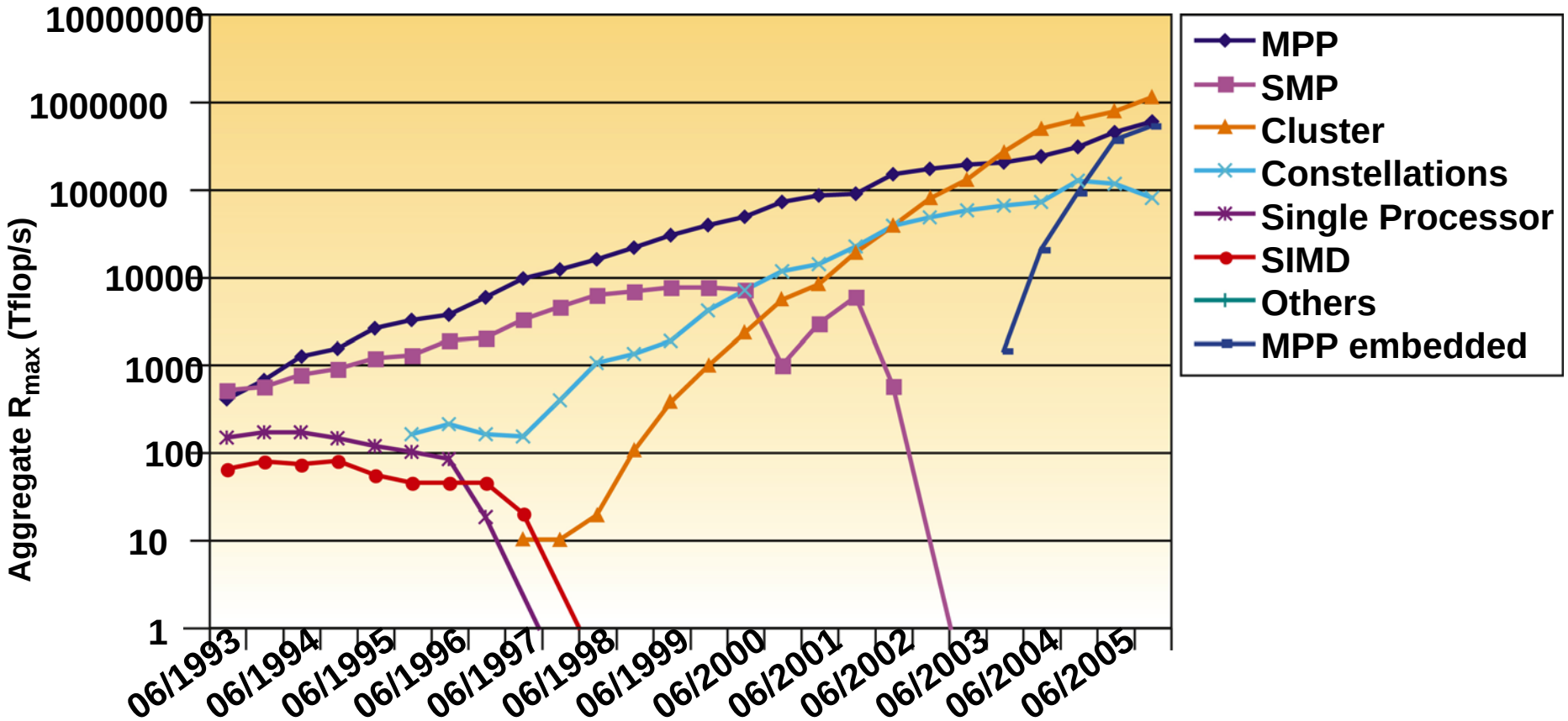
The Transition to Low-Power Technology is Inevitable

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 3 - 4 years
- Embedded processors or game processors will be the next step (BG, Cell, Nvidia, SiCortex. Tensilica)

BG/L—the Rise of the Embedded Processor

TOP 500 Performance by Architecture



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