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

Festkolloquium: Prof. Dr. Arndt Bode 60 Jahre Parallelrechner – Hochleistung für Wissenschaft und Forschung,
München, October 10, 2008

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

HPC User Forum, Stuttgart, Germany
October 13th and 14th, 2008

BERKELEY LAB NERSC Office of Science U.S. DEPARTMENT OF ENERGY

Acknowledgements

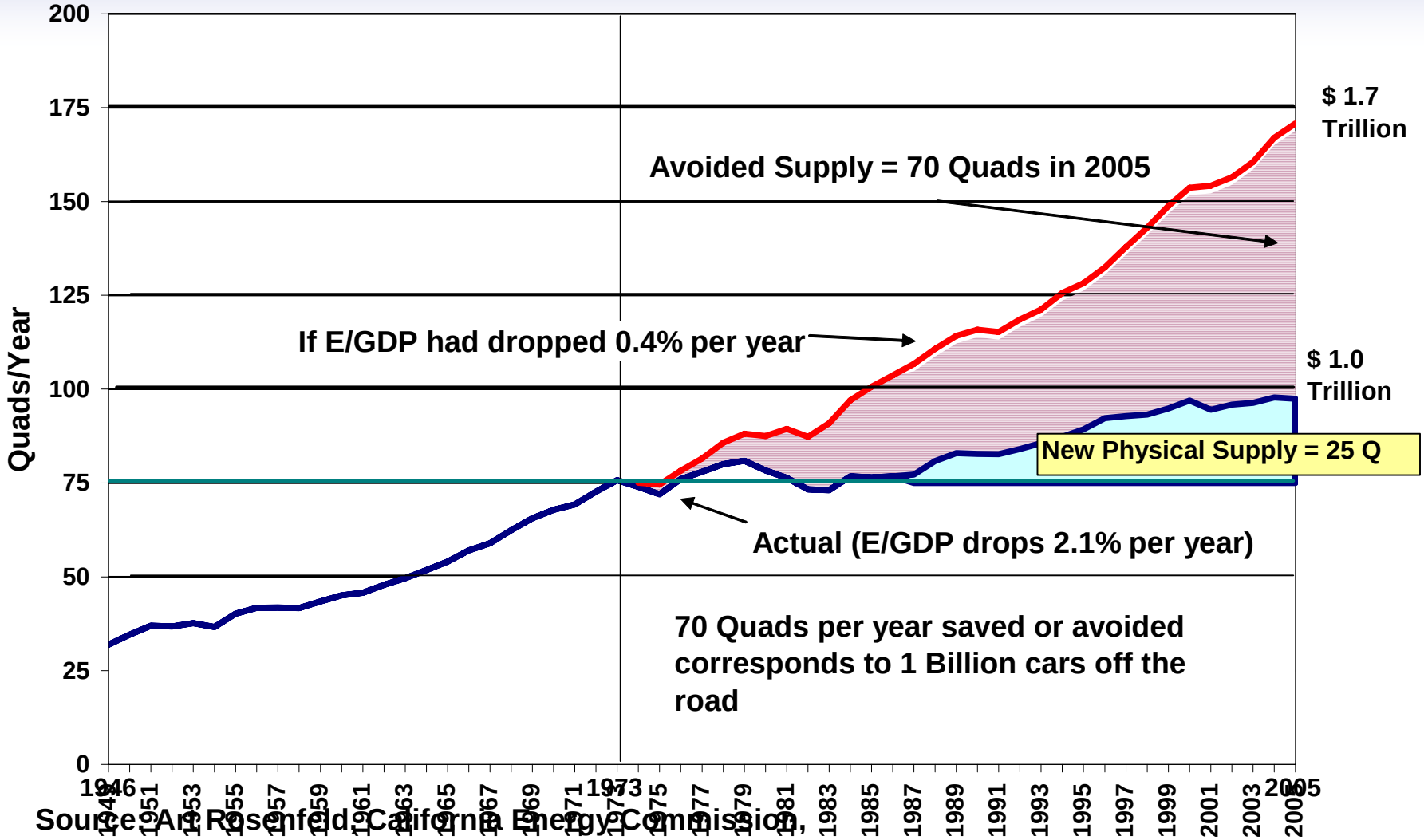
A large number of individuals have contributed to energy efficiency in computing at the Lab and to this presentation:

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 energy 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. **Building and computer room energy efficiency**
3. **Computer architecture for energy efficiency- the Green Flash project**
4. **Future Direction**

Outline

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

   Two interrelated issues:

   - **Building and infrastructure problem**
   - **Computer architecture problem**

The Problem

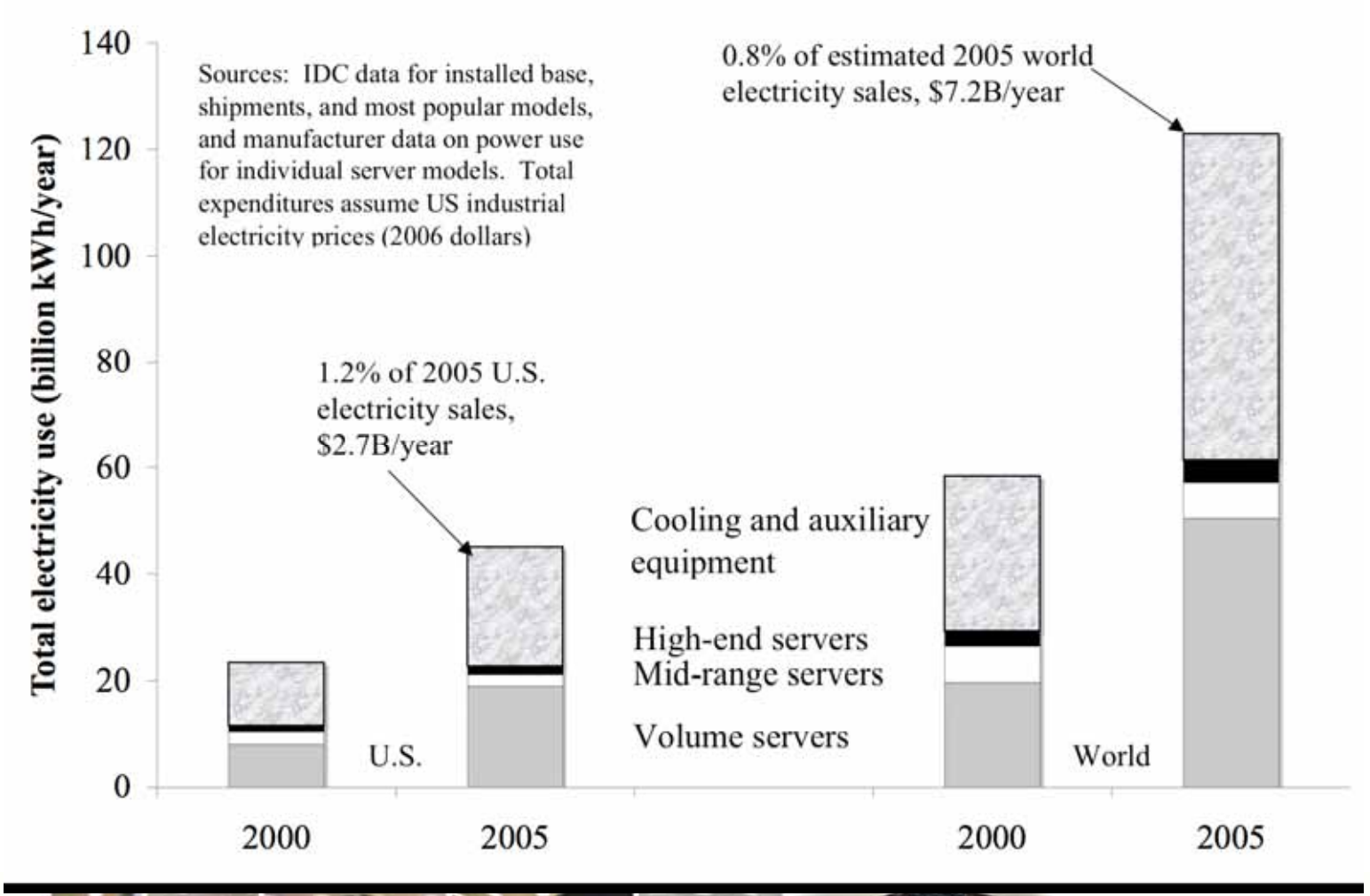
- **"Big IT" – all electronics**
  - **PCs / etc., consumer electronics, telephony**
    - Residential, commercial, industrial
  - **More than 200 TWh/year**
  - **$16 billion/year**
    - Based on .08$/KWh
  - **Nearly 150 million tons of CO_2 per year**
    - Roughly equivalent to 30 million cars!

Numbers represent U.S. only

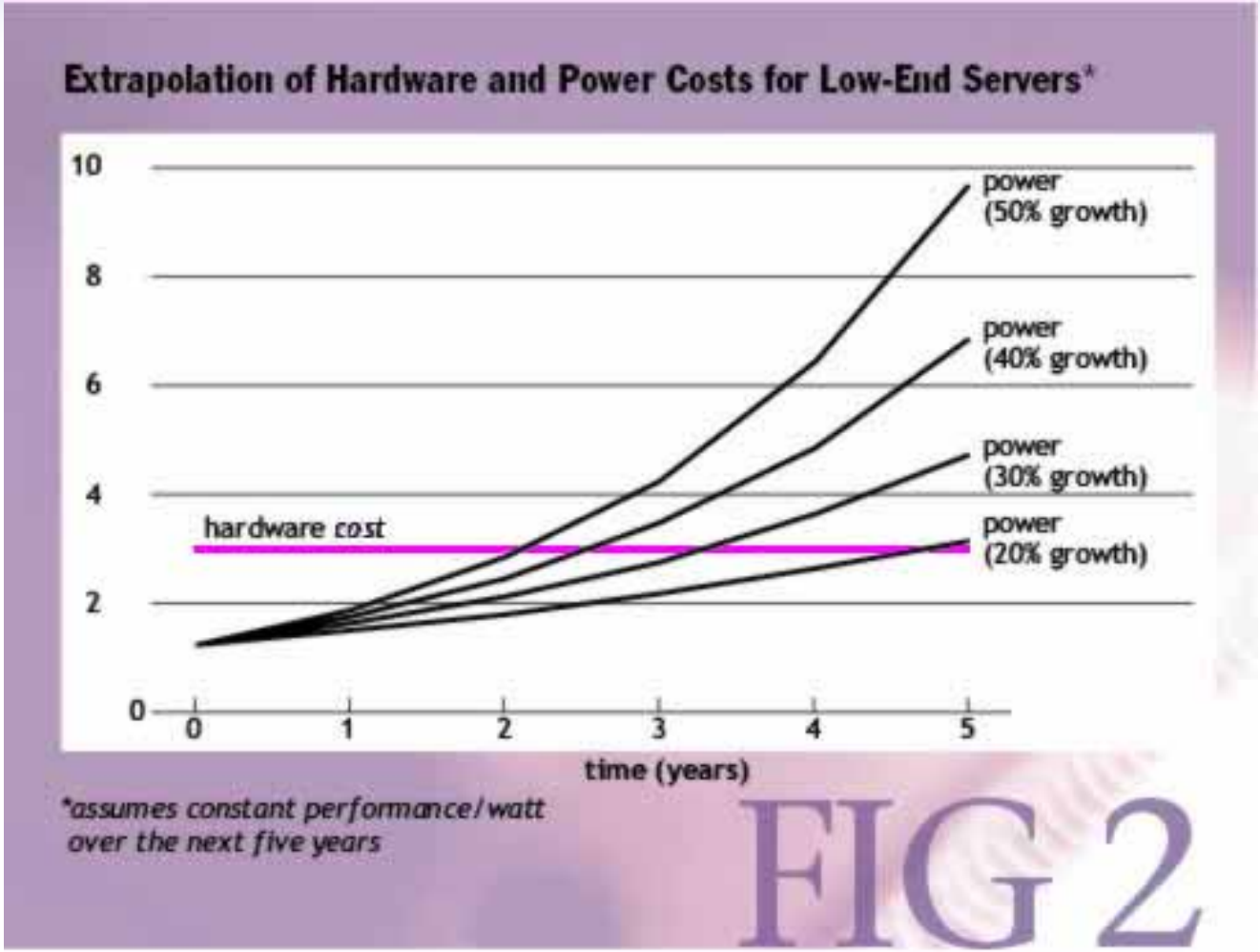
One central baseload power plant (about 7 TWh/yr)

… 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

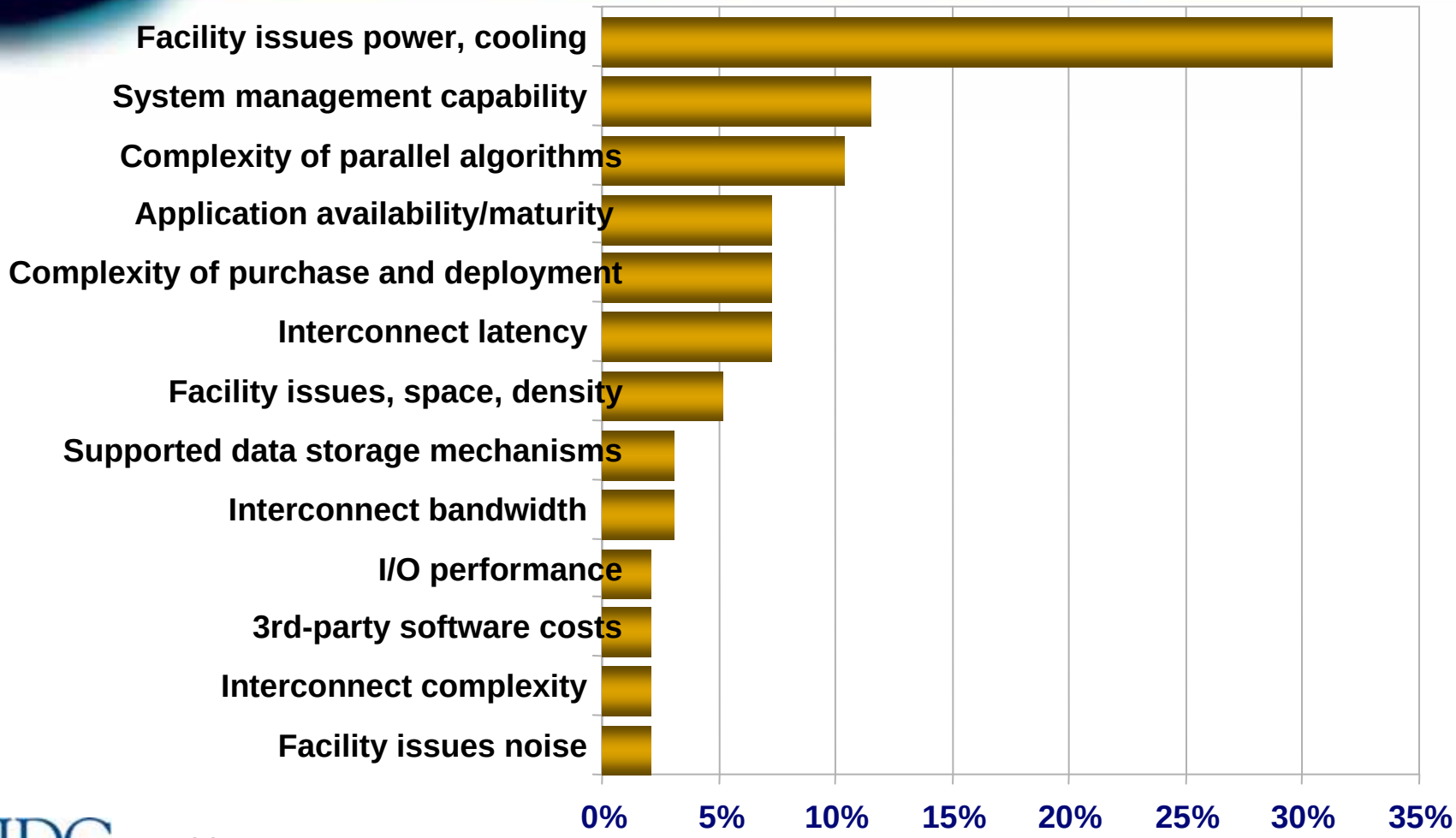


Unrestrained IT power consumption could eclipse hardware costs and put great pressure on affordability, data center infrastructure, and the environment.

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

NERSC
Berkeley Lab
Office of Science
U.S. Department of Energy

Top Challenges to Clusters



IDC Analyze the Future

n = 96

Combined facilities score — 38.6

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

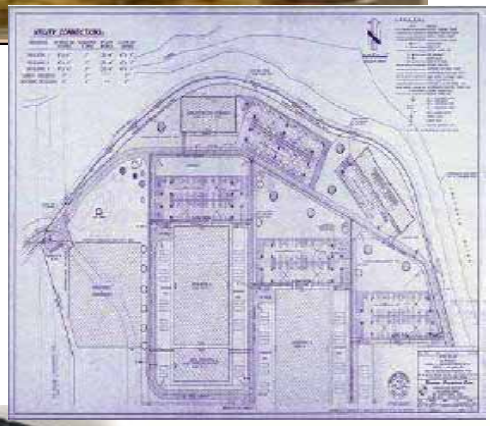
- **June 2006 - Google begins building a new data center near the Columbia River on the border between Washington and Oregon**
  - Because the location is “at the intersection of cheap electricity and readily accessible data networking”

    “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

NERSC

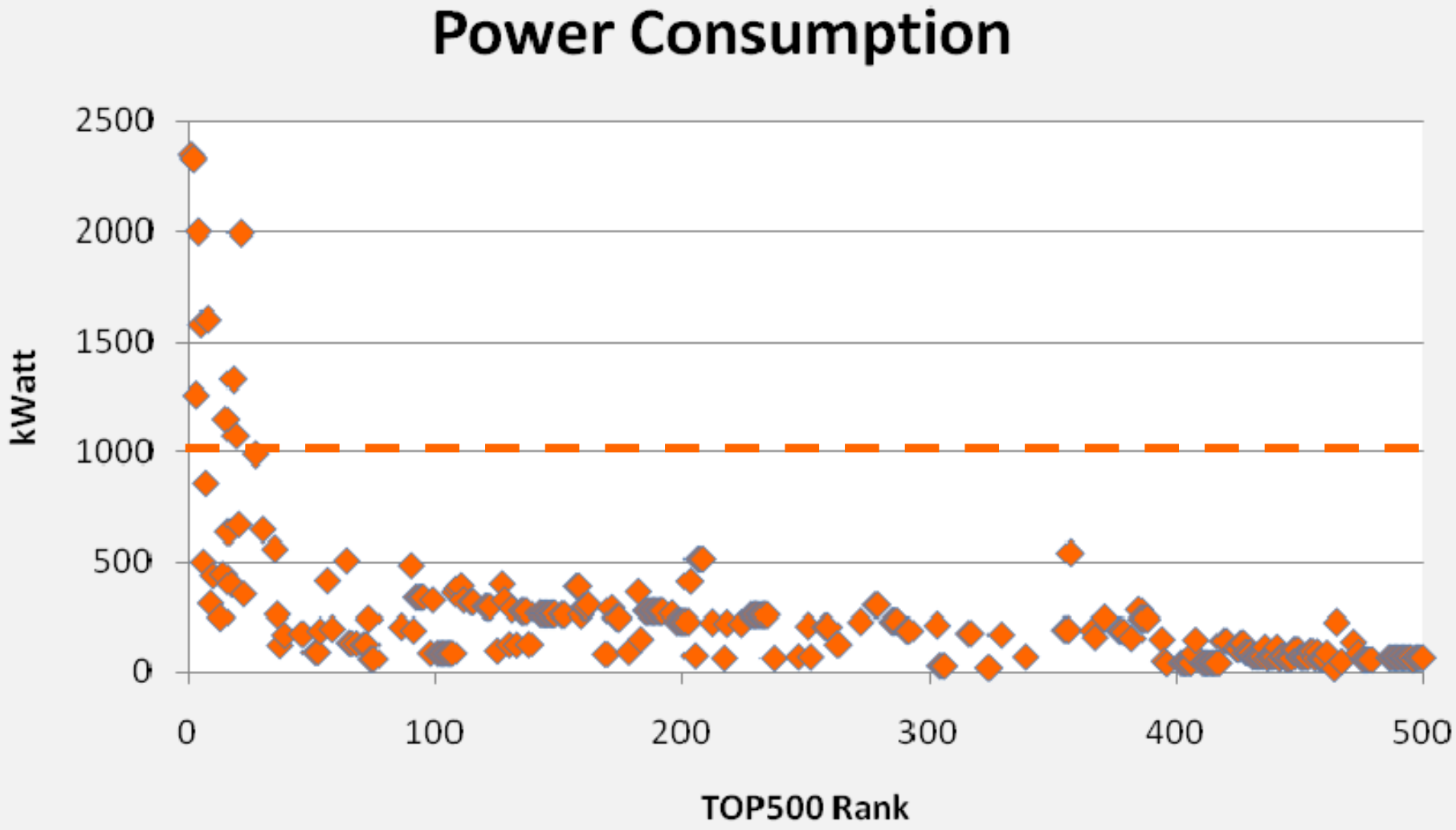
BERKELEY LAB

Office of Science
U.S. DEPARTMENT OF ENERGY

Microsoft Quincy, Wash.
470,000 Sq Ft, 47MW!

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

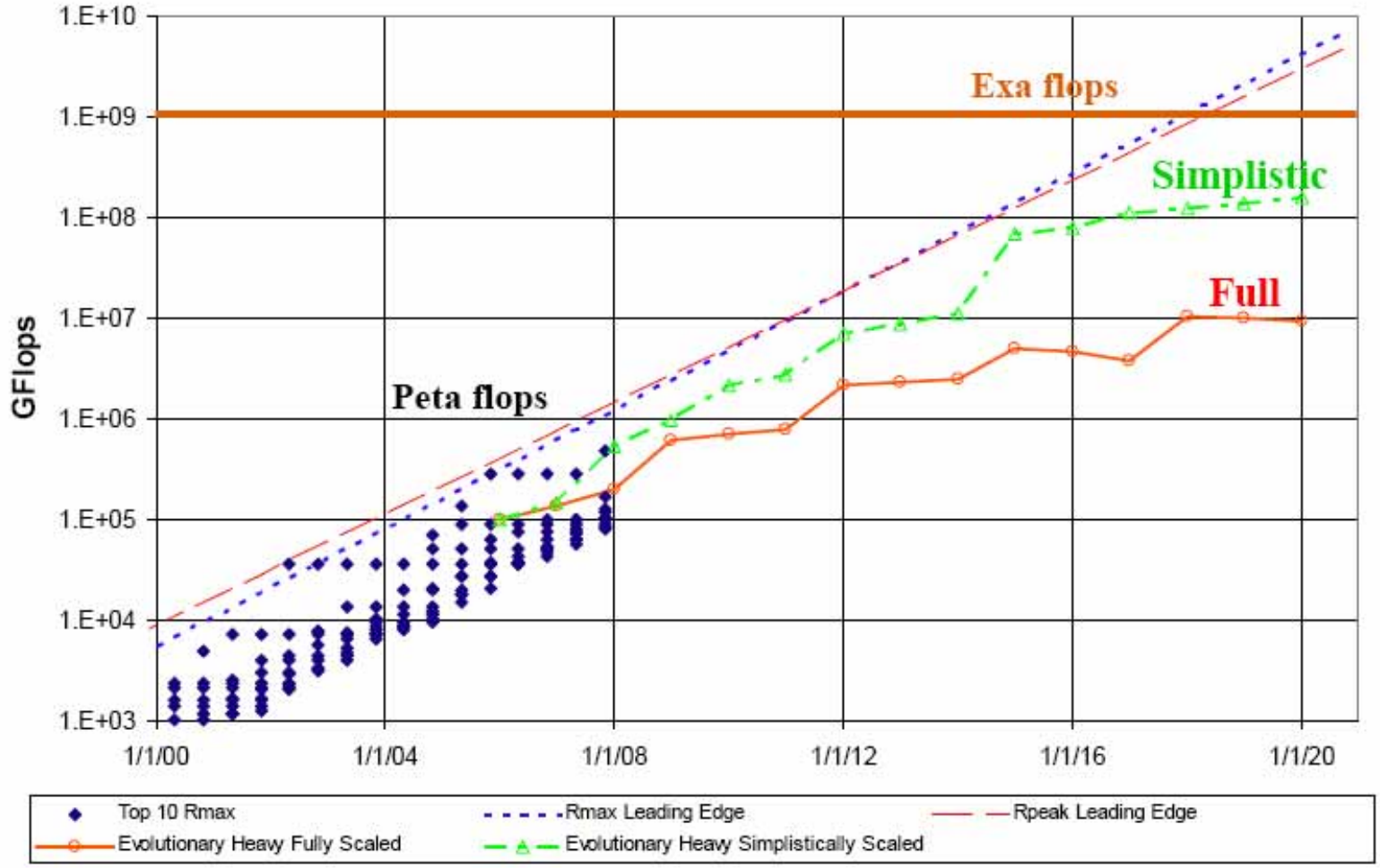
Absolute Power Levels



DARPA Exascale Study

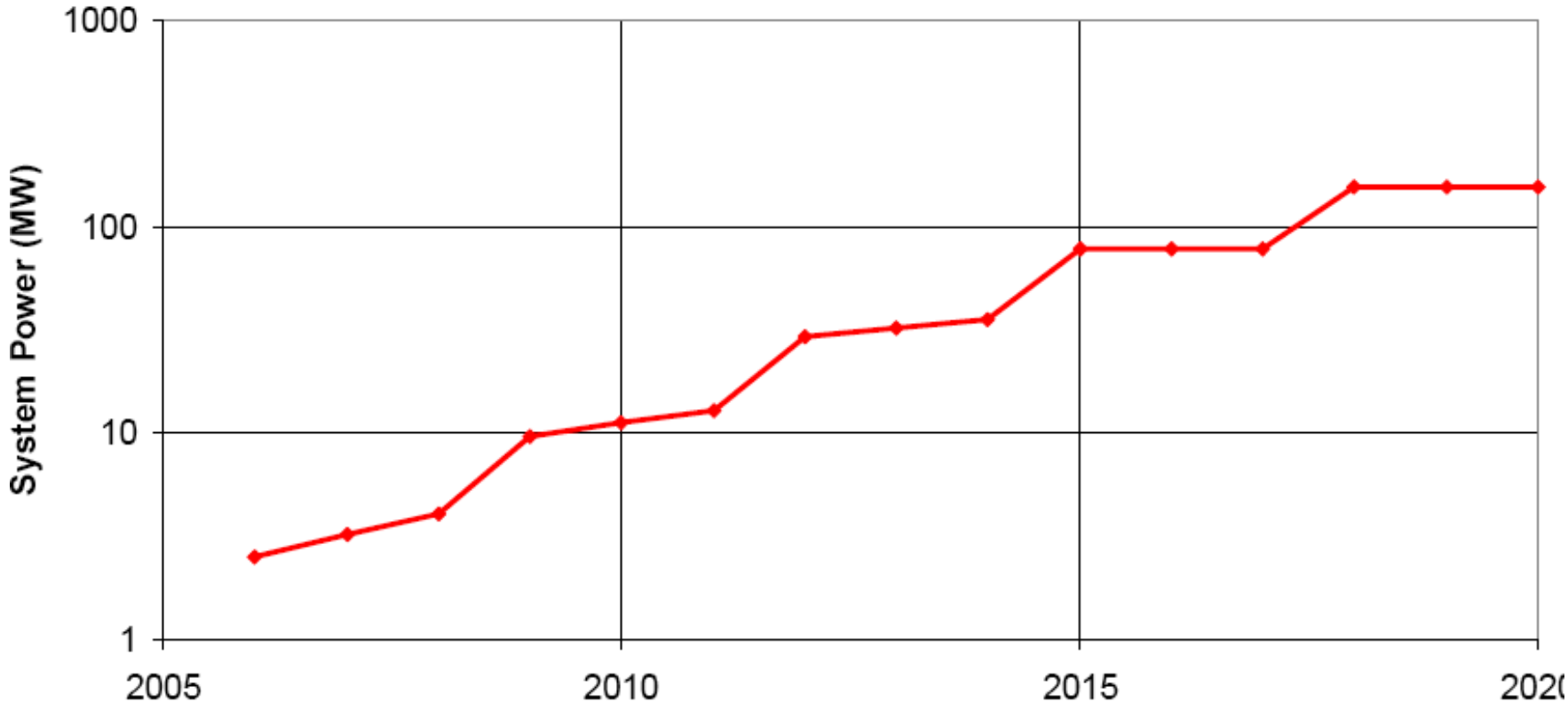
- Commissioned by DARPA to explore the challenges for Exaflop computing
- Two model for future performance growth
  - Simplistic: ITRS roadmap; power for memory grows linear with #of chips; power for interconnect stays constant
  - Fully scaled: same as simplistic, but memory and router power grow with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

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!

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 Direction

Understanding Power Consumption in HPC Computer Room Environment

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

- LBNL has long-term experience in computer room energy efficiency for data centers (power distribution, air flow, cooling technology)
- Usage patterns are significantly different between IT and HPC centers
- Need to understand and improve computer room issues for HPC centers

Focus on PUE

- PUE = “power usage effectiveness” metric promoted by “Green Grid”
- PUE = total facility power/ computer equipment power
- Reduce PUE by consistent application of facilities improvements

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

PERKINS
+WILL

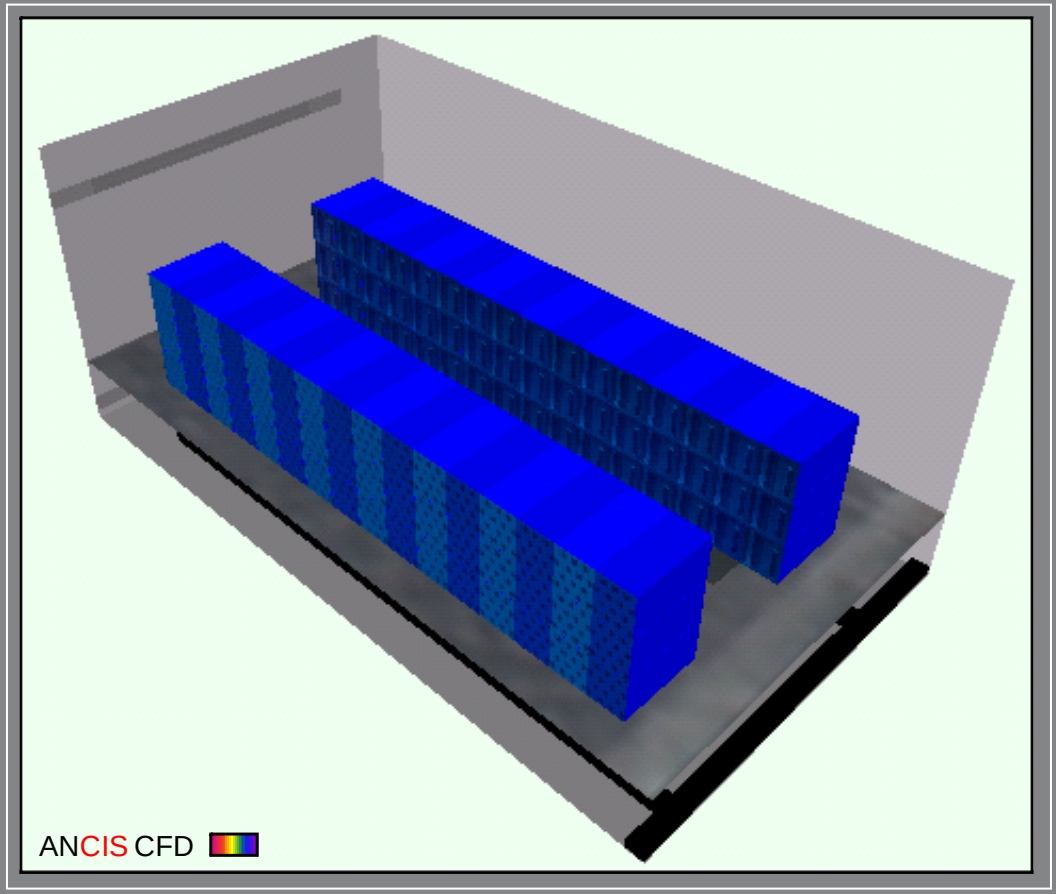
Ideas + buildings
that honor the broader
goals of society

Computational Research & Theory Facility - LBNL

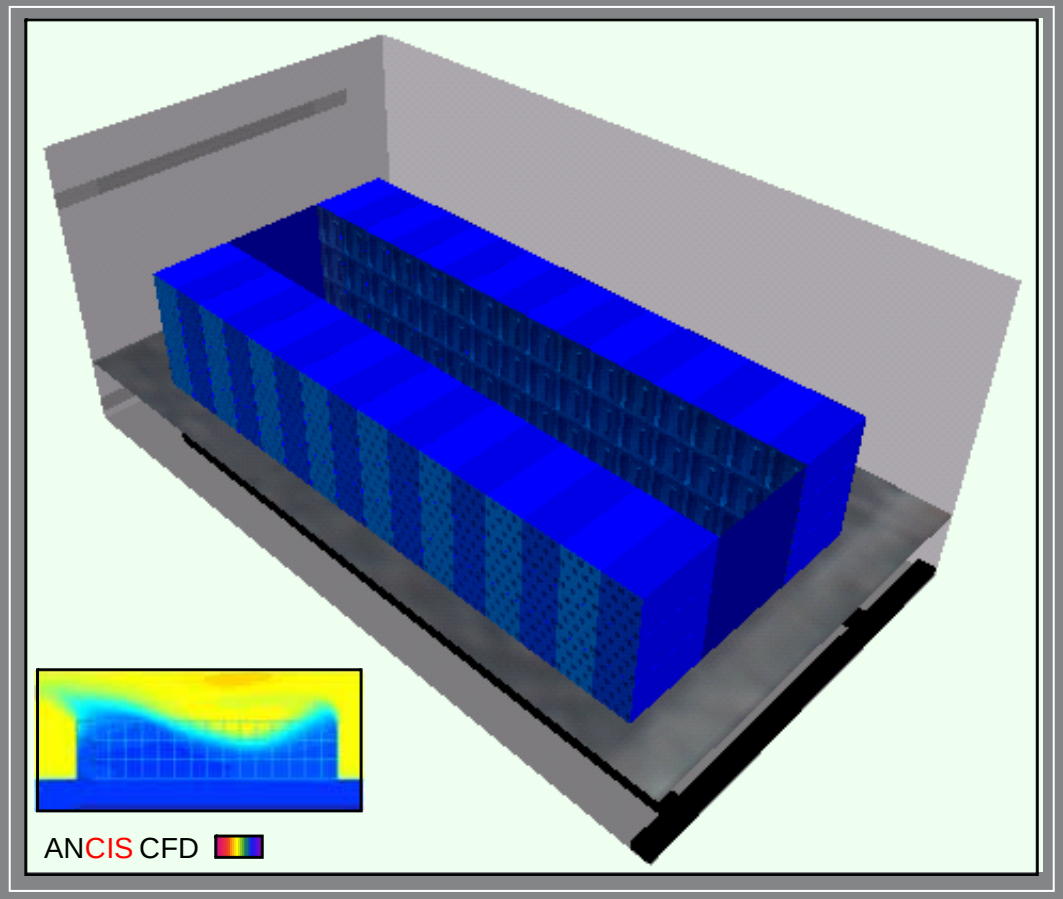
50% Schematic Design 09.14.07

eyp mission critical facilities

BERKELEY LAB



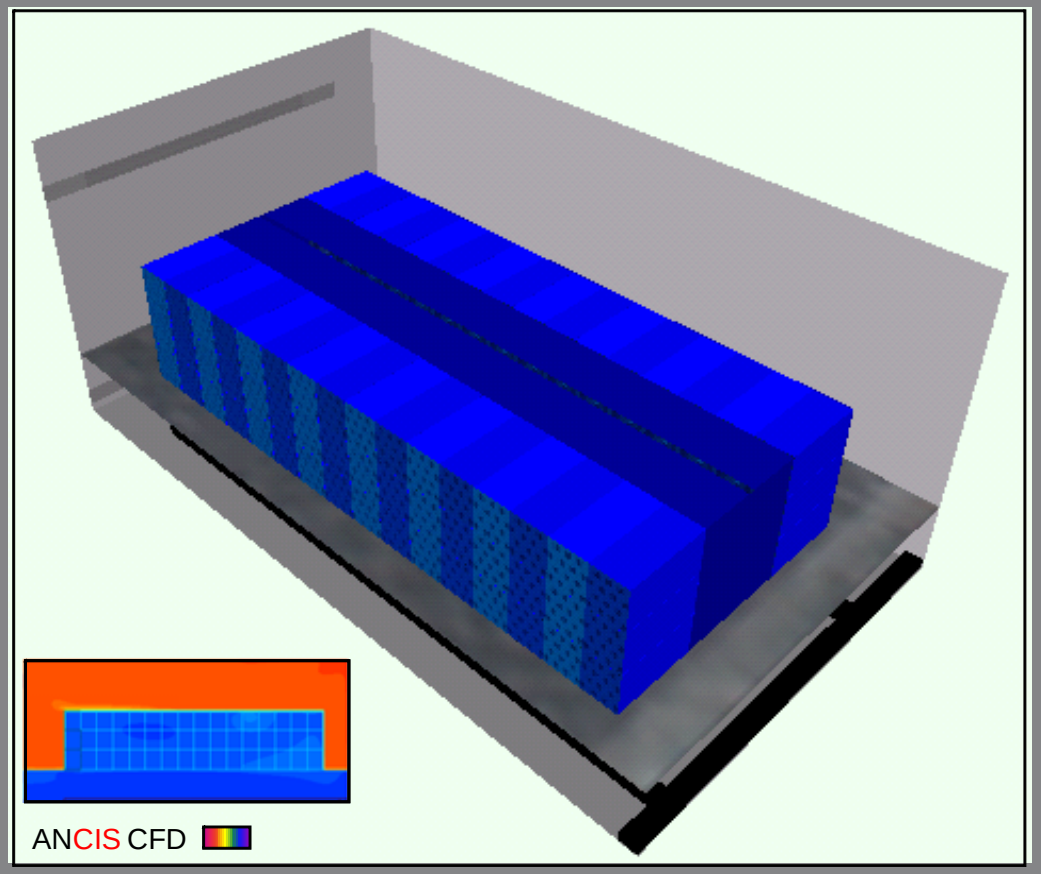
Proof of Concept Simulations



Cold-Aisle Doors

eyp mission critical facilities

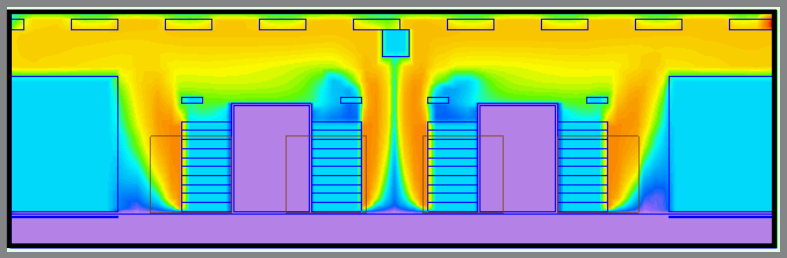
BERKELEY LAB



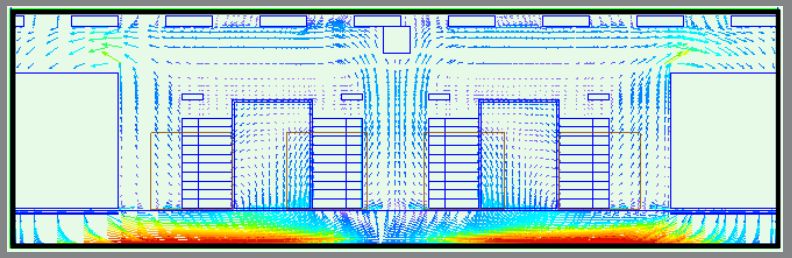
Cold-Aisle Enclosure

CFD Modeling of Alternatives

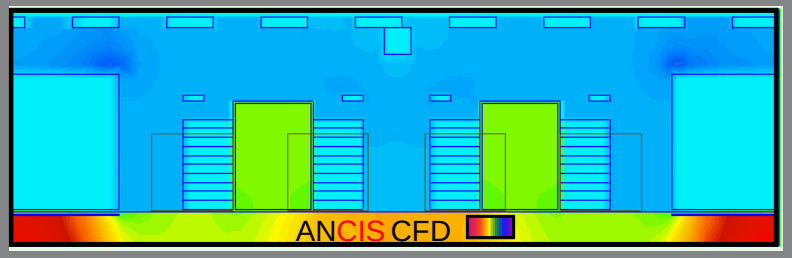
Example:
Enclosed cold equipment aisles



Temperature

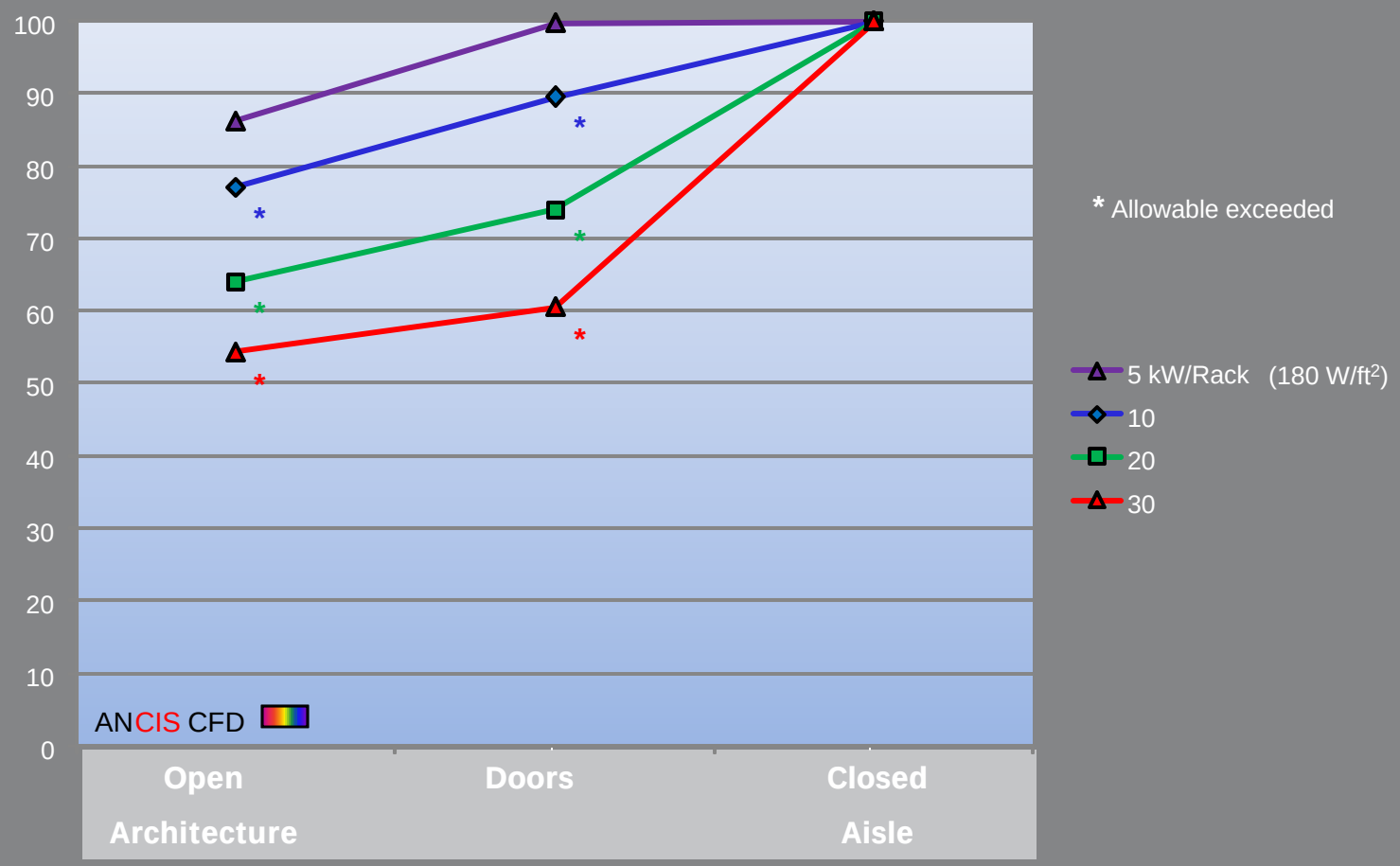


Velocity



Pressure

RCI vs. Architecture



PERKINS
+WILL

Ideas + buildings
that honor the broader
goals of society

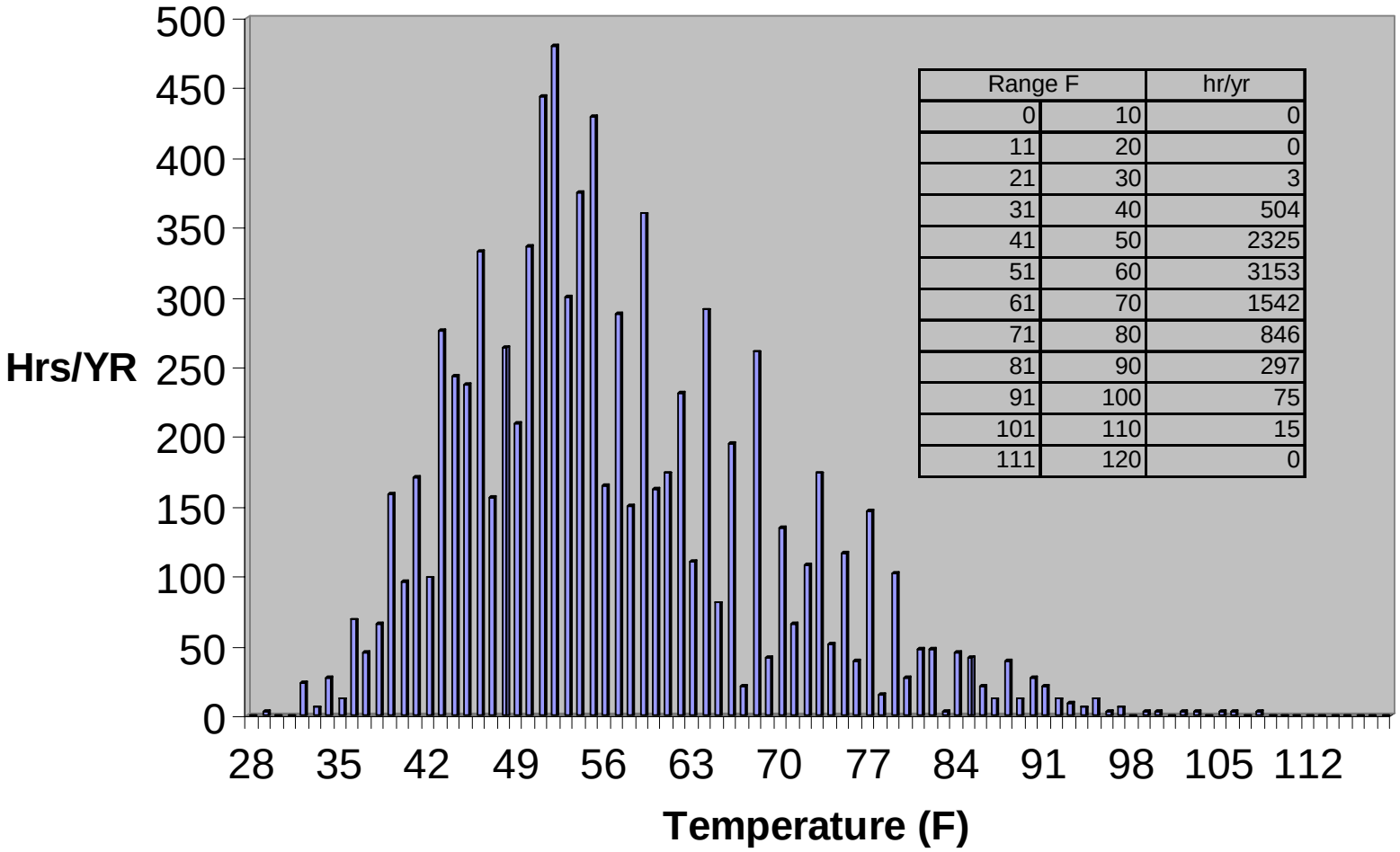
Computational Research & Theory Facility - LBNL

50% Schematic Design 09.14.07

eyp mission critical facilities

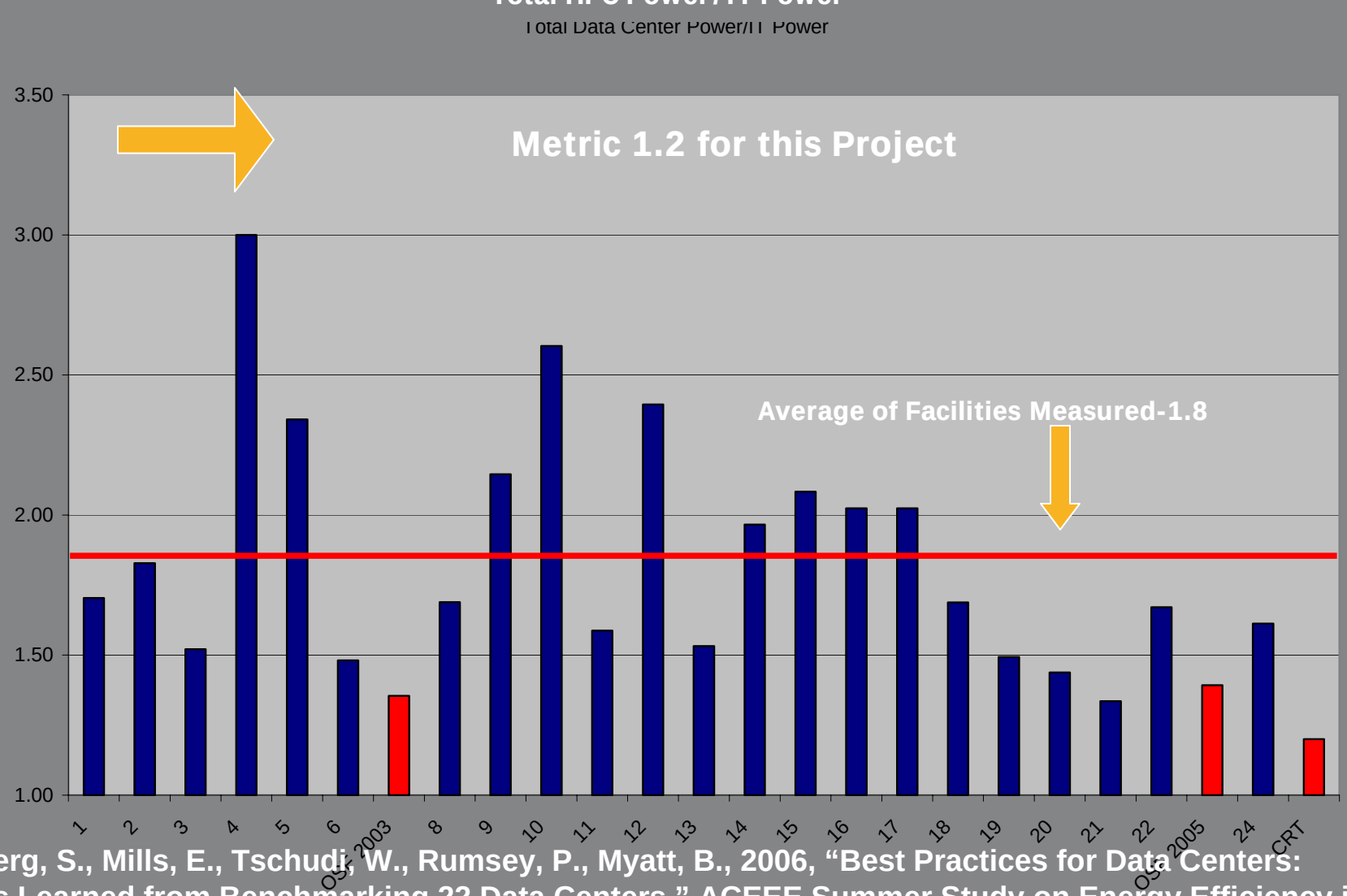
BERKELEY LAB

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



Greenberg, S., Mills, E., Tschudi, W., Rumsey, P., Myatt, B., 2006, "Best Practices for Data Centers: Lessons Learned from Benchmarking 22 Data Centers," ACEEE Summer Study on Energy Efficiency in Buildings, http://eetd.lbl.gov/emills/PUBS/PDF/ACEEE-datacenters.pdf.

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 Direction

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: more than **120 MW**
- **The current approach is not sustainable!**

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

NERSC
BERKELEY LAB
Office of Science
U.S. DEPARTMENT OF ENERGY

Path to Power Efficiency
Reducing Waste in Computing

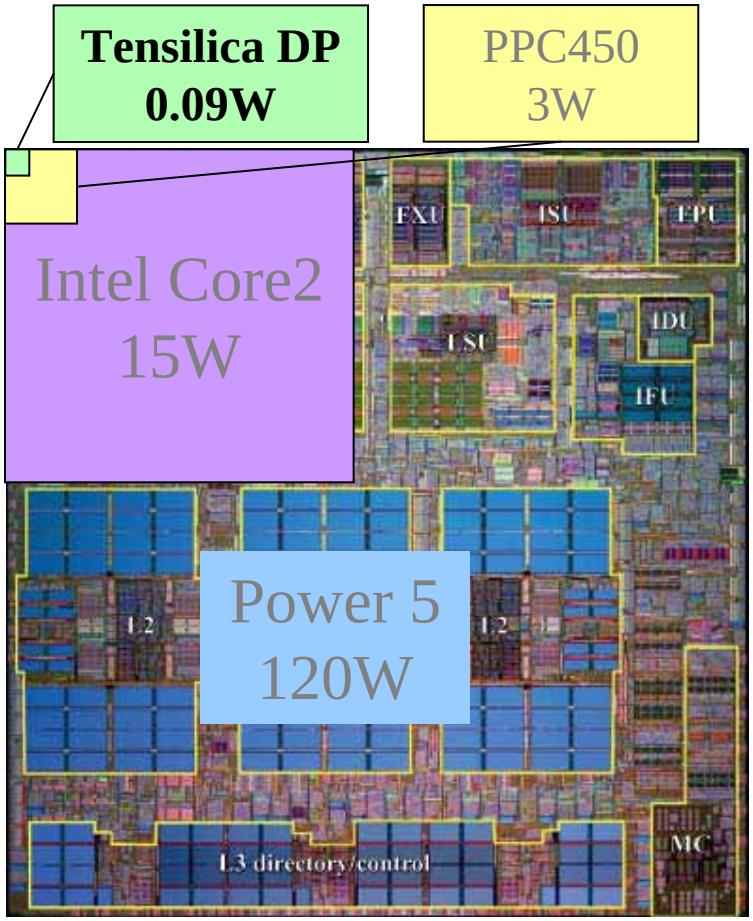
- Examine methodology of low-power embedded computing market
  - optimized for low power, low cost and high computational efficiency

"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

Design for Low Power: More Concurrency

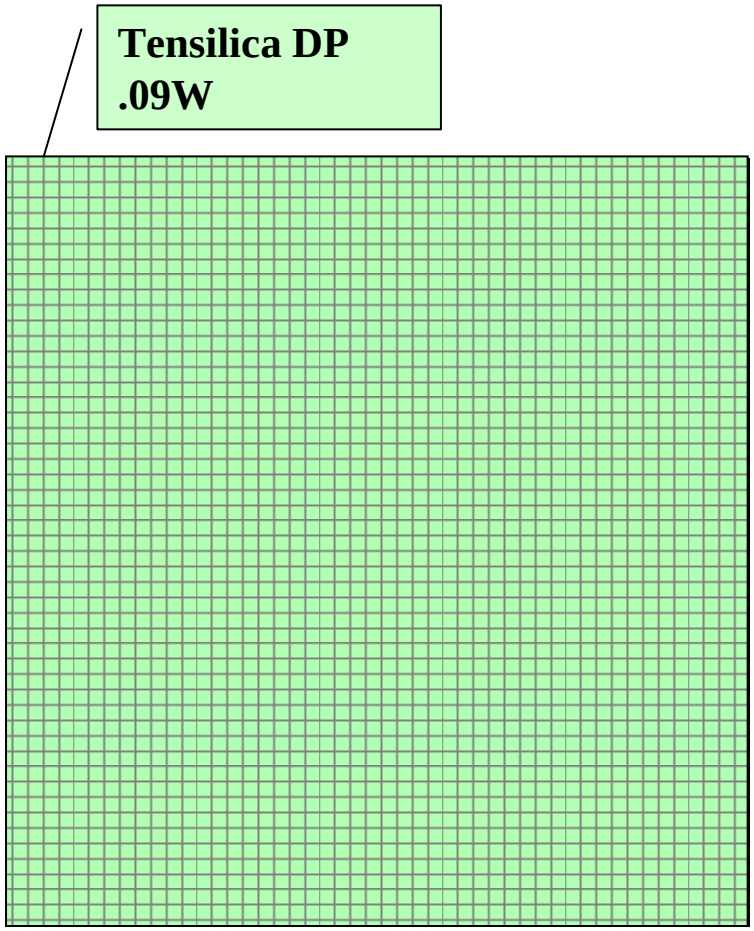


- Cubic power improvement with lower clock rate due to V^2F
- Slower clock rates enable use of simpler cores
- Simpler cores use less area (lower leakage) and reduce cost
- Tailor design to application to reduce waste$ $

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

NERSC
BERKELEY LAB
Office of Science
U.S. DEPARTMENT OF ENERGY

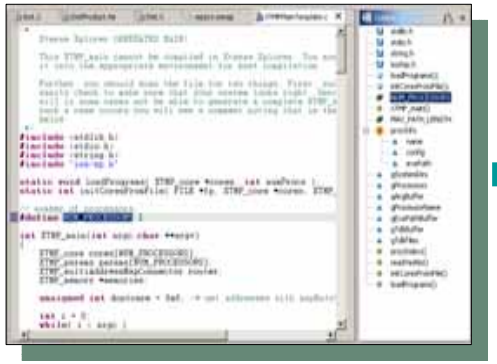
Low Power Design Principles



- 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

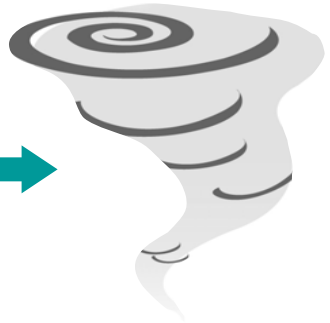
Even if each core operates at 1/3 to 1/10th efficiency of largest chip, you can pack 100s more cores onto a chip and consume 1/20 the power

Embedded Design Automation
(Example from Existing Tensilica Design Flow)



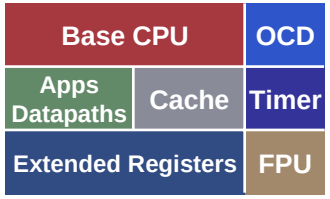
Processor configuration

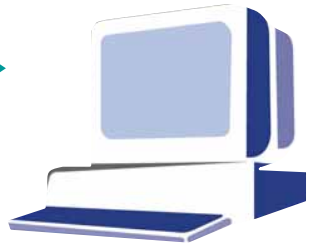
1. Select from menu
2. Automatic instruction discovery (XPRES Compiler)
3. Explicit instruction description (TIE)



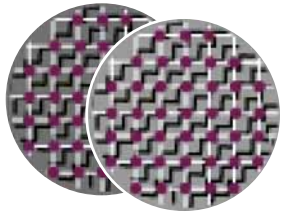
Processor Generator (*Tensilica*)

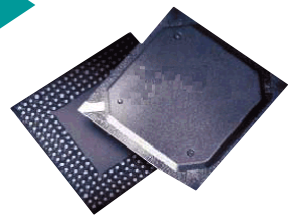
Application-optimized processor implementation (RTL/Verilog)





Tailored SW Tools: Compiler, debugger, simulators, Linux, other OS Ports *(Automatically generated together with the Core)*





Build with any process in any fab

Advanced Hardware Simulation (RAMP)

- **Research Accelerator for Multi-Processors (RAMP)**
  - **Utilize FGPA boards to emulate large-scale multicore systems**
  - **Simulate hardware before it is built**
  - **Break slow feedback loop for system designs**
  - **Allows fast performance validation**
  - **Enables tightly coupled hardware/software/science co-design** *(not possible using conventional approach)*

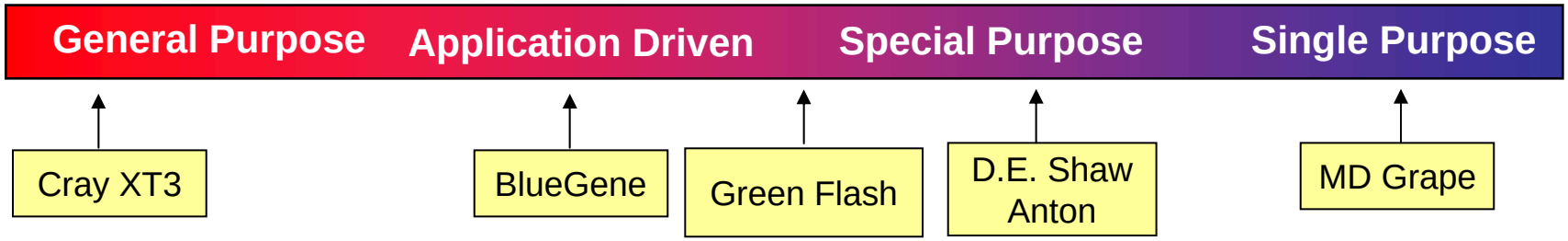
- **Technology partners:**
  - **UC Berkeley: John Wawrzynek, Jim Demmel, Krste Asanovic, Kurt Keutzer**
  - **Stanford University / Rambus Inc.: Mark Horowitz**
  - **Tensilica Inc.: Chris Rowen**

NERSC

BERKELEY LAB

Office of Science
U.S. DEPARTMENT OF ENERGY

Customization Continuum: Green Flash



- **Application-driven does NOT necessitate a special purpose machine**
- **MD-Grape: Full custom ASIC design**
  - **1 Petaflop performance for one application using 260 kW for $9M**
- **D.E. Shaw Anton System: Full and Semi-custom design**
  - **Simulate 100x–1000x timescales vs any existing HPC system (~200kW)**
- **Application-Driven Architecture (Green Flash): Semicustom design**
  - **Highly programmable core architecture using C/C++/Fortran**
  - **Goal of 100x power efficiency improvement vs general HPC approach**
  - **Better understand how to build/buy application-driven systems**
  - **Potential: 1km-scale model (~200 Petaflops peak) running in O(5 years)**

NERSC
BERKELEY LAB
Office of Science
U.S. DEPARTMENT OF ENERGY

Green Flash Strawman System Design

We examined three different approaches (in 2008 technology)

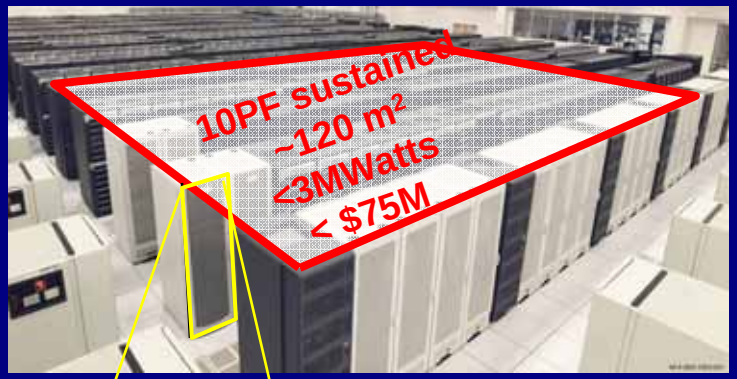
Computation .015°X.02°X100L: 10 PFlops sustained, ~200 PFlops peak

- 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

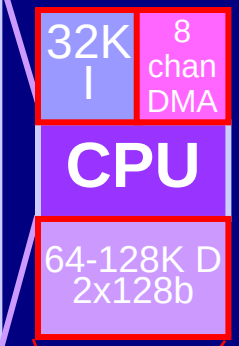
Climate System Design Concept

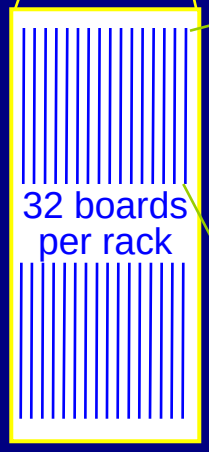
Strawman Design Study



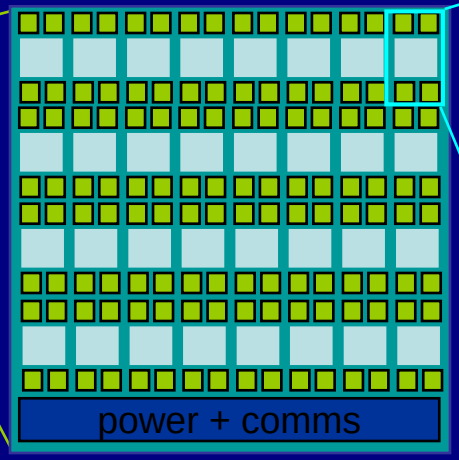
VLIW CPU:

- 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

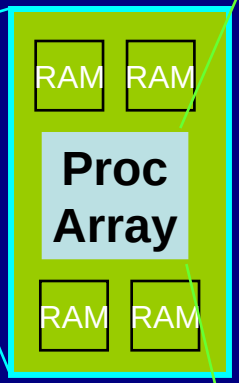




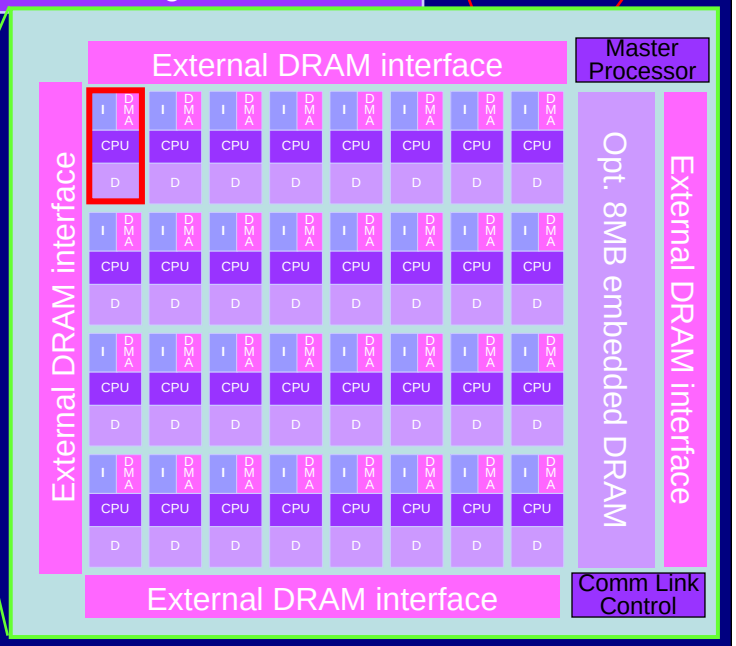
100 racks @ ~25KW



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

Portable Performance for Green Flash

- Challenge: Our approach would produce multiple architectures, each different in the details
  - **Labor-intensive user optimizations for each specific architecture**
  - **Different architectural solutions require vastly different optimizations**
  - **Non-obvious interactions between optimizations & HW yield best results**

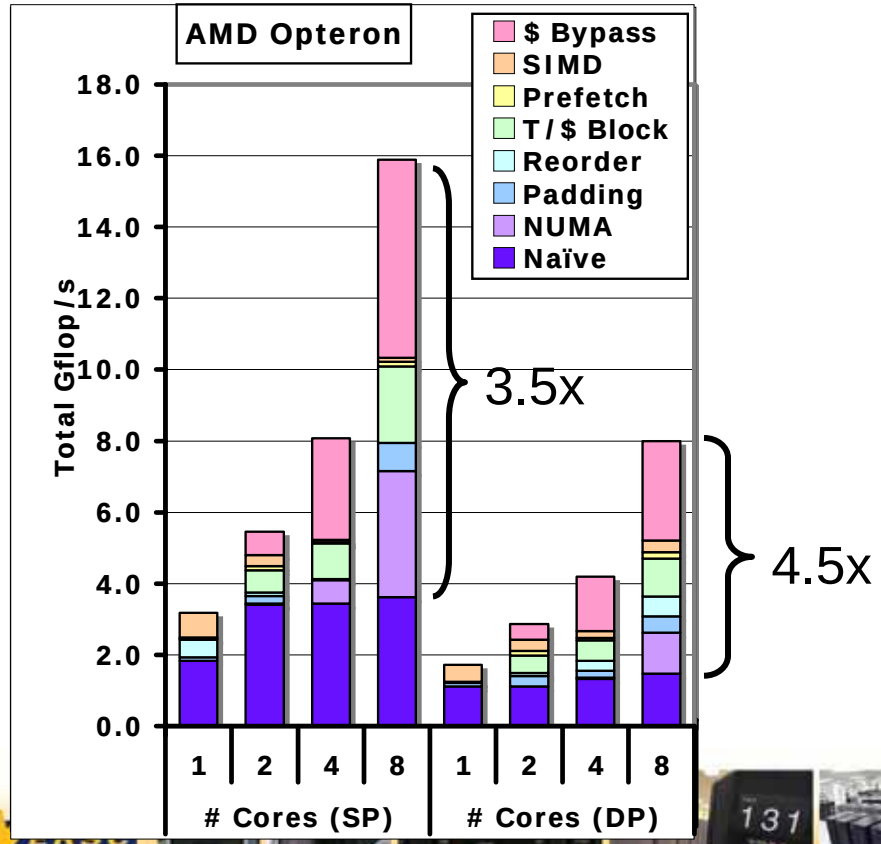
- Our solution: Auto-tuning
  - Automate search across a complex optimization space
  - Achieve performance far beyond current compilers
  - Attain performance portability for diverse architectures

Auto-Tuning for Multicore

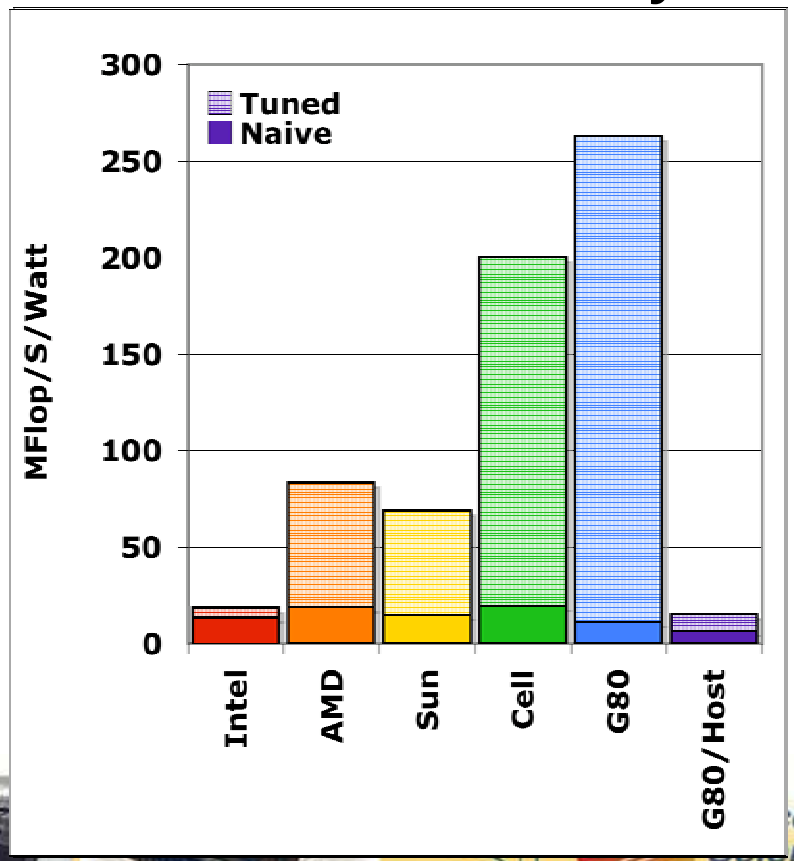
(finite-difference computation)

- **Take advantage of unique multicore features via auto-tuning**
- **Attains performance portability across different designs**
- **Only requires basic compiling technology**
- **Achieve high serial performance, scalability, and optimized power efficiency**

Performance Scaling



Power Efficiency

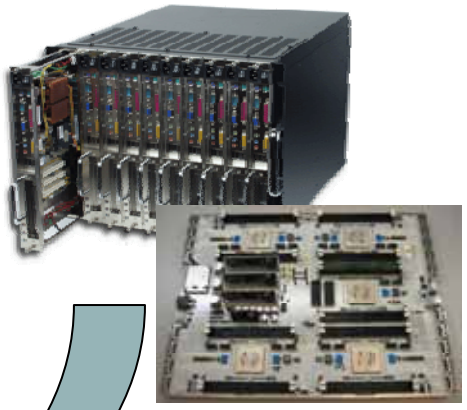


Traditional New Architecture Hardware/Software Design

How long does it take for a full scale application to influence architectures?

Design New System (2 year concept phase)

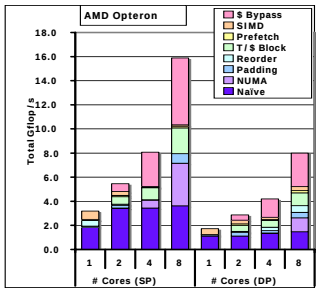
Build Hardware (2 years)



Cycle Time 4-6+ years

Port Application

Tune Software (2 years)

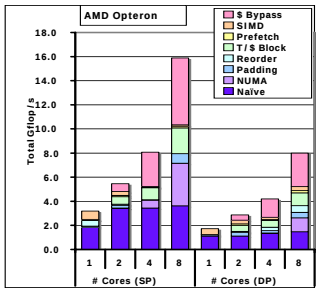


Proposed New Architecture Hardware/Software Co-Design

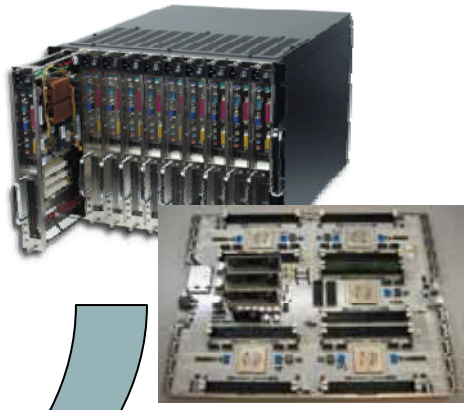
How long does it take for a full scale application to influence architectures?

Synthesize SoC (hours)

Autotune Software (Hours)



Cycle Time 1-2 days



Emulate Hardware (RAMP) (hours)

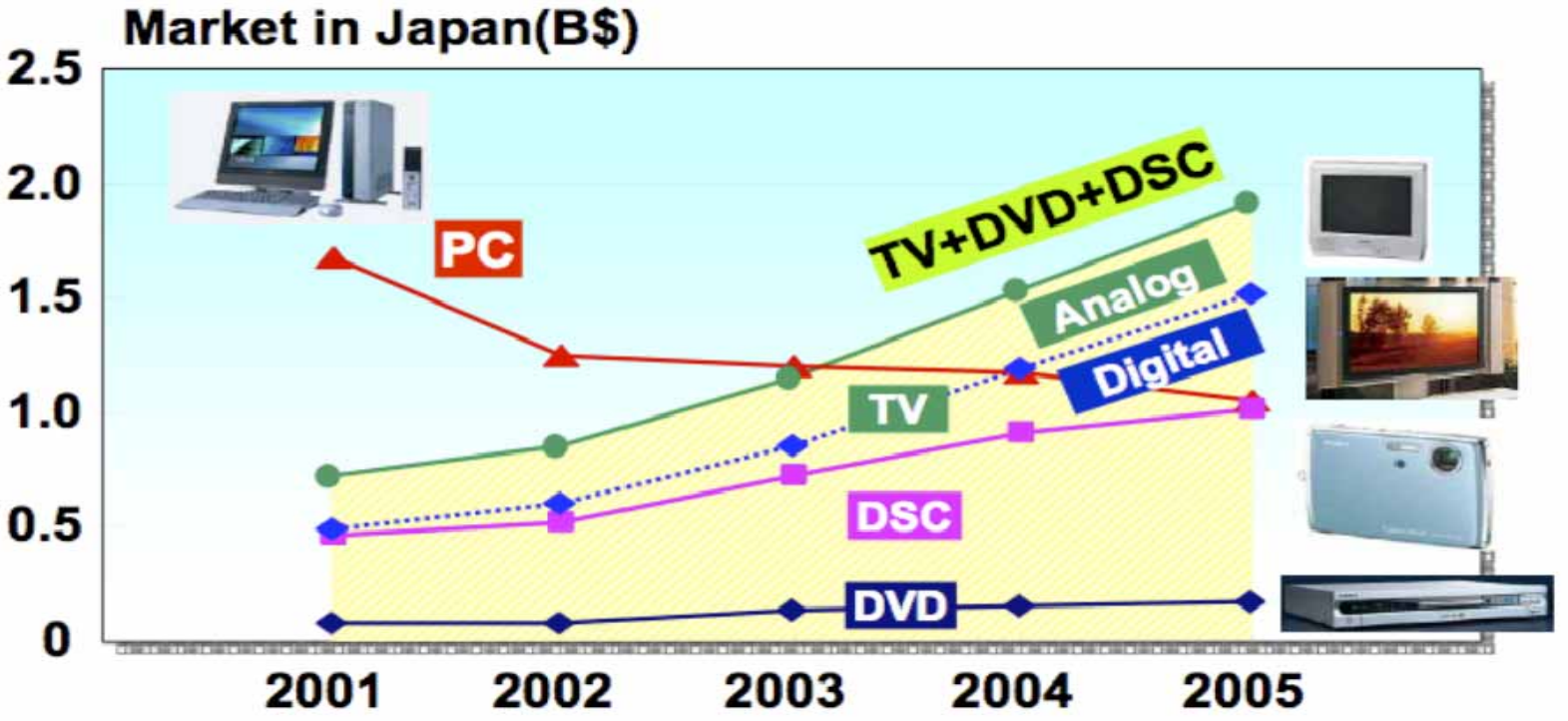
Build application

Outline

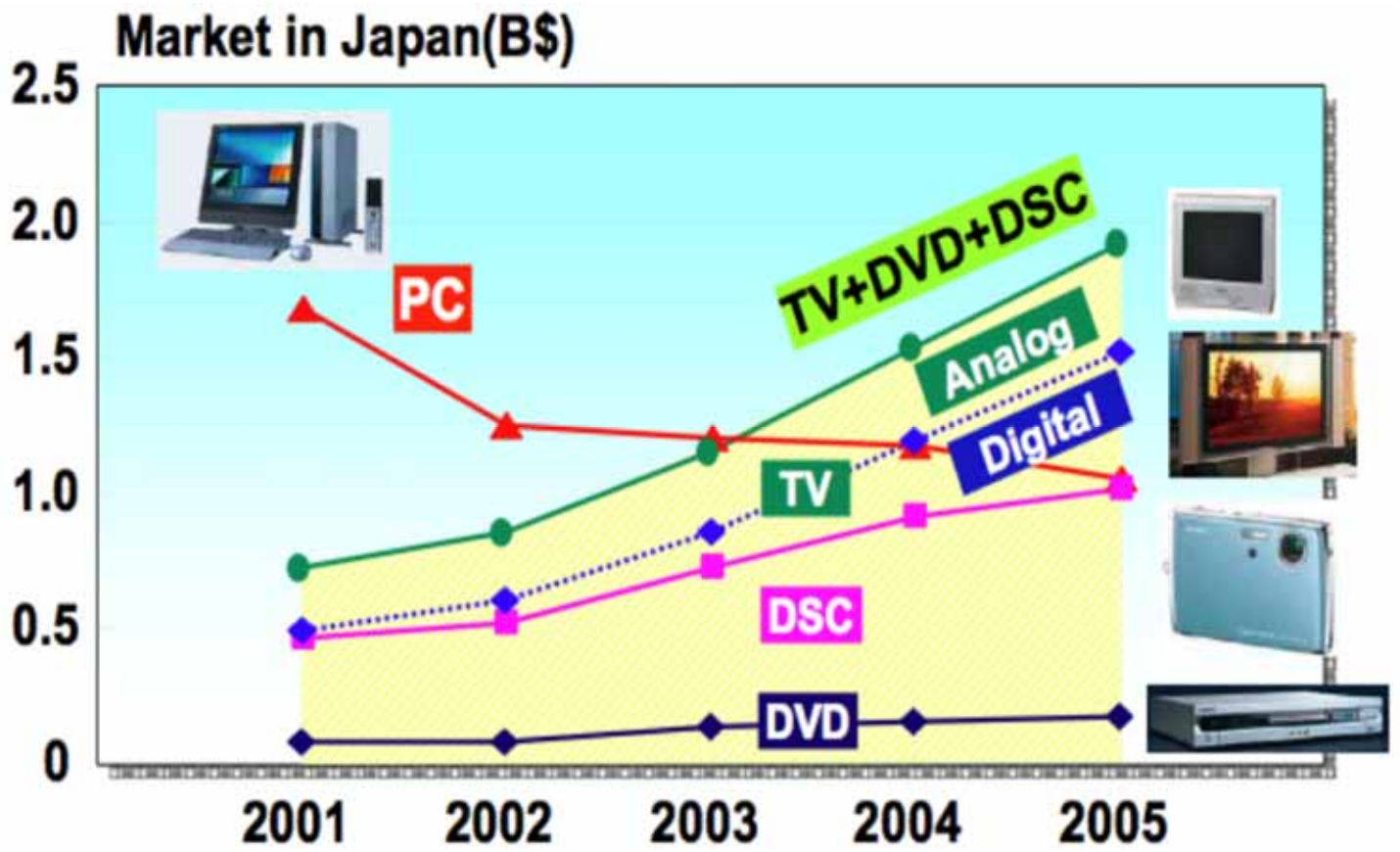
1. Power consumption has become an industry-wide issue for computing
2. Building and computer room energy efficiency
3. Computer architecture for energy efficiency- the Green Flash project
4. Future

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 Convergence

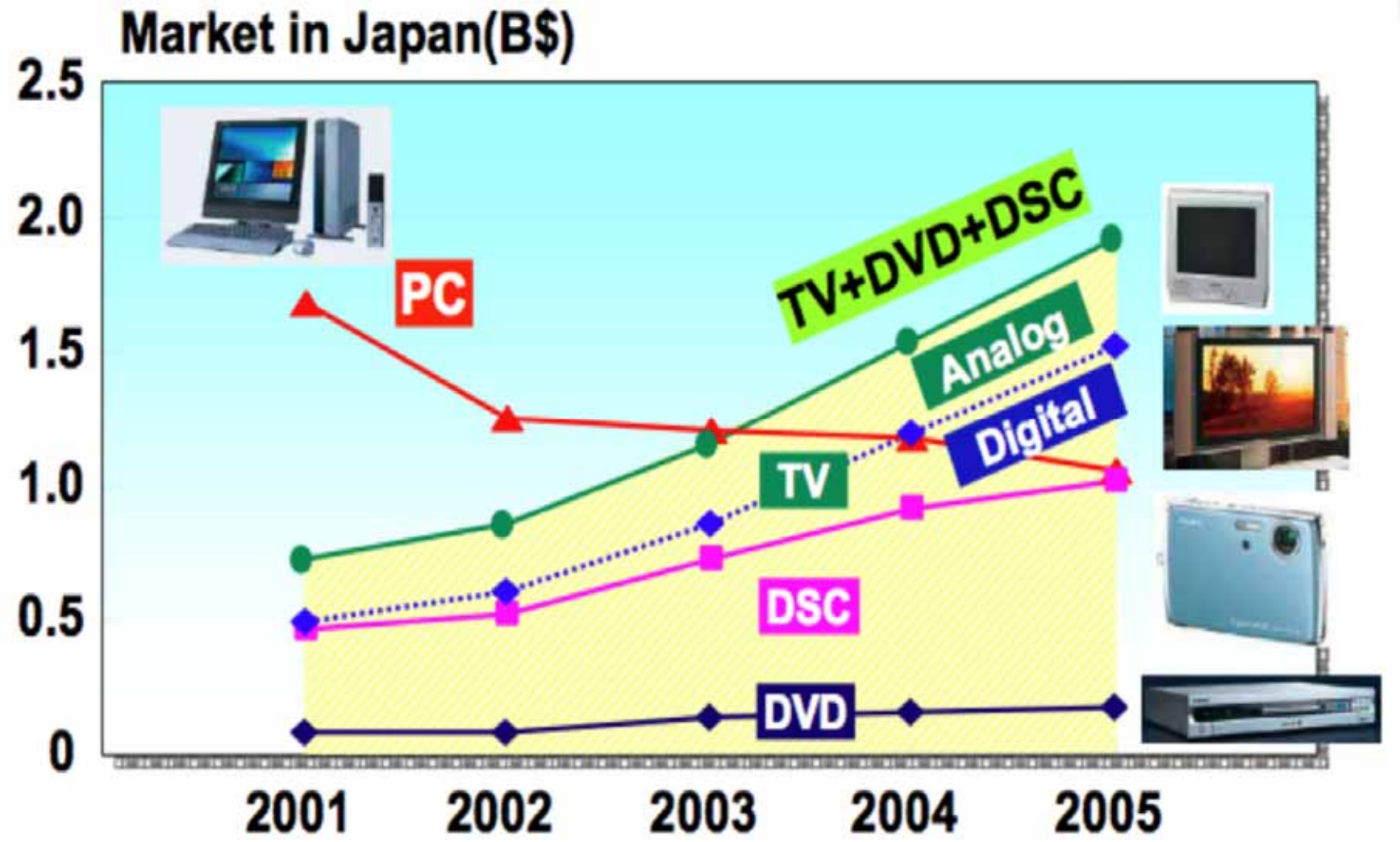


NERSC

Berkeley Lab

Office of Science
U.S. Department of Energy

Consumer Electronics Convergence

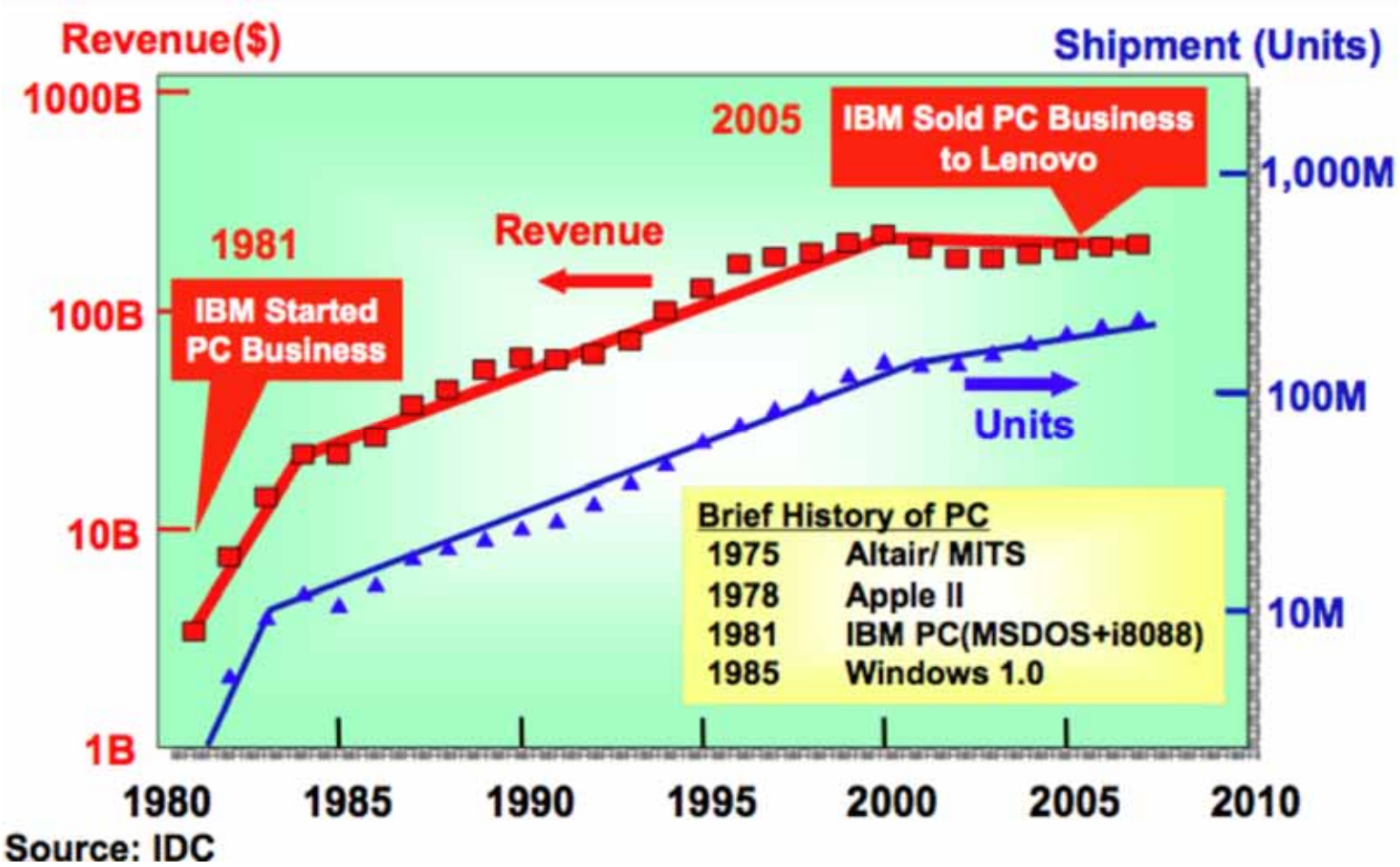


NERSC

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

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

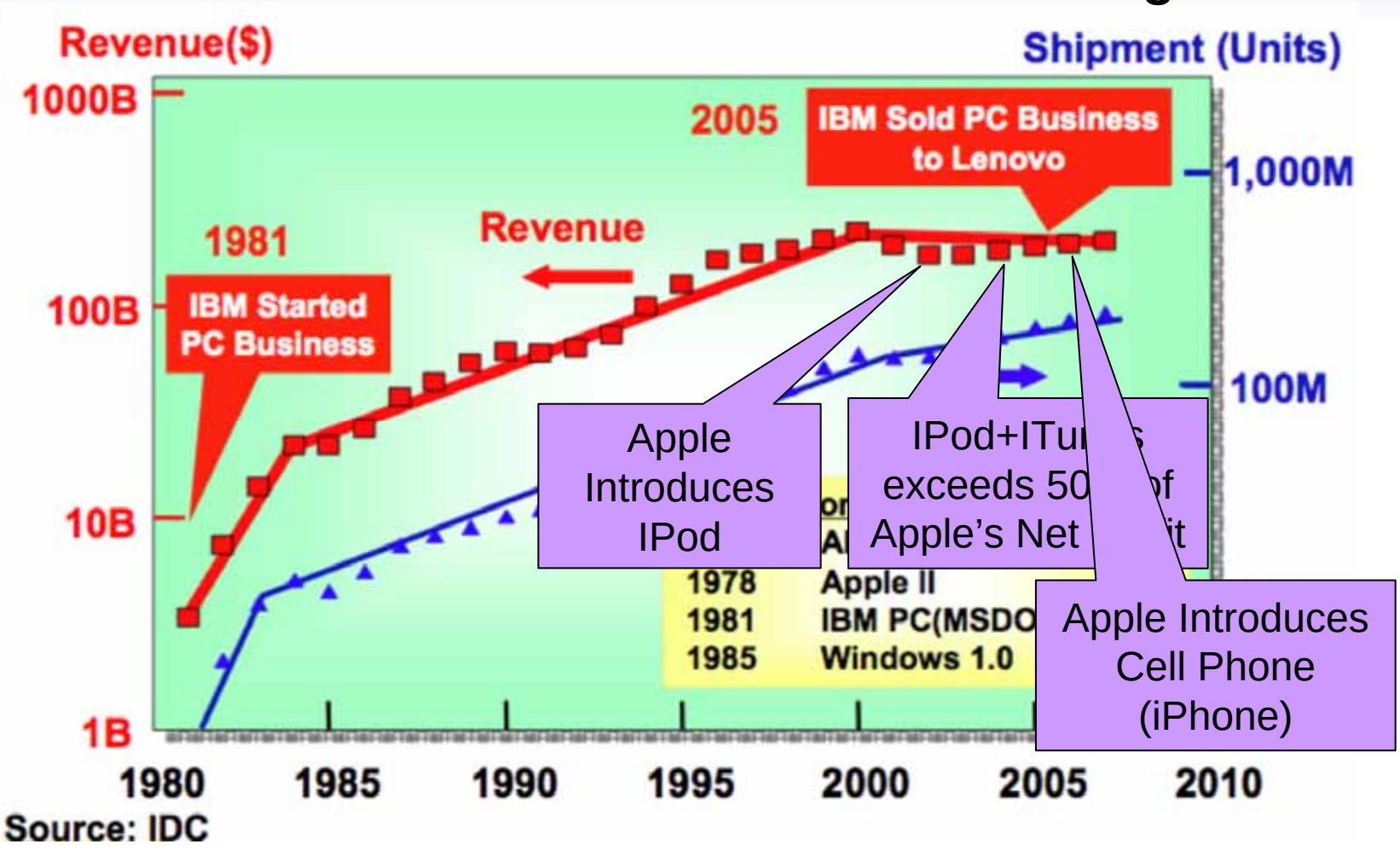


NERSC
BERKELEY LAB
Office of Science
U.S. DEPARTMENT OF ENERGY

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

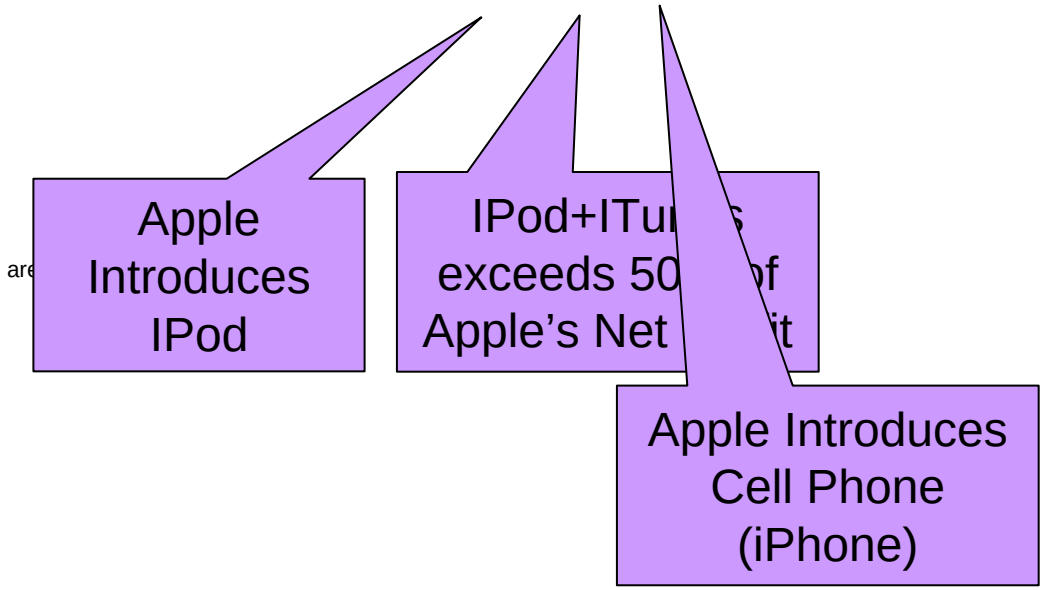
QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

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

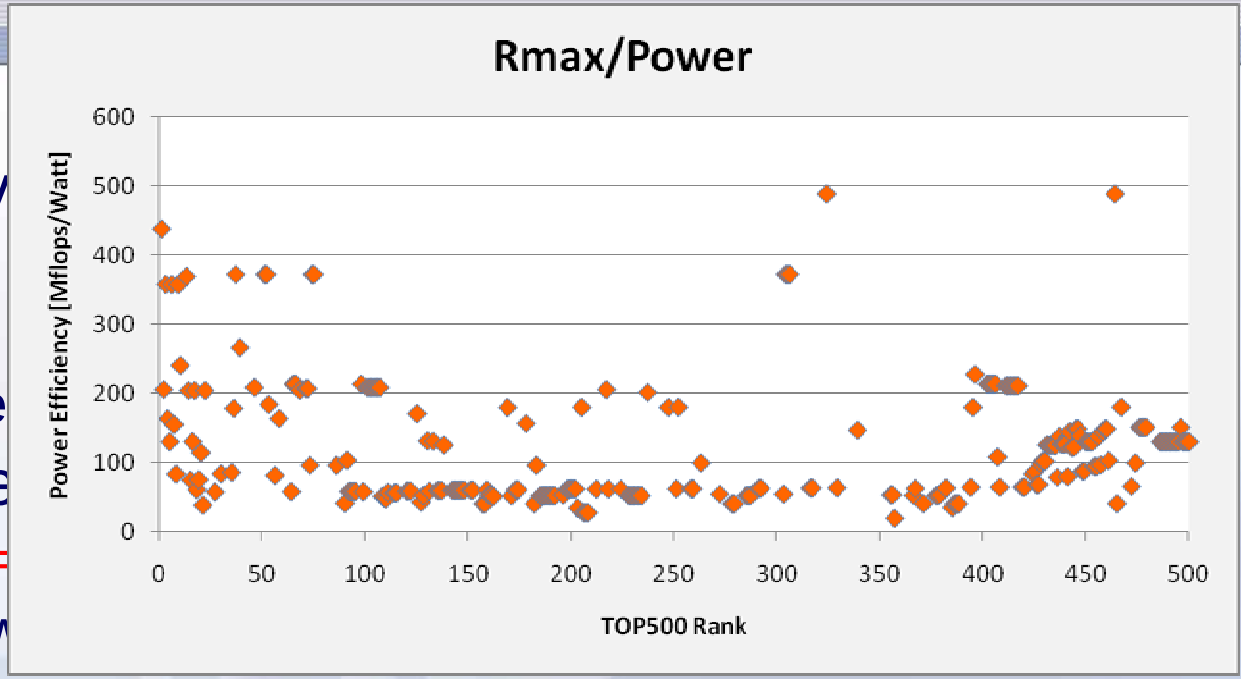


NERSC
Office of Science
U.S. DEPARTMENT OF ENERGY
BERKELEY LAB

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



Power Ranking and How Not to do it!



- To rank objects by
  - Weight or Volume
  - Rmax (TOP500)
    - A 'larger' syste
- The ratio of 2 exte
  - (weight/volumne =
  - Performance / Pow
- One *can-not* 'rank' objects with densities BY SIZE:
  - Density does not tell anything about size of an object
  - A piece of lead is not heavier or larger than one piece of wood.
- Linpack (sub-linear) / Power (linear)
  will always sort smaller systems before larger ones!

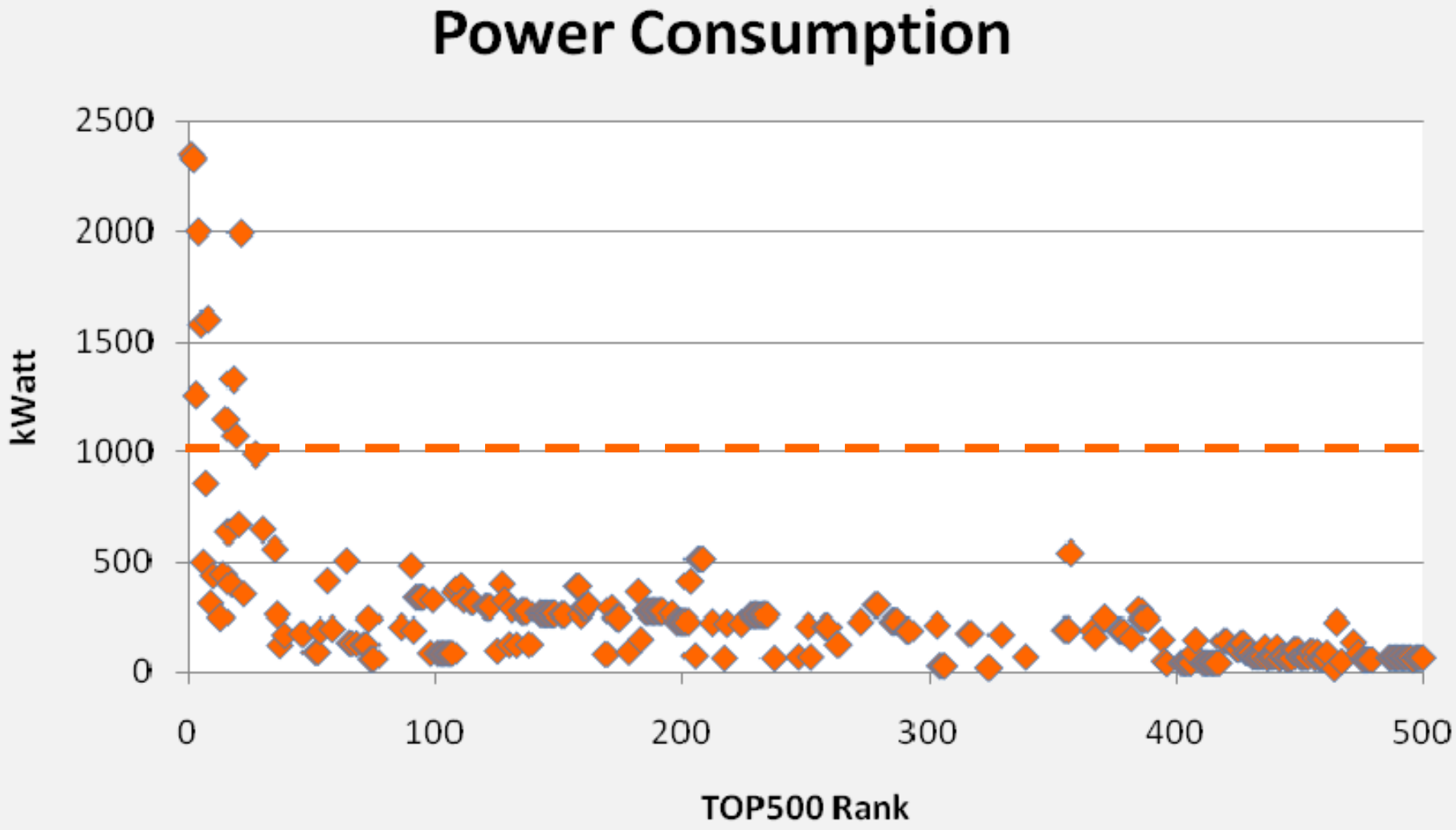
The Transition to Low-Power Technology is Inevitable

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

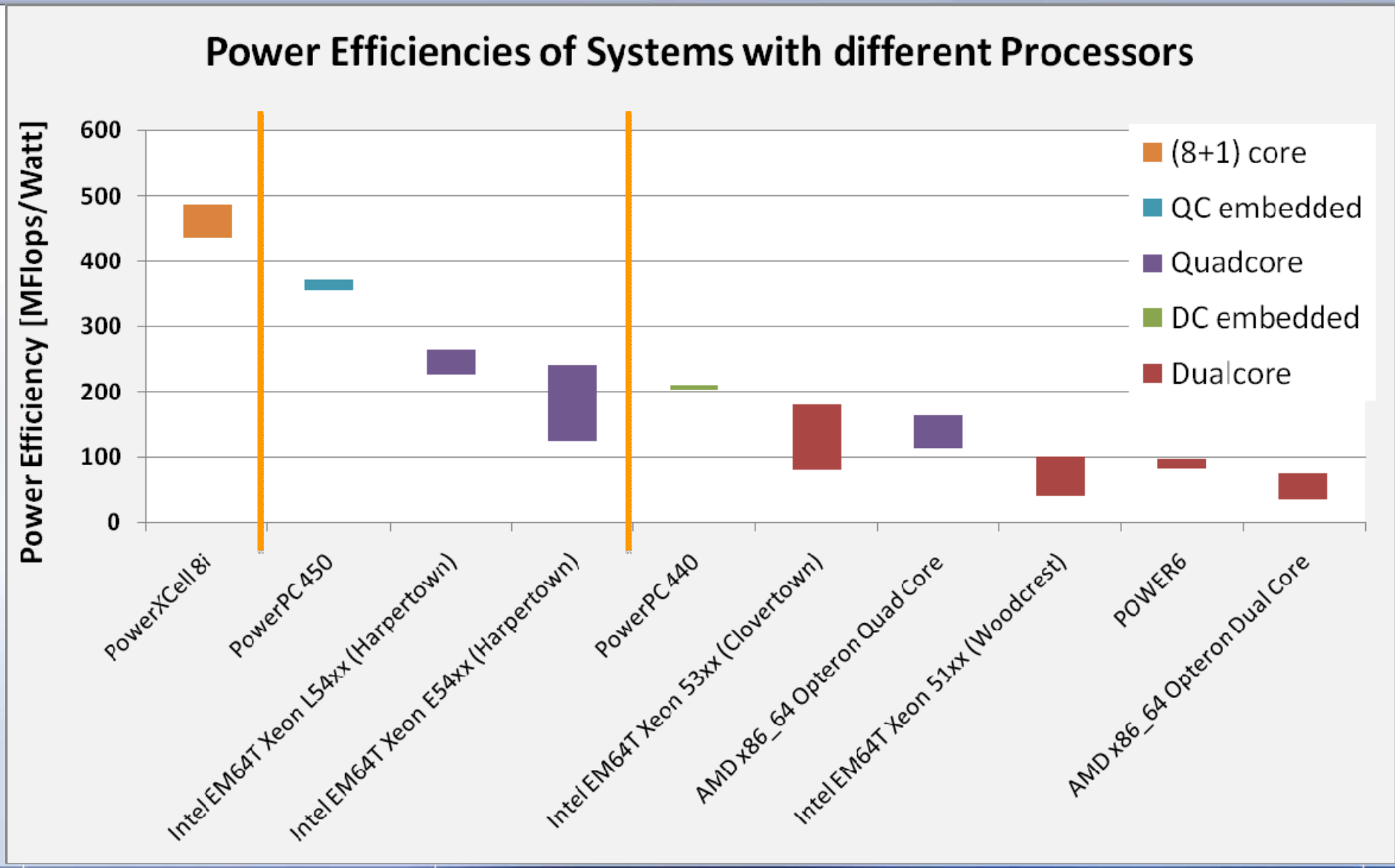
- Information "factories" are only affordable for a few government labs and large commercial companies (Google, MSN, Yahoo …)
  - Midrange installations will soon hit the 1 - 2 MW wall, requiring costly new installations
  - Economics will change if operating expenses of a server exceed acquisition cost
- 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

NERSC
BERKELEY LAB
Office of Science
U.S. DEPARTMENT OF ENERGY

Absolute Power Levels

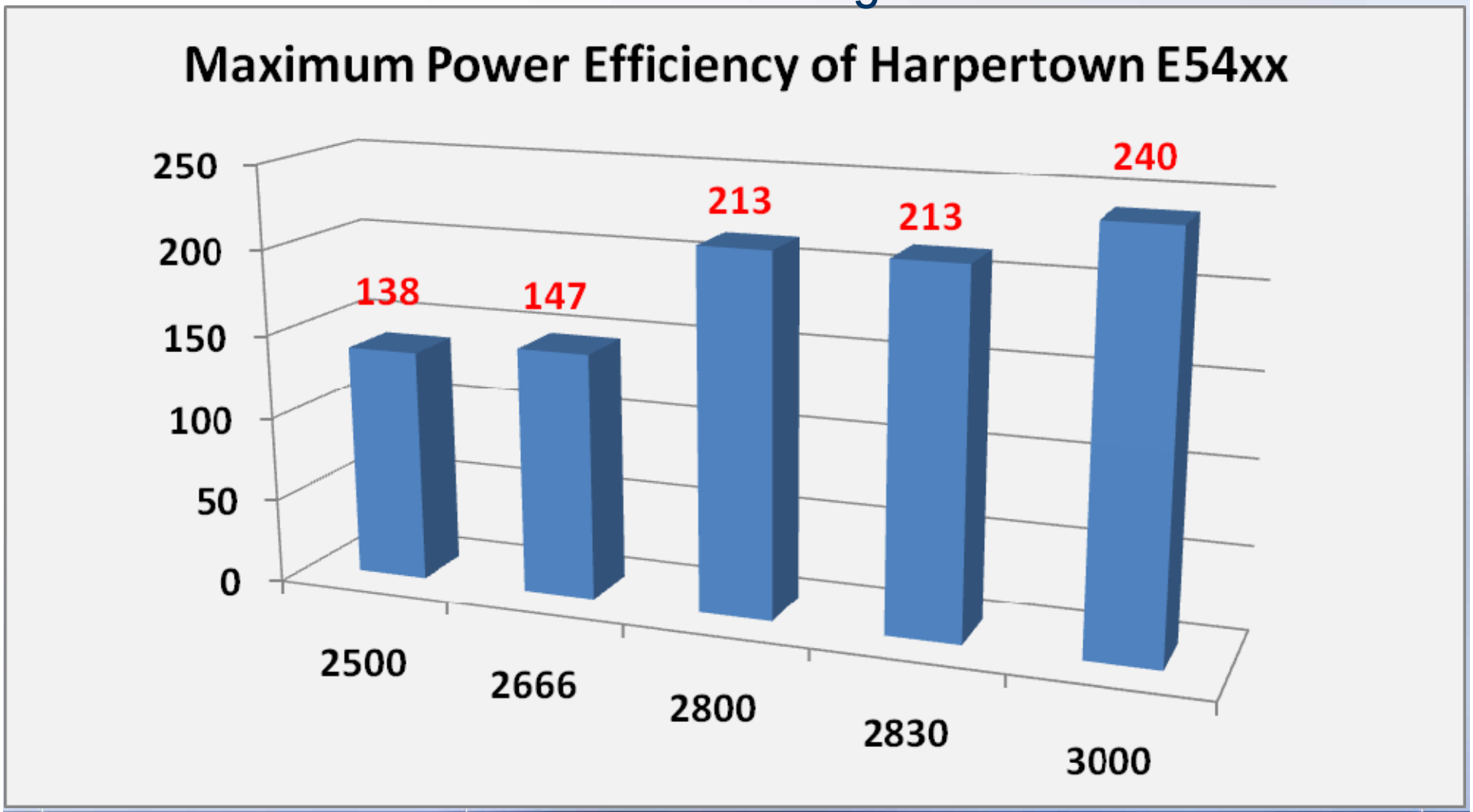


Power Efficiency related to Processors

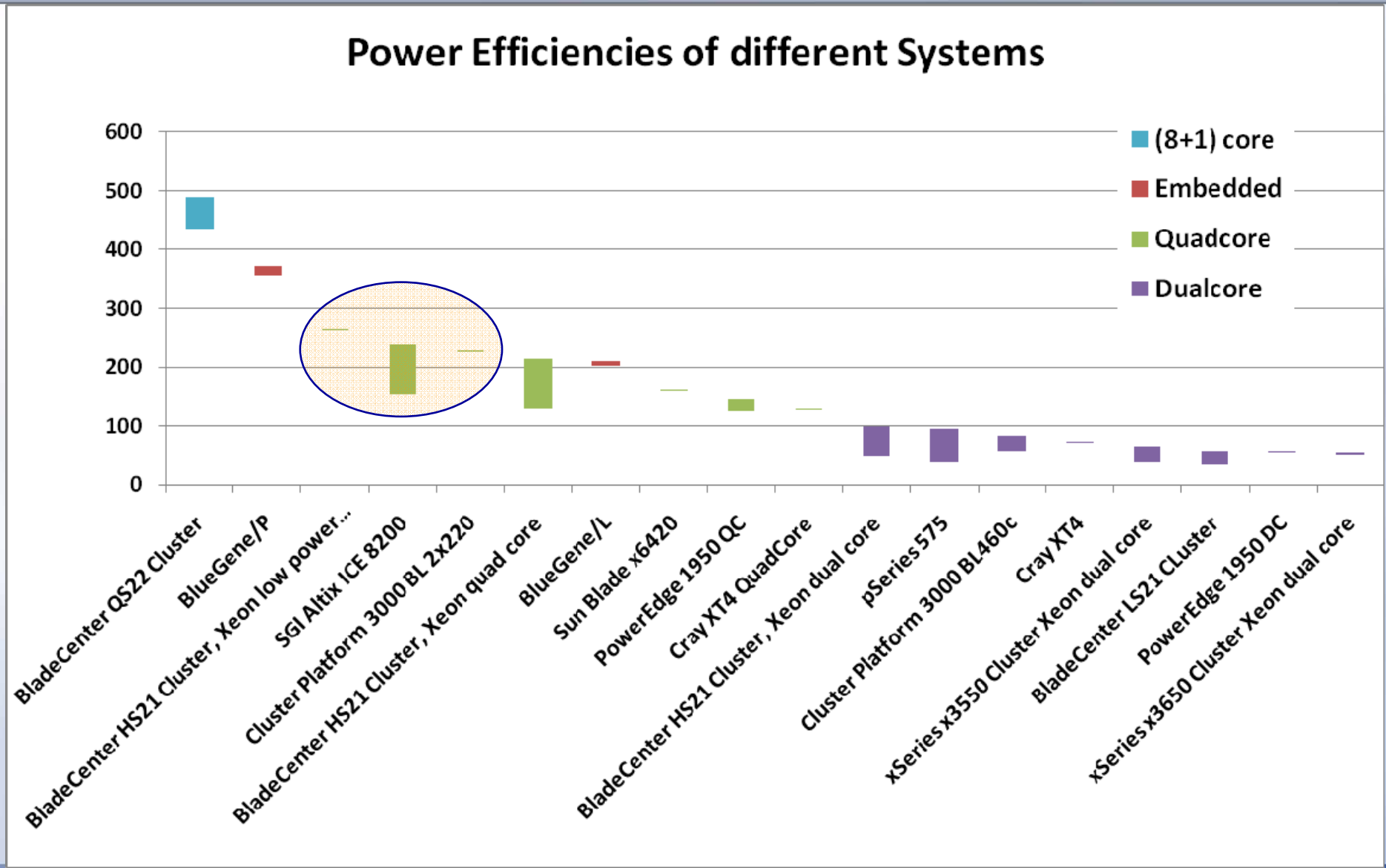


Frequencies and Power Efficiency

Power rating is 80 Watts each!

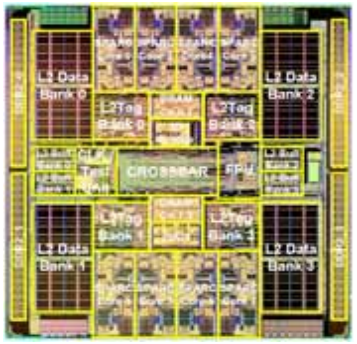


Most Power Efficient Systems



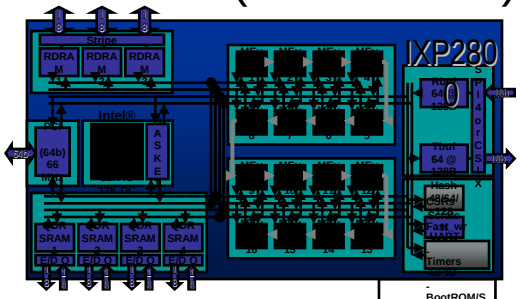
Convergence of Platforms

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

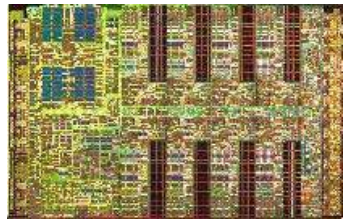


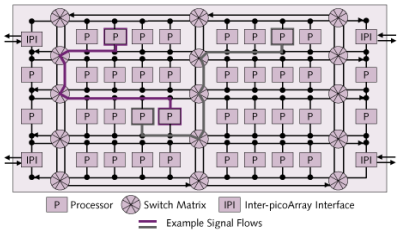
Sun Niagara
8 GPP cores (32 threads)

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

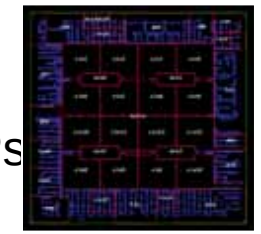


IBM Cell
1 GPP (2 threads)
8 ASPs

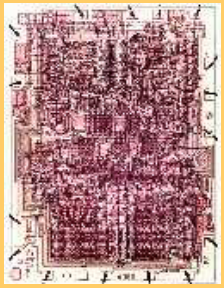




Picochip DSP
1 GPP core
248 ASPs



Cisco CRS-1
188 Tensilica GPPs



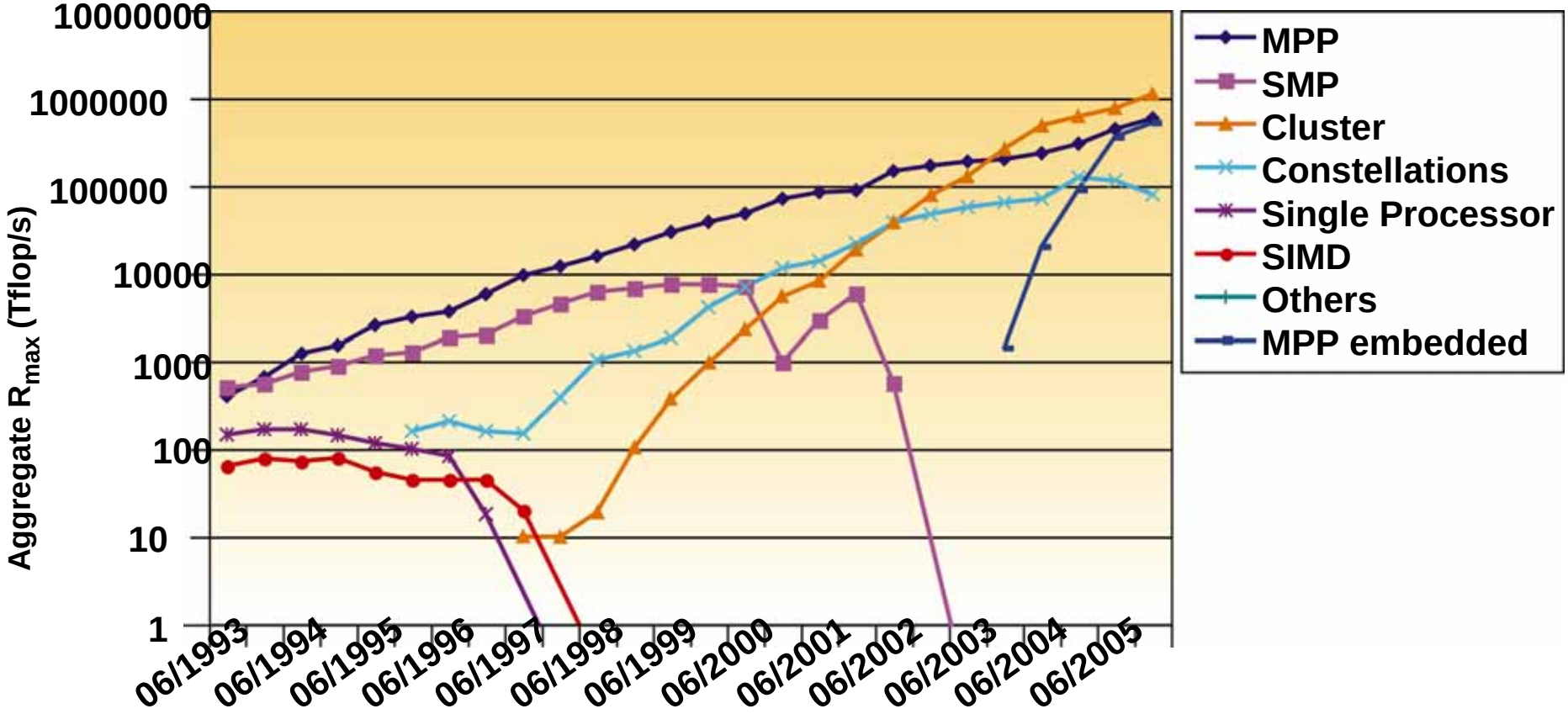
Intel 4004 (1971):
4-bit processor,
2312 transistors,
~100 KIPS,
10 micron PMOS,
11 mm² chip

1000s of processor cores per die

"The Processor is the new Transistor" [Rowen]

BG/L—the Rise of the Embedded Processor

TOP 500 Performance by Architecture



NERSC
BERKELEY LAB
Office of Science
U.S. DEPARTMENT OF ENERGY

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**

Happy 60th Birthday!

QuickTime™ and a
decompressor
are needed to see this picture.

QuickTime™ and a
decompressor
are needed to see this picture.

… and keep up with "green" computing and commuting

NERSC

BERKELEY LAB

Office of Science
U.S. DEPARTMENT OF ENERGY