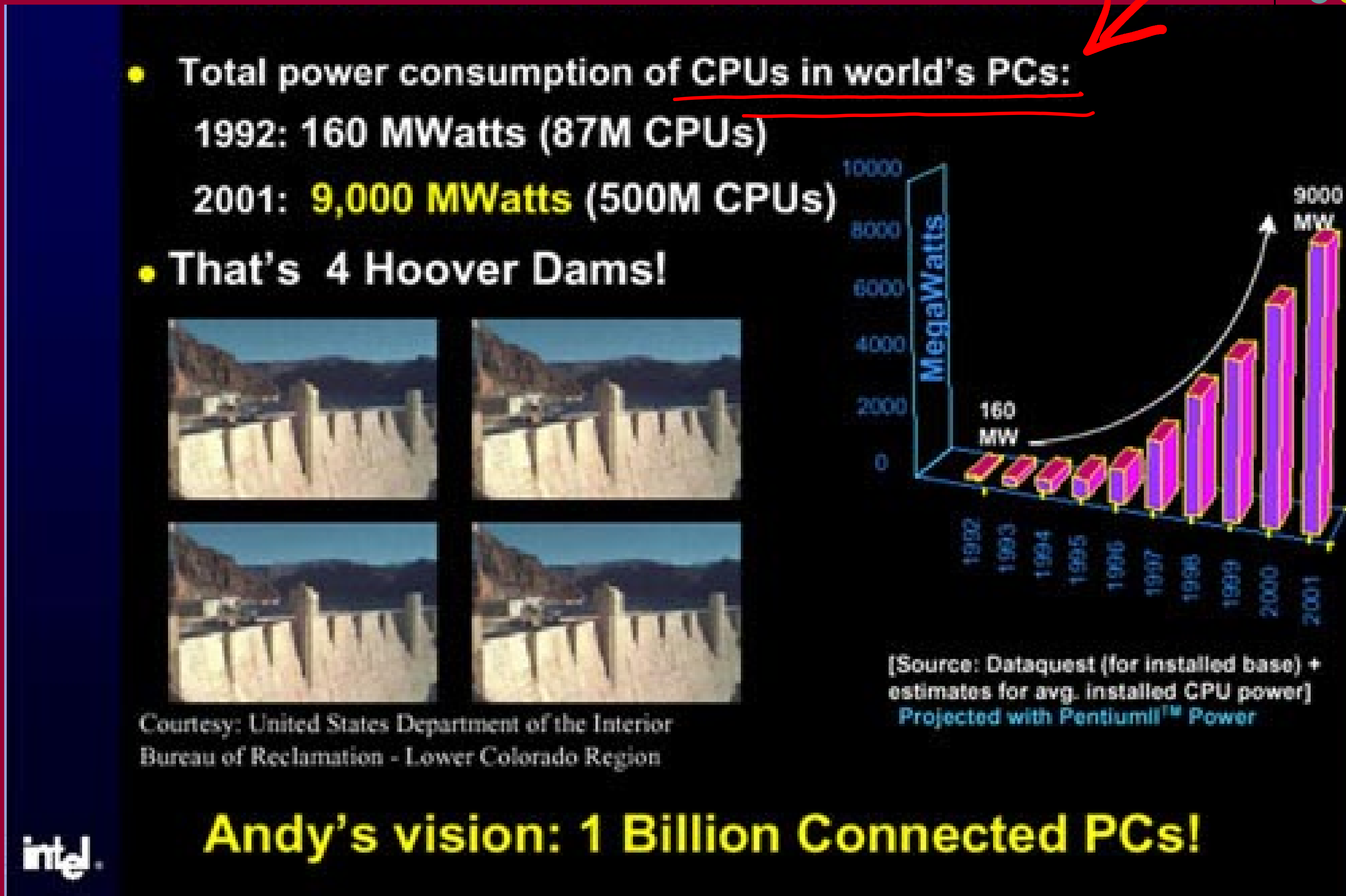
Global Climate Warming? Yes … In The Machine Room

Wu FENG
feng@cs.vt.edu

Departments of Computer Science and
Electrical & Computer Engineering

Time step: 1019

Environmental Burden of PC CPUs



Source: Cool Chips & Micro 32

Power Consumption of World's CPUs

Year	Power (in MW)	# CPUs (in millions)
1992	180	87
1994	392	128
1996	959	189
1998	2,349	279
2000	5,752	412
2002	14,083	607
2004	34,485	896
2006	87,439	1,321

Power Consumption of World's CPUs

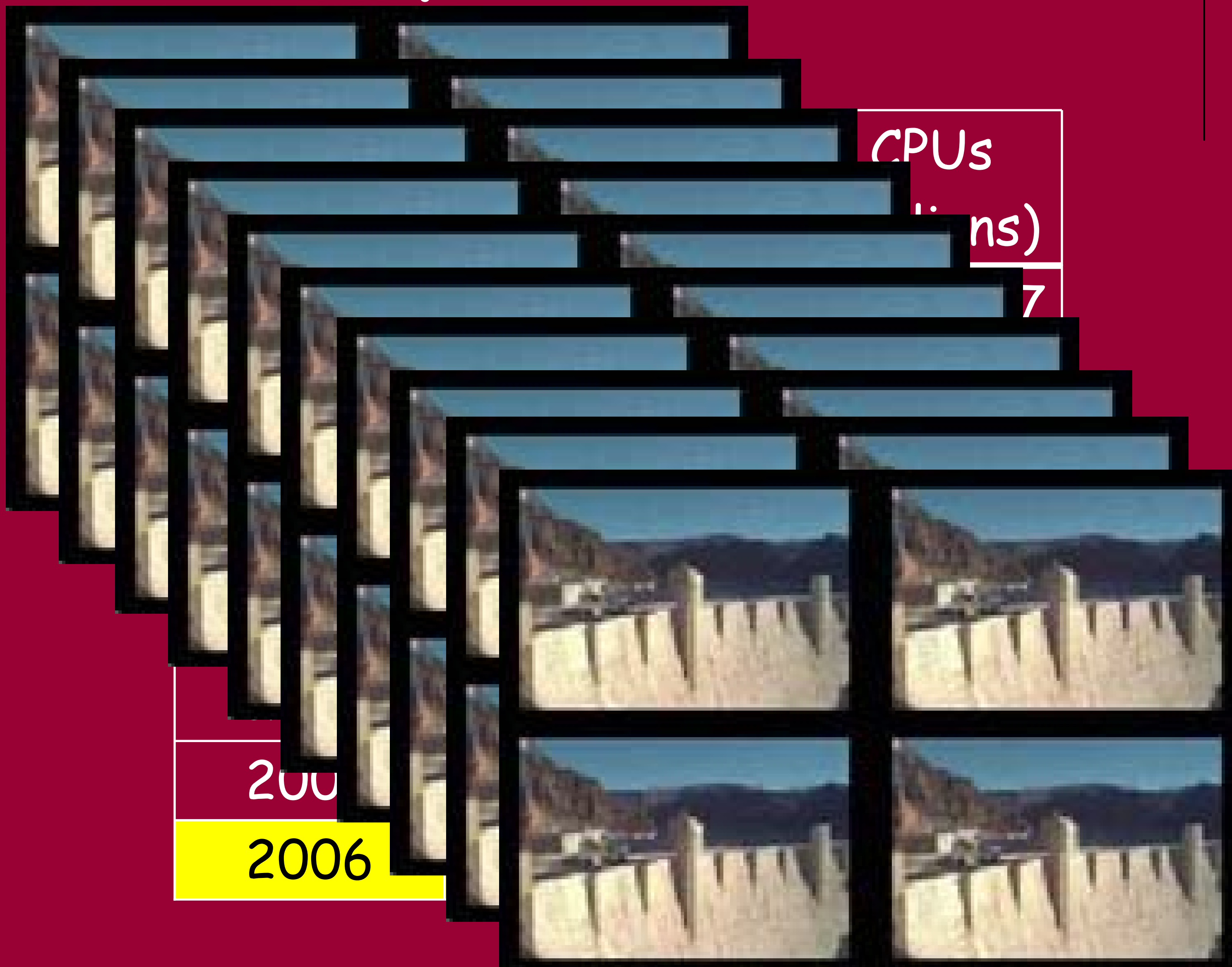
CPUs

...ns)

7

200...

2006



W. Feng, feng@cs.vt.edu, (540) 231-1192

And Now We Want Petascale …

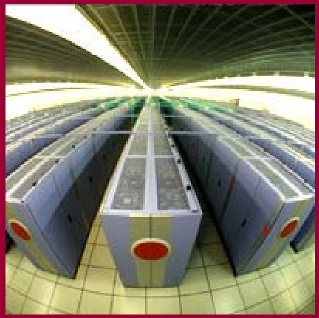
High-Speed Train
10 Megawatts

Conventional Power Plant
300 Megawatts

- What is a conventional petascale machine?
  - Many high-speed bullet trains … a significant start to a conventional power plant.
  - "Hiding in Plain Sight, Google Seeks More Power," *The New York Times*, June 14, 2006.

Top Three Reasons for "Eliminating" Global Climate Warming in the Machine Room

3. HPC "Contributes" to Global Climate Warming :-)
   - "I worry that we, as HPC experts in global climate modeling, are contributing to the very thing that we are trying to avoid: the generation of greenhouse gases." - Noted Climatologist
2. Electrical Power Costs $$$.
   - Japanese Earth Simulator
     - Power & Cooling: 12 MW/year → $9.6 million/year?
   - Lawrence Livermore National Laboratory
     - Power & Cooling of HPC: $14 million/year
     - Power-up ASC Purple → "Panic" call from local electrical company.



1. Reliability & Availability Impact Productivity
   - California: State of Electrical Emergencies (July 24-25, 2006)
     - 50,538 MW: A load not expected to be reached until **2010**!

Reliability & Availability of HPC

Systems	CPUs	Reliability & Availability
ASCI Q	**8,192**	**MTBI: 6.5 hrs.** 114 unplanned outages/month. ❖ HW outage sources: storage, CPU, memory.
ASCI White	**8,192**	**MTBF: 5 hrs. (2001) and 40 hrs. (2003).** ❖ HW outage sources: storage, CPU, 3rd-party HW.
NERSC Seaborg	**6,656**	**MTBI: 14 days. MTTR: 3.3 hrs.** ❖ SW is the main outage source. **Availability: 98.74%.**
PSC Lemieux	**3,016**	**MTBI: 9.7 hrs.** **Availability: 98.33%.**
Google	**~15,000**	**20 reboots/day; 2-3% machines replaced/year.** ❖ HW outage sources: storage, memory. **Availability: ~100%.**

MTBI: mean time between interrupts; MTBF: mean time between failures; MTTR: mean time to restore

Source: Daniel A. Reed, RENCI

Reliability & Availability of HPC

Systems	CPUs	Reliability & Availability
ASCI Q	8,192	**MTBI: 6.5 hrs.** 114 unplanned outages/month. …mory.
ASCI White		…W.
NER… Sea…		
PSC Lemieux		
Google	~15,000	**20 reboo… …% machines replaced/year.** HW outage sources: storage, memory. **Availability: ~100%.**

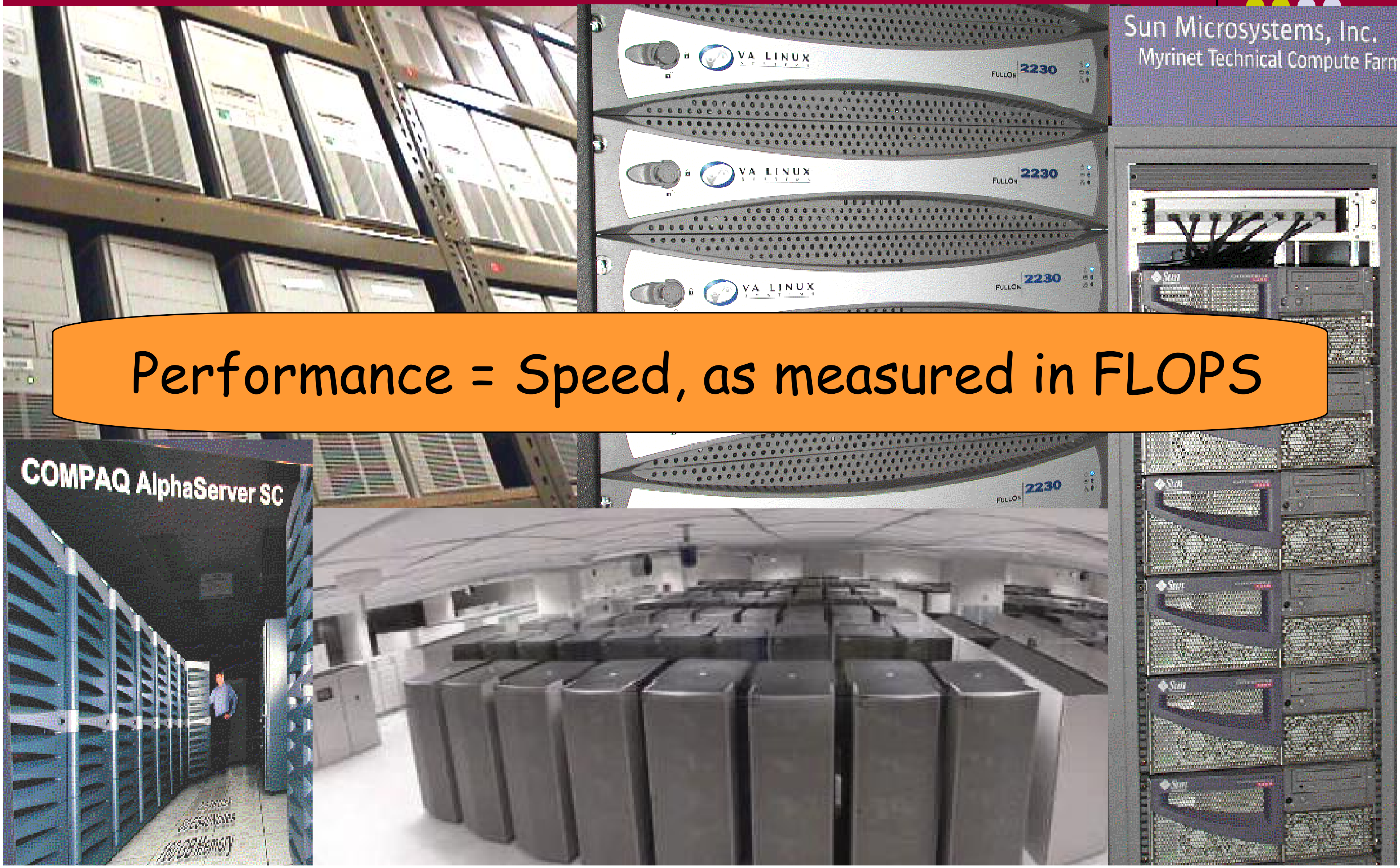
How in the world did we end up in this "predicament"?

MTBI: me… …etween interrupts; MTBF: mean time between failures; MTTR: mean time to restore

Source: Daniel A. Reed, RENCI

What Is Performance?

(Picture Source: T. Sterling)



Performance = Speed, as measured in FLOPS

Unfortunate Assumptions in HPC

Adapted from David Patterson, UC-Berkeley

- Humans are largely infallible.
  - Few or no mistakes made during integration, installation, configuration, maintenance, repair, or upgrade.
- Software will eventually be bug free.
- Hardware MTBF is already very large (~100 years between failures) and will continue to increase.
- Acquisition cost is what matters; maintenance costs are irrelevant.

- These assumptions are arguably at odds with what the traditional Internet community assumes.
  - Design robust software under the assumption of hardware unreliability.

Unfortunate Assumptions in HPC

Adapted from David Patterson, UC-Berkeley

- Humans are largely infallible.
  - Few or no mistake… …rin… …n,
- S…

… proactively address issues of continued hardware unreliability via lower-power hardware and/or robust software *transparently*.

- These assumptions ar… …at odds with what th… …al Internet community assumes.
  - Design robust software under the assumption of hardware …reliability.

Supercomputing in Small Spaces
(Established 2001)

- Goal
  - Improve efficiency, reliability, and availability (ERA) in large-scale computing systems.
    - Sacrifice a little bit of raw performance.
    - Improve overall system throughput as the system will "always" be available, i.e., effectively no downtime, no HW failures, etc.
  - Reduce the total cost of ownership (TCO). Another talk …

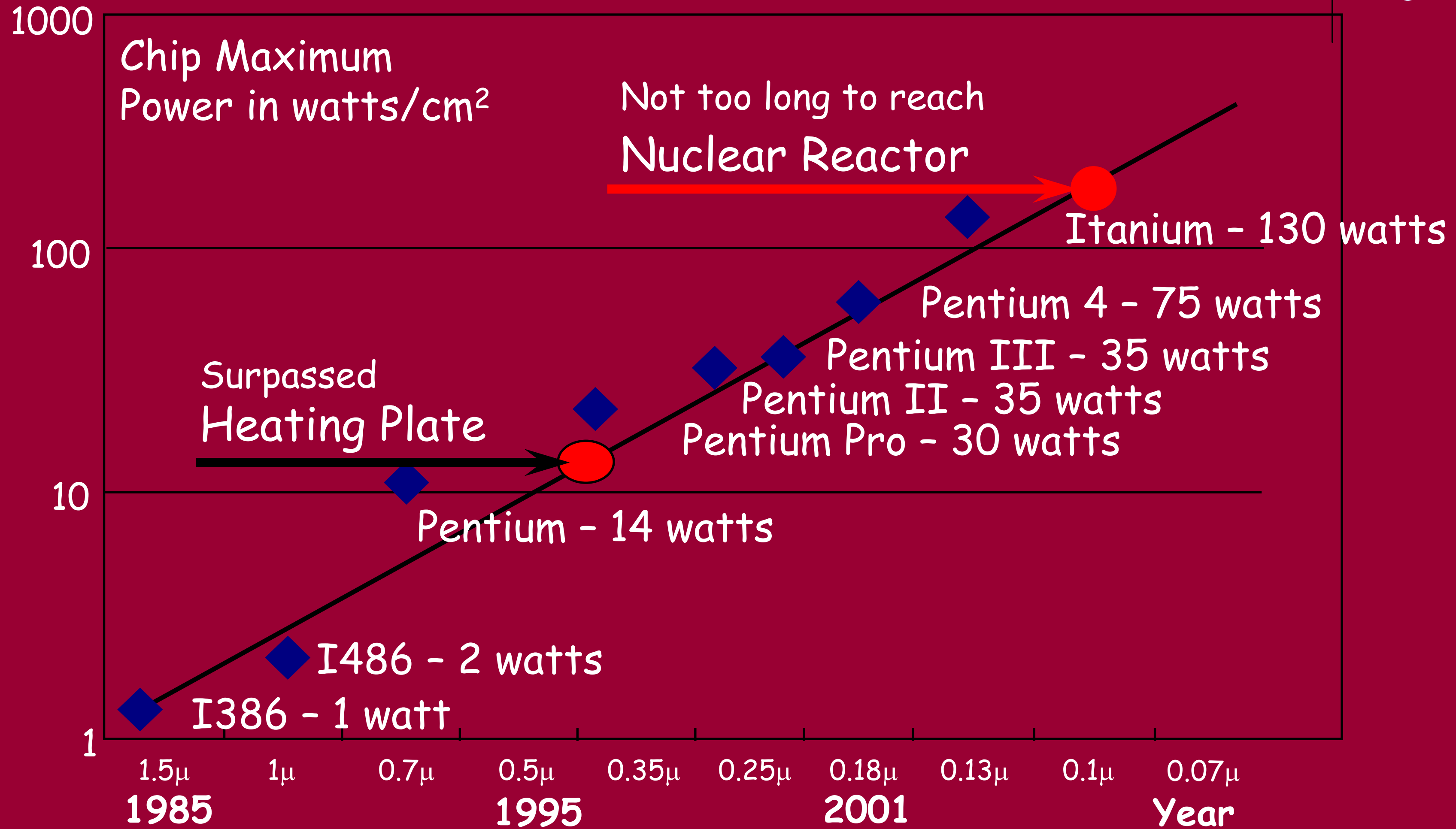
- Crude Analogy
  - Formula One Race Car: Wins raw performance but reliability is so poor that it requires frequent maintenance. Throughput low.
  - Toyota Camry V6: Loses raw performance but high reliability results in high throughput (i.e., miles driven/month → answers/month).

Improving Reliability & Availability
(Reducing Costs Associated with HPC)

- Observation
  - High speed α high power density α high temperature α low reliability
  - Arrhenius' Equation*
    (circa 1890s in chemistry → circa 1980s in computer & defense industries)
    - As temperature increases by 10° C …
      - ✓ The failure rate of a system doubles.
    - Twenty years of unpublished empirical data .

* The time to failure is a function of $e^{-Ea/kT}$ where Ea = activation energy of the failure mechanism being accelerated, k = Boltzmann's constant, and T = absolute temperature

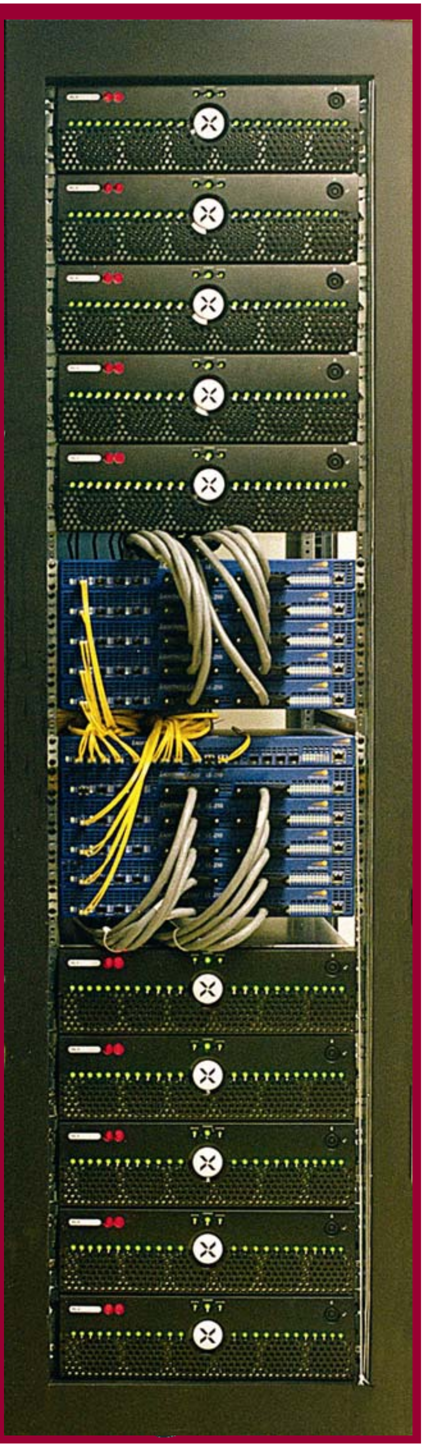
Moore's Law for Power ($P \propto V^2f$)

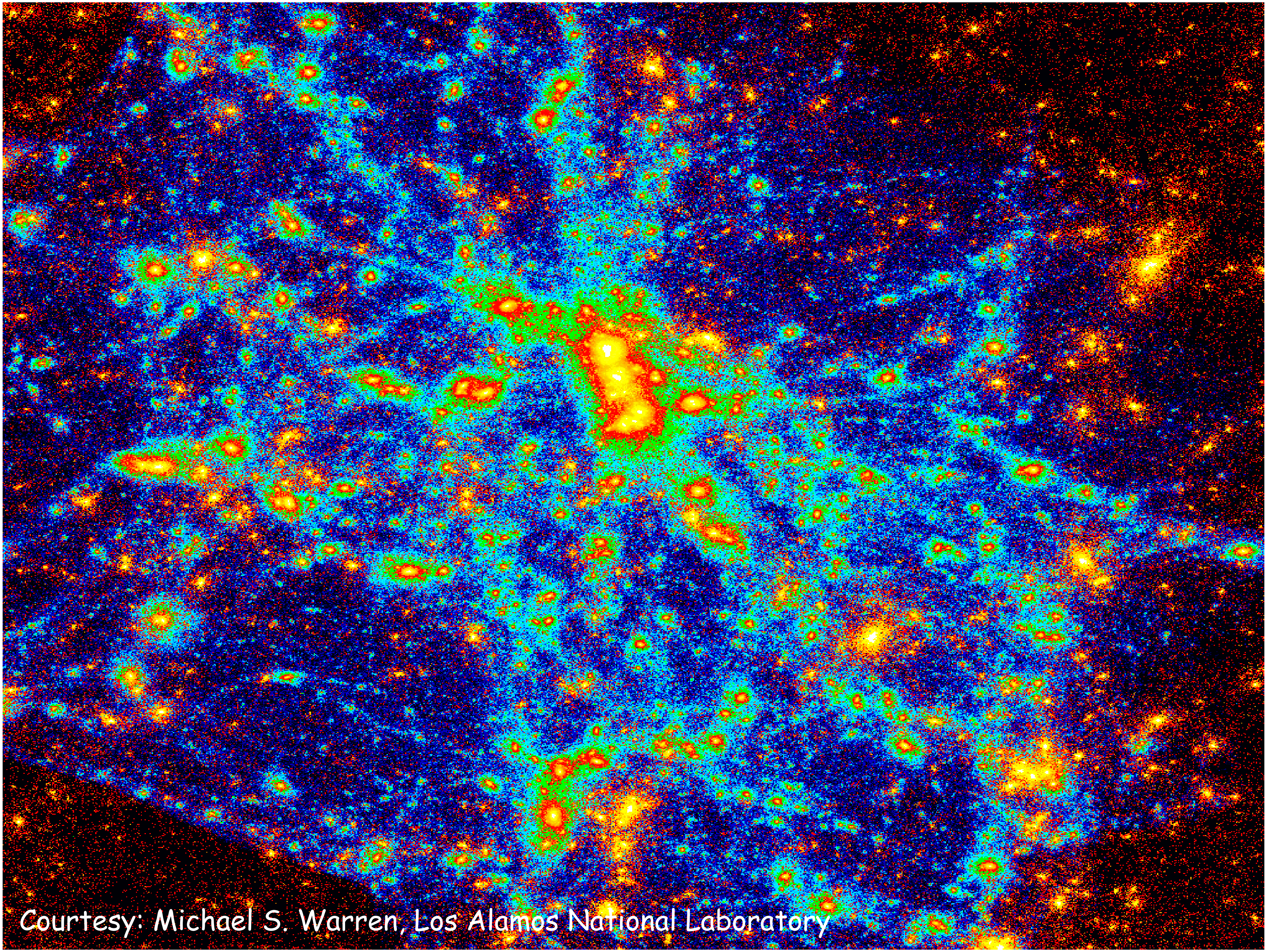


Source: Fred Pollack, Intel. New Microprocessor Challenges in the Coming Generations of CMOS Technologies, MICRO32 and Transmeta

"Green Destiny" Bladed Beowulf
(circa February 2002)

- A 240-Node Beowulf in Five Square Feet
- Each Node
  - 1-GHz Transmeta TM5800 CPU w/ High-Performance Code-Morphing Software running Linux 2.4.x
  - 640-MB RAM, 20-GB hard disk, 100-Mb/s Ethernet (up to 3 interfaces)
- Total
  - 240 Gflops peak (Linpack: 101 Gflops in March 2002.)
  - 150 GB of RAM (expandable to 276 GB)
  - 4.8 TB of storage (expandable to 38.4 TB)
  - *Power Consumption: Only 3.2 kW.*
- Reliability & Availability
  - *No unscheduled downtime in 24-month lifetime.*
    - Environment: A dusty 85°-90° F warehouse!




Courtesy: Michael S. Warren, Los Alamos National Laboratory

Parallel Computing Platforms
(An "Apples-to-Oranges" Comparison)

- Avalon (1996)
  - 140-CPU Traditional Beowulf Cluster
- ASCI Red (1996)
  - 9632-CPU MPP
- ASCI White (2000)
  - 512-Node (8192-CPU) Cluster of SMPs
- Green Destiny (2002)
  - 240-CPU Bladed Beowulf Cluster

- Code: N-body gravitational code from Michael S. Warren, Los Alamos National Laboratory

Parallel Computing Platforms Running the N-body Gravitational Code

Machine	Avalon Beowulf	ASCI Red	ASCI White	Green Destiny
Year	1996	1996	2000	2002
Performance (Gflops)	18	600	2500	*58*
Area (ft^2)	120	1600	9920	5
Power (kW)	18	1200	2000	5
DRAM (GB)	36	585	6200	150
Disk (TB)	0.4	2.0	160.0	4.8
DRAM density (MB/ft^2)	300	366	625	30000
Disk density (GB/ft^2)	3.3	1.3	16.1	960.0
Perf/Space (Mflops/ft^2)	150	375	252	11600
Perf/Power (Mflops/watt)	1.0	0.5	1.3	11.6

Parallel Computing Platforms Running the N-body Gravitational Code

Machine	Avalon Beowulf	ASCI Red	ASCI White	Green Destiny
Year	1996	1996	2000	2002
Performance (Gflops)	18	600	2500	*58*
Area (ft^2)	120	1600	9920	5
Power (kW)	18	1200	2000	5
DRAM (GB)	36	585	6200	150
Disk (TB)	0.4	2.0	160.0	4.8
DRAM density (MB/ft^2)	300	366	625	3000
Disk density (GB/ft^2)	3.3	1.3	16.1	960.0
Perf/Space ($Mflops/ft^2$)	150	375	252	11600
Perf/Power (Mflops/watt)	1.0	0.5	1.3	11.6

Yet in 2002 …

www.nytimes.com
The New York Times ON THE WEB

June 25, 2002

At Los Alamos, Two Visions of Supercomputing

- "Green Destiny is *so* low power that it runs just as fast when it is unplugged."
- "The slew of expletives and exclamations that followed Feng's description of the system …"
- "In HPC, no one cares about power & cooling, and no one ever will …"
- "Moore's Law for Power will stimulate the economy by creating a new market in cooling technologies."

InfoWorld HOME NEWS TEST CENTER

Green Destiny draws cheers and jeers

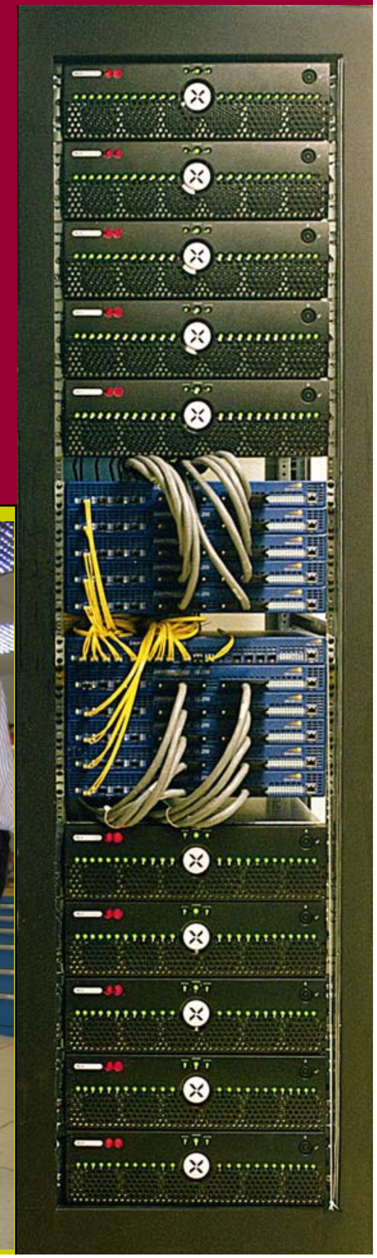
For many of the Los Alamos scientists, the unveiling of Green Destiny was their first introduction to blade servers -- never mind blade servers being used to build a supercomputer. The slew of expletives and exclamations that followed Feng's description of the system made it clear that the blades had captured the audience's attention. Some murmured, "Wow," while others let out multiple shouts of, "Jesus!" as their jaws dropped.

Several scientists here did not share the enthusiasm for Green Destiny, however. Los Alamos, after all, is the home to several massive supercomputers that take up entire floors of buildings and require several cooling systems shaped like mini-nuclear reactors to keep them running. These "real" supercomputers handle serious work, and some of the people running them consider Green Destiny a joke. One scientist walked out of Feng's presentation, making his feelings clear.

Today: Recent Trends in HPC

- Low(er)-Power Multi-Core Chipsets
  - AMD: Athlon64 X2 (2) and Opteron (2)
  - ARM: MPCore (4)
  - IBM: PowerPC 970 (2)
  - Intel: Woodcrest (2) and Cloverton (4)
  - PA Semi: PWRficient (2)
- Low-Power Supercomputing
  - Green Destiny (2002)
  - Orion Multisystems (2004)
  - BlueGene/L (2004)
  - MegaProto (2004)

October 2003
BG/L half rack prototype
500 Mhz
512 nodes/1024 proc.
2 TFlop/s peak
1.4 Tflop/s sustained



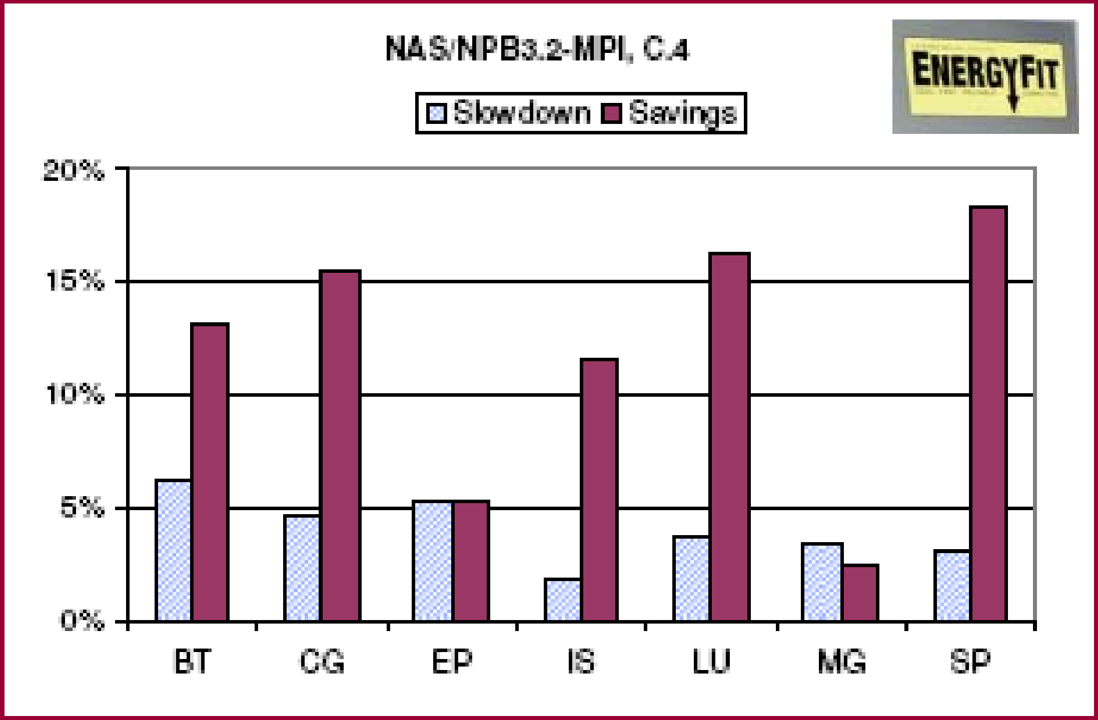
SPEC95 Results on an AMD XP-M

program	β	*2step*	*nqPID*	*freq*	*mips*	*beta*
swim	0.02	1.00/1.00	1.04/0.70	1.00/0.96	1.00/1.00	1.04/0.61
tomcatv	0.24	1.00/1.00	1.03/0.69	1.00/0.97	1.03/0.83	1.00/0.85
su2cor	0.27	0.99/0.99	1.05/0.70	1.00/0.95	1.01/0.96	1.03/0.85
compress	0.37	1.02/1.02	1.13/0.75	1.02/0.97	1.05/0.92	1.01/0.95
mgrid	0.51	1.00/1.00	1.18/0.77	1.01/0.97	1.00/1.00	1.03/0.89
vortex	0.65	1.01/1.00	1.25/0.81	1.01/0.97	1.07/0.94	1.05/0.90
turb3d	0.79	1.00/1.00	1.29/0.83	1.03/0.97	1.01/1.00	1.05/0.94
go	1.00	1.00/1.00	1.37/0.88	1.02/0.99	0.99/0.99	1.06/0.96

relative time / relative energy
with respect to total execution time and system energy usage

- Results on newest SPEC are even better …

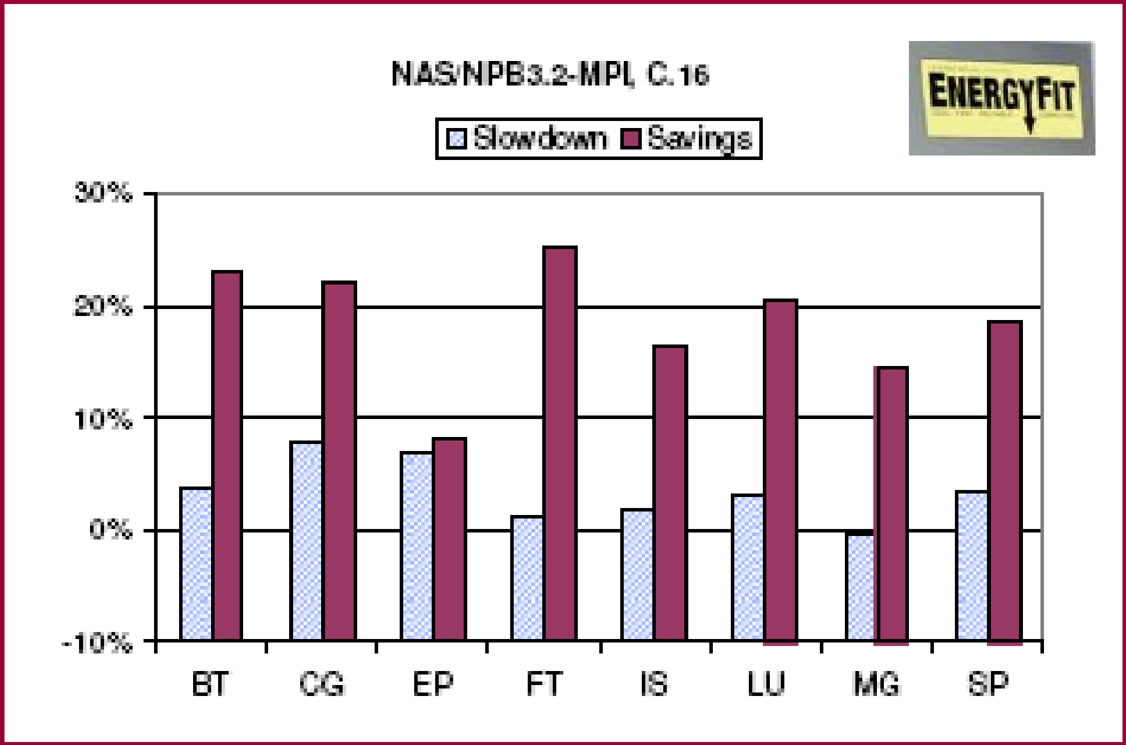
NAS Parallel on an Athlon-64 Cluster



AMD Athlon-64 Cluster

"A Power-Aware Run-Time System for High-Performance Computing," *SC|05,* Nov. 2005.

NAS Parallel on an Opteron Cluster



AMD Opteron Cluster

"A Power-Aware Run-Time System for High-Performance Computing," *SC|05,* Nov. 2005.

HPC Should Care About Electrical Power Usage

Perspective

- FLOPS Metric of the TOP500
  - Performance = Speed (as measured in FLOPS with Linpack)
  - May not be "fair" metric in light of recent low-power trends to help address efficiency, usability, reliability, availability, and total cost of ownership.
- The Need for a Complementary Performance Metric?
  - Performance = *f(* speed, "time to answer", power consumption, "up time", total cost of ownership, usability, …*)*
  - Easier said than done …
    - Many of the above dependent variables are difficult, if not impossible, to quantify, e.g., "time to answer", TCO, usability, etc.
- The Need for a Green500 List
  - Performance = *f(* speed, power consumption*)* as speed and power consumption can be quantified.

Challenges for a Green500 List

- What Metric To Choose?
  - *ED^n*: Energy-Delay Products, where *n* is a non-negative int. (borrowed from the circuit-design domain)
  - *Speed / Power Consumed*
    - FLOPS / Watt, MIPS / Watt, and so on
  - *SWaP: Space, Watts and Performance Metric* (Courtesy: Sun)
- What To Measure? Obviously, energy or power … but
  - Energy (Power) consumed by the computing system?
  - Energy (Power) consumed by the processor?
  - Temperature at specific points on the processor die?
- How To Measure Chosen Metric?
  - Power meter? But attached to what? At what time granularity should the measurement be made?

"Making a Case for a Green500 List" (Opening Talk)
IPDPS 2005, Workshop on High-Performance, Power-Aware Computing.

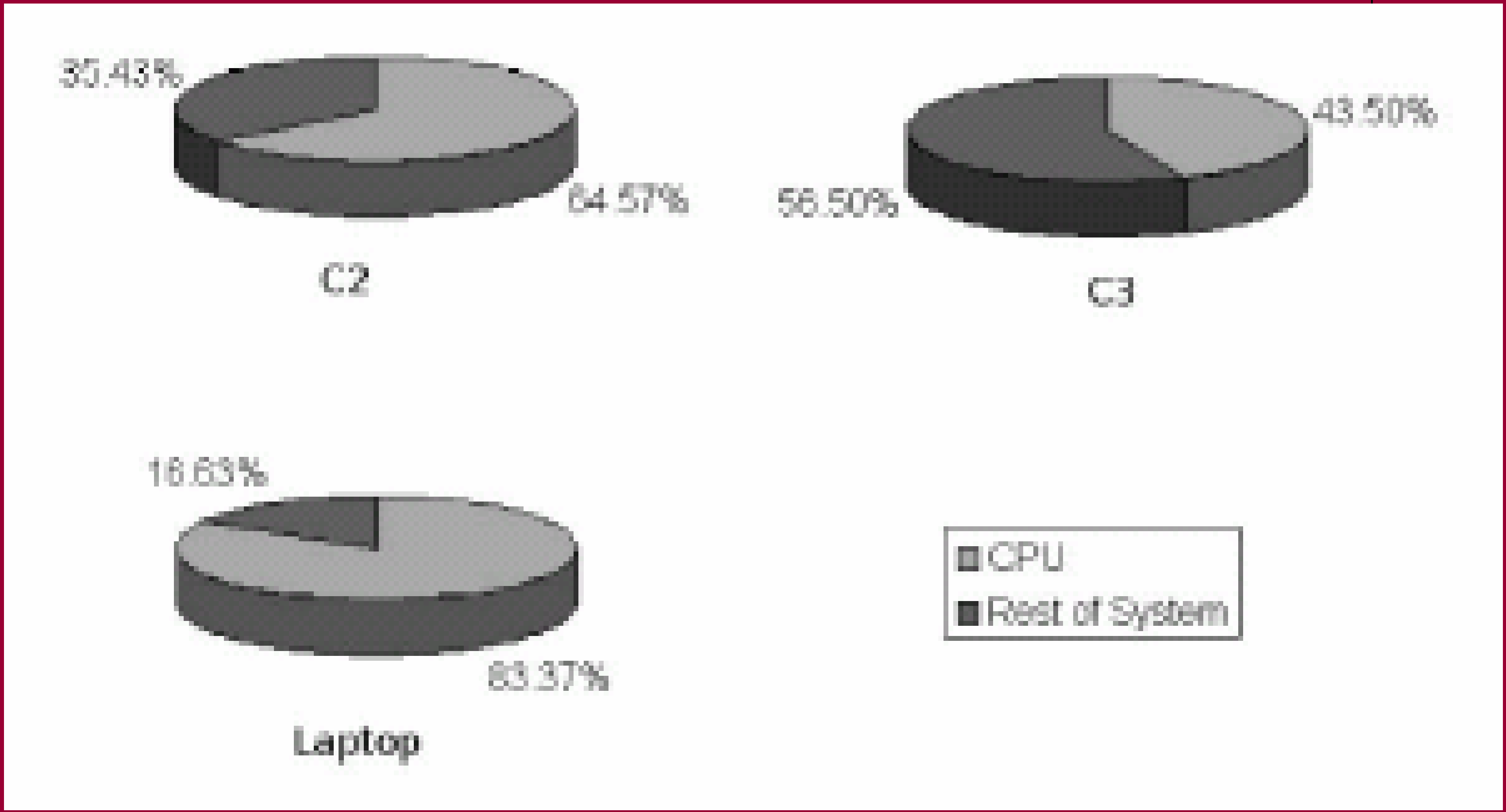
Challenges for a Green500 List

- What Metric To Choose?
  - *ED ^{n}*: Energy-Delay Products, where *n* is a non-negative int. (borrowed from the circuit-design domain)
  - *Speed / Power Consumed*
    - FLOPS / Watt, MIPS / Watt, and so on
  - *SWaP: Space, Watts and Performance Metric* (Courtesy: Sun)
- What To Measure? Obviously, energy or power … but
  - Energy (Power) consumed by the computing system?
  - Energy (Power) consumed by the processor?
  - Temperature at specific points on the processor die?
- How To Measure Chosen Metric?
  - Power meter? But attached to what? At what time granularity should the measurement be made?

Hmm…

"Making a Case for a Green500 List" (Opening Talk)
IPDPS 2005, Workshop on High-Performance, Power-Aware Computing.

Power: CPU or System?



Efficiency of Four-CPU Clusters

Name	CPU	LINPACK (Gflops)	Avg Pwr (Watts)	Time (s)	ED (*10^6)	ED^2 (*10^9)	Flops/W	$V_{\partial=-0.5}$
C1	3.6G P4	19.55	713.2	315.8	71.1	22.5	27.4	33.9
C2	2.0G Opt	12.37	415.9	499.4	103.7	51.8	29.7	47.2
C3	2.4G Ath64	14.31	668.5	431.6	124.5	53.7	21.4	66.9
C4	2.2G Ath64	13.40	608.5	460.9	129.3	59.6	22.0	68.5
C5	2.0G Ath64	12.35	560.5	499.8	140.0	70.0	22.0	74.1
C6	2.0G Opt	12.84	615.3	481.0	142.4	64.5	20.9	77.4
C7	1.8G Ath64	11.23	520.9	549.9	157.5	86.6	21.6	84.3

W. Feng, feng@cs.vt.edu, (540) 231-1192

Efficiency of Four-CPU Clusters

Name	CPU	LINPACK (Gflops)	Avg Pwr (Watts)	Time (s)	ED ($*10^6$)	ED2 ($*10^9$)	Flops/ W	$V_{\partial=-0.5}$
C1	3.6G P4	19.55	713.2	315.8	71.1	22.5	27.4	33.9
C2	2.0G Opt	12.37	415.9	499.4	103.7	51.8	29.7	47.2
C3	2.4G Ath64	14.31	668.5	431.6	124.5	53.7	21.4	66.9
C4	2.2G Ath64	13.40	608.5	460.9	129.3	59.6	22.0	68.5
C5	2.0G Ath64	12.35	560.5	499.8	140.0	70.0	22.0	74.1
C6	2.0G Opt	12.84	615.3	481.0	142.4	64.5	20.9	77.4
C7	1.8G Ath64	11.23	520.9	549.9	157.5	86.6	21.6	84.3

1x4P

2x2P

TOP500 as Green500?

TOP500 Power Usage (Source: J. Dongarra)

Name	Peak Perf	Peak Power	MFLOPS/W	TOP500 Rank
BlueGene/L	367,000	2,500	146.80	1
ASC Purple	92,781	7,600	12.20	3
Columbia	60,960	3,400	17.93	4
Earth Simulator	40,960	11,900	3.44	10
MareNostrum	42,144	1,071	39.35	11
Jaguar-Cray XT3	24,960	1,331	18.75	13
ASC Q	20,480	10,200	2.01	25
ASC White	12,288	2,040	6.02	60

TOP500 as Green500

Relative Rank	TOP500	Green500
1	BlueGene/L (IBM)	BlueGene/L (IBM)
2	ASC Purple (IBM)	MareNostrum (IBM)
3	Columbia (SGI)	Jaguar-Cray XT3 (Cray)
4	Earth Simulator (NEC)	Columbia (SGI)
5	MareNostrum (IBM)	ASC Purple (IBM)
6	Jaguar-Cray XT3 (Cray)	ASC White (IBM)
7	ASC Q (HP)	Earth Simulator (NEC)
8	ASC White (IBM)	ASC Q (HP)

TOP500 as Green500

Relative Rank	TOP500	Green500
1	BlueGene/L (IBM)	BlueGene/L (IBM)
2	ASC Purple (IBM)	MareNostrum (IBM)
3	Columbia (SGI)	Jaguar-Cray XT3 (Cray)
4	Earth Simulator (NEC)	Columbia (SGI)
5	MareNostrum (IBM)	ASC Purple (IBM)
6	Jaguar-Cray XT3 (Cray)	ASC White (IBM)
7	ASC Q (HP)	Earth Simulator (NEC)
8	ASC White (IBM)	ASC Q (HP)

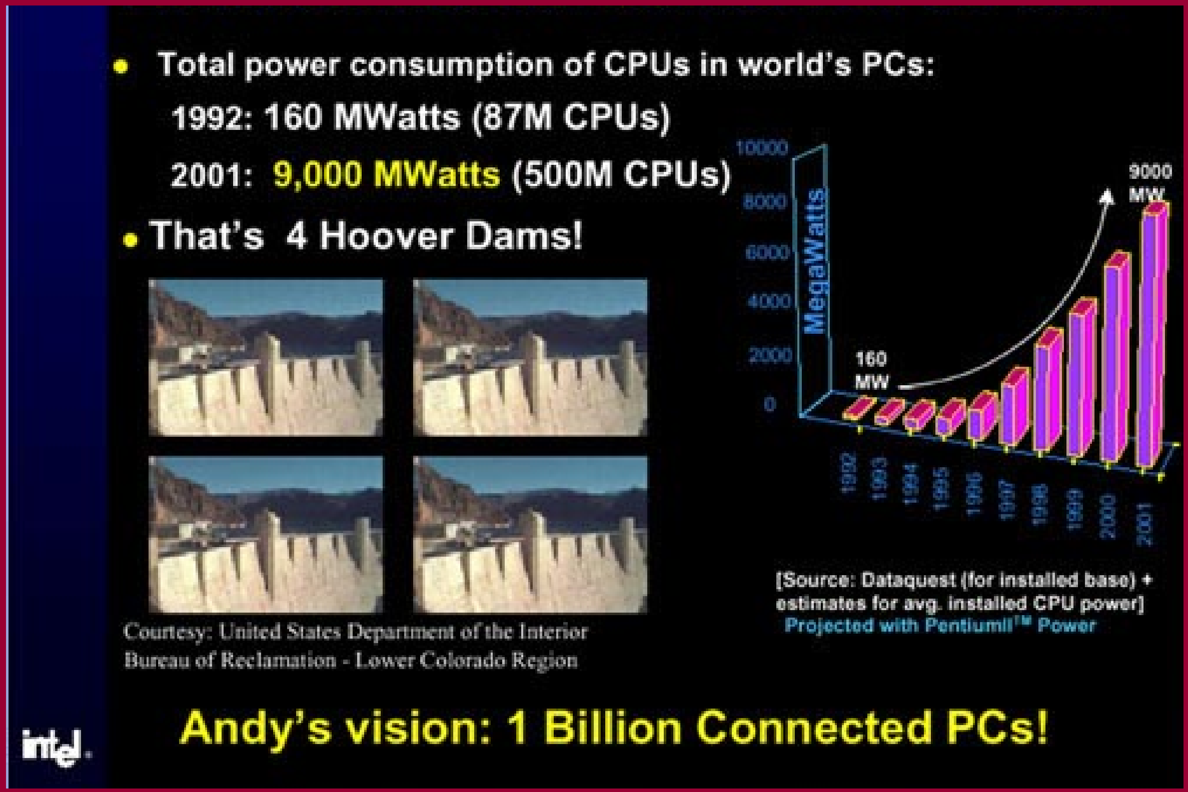
"A Call to Arms"

- Constructing a Green500 List
  - Required Information
    - Performance, as defined by Speed — Hard
    - Power — Hard
    - Space (optional) — Easy
- What Exactly to Do?
- How to Do It?
- Solution: Related to the purpose of CCGSC … :-)
  - Doing the above "TOP500 as Green500" exercise leads me to the following solution.

Talk to Jack …

- We already have LINPACK and the TOP500
  Plus
- Space (in square ft. or in cubic ft.)
- Power
  - Extrapolation of reported CPU power?
  - Peak numbers for each compute node?
  - Direct measurement? Easier said than done?
    - Force folks to buy industrial-strength multimeters or oscilloscopes. Potential barrier to entry.
  - Power bill?
    - Bureaucratic annoyance. Truly representative?

Let's Make Better Use of Resources



Source: Cool Chips & Micro 32

… and Reduce Global Climate Warming in the Machine Room …

For More Information

- Visit "Supercomputing in Small Spaces" at http://sss.lanl.gov
  - Soon to be re-located to Virginia Tech
- Affiliated Web Sites
  - http://www.lanl.gov/radiant enroute to http://synergy.cs.vt.edu
  - http://www.mpiblast.org
- Contact me (a.k.a. "Wu")
  - E-mail: feng@cs.vt.edu
  - Phone: (540) 231-1192