Setting Trends for Energy-Efficient Supercomputing

Natalie Bates, EE HPC Working Group

Tahir Cader, HP & The Green Grid

Wu Feng, Virginia Tech & Green500

John Shalf, Berkeley Lab & NERSC

Horst Simon, Berkeley Lab & TOP500

Erich Strohmaier, Berkeley Lab & TOP500

ISC BoF; May 31, 2010; Hamburg, Germany

Why We Are Here

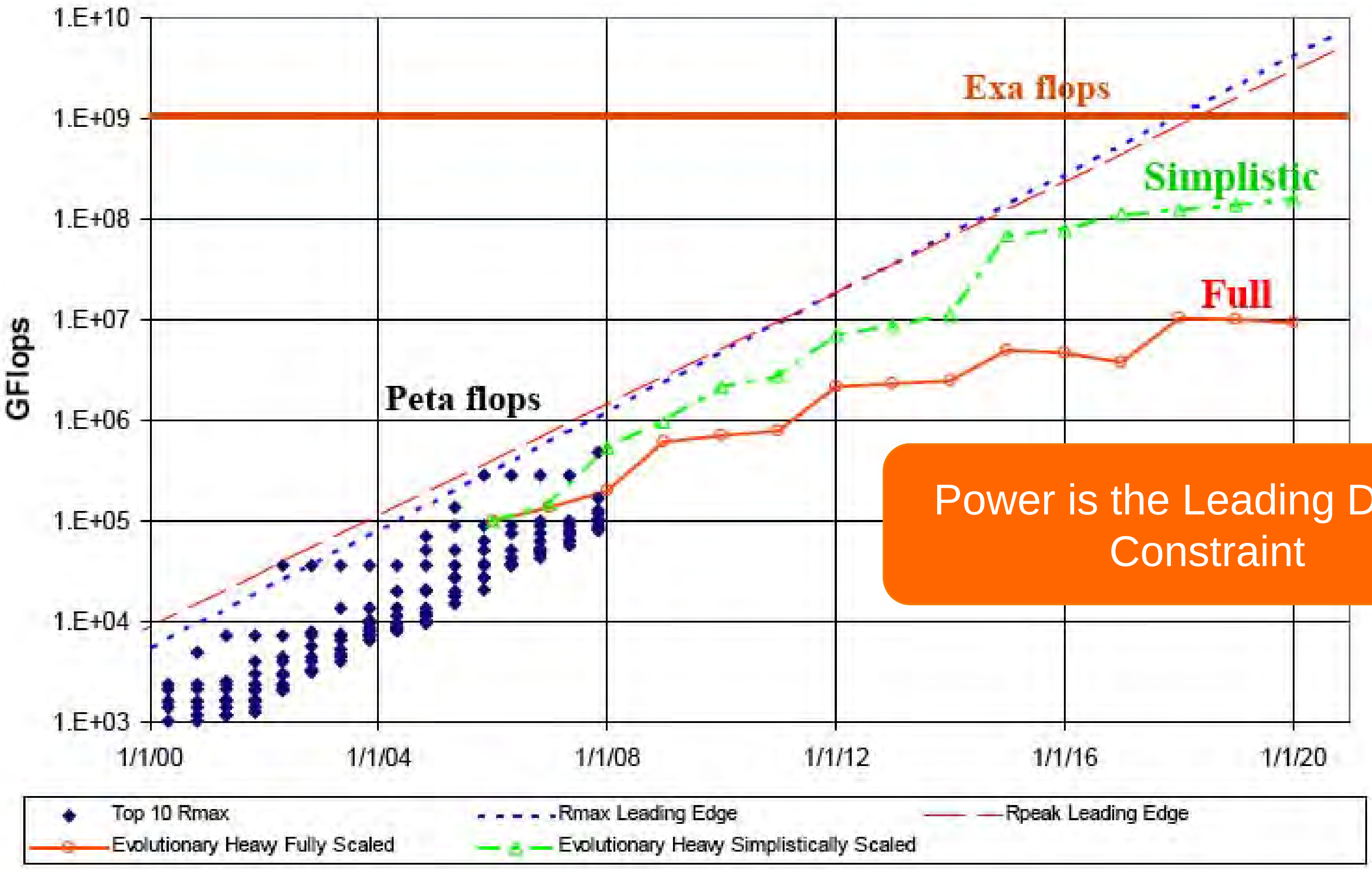
- **“Can only improve what you can measure”**

- **Context**
  - Power consumption of HPC and facilities cost are increasing

- **What is needed?**
  - Converge on a common basis between different research and industry groups for:
    - metrics
    - methodologies
    - workloads

    for energy-efficient supercomputing, so we can make progress towards solutions.

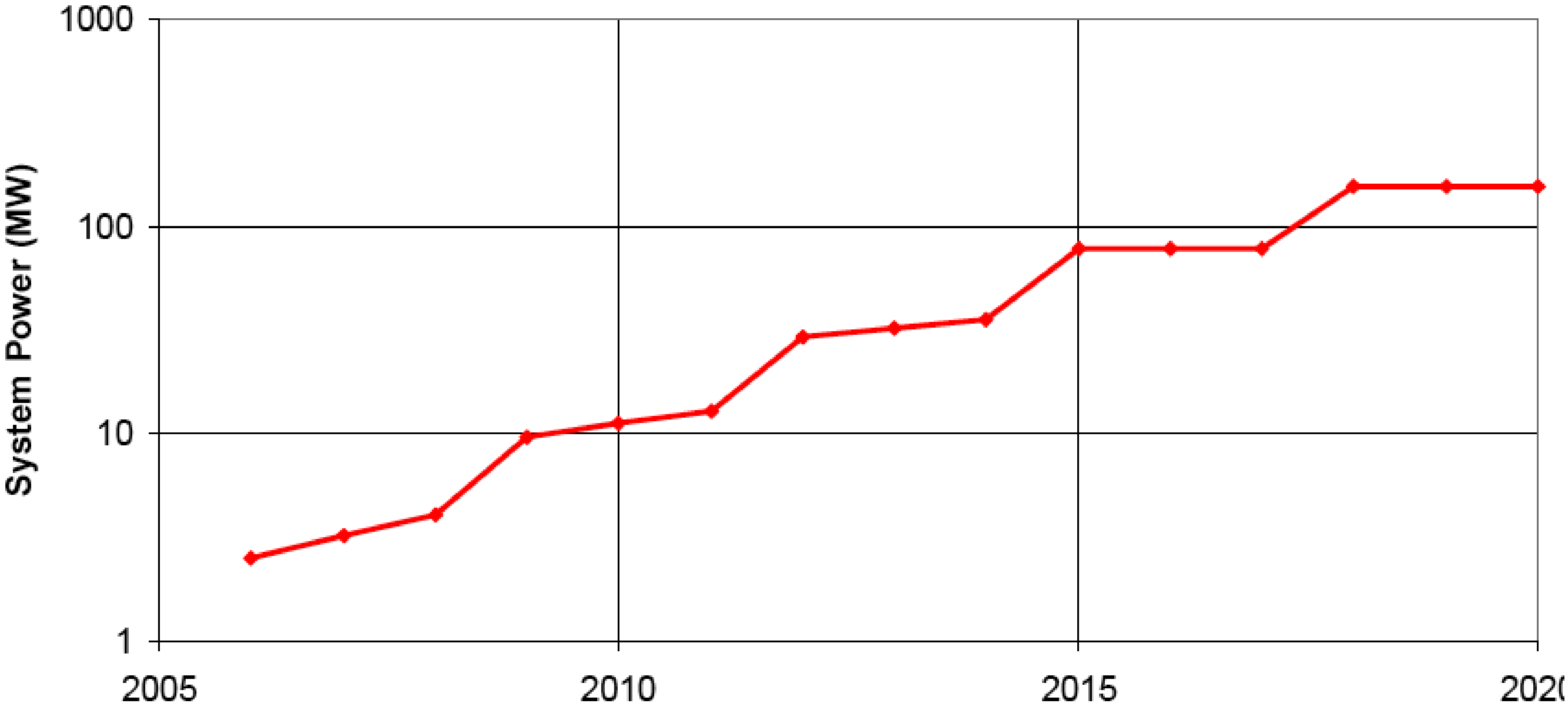
Current Technology Roadmaps will Depart from Historical Gains



Power is the Leading Design Constraint

From Peter Kogge, DARPA Exascale Study

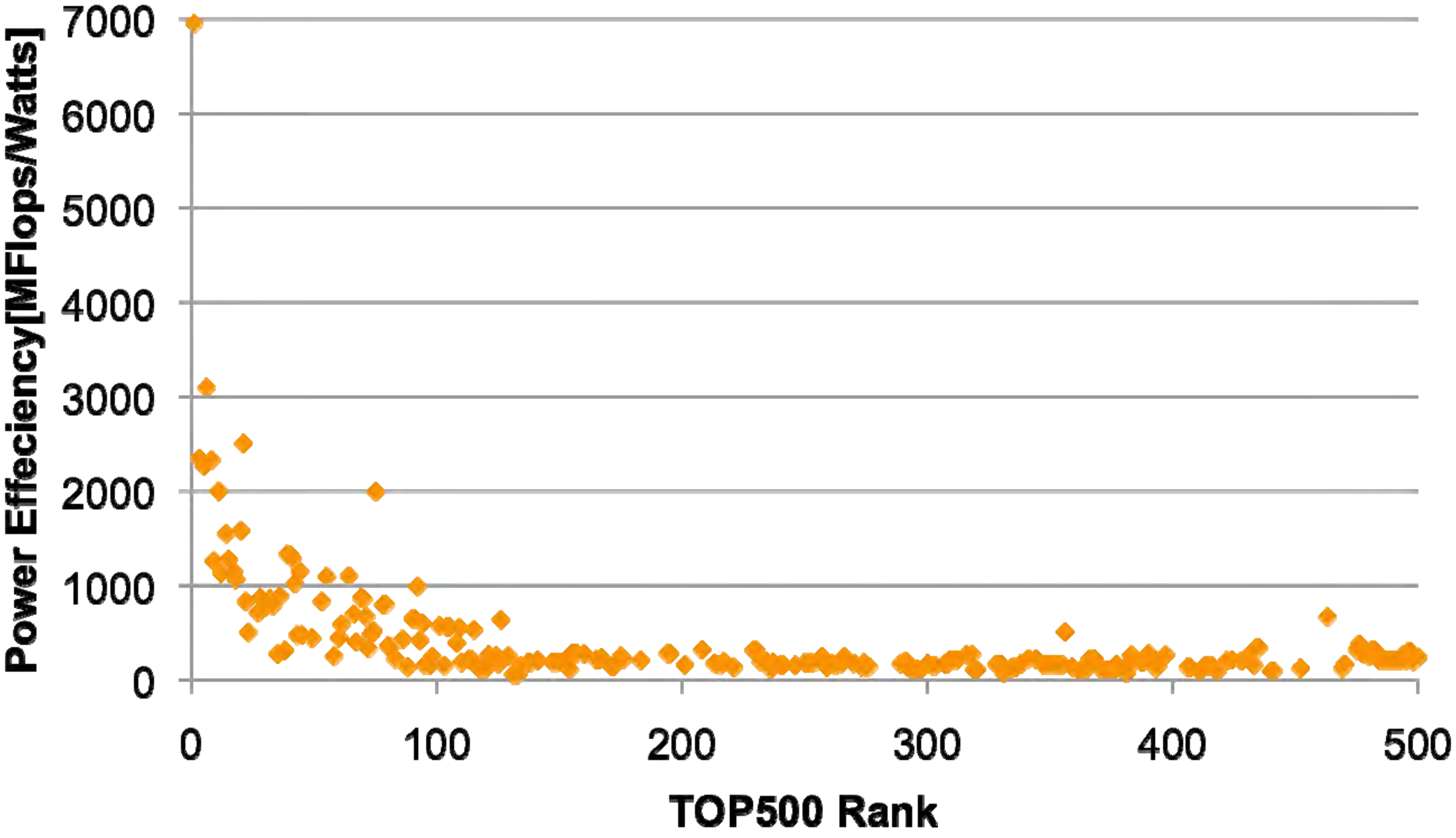
… and the power costs will still be staggering



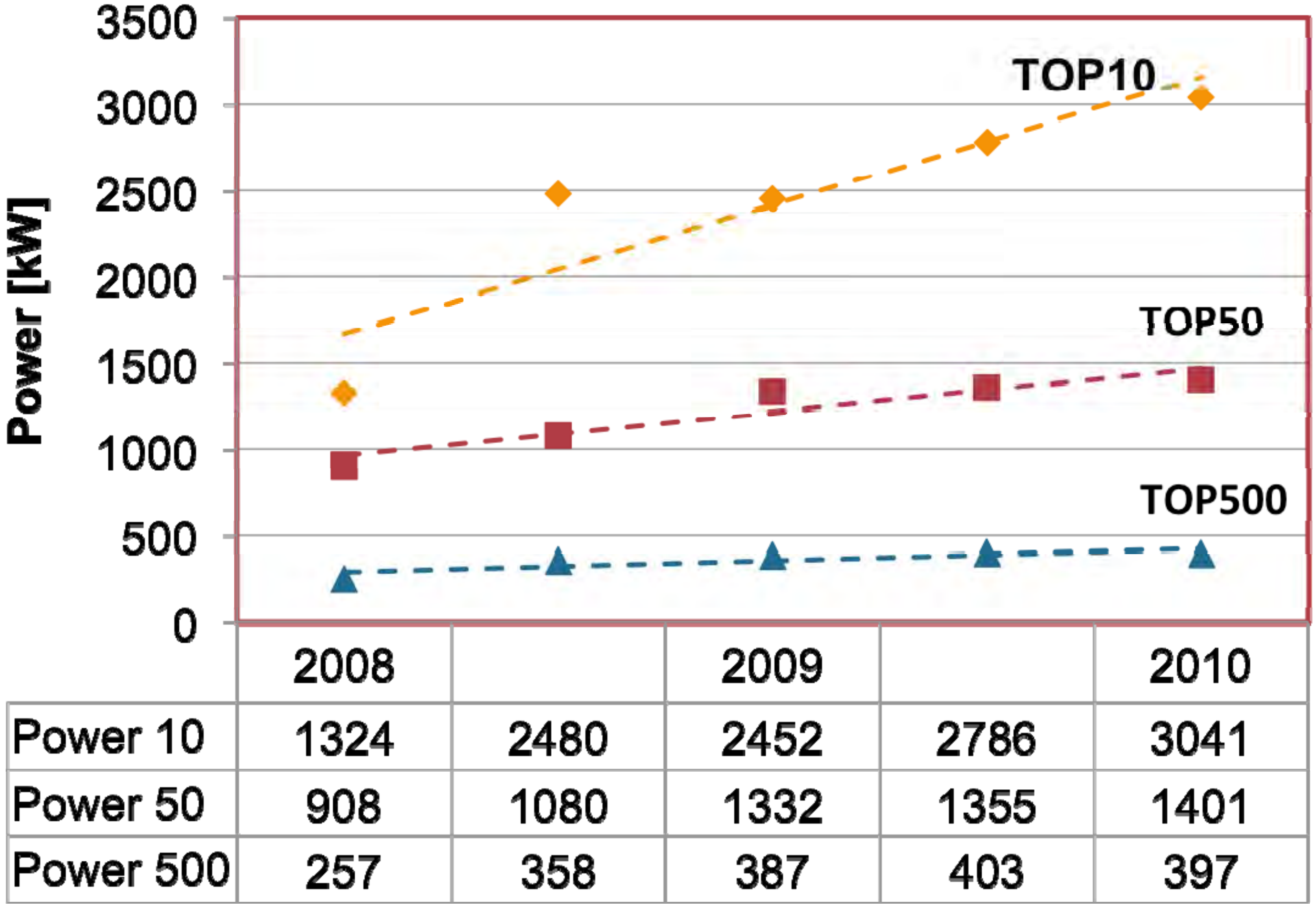
From Peter Kogge,
DARPA Exascale Study

$1M per megawatt per year! (with CHEAP power)

Absolute Power Levels

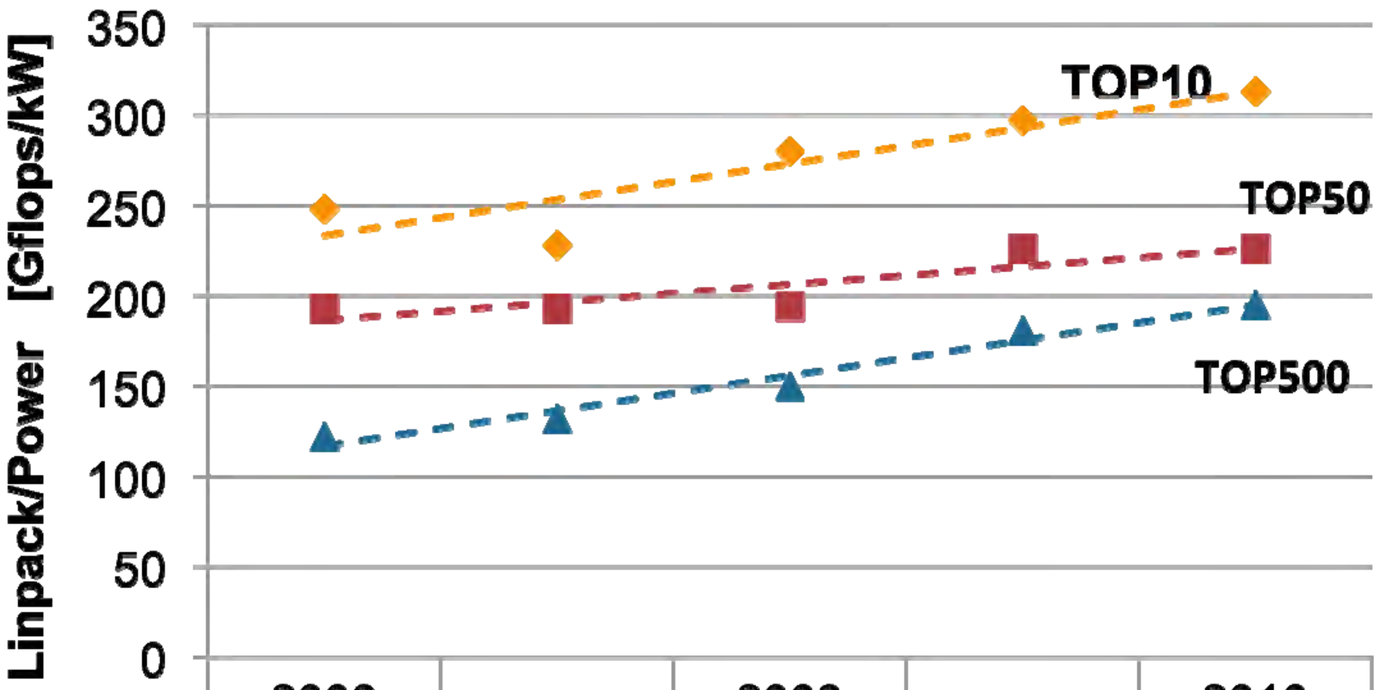


Power Consumption



	2008		2009		2010
Power 10	1324	2480	2452	2786	3041
Power 50	908	1080	1332	1355	1401
Power 500	257	358	387	403	397

Power Efficiency



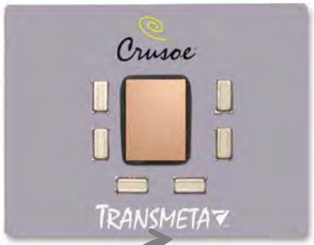
	2008		2009		2010
p-eff 10	248	228	280	297	313
p-eff 50	193	193	194	226	226
p-eff 500	122	132	150	181	195

What We Have Done

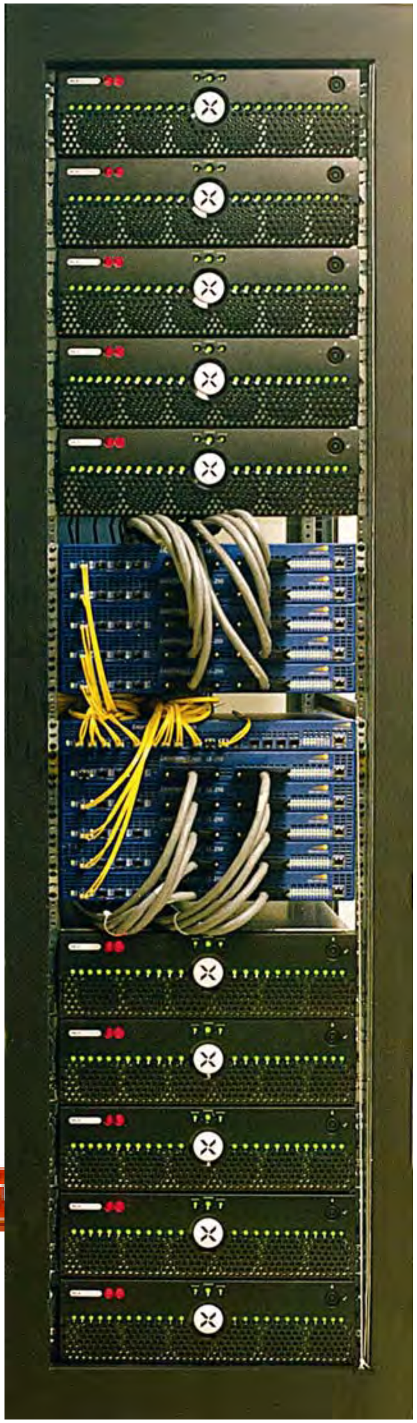
- **Stages of Green Supercomputing**
  - Denial
  - Awareness
  - Hype
  - Substance

The Denial Phase (2001 – 2004)

- Green Destiny
  - A 240-Node Supercomputer in 5 Sq. Ft.
  - LINPACK Performance: 101 Gflops
  - Power Consumption: 3.2 kW



embedded processor



- Prevailing Views
  - “Green Destiny is so low power that it runs just as fast when it is unplugged.”
  - “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.”

Green Destiny draws cheers and jeers

The New York Times

At Los Alamos, Two Visions of Supercomputing

The Awareness Phase (2004 – 2008)

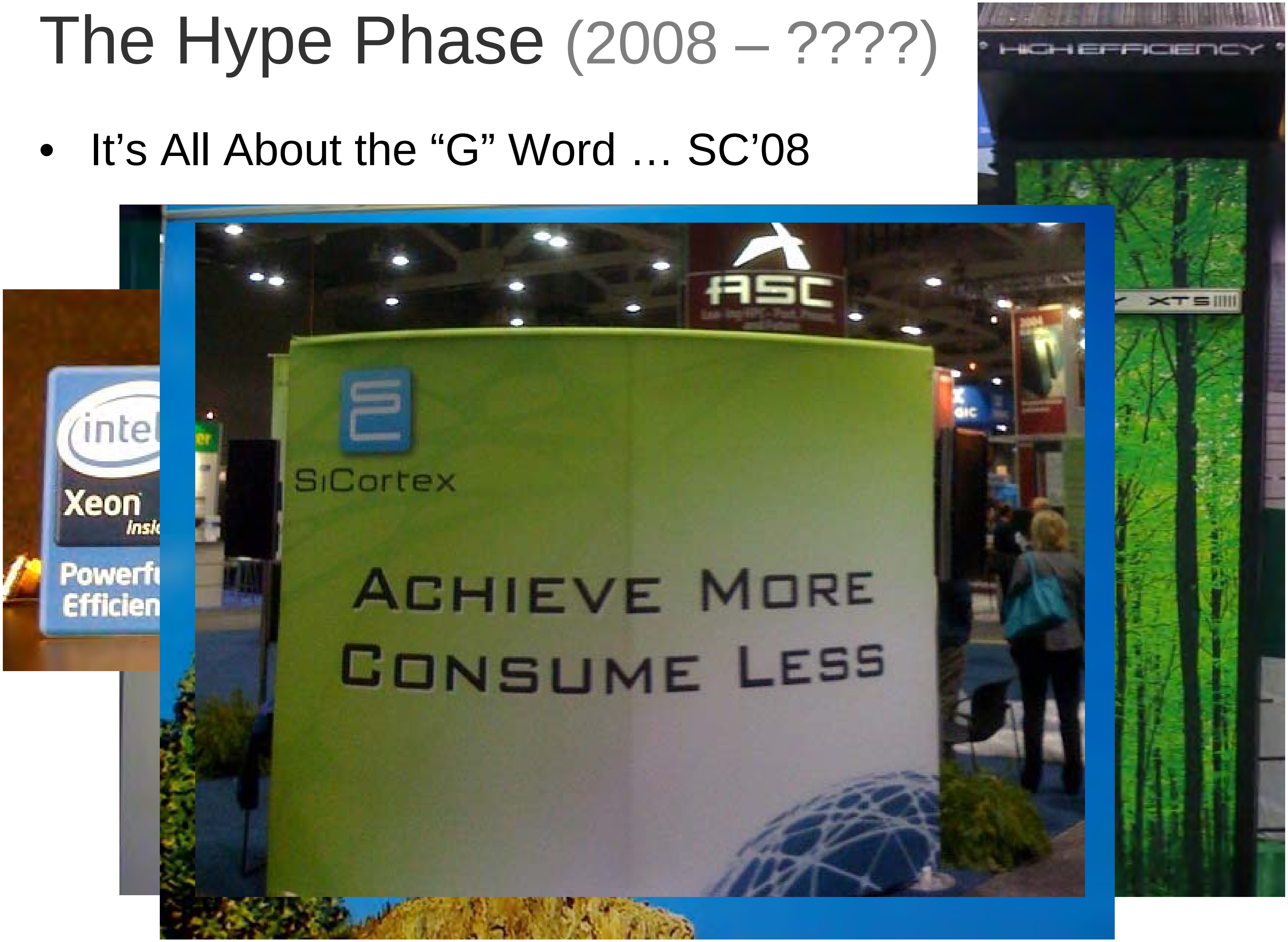
- Green Movements & Studies
  - IEEE Int'l Parallel & Distributed Processing Symp. (2005)
    - Workshop on High-Performance, Power-Aware Computing (HPPAC) → *Green500*
    - Metrics: Energy-Delay Product and FLOPS/Watt → FLOPS/watt
  - Green Grid (2007)
    - Industry-driven consortium of all the top system vendors
    - Metric: Power Usage Efficiency (PUE)
  - Kogge et al., "ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems, DARPA ITO, AFRL, 2008.
- Green IT Companies

Orion Multisystems (2004 – 2005)

SiCortex
(2003 – 2009)

The Hype Phase (2008 – ????)

- It’s All About the “G” Word … SC’08



The Substance Phase (2010 – ???)

- Current Lists & Consortia
  - Green500 (FLOPS/Watt)
    - Measured/Reported & Derived (not necessarily peak)
      - New Wrinkle: Measured Linpack *but* optimized relative to FLOPS/Watt
    - Exploratory Lists: *Little Green500 and HPCC Green500*
  - TOP500 Power (FLOPS/Watt)
    - Measured/Reported
      - Power measurement when running optimized Linpack
  - Green Grid (PUE and Productivity Proxy)
    - Power Usage Effectiveness (Note: Workload Independent)
    - Proxy Proposals for Measuring Data Center Productivity

“Can’t improve what you can’t measure.”

Why We Are Here

- **“Can only improve what you can measure”**

- **Context**
  - Power consumption of HPC and facilities cost are increasing

- **What is needed?**
  - Converge on a common basis between different research and industry groups for:
    - metrics
    - methodologies
    - workloads

    for energy-efficient supercomputing, so we can make progress towards solutions.

Metrics:
Can’t improve what you don’t measure

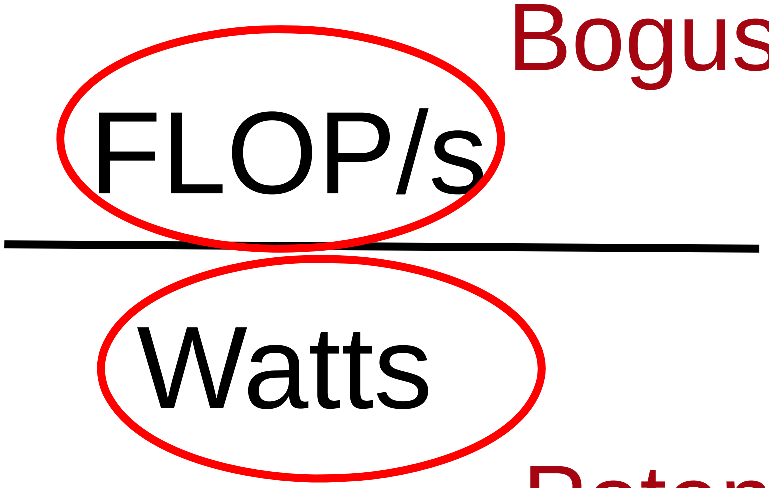
- **Collecting Metrics for HPC Power Usage (Green500, Top500, SpecHPC)**
  – Raise Community Awareness of HPC System Power Efficiency
  – Push vendors toward more power efficient solutions (shine a light on inefficiency)

- **Choice of measurement has a dramatic effect on the outcome *(Law of unintended consequences)***
  – Suddenly everything is “green”
  – But is anything ***really*** getting better? (everything looks better on an exponential curve)

Anatomy of a “Value” Metric

$$\frac{\text{Good Stuff}}{\text{Bad Stuff}}$$

Anatomy of a “Value” Metric



$$\frac{\text{FLOP/s}}{\text{Watts}}$$

Bogus!!!

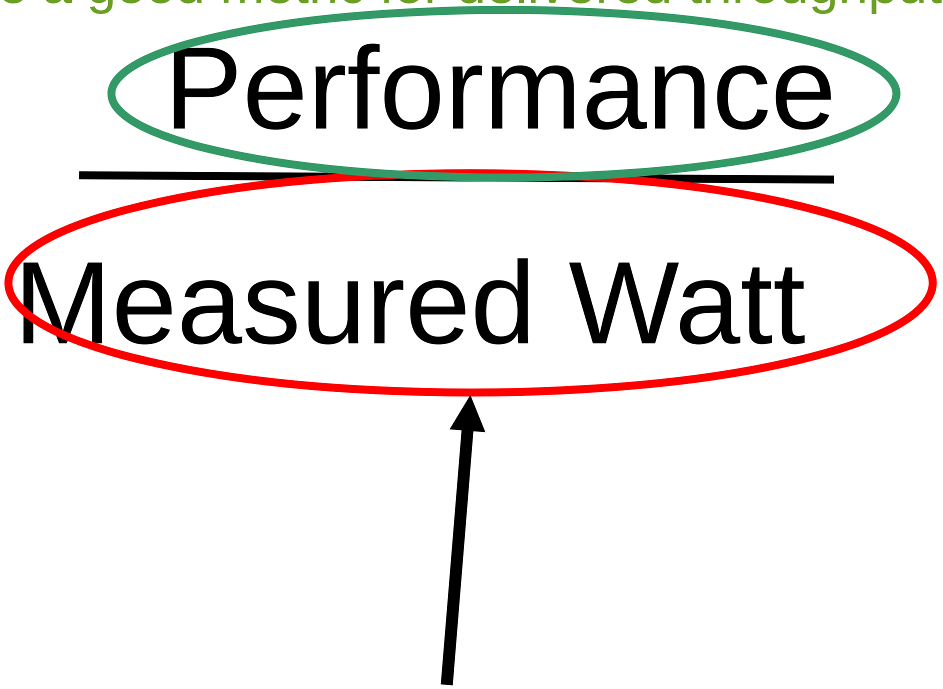
Potentially
Bogus!!

Anatomy of a “Value” Metric

Choose your own metric for performance!

(doesn’t need to be HPL, or FLOPS)

(choose a good metric for delivered throughput on workload)



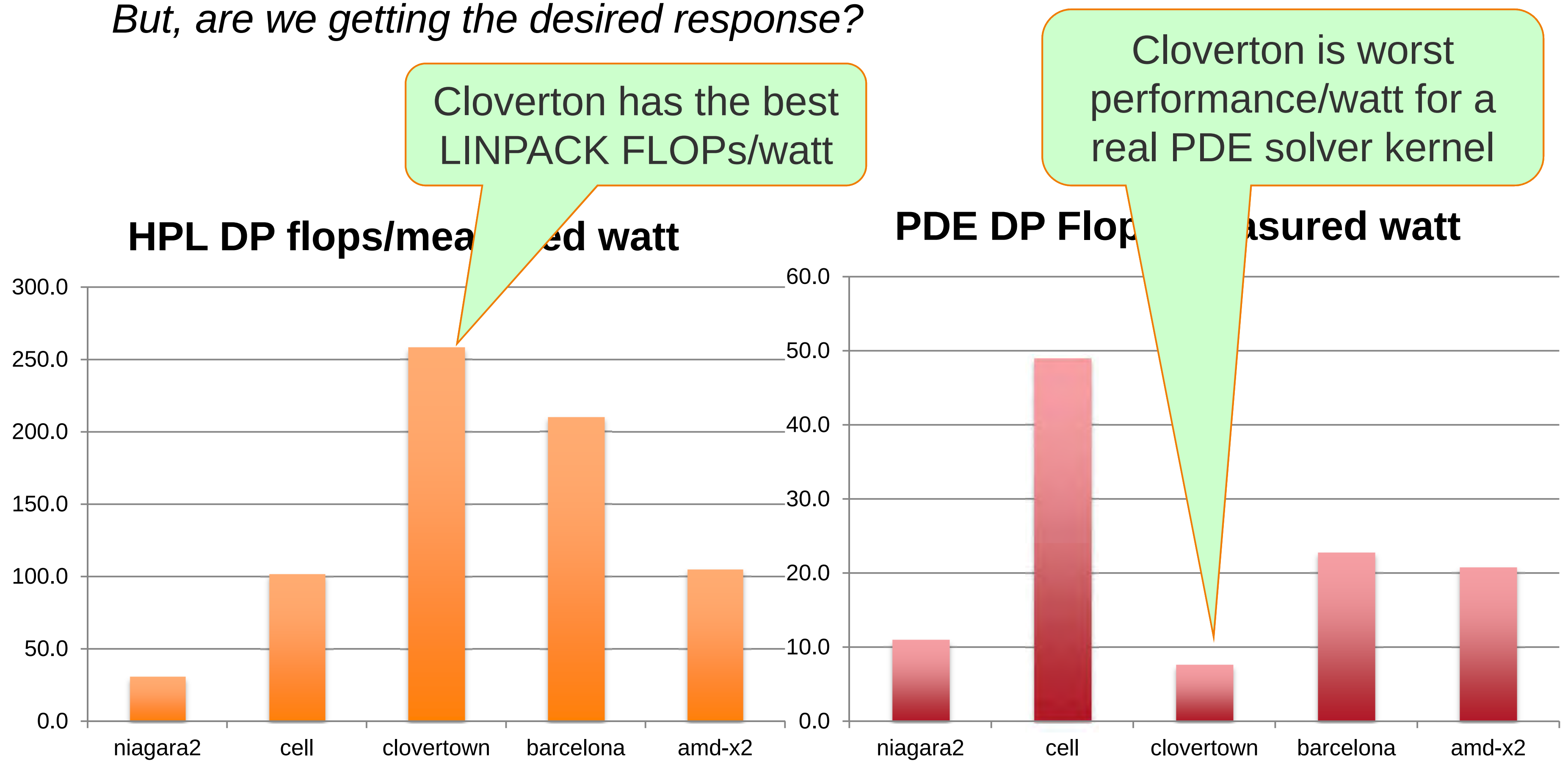
Performance

Measured Watt

Formal process for collecting this data emerging
(Green500, Top500, and eventually SpecPowerHPC)

Are We Really Improving?

Performance/measured_watt
is much more useful than
FLOPs/peak_watt
But, are we getting the desired response?



Cloverton has the best LINPACK FLOPs/watt

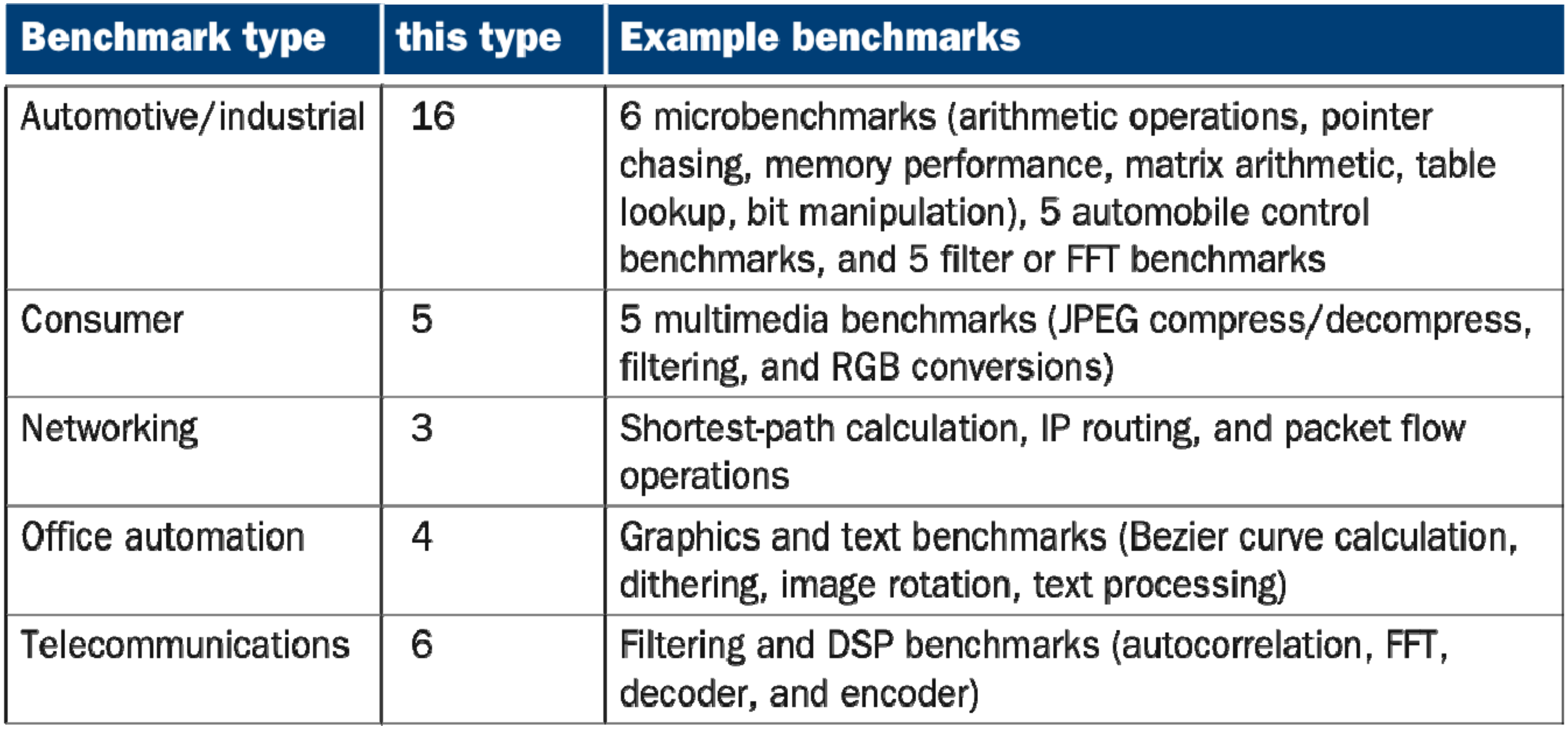
Cloverton is worst performance/watt for a real PDE solver kernel

Workload-based Metric

- **Workload-based benchmarks are a model of intended workload (throughput on “real problem)**
  - “Benchmarks are only useful insofar as they reflect the intended workload.” Ingrid Bucher, LANL, 1985
  - FLOPs and HPL are not a workload – need application

- **Examples**
  - EMBCC Embedded benchmarks (34 kernels grouped into 6 distinct workload Categories)
  - SPEC: 12 kernels grouped into two workload categories
  - NERSC SSP: 6 applications extracted from DOE workload
  - DoD-MOD TI-0x benchmarks: 8 applications extracted from DoD engineering workload

EMBCC Benchmark Suite Example



Benchmark type	this type	Example benchmarks
Automotive/industrial	16	6 microbenchmarks (arithmetic operations, pointer chasing, memory performance, matrix arithmetic, table lookup, bit manipulation), 5 automobile control benchmarks, and 5 filter or FFT benchmarks
Consumer	5	5 multimedia benchmarks (JPEG compress/decompress, filtering, and RGB conversions)
Networking	3	Shortest-path calculation, IP routing, and packet flow operations
Office automation	4	Graphics and text benchmarks (Bezier curve calculation, dithering, image rotation, text processing)
Telecommunications	6	Filtering and DSP benchmarks (autocorrelation, FFT, decoder, and encoder)

Workload-based Metrics

- **Proposal: Use Workload-based Metrics to Represent HPC Energy Efficiency**
  - Use workload-based metrics for numerator of value and “measured power” for denominator
  - Define distinct workload categories for HPC
- **Examples**
  - SPEC-FP/measured-watt: for embarrassingly parallel workloads (e.g. seismic processing)
  - NERSC SSP/measured-watt: For scalable MPI workloads
  - Green Grid “Productivity Proxies”
- ***Issues to resolve in this BoF discussion***
  - Identifying workload categories
  - What to include in measurement (facility/cooling, or just equipment)
  - How to measure (what methodology)

Global Harmonization of Metrics

- Vast proliferation of energy efficiency metrics in the commercial space has confused and paralyzed data center owners

- The Green Grid's Power Usage Effectiveness (PUE) metric has achieved significant worldwide adoption
  - PUE = Total Facility Power / IT Power

- In February of 2010, The US DOE hosted a meeting targeted at the Global Harmonization of Metrics
  - Attendees included The Green Grid, US DOE, US EPA, EU Code of Conduct, Japan's METI, and Japan's GIPC
  - At this meeting, the group agreed to "harmonize" on the PUE

- As a next step, the group agreed on the need for immediate work on a data center productivity metric

Data Center Productivity Proxies

#1: Useful Work Self-Assessment and Reporting

- Aggregate units of useful work, during an assessment window, divided by cumulative energy consumption.

#2: DCeP Subset by Productivity Link

- Aggregate units of useful work reported by a subset of the IT infrastructure, during an assessment window, divided by the total energy consumed (scalable to full system). Work is reported by Intel’s Productivity Link.

#3: DCeP Subset by Sample Load

- Aggregate units of useful work reported by a subset of the IT infrastructure, during an assessment window, divided by the total energy consumed (scalable to full system). Work is reported by custom code.

#4: Bits per Kilowatt-hour

- Ratio of the total bit volume from every outbound router on the data center network divided by the total energy consumed.

Data Center Productivity Proxies

#5: Weighted CPU Utilization – SPECint_rate

- Aggregate useful work derived from average CPU utilization and frequency, using SPECint_rate. Useful work is divided by total energy usage.

#6: Weighted CPU Utilization – SPECpower

- Uses published SPECPower results in conjunction with measurements of CPU utilization to estimate efficiency for a number of servers. Results scaled to the data center are divided by total energy usage.

#7: Compute Units per Second (CUPS)

- Utilizes the year servers were purchased, the estimated CUPS, and average CPU utilization over an assessment window to determine work. Divide the result by the total energy consumed.

#8: Operating System Workload Efficiency

- Ratio of total number of operating system instances running in the facility during the assessment window divided by the total facilities power.

Data Center Productivity Metric

Productivity Proxies

Proxy	Description
#1: Useful Work Self-Assessment and Reporting	Measure the aggregate amount of useful work that a data center produces during an assessment window and divide by the total amount of energy consumed during this time
#2: DCeP Subset by Productivity Link	Aggregate units of useful work reported by a subset of the IT infrastructure in a data center during an assessment window, scaling this number so that it represents the entire data center, and then divide the result by the total energy consumed. The units of work are reported by an API that runs in conjunction with each application running on the subset.
#3: DCeP Subset by Sample Load	Obtain "useful work number" from an instrumented subset of servers running a sample workload during an assessment window, scale to represent the entire data center and divide by the total energy consumed by the data center.
#4: Bits per Kilowatt-hour	Ratio of the total bit volume from every outbound router on the data center network divided by the total energy consumed.
#5: Weighted CPU Utilization – SPECint_rate	The amount of useful work produced in the data center is derived from the average CPU utilization, processor frequency and SPECint_rate2006 result for each server in the data center. Work is then divided by the total energy drawn by the data center during the assessment window.
#6: Weighted CPU Utilization – SPECpower	Utilizes published data from SPECpower benchmark results, in conjunction with a direct measurement of CPU utilization, to estimate the work efficiency of an individual server or groups of servers. An efficiency number for the entire data center can then be obtained by correlating CPU utilization to server models and SPECpower scores. Divide the result by the energy consumed by the data center during the assessment window.
#7: Compute Units per Second (CUPS)	Utilizes the year servers were purchased, the estimated CUPS, and average CPU utilization over an assessment window to determine work. Divide the result by the total energy consumed.
#8: Operating System Workload Efficiency	Ratio of total number of operating system instances running in the facility during the assessment window divided by the total facilities power.

Source: The Green Grid White Paper #17: Proxy Proposals for Measuring Data Center Productivity

Why We Are Here

- **“Can only improve what you can measure”**

- **Context**
  - Power consumption of HPC and facilities cost are increasing

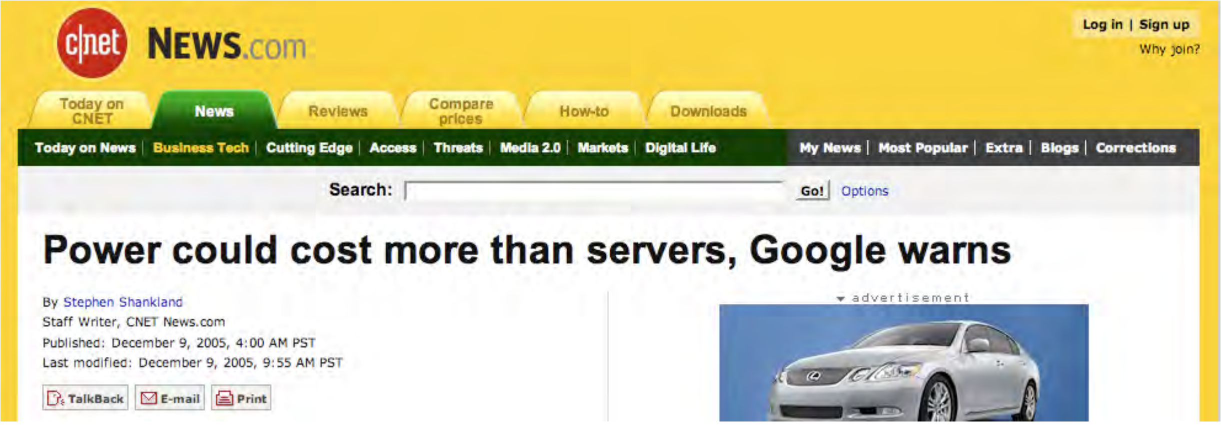
- **What is needed?**
  - Converge on a common basis between different research and industry groups for:
    - metrics
    - methodologies
    - workloads

    for energy-efficient supercomputing, so we can make progress towards solutions.

EXTRA SLIDES

Power is an Industry Wide Problem

(2% of US power consumption and growing)


c|net NEWS.com
Log in | Sign up
Why join?
Today on CNET
News
Reviews
Compare prices
How-to
Downloads
Today on News | Business Tech | Cutting Edge | Access | Threats | Media 2.0 | Markets | Digital Life
My News | Most Popular | Extra | Blogs | Corrections
Search:
Go!
Options
TalkBack
E-mail
Print
advertisement

Power could cost more than servers, Google warns

By Stephen Shankland
Staff Writer, CNET News.com
Published: December 9, 2005, 4:00 AM PST
Last modified: December 9, 2005, 9:55 AM PST

The New York Times

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

New Google Plant in The Dulles, Oregon,
from NYT, June 14, 2006

Relocate to Iceland?

Traditional Sources of Performance Improvement are Flat-Lining

- **New Constraints**
  - 15 years of *exponential* clock rate growth has ended
- **But Moore’s Law continues!**
  - 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!

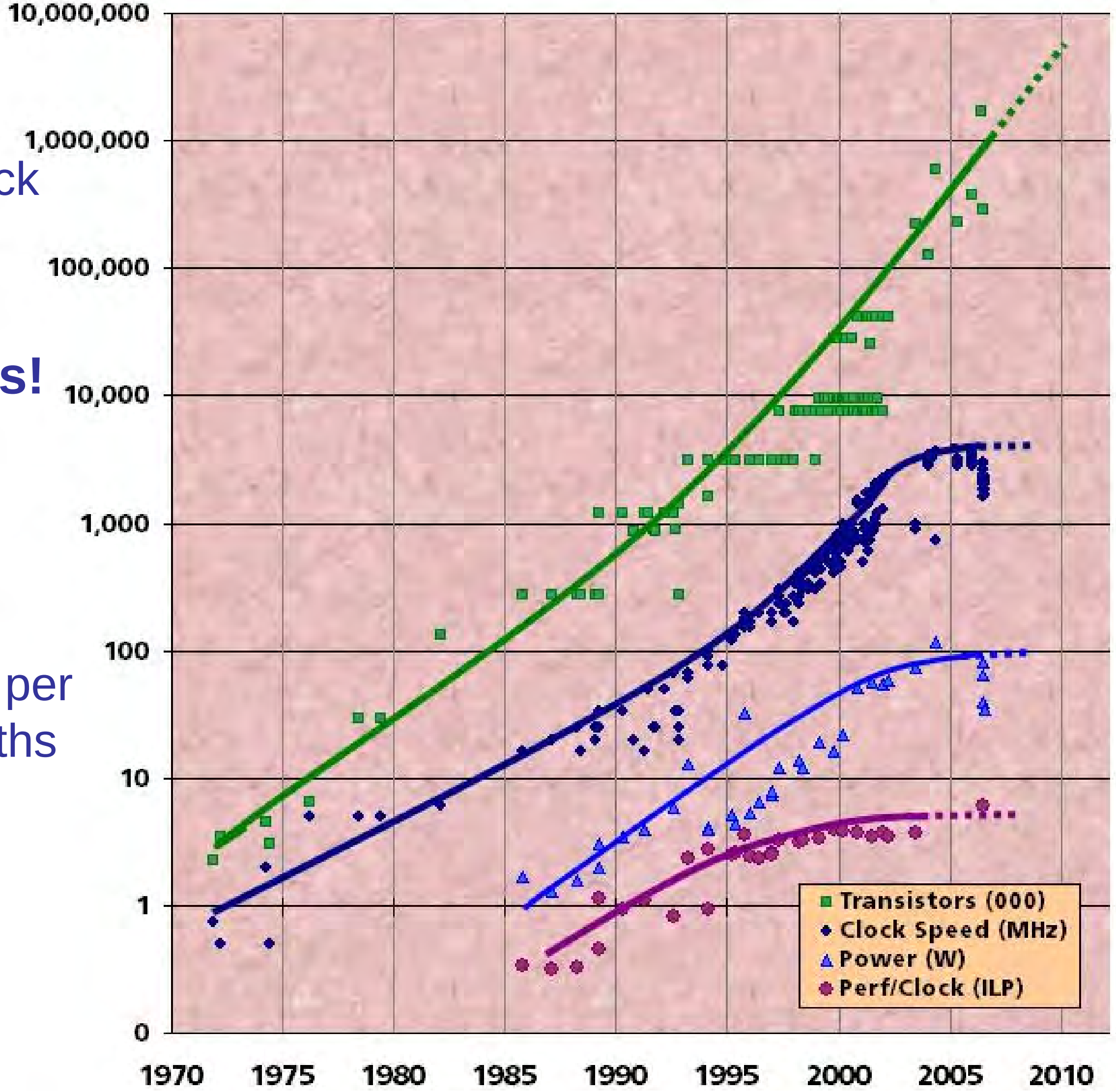


Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

Agreement in principal

- Collaboration between Top500, Green500, Green Grid and EE HPC WG
- Improve methodology, metrics, instrumentation and testing
- Evaluate new technologies for HPC compute system energy efficiency
- Report progress at SC and ISC