Power Efficiency and the Top500

John Shalf, Shoaib Kamil, Erich Strohmaier, David Bailey

Lawrence Berkeley National Laboratory (LBNL)
National Energy Research Supercomputing Center (NERSC)

Top500 Birds of a Feather
SC2007, Reno Nevada
November 14, 2007

New Design Constraint: *POWER*

- **Transistors still getting smaller**
  - Moore's Law is alive and well
- **But Denard scaling is dead!**
  - No power efficiency improvements with smaller transistors
  - No clock frequency scaling with smaller transistors
  - All "magical improvement of silicon goodness" has ended
- **Traditional methods for extracting more performance are well-mined**
  - Cannot expect exotic architectures to save us from the "power wall"
  - Even resources of DARPA can only accelerate existing research prototypes (not "magic" new technology)!

Estimated Exascale Power Requirements

- **Recent Baltimore Sun Article on NSA system in Maryland**
  - Consuming 75MW and growing up to 15MW/year
  - Not enough power left for city of Baltimore!

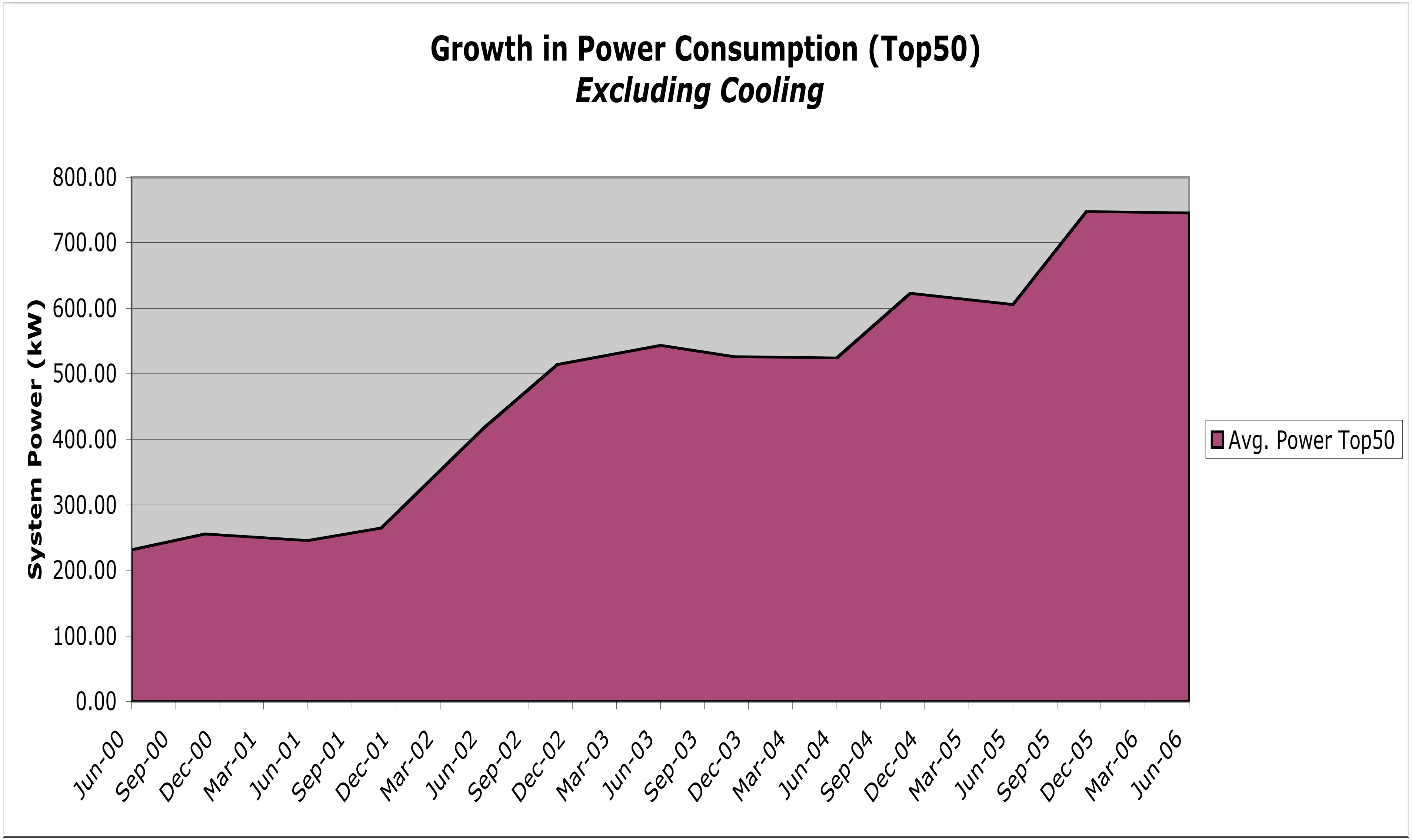
- **LBNL IJHPCA Study for ~1/5 Exaflop for Climate Science in 2008**
  - Extrapolation of Blue Gene and AMD design trends
  - Estimate: **20 MW** for BG and **179 MW** for AMD

- **DOE E3 Report**
  - Extrapolation of existing design trends to exascale in 2016
  - Estimate: **130 MW**

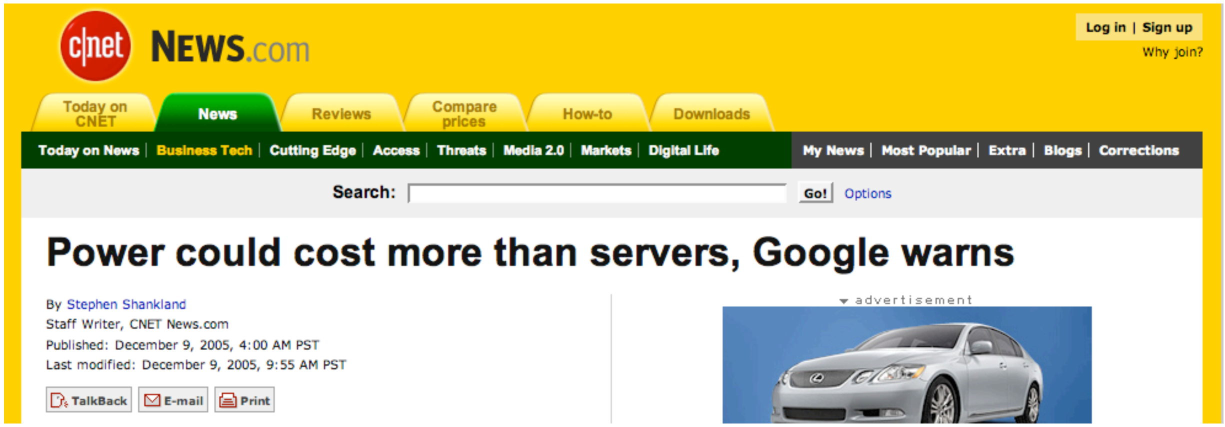
- **DARPA Study**
  - More detailed assessment of component technologies
  - Estimate: **20 MW** just for memory alone, **60 MW** aggregate extrapolated from current design trends

The current approach is not sustainable!

Top500 Estimated Power Requirements



Power is an Industry Wide Problem


cnet 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

Broader Objective

- **Role of Top500**
  - Collected history of largest HEC investments in the world
  - Archive of system metrics plays important role in analyzing industry trends (architecture, interconnects, processors, OS, vendors)
  - Can play an important role in collecting data necessary to understand power efficiency trends
  - Feed data to studies involving benchmarks other than LINPACK as well (you can do your own rankings!)
- **Objectives**
  - Use Top500 List to track power efficiency trends
  - Raise Community Awareness of HPC System Power Efficiency
  - Push vendors toward more power efficient solutions by providing a venue to compare their power consumption

Recap *(from SC06 and ISC07)*

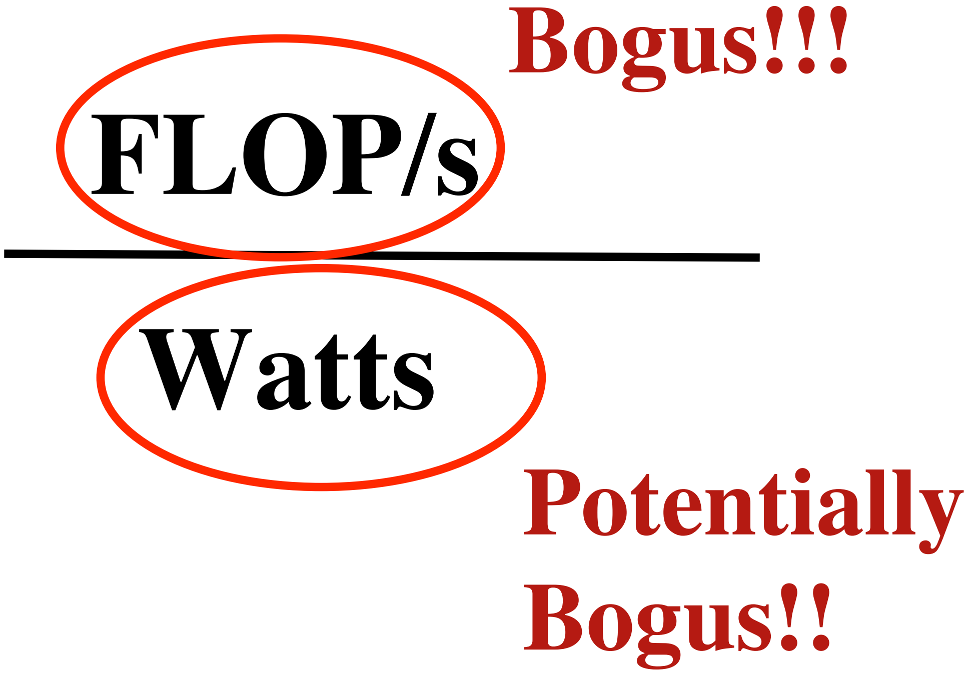
- **Last year we presented the following**
  - Introduced Looming Power Crisis in HPC
  - Described Top500's interest in collecting power efficiency data
  - Estimated Top500 power trends from vendor specs.

- **Questions about measurement**
  - *How far is manufacturer's specs from actual power?*
  - *If I have to measure, how do I do it?*
  - *Does power consumed by HPL reflect power consumed by regular workload? (is this data relevant?)*
  - *How do I collect the data without taking down my system for an entire day? (making it non-invasive)*
  - *What are some real-world experiences with collecting this data?*
  - *What metric should be used for ranking systems?*

Anatomy of a “Value” Metric

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

Anatomy of a "Value" Metric



Bogus!!!

FLOP/s

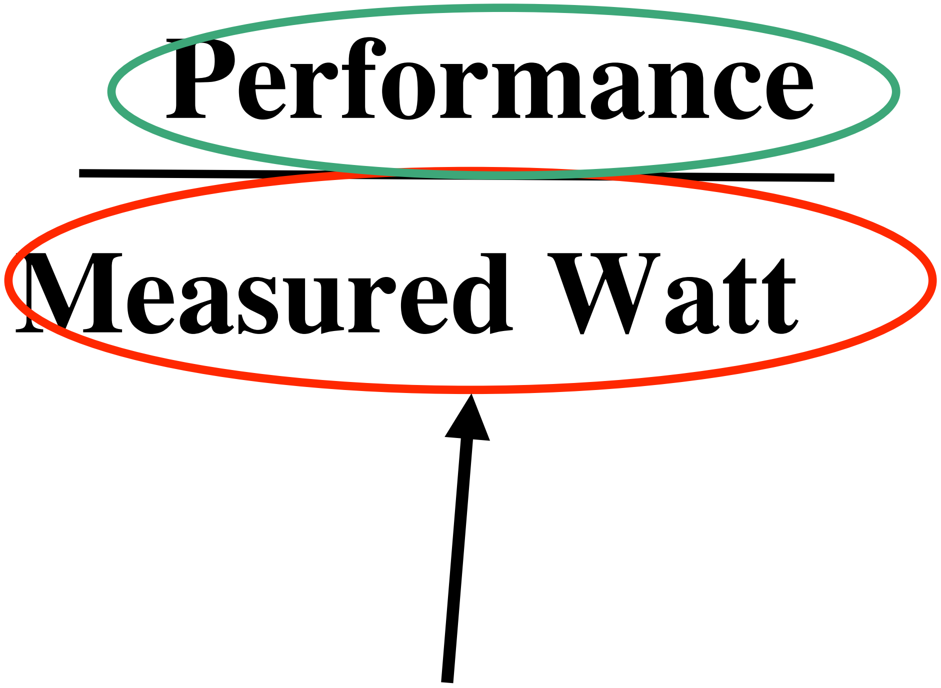
Watts

Potentially Bogus!!

Anatomy of a “Value” Metric

Choose your own metric for performance!

(doesn’t need to be HPL, or FLOPS)



Performance

Measured Watt

This is what Top500 is interested in collecting

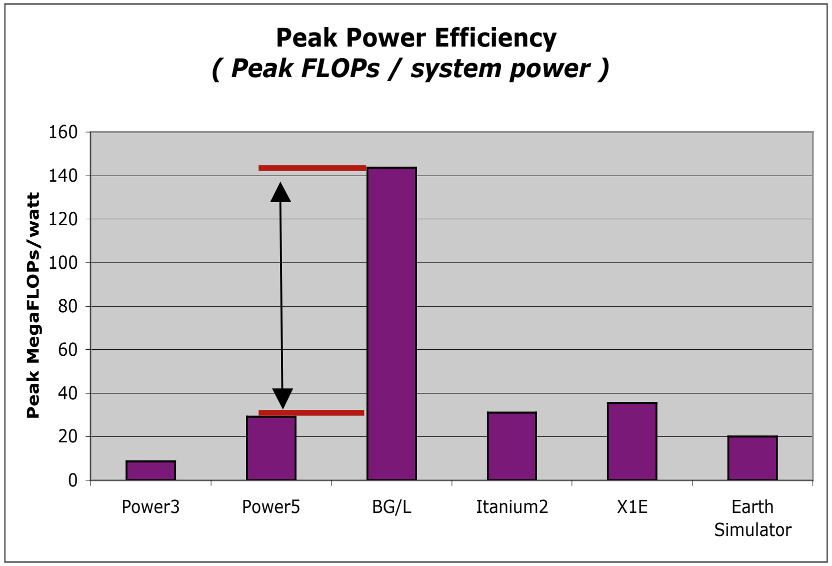
(but is it applicable to other workloads?)

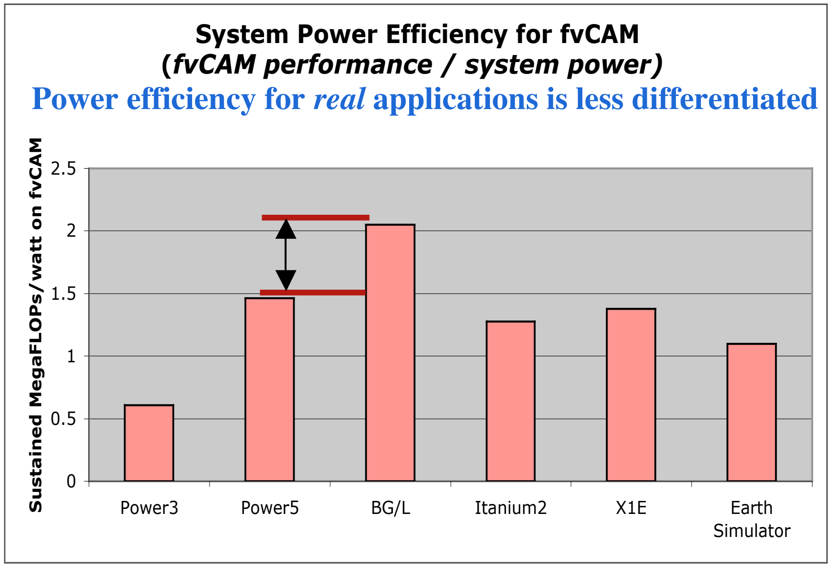
Power Efficiency running fvCAM

Performance/measured_watt

is much more useful than

FLOPs/peak_watt





Power efficiency for *real* applications is less differentiated

Benchmark results from Michael Wehner, Art Mirin, Patrick Worley, Leonid Oliker

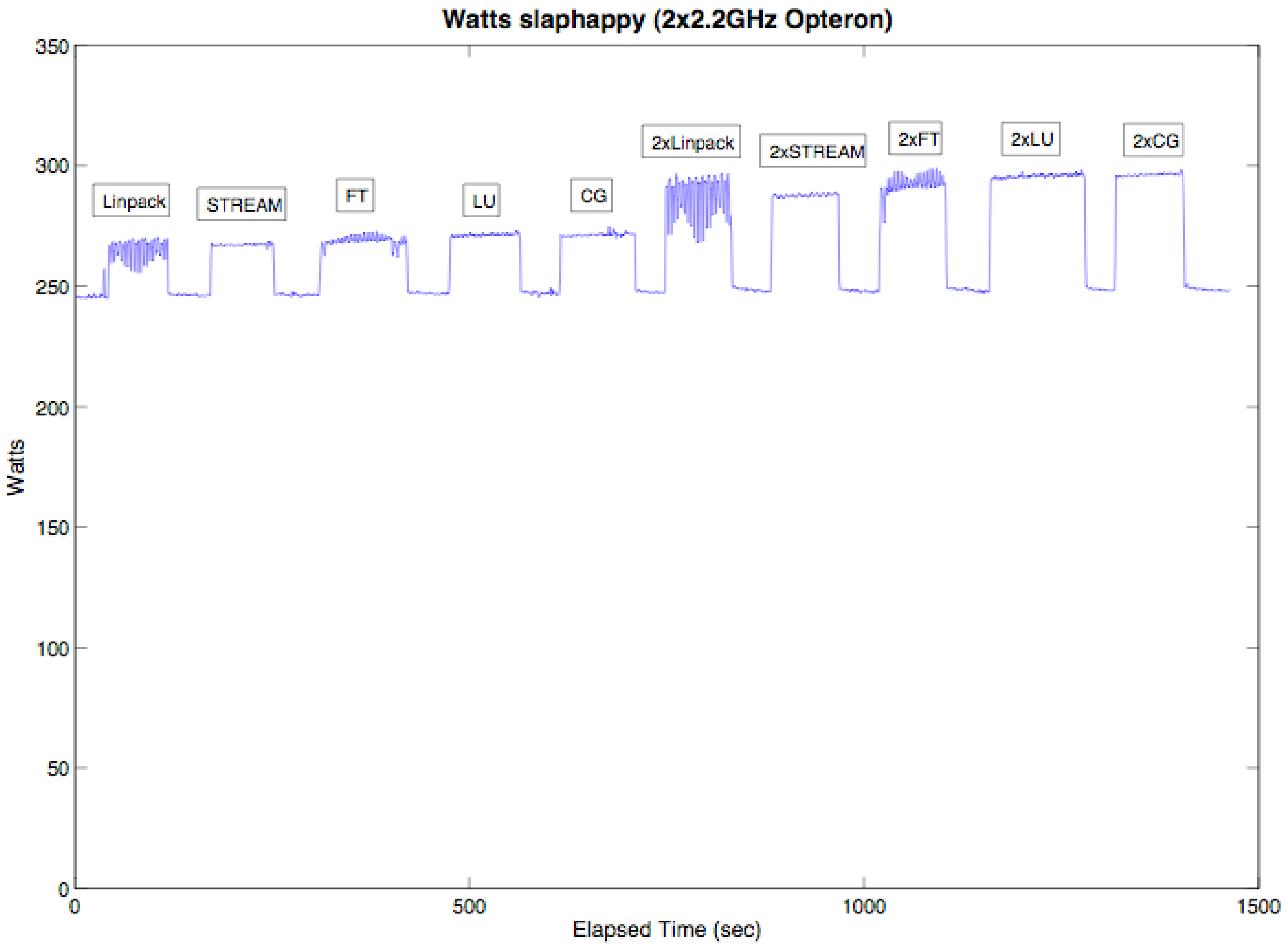
Questions

- How do I measure power?
- Does Power Consumed when running LINPACK reflect power consumption when running normal system load (or other benchmarks)?
- Is there a sane way to estimate power consumption under LINPACK/HPL without taking down my system to perform the measurement?
- What should be included/excluded from measurement?

Compared Methods Measure Power

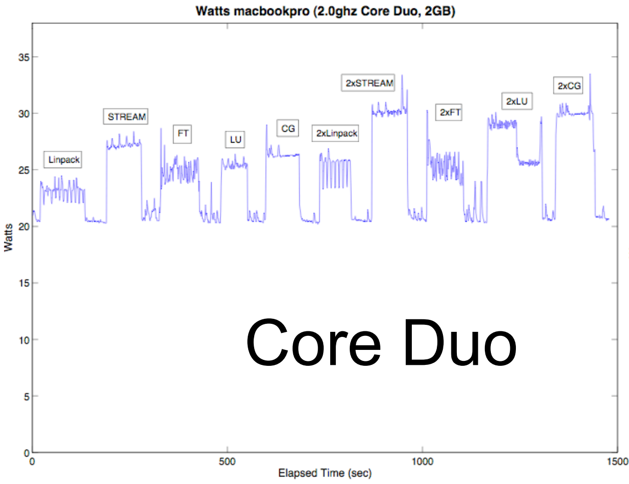
- **Clamp meters**
  - +: easy to use, don’t need to disconnect test systems, wide variety of voltages
  - -: very inaccurate for more than one wire
- **Inline meters**
  - +: accurate, easy to use, can output over serial
  - -: must disconnect test system, limited voltage, limited current
- **Power panels / PDU panels**
  - Unknown accuracy, must stand and read, usually coarse-grained (unable to differentiate power loads)
  - Sometimes the best or only option: can get numbers for an entire HPC system
- **Integrated monitoring in system power supplies (Cray XT)**
  - +: accurate, easy to use
  - - : only measures single cabinet. Must know power supply conversion efficiency to project nominal power use at wall socket

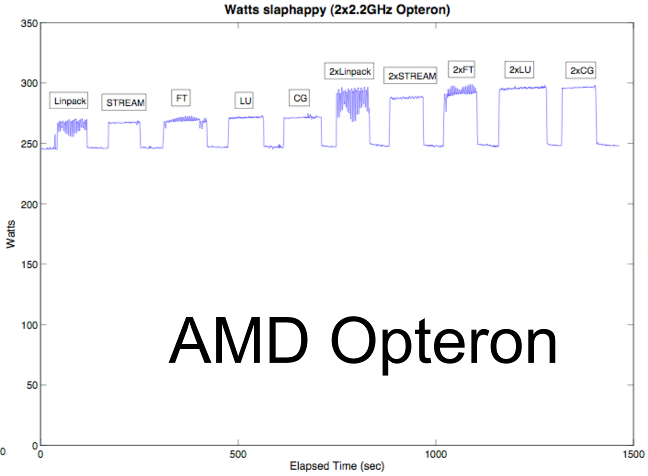
Single Node Tests: AMD Opteron

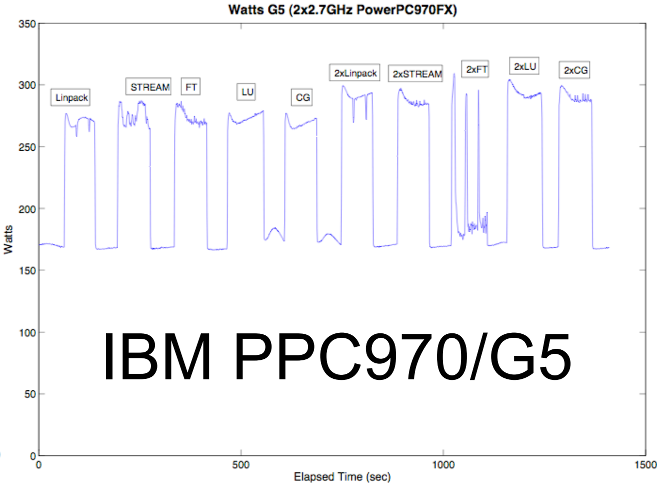


- Highest power usage is 2x NAS FT and LU

Similar Results when Testing Other CPU Architectures

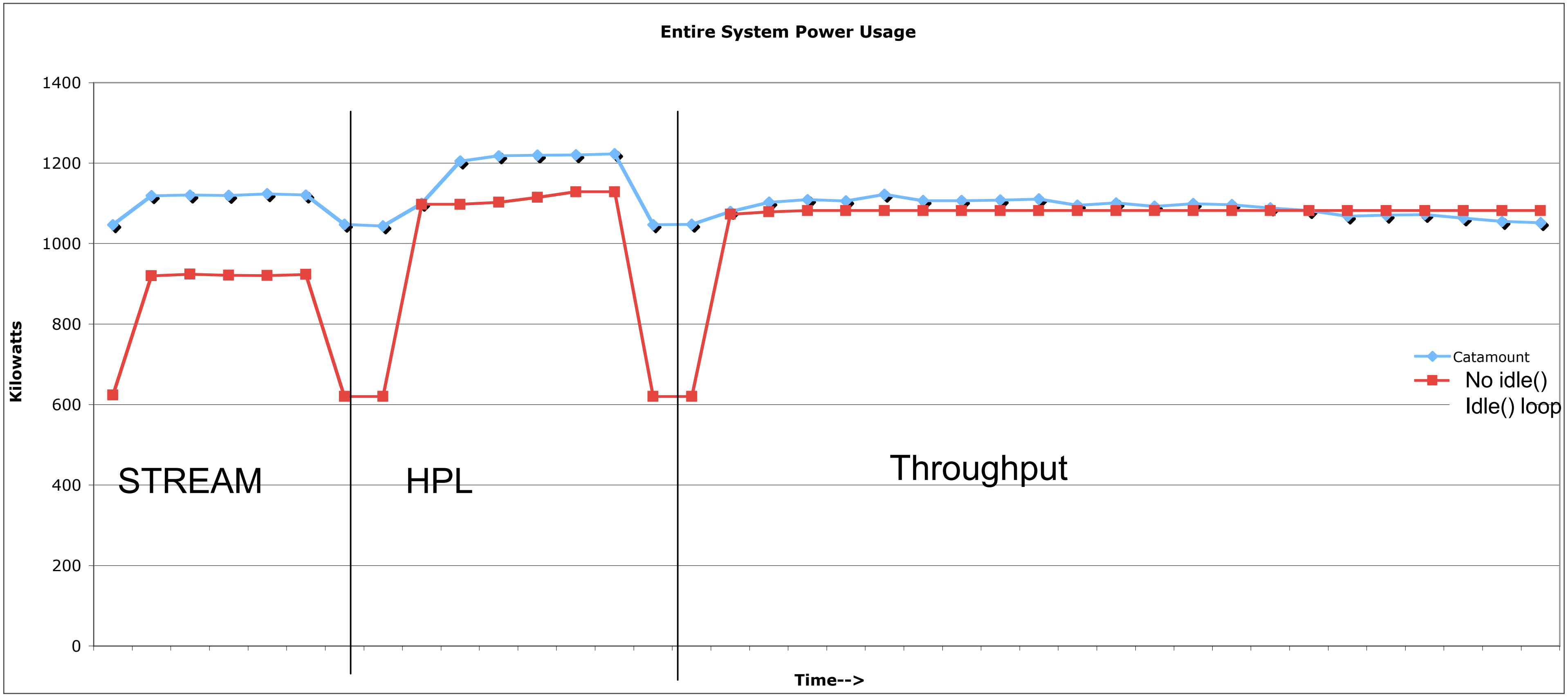






- Power consumption far less than manufacturer' estimated "nameplate power"
- Idle power much lower than active power
- Power consumption when running LINPACK is very close to power consumed when running other compute intensive applications

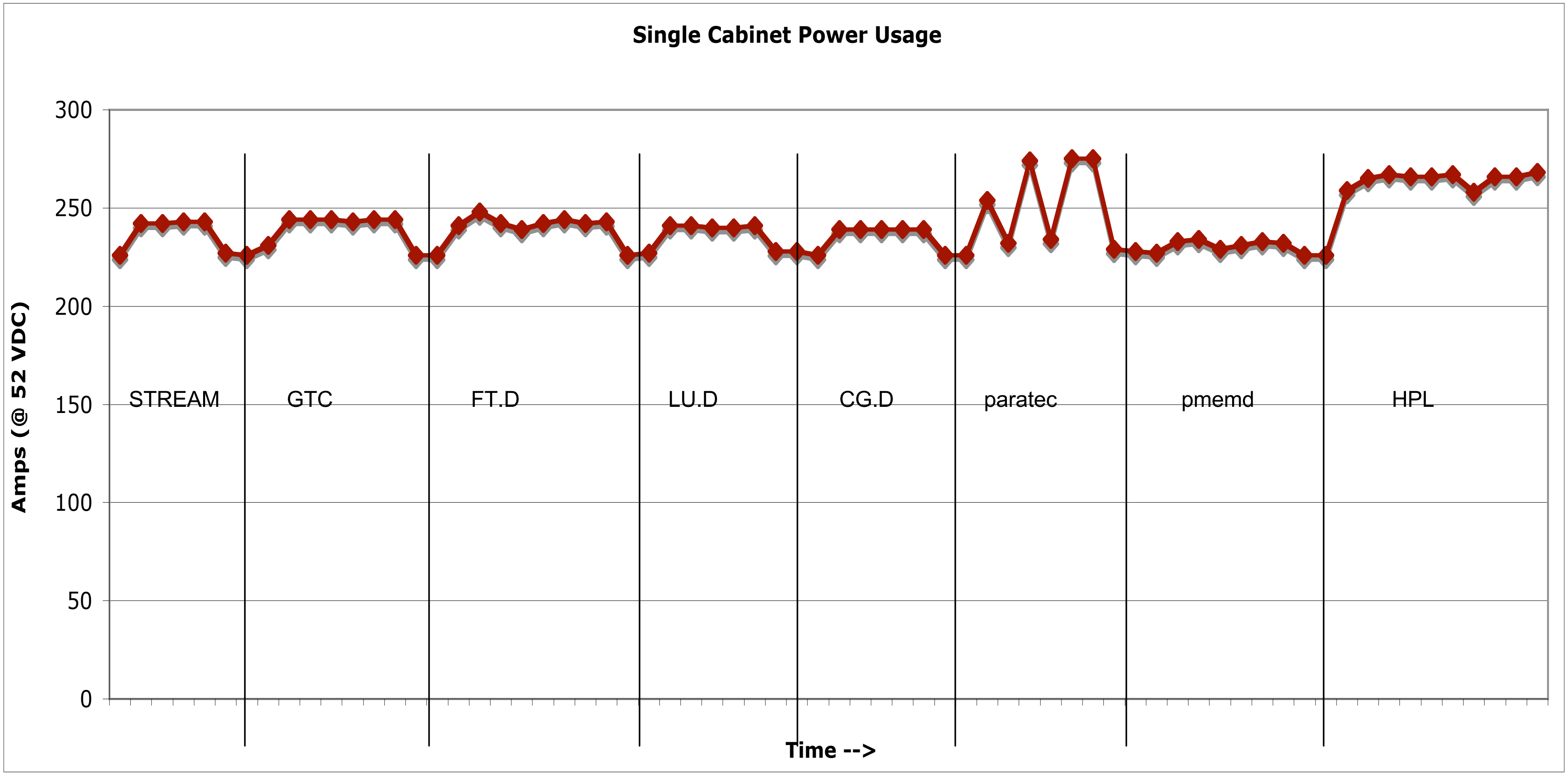
Full System Test on Cray XT4 (franklin.nersc.gov)



- Tests run across all 19,353 compute cores
- Throughput: NERSC “realistic” workload composed of full applications
- idle() loop allows powersave on unused processors; (generally more efficient)

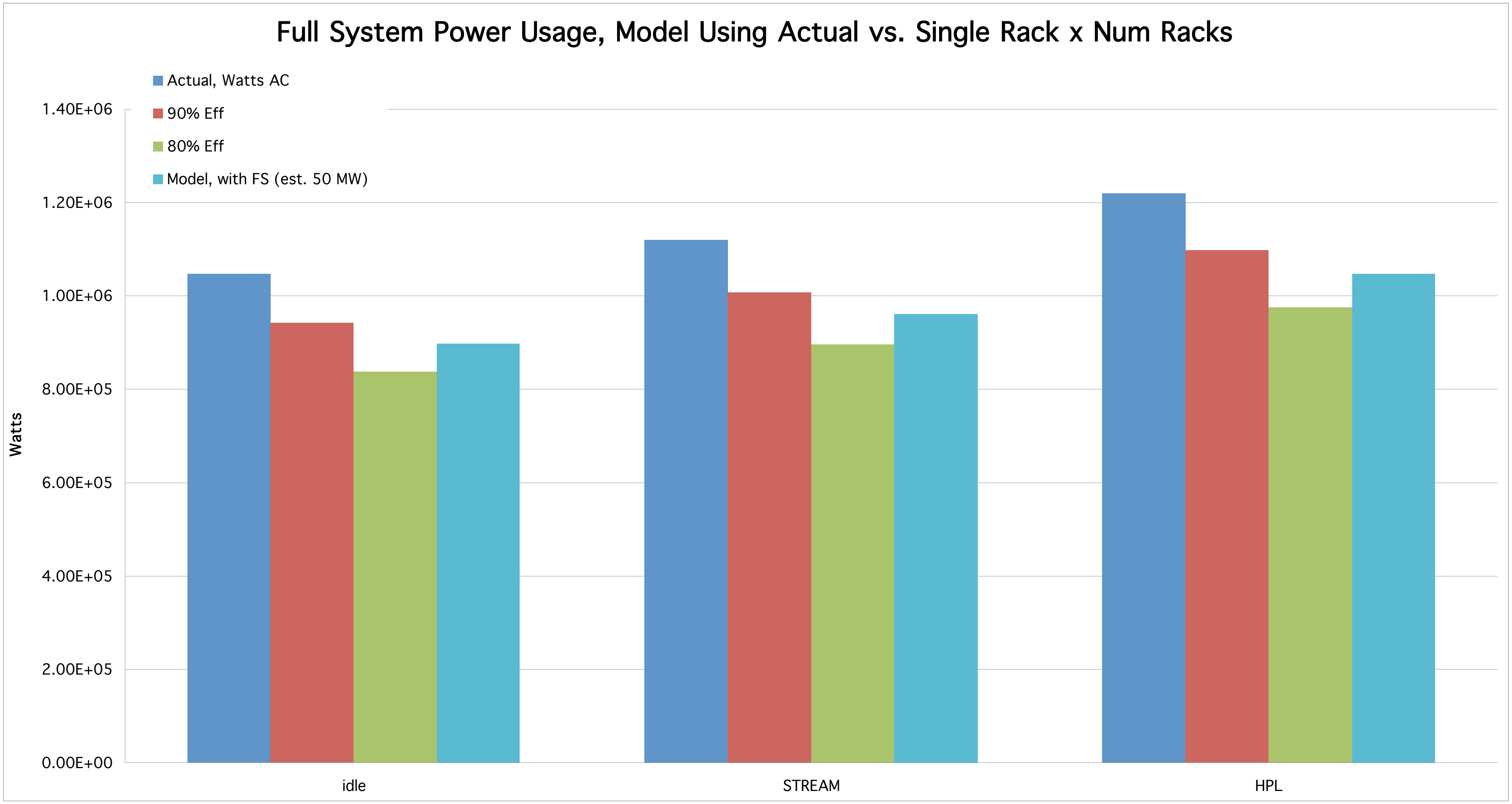
Single Rack Tests

(extrapolating power consumption)



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

Modeling the Entire System



- Error factor is 0.05 if we assume 90% efficiency

Power Measurement

(what to include/exclude)

- **Facility Cooling: NO!**
- **Compute Nodes: Yes!**
  - Collect power consumption when running HPL
  - Can accurately extrapolate from HPL running on any subset of the system
- **Switches (infiniband or Ethernet) Yes!**
  - Are not generally stressed by HPL
  - Appear to consume nearly the same power under any load *(may change in future with IEEE stds.)*
- **I/O: NO!**
  - Probably should exclude if possible
  - Ambiguous contribution because of increasing use of facility-wide filesystems
  - Not stressed by HPL

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-sys 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!
- **Need to come up with a better NUMERATOR for our "power efficiency" metric**
  - Top500 will be able to collect the data
  - But ranking by **HPL_FLOPs/measure_watts** probably not the right message to the HPC industry!