Power Efficiency and the Top500

John Shalf and David Bailey

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

December 7, 2006

BIPS

BERKELEY LAB

NERSC

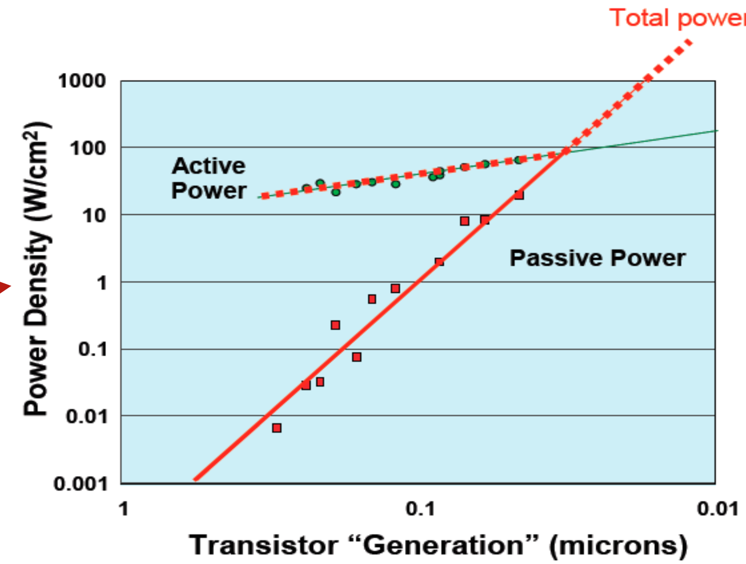
Office of Science
U.S. DEPARTMENT OF ENERGY

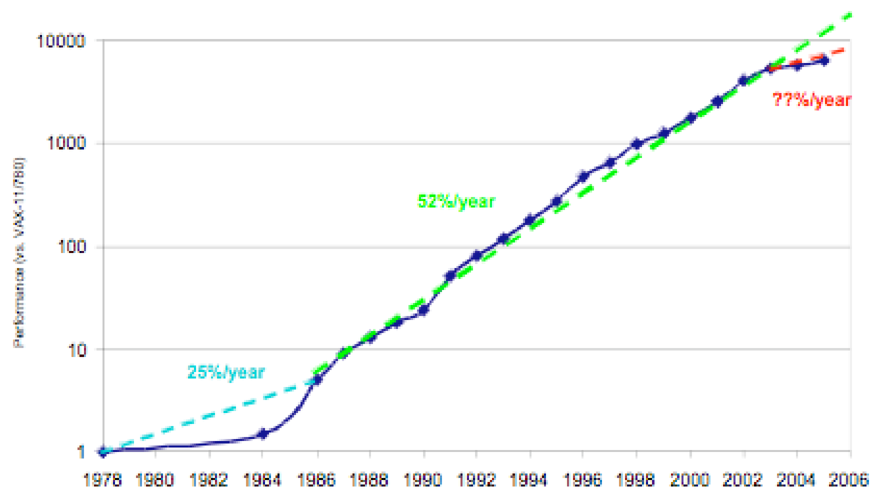
What is Happening Now?

- Moore's Law
  - Silicon lithography will improve by 2x every 18 months
  - Double the number of transistors per chip every 18mo.
- CMOS Power

  Total Power = $\underbrace{V^2 * f * C}_{\text{active power}} + \underbrace{V * I_{leakage}}_{\text{passive power}}$

  - As we reduce feature size Capacitance (C) decreases proportionally to transistor size
  - Enables increase of clock frequency (f) proportionally to Moore's law lithography improvements, with same power use
  - This is called "Fixed Voltage Clock Frequency Scaling" *(Borkar `99)*
- Since ~90nm
  - $V^2 * f * C \; {\sim}{=} \; V * I_{leakage}$
  - Can no longer take advantage of frequency scaling because passive power ($V * I_{leakage}$) dominates
  - Result is recent clock-frequency stall reflected in Patterson Graph at right





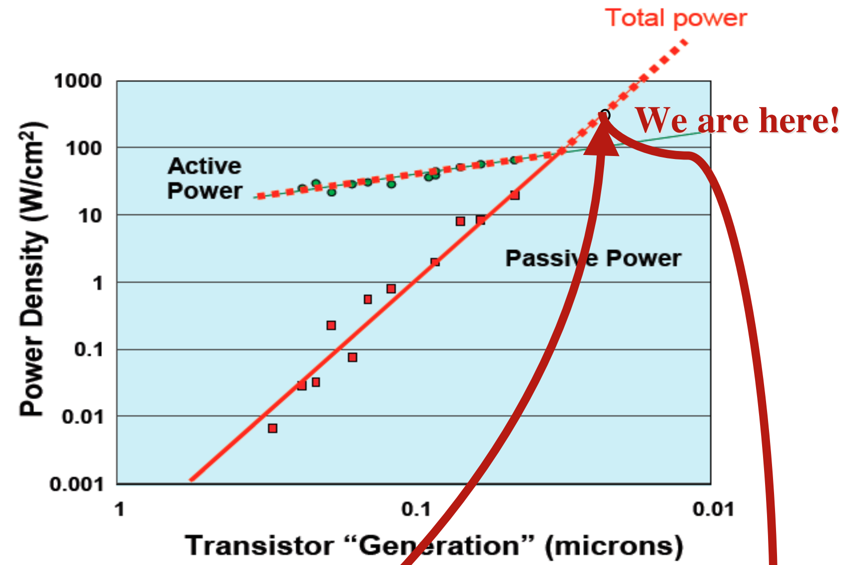
SPEC_Int benchmark performance since 1978 from Patterson & Hennessy Vol 4.

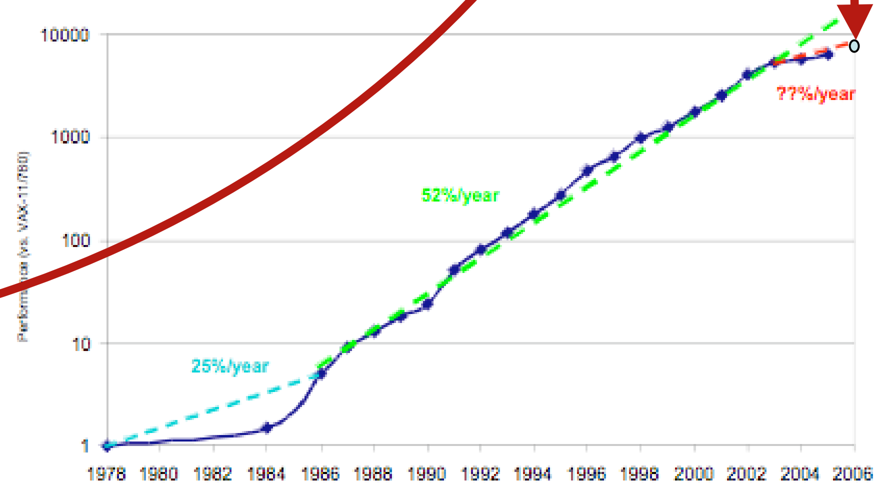
What is Happening Now?

- Moore's Law
  - Silicon lithography will improve by 2x every 18 months
  - Double the number of transistors per chip every 18mo.
- CMOS Power

  Total Power = $\underbrace{V^2 * f * C}_{\text{active power}} + \underbrace{V * I_{leakage}}_{\text{passive power}}$

  - As we reduce feature size Capacitance (C) decreases proportionally to transistor size
  - Enables increase of clock frequency (f) proportionally to Moore's law lithography improvements, with same power use
  - This is called "Fixed Voltage Clock Frequency Scaling" *(Borkar `99)*
- Since ~90nm
  - $V^2 * f * C \;\sim= V * I_{leakage}$
  - Can no longer take advantage of frequency scaling because passive power ($V * I_{leakage}$) dominates
  - Result is recent clock-frequency stall reflected in Patterson Graph at right





SPEC_Int benchmark performance since 1978 from Patterson & Hennessy Vol 4.

What is *Going* to Happen?

- New Constraints
  - Power limits clock rates
  - Cannot squeeze more performance from ILP *(complex cores)* either!
- But Moore's Law continues!
  - What to do with all of those transistors if everything else is flat-lining?
  - Now, #cores per chip doubles every 18 months *instead* of clock frequency!
- **Power Consumption** is chief concern for system architects
- **Power-Efficiency** is the primary concern of consumers of computer systems!

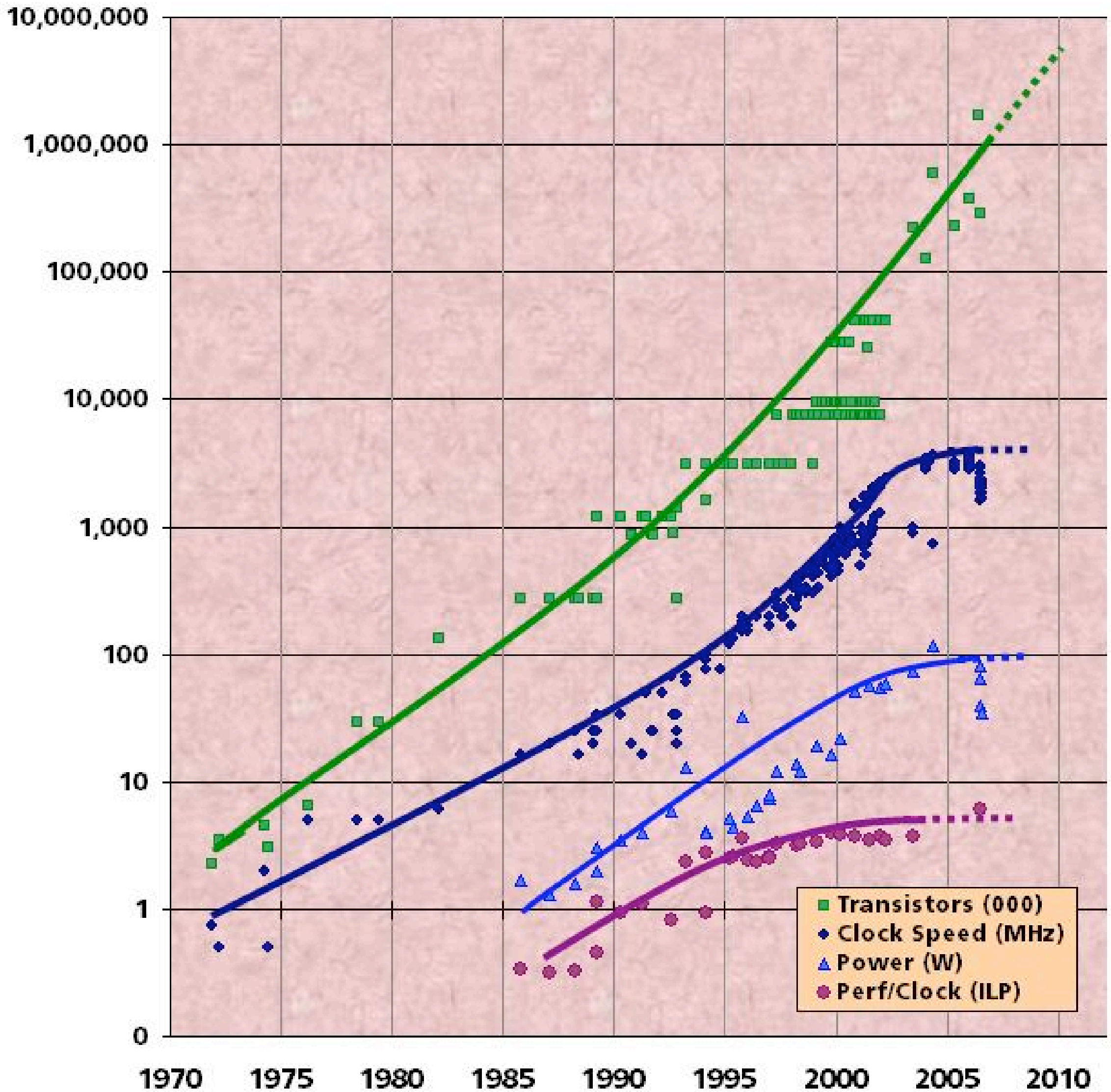
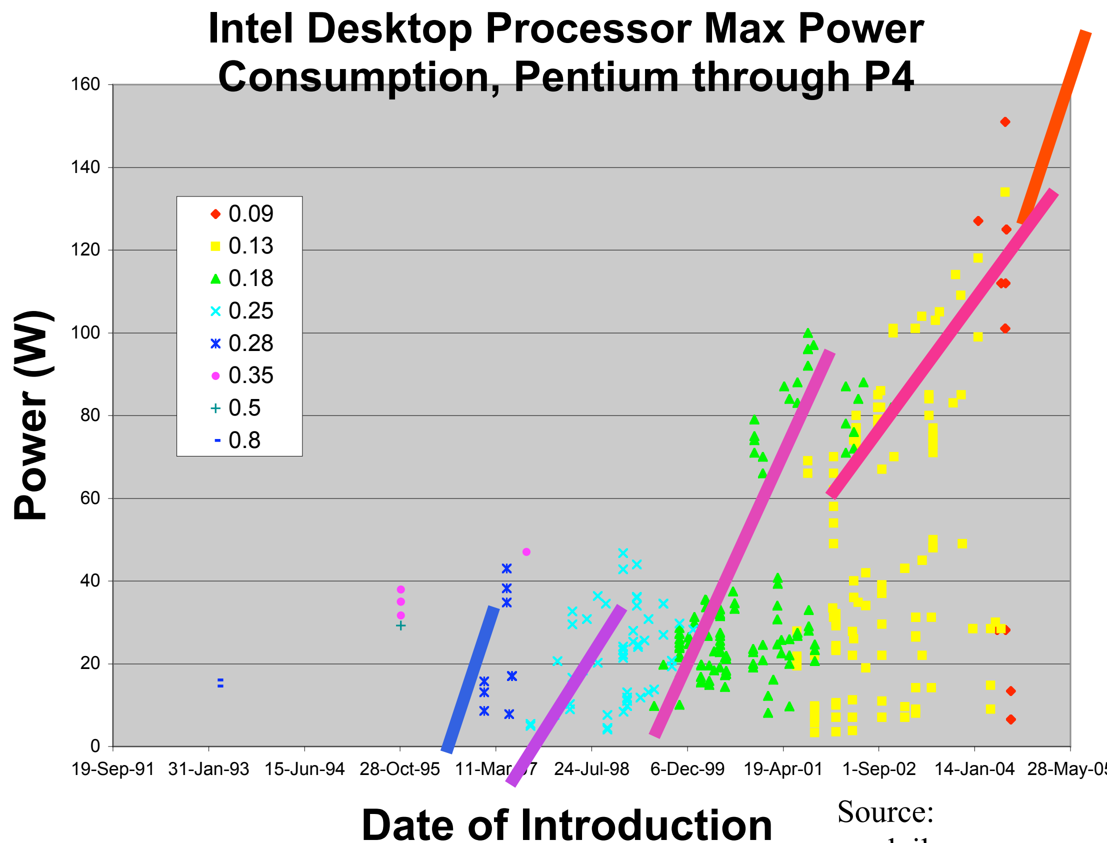


Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

Microprocessors: Up Against the Wall(s)

From Joe Gebis

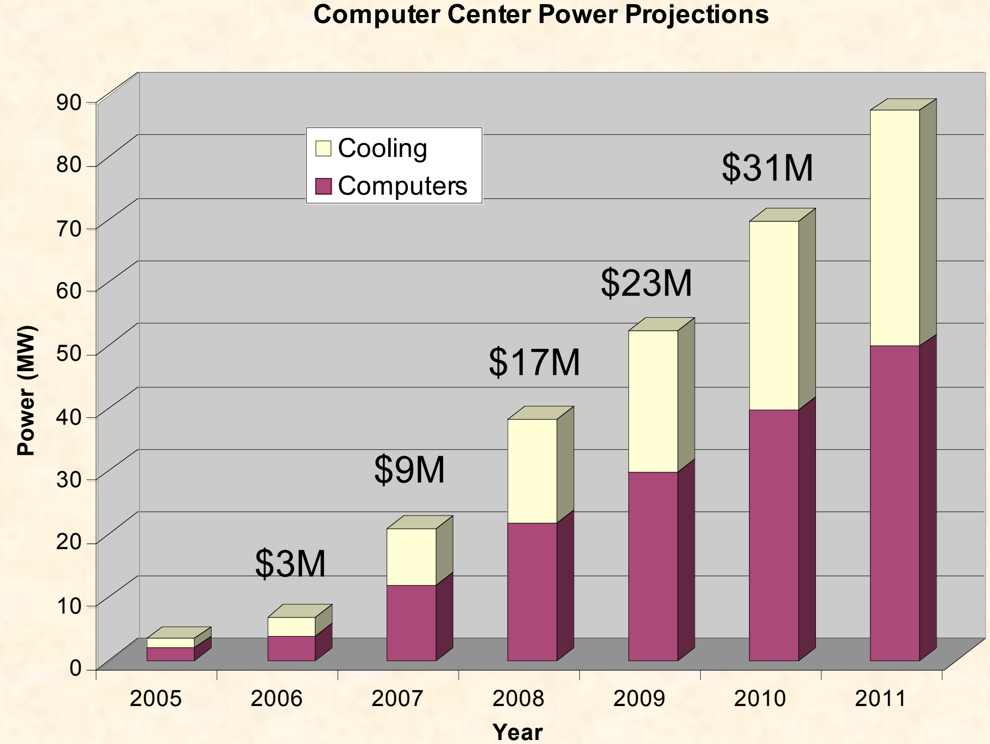
- Microprocessors are hitting a power wall
  - Higher clock rates and greater leakage increasing power consumption
- Reaching the limits of what non-heroic heat solutions can handle
- Newer technology becoming more difficult to produce, removing the previous trend of "free" power improvement



Source: sandpile.org

ORNL Computing Power and Cooling 2006 - 2011

- Immediate need to add 8 MW to prepare for 2007 installs of new systems
- NLCF petascale system could require an additional 10 MW by 2008
- Need total of 40-50 MW for projected systems by 2011
- Numbers just for computers: add 75% for cooling
- Cooling will require 12,000 – 15,000 tons of chiller capacity



Cost estimates based on $0.05 kW/hr

Annual Average Electrical Power Rates $/MWh

Site	FY 2005	FY 2006	FY 2007	FY 2008	FY 2009	FY 2010
LBNL	43.70	50.23	53.43	57.51	58.20	56.40 *
ANL	44.92	53.01				
ORNL	46.34	51.33				
PNNL	49.82	N/A				

Data taken from Energy Management System-4 (EMS4). EMS4 is the DOE corporate system for collecting energy information from the sites. EMS4 is a web-based system that collects energy consumption and cost information for all energy sources used at each DOE site. Information is entered into EMS4 by the site and reviewed at Headquarters for accuracy.

Tension Between Commodity and Specialized Architecture

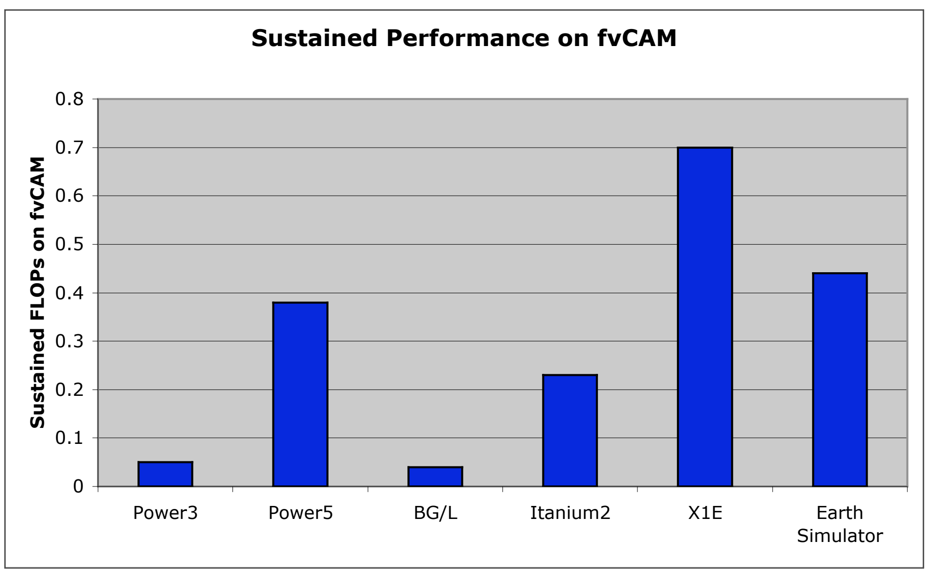
- Commodity Components
  - Amortize high development costs by sharing costs with high volume market
  - Accept lower computational efficiency for much lower capital equipment costs!
- Specialization
  - Specialize to task in order to improve computational efficiency.
  - Specialization used very successfully by embeded processor community
  - Not cost effective if volume is too low.
- When cost of power exceeds capital equipment costs
  - Commodity clusters are optimizing wrong part of the cost model
  - Will need for higher computational efficiency drive more specialization? *(look at embedded market… lots of specialization*)

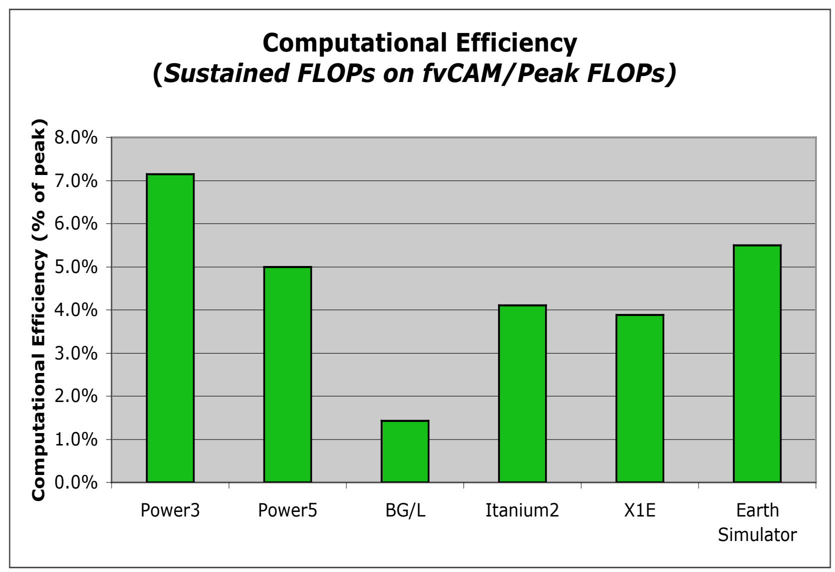
Tension between concurrency and power efficiency

- Highly concurrent systems can be more power efficient
  - *Dynamic power is proportional to V^2fC*
  - *Build systems with even higher concurrency?*
- However, many algorithms are unable to exploit massive concurrency yet
  - *If higher concurrency cannot deliver faster time to solution, then power efficiency benefit wasted*
  - *So we should build fewer/faster processors?*

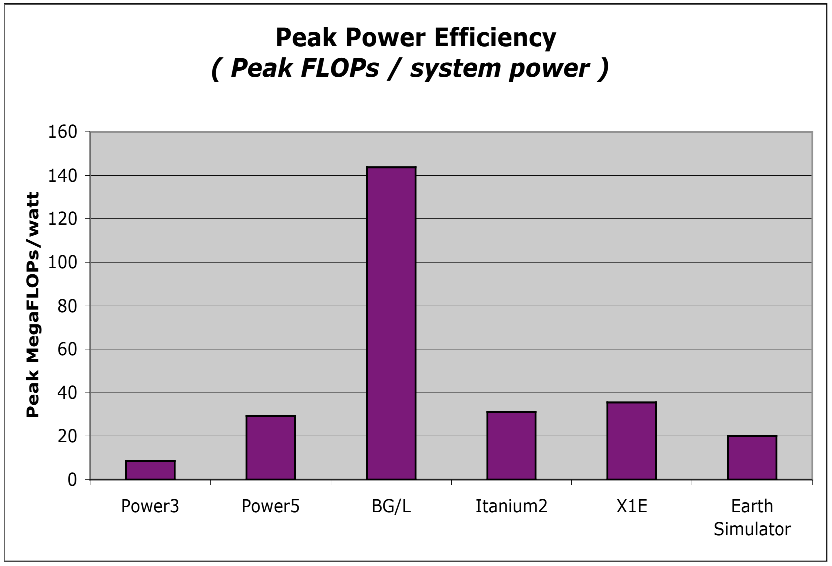
Power Efficiency vs. Power Consumption

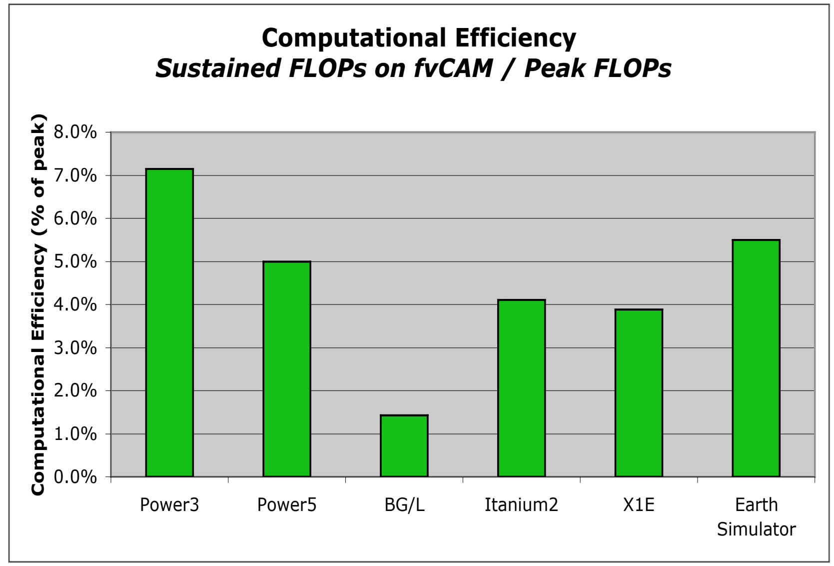
- Vendor Focus has been driven by Peak FLOPs/watt or reducing idle-power consumption using Dynamic Frequency/Voltage Scaling
  - Good for Consumer electronics which are idle most of the time
  - Marginal Benefit for HPC
    - Run ~100% loads
    - Time to solution is important
    - Effective/sustained performance is more important than peak
- Need a good metric for computational efficiency in order to influence industry
  - Example with Climate Code (fvCAM) to show how easy it is to mislead

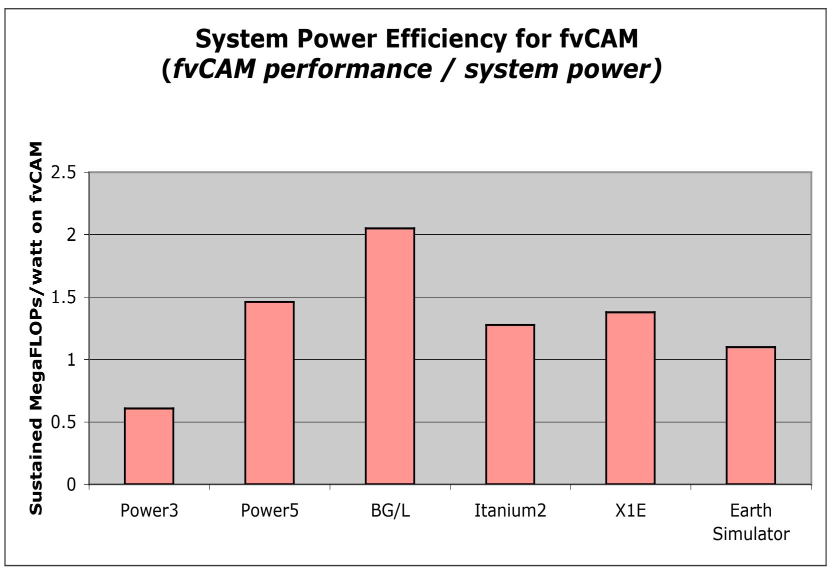


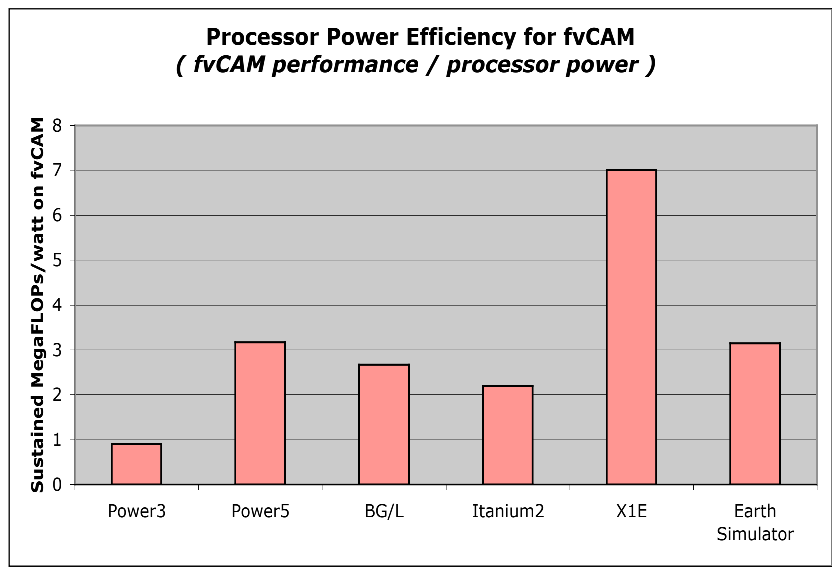


Power Efficiency running fvCAM



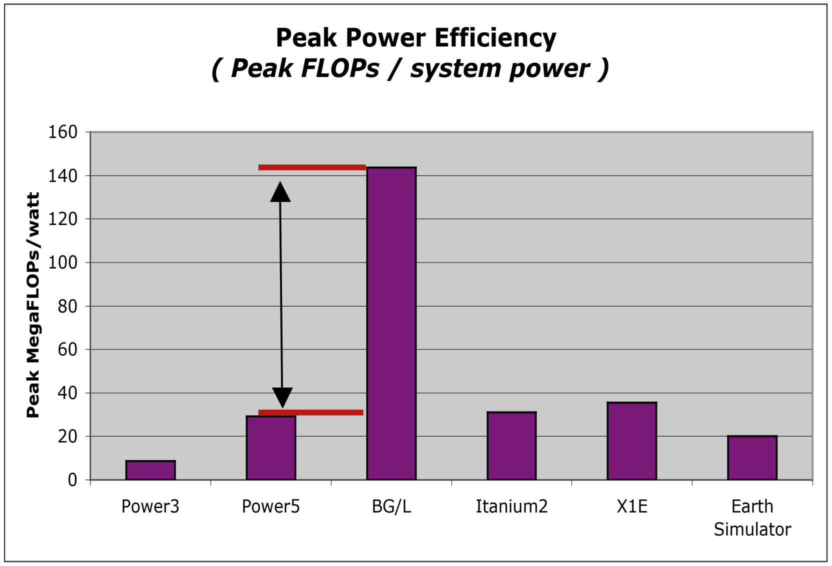


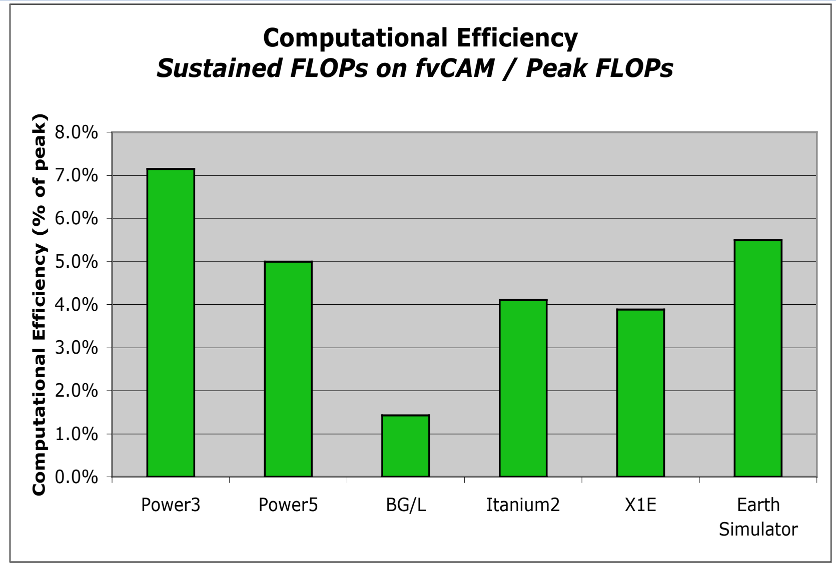


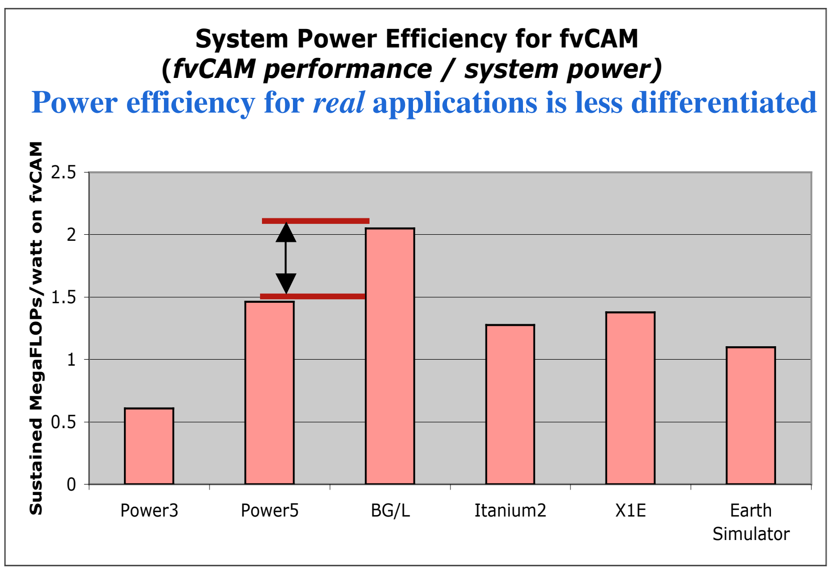


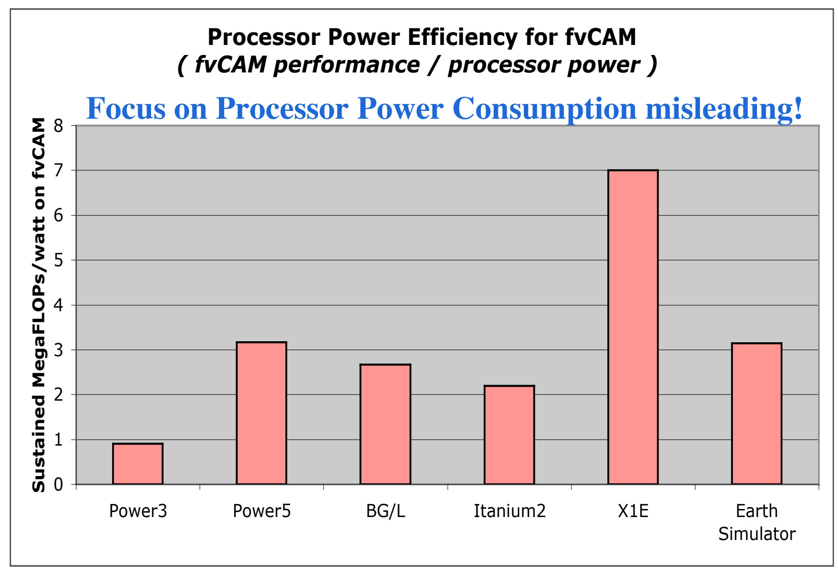
Benchmark results from Michael Wehner, Art Mirin, Patrick Worley, Leonid Oliker

Power Efficiency running fvCAM









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

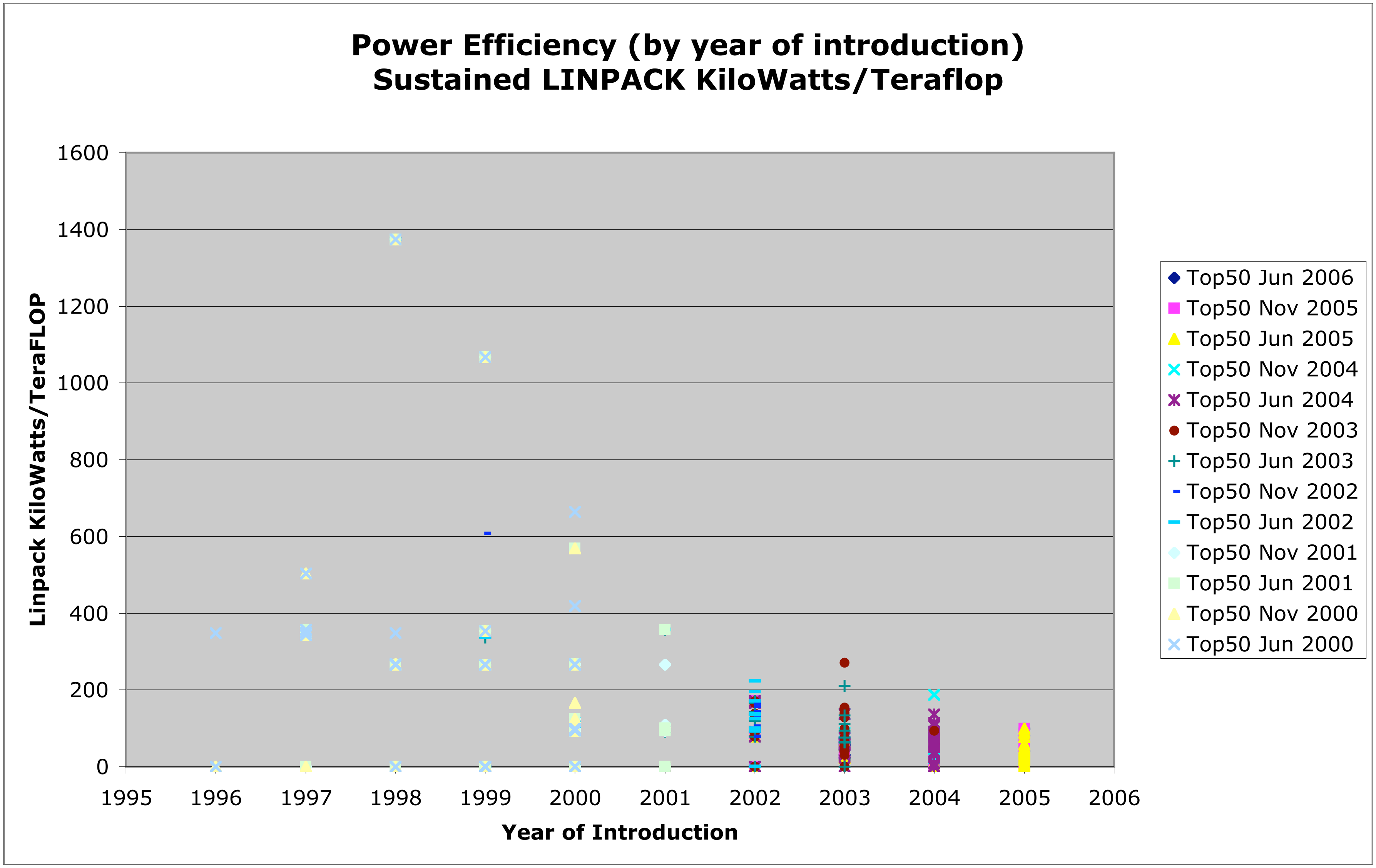
Need Power Efficiency Metrics based on Effective Performance

- We want to push industry in the *right* direction
- Leverage *established* performance benchmarks to serve as numerator for “power efficiency” ratio
- Segregate by workload
  - Transactional Workload: *EnergyStar Server Metrics* (Koomey 2006)
  - Small/Workstation: *Spec2006/Watt*
  - Midrange Cluster: *NAS Parallel Benchmarks MOPS/Watt*
  - HEC/Top500: *LINPACK/Watt? HPCC/Watt? SSP/Watt?*
- Role of Top500
  - Collected history of largest HEC investments in the world
  - Archive of system metrics plays important role in analyzing industry trends
  - 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

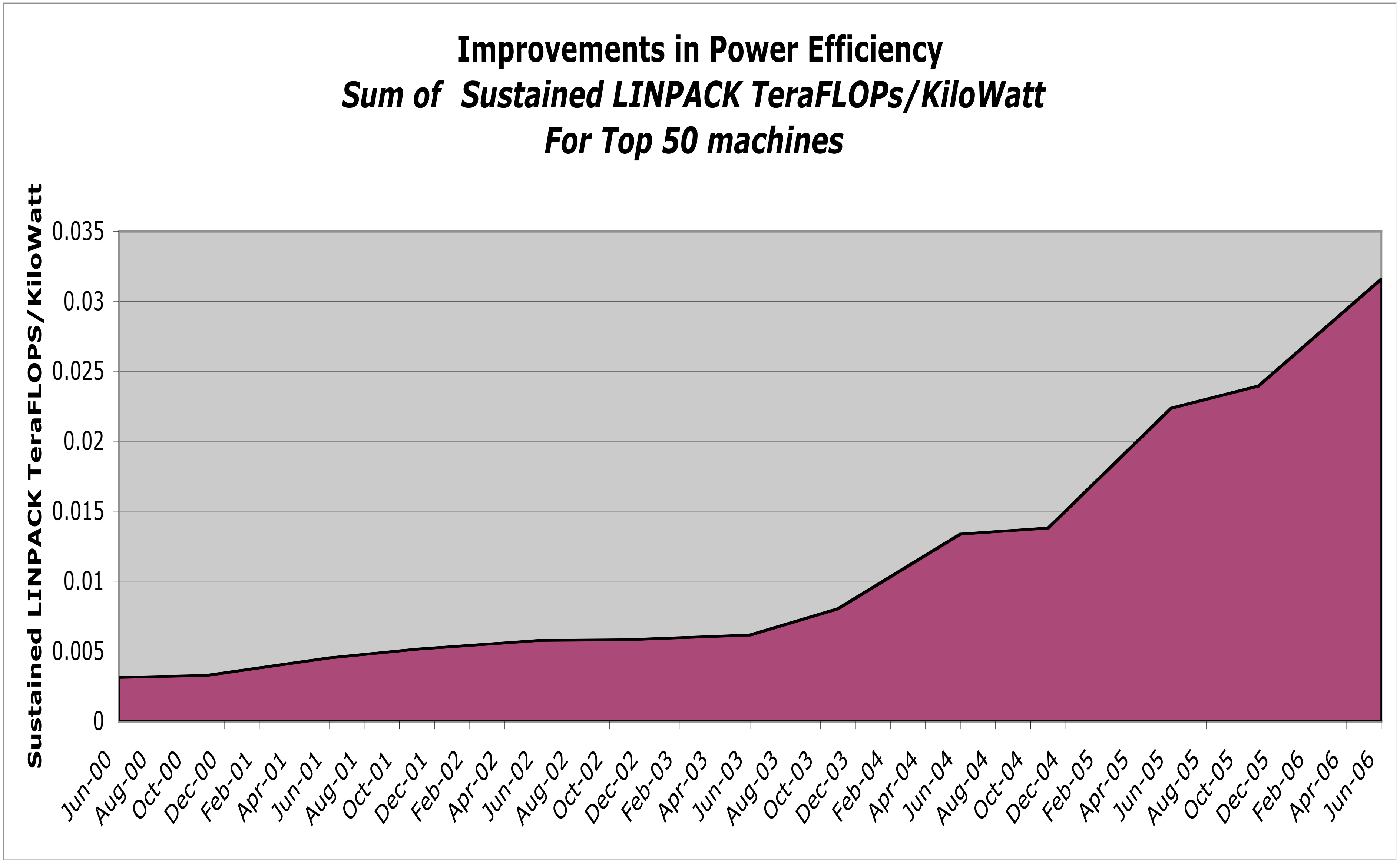
A Call to Action

- Please provide power consumption parameters to Top500 as part of machine configuration
- Segregate into 3 primary areas
  - System power consumption: *All system components excluding facility cooling and disk subsystem, SAN or archival storage*
  - Facility cooling power requirements: *Air handlers, chillers etc…*
  - Disk power requirements: *All mounted filesystems that are served locally excluding archival/tertiary storage*
- Data collection
  - Worst Case: Max rated power consumption (*mfr. specs.)*
  - Better: Measured power under full load (*inductive clamp mtr)*
  - Best: Measured Power running LINPACK *(realtime measure)*
- Over time, we will be able to determine if we are doing better or worse on these metrics
- Check out http://www.green500.org/ !!!

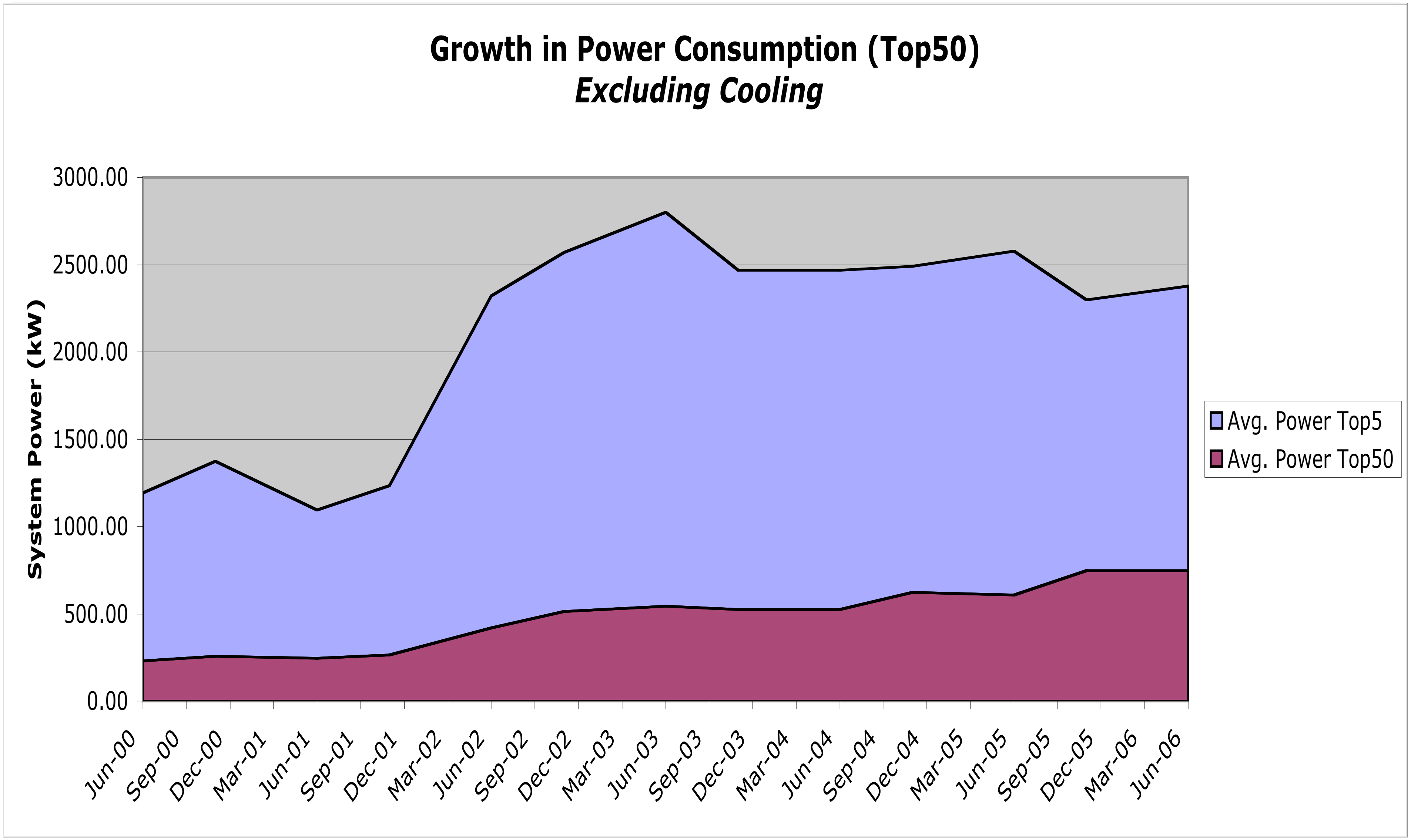
Power Efficiency of Top50 for 5 years



Good News
(Power Efficiency is Still Improving)

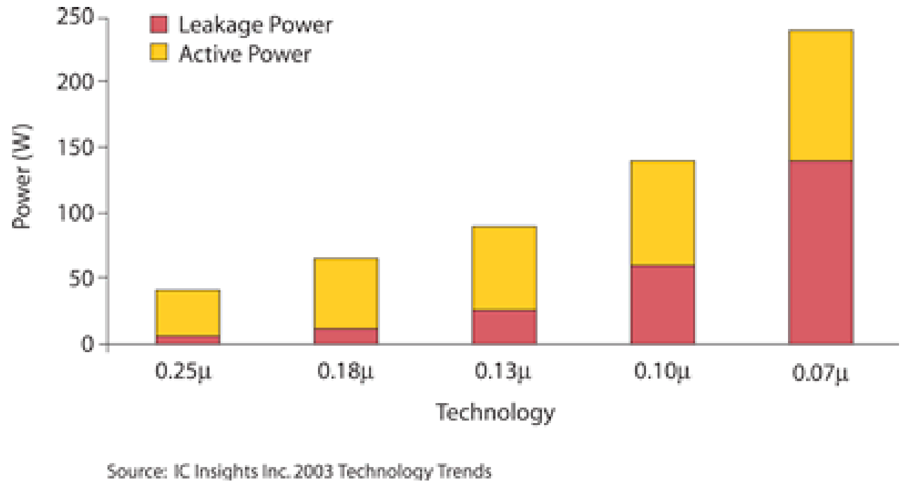


Bad News
(Power Requirements are Growing)



Bonus Material on Power Trends From IBM Journal of Research

http://www.research.ibm.com/journal/rd/504/haensch.html



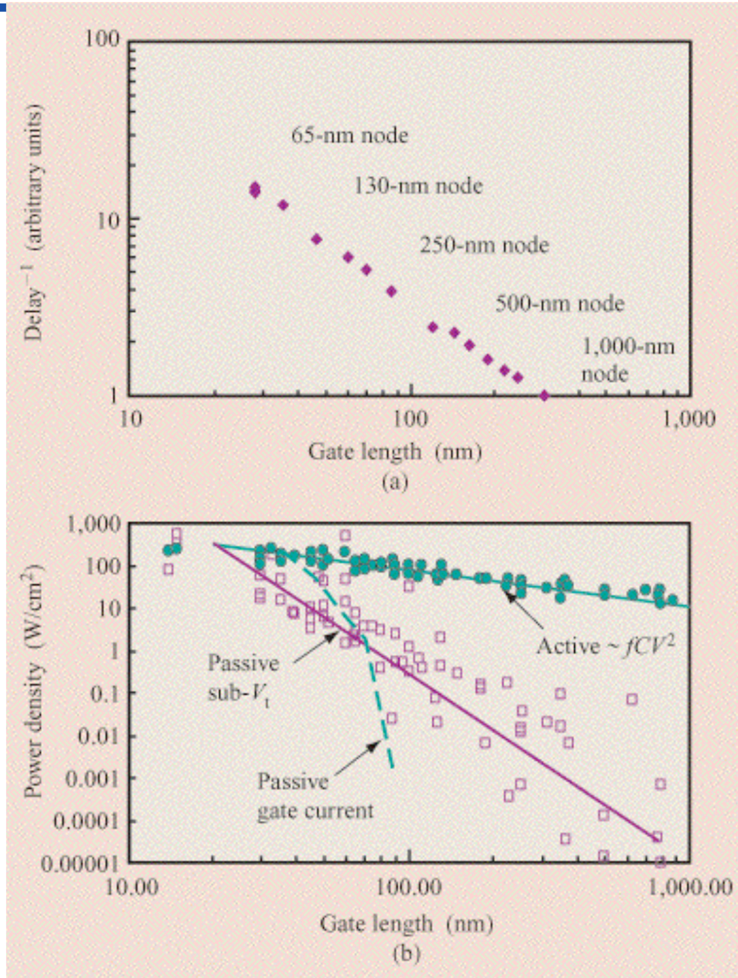


Figure 1

(a) MOSFET performance vs. gate length; normalized MOSFET intrinsic device delay (CV/I_{eff}) vs. gate length. (b) Power density vs. gate length; data collected from literature for active power density and passive power density. Lines are intended to show trend. (fCV^2 = frequency × capacitance × voltage^2.)