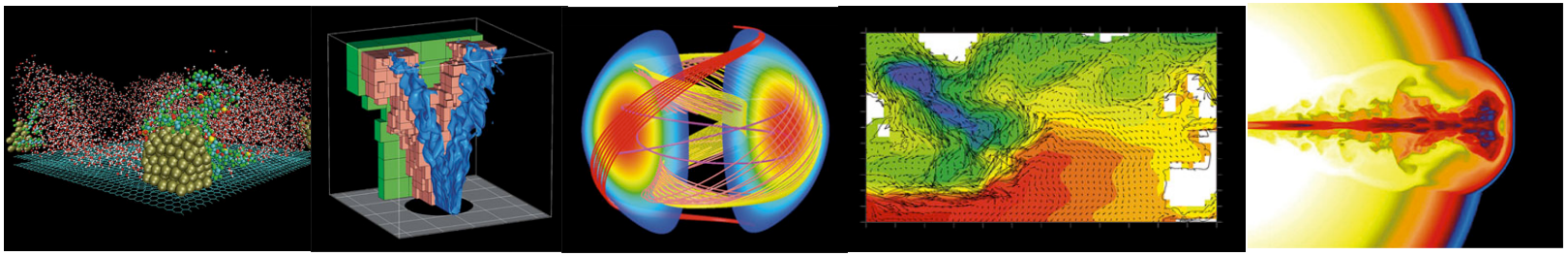


NERSC Power Efficiency Analysis

John Shalf
SDSA Team Leader
jshalf@lbl.gov

NERSC User Group Meeting
September 17, 2007

BERKELEY LAB

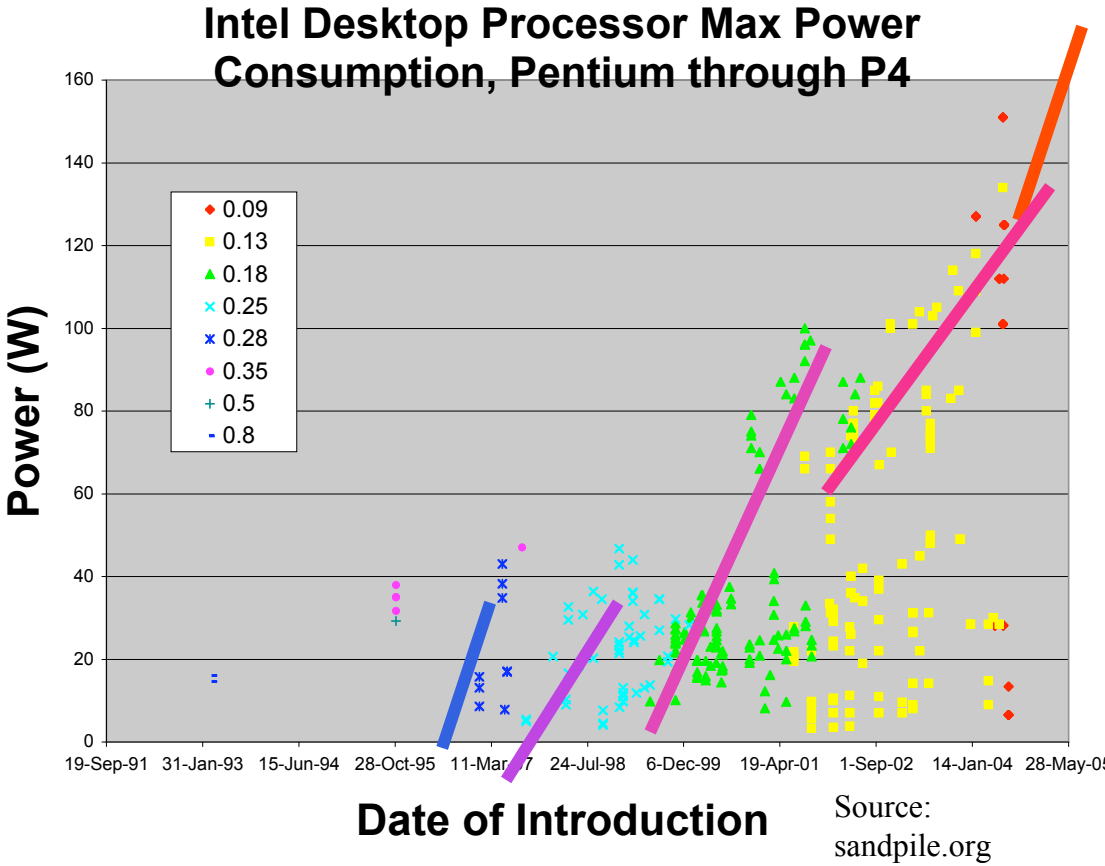
NERSC

Office of Science
U.S. DEPARTMENT OF ENERGY

Microprocessors: Up Against the Wall(s)

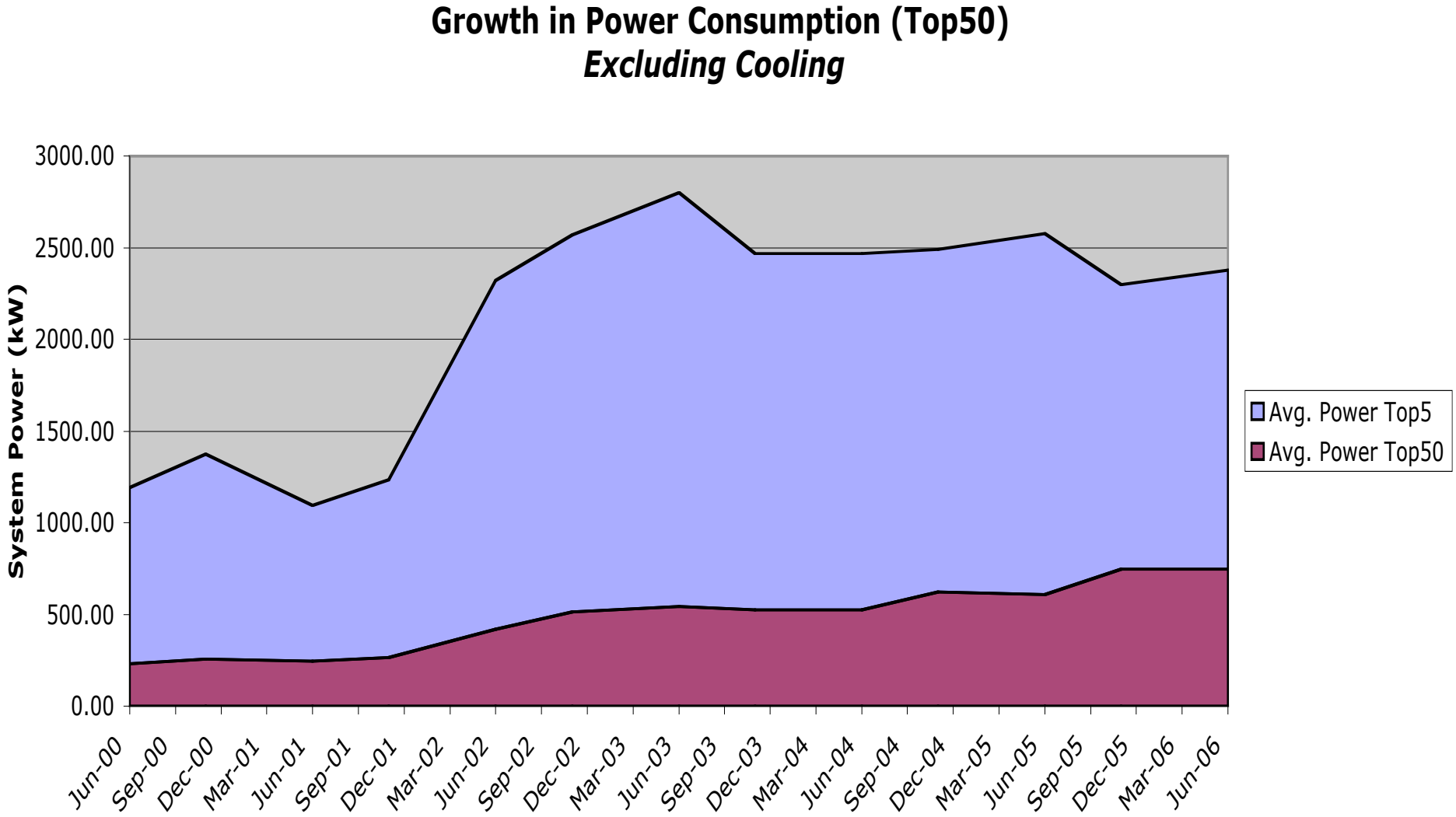
From Joe Gebis

- 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



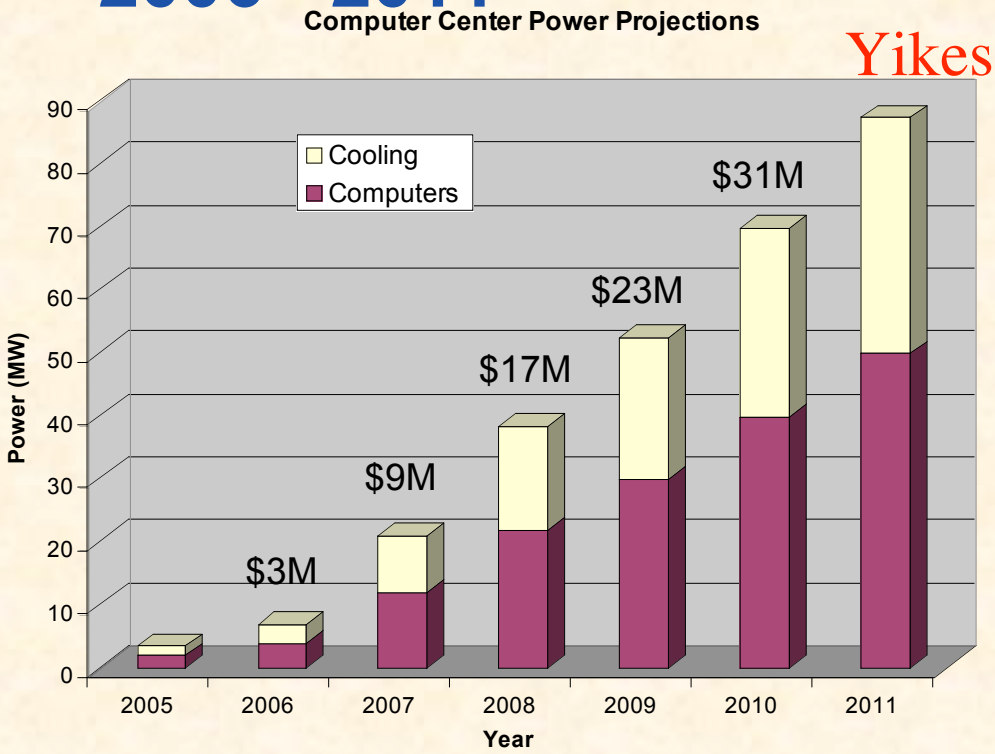
Source: sandpile.org

HPC Power Draw on the Rise



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.

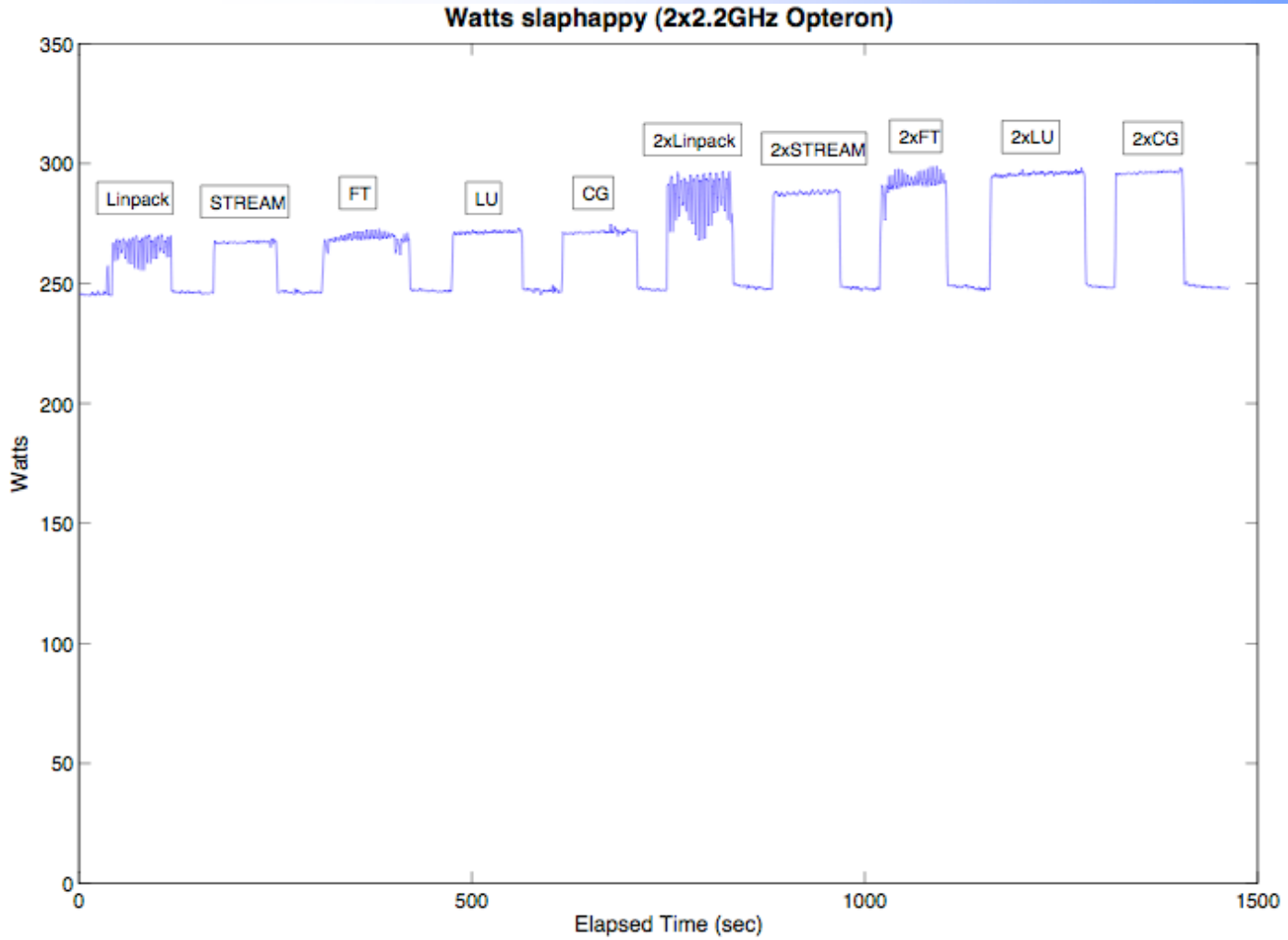
Need Power Efficiency Metrics based on Effective Performance

- 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

Broad Objective for Top500

- 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

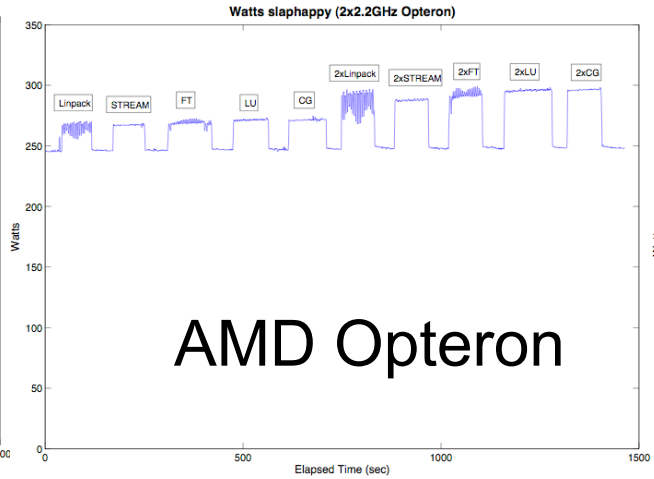
Single Node Tests: AMD Opteron



- **Highest power usage is 2x NAS FT and LU**

Similar Results when Testing Other CPU Architectures

Core Duo

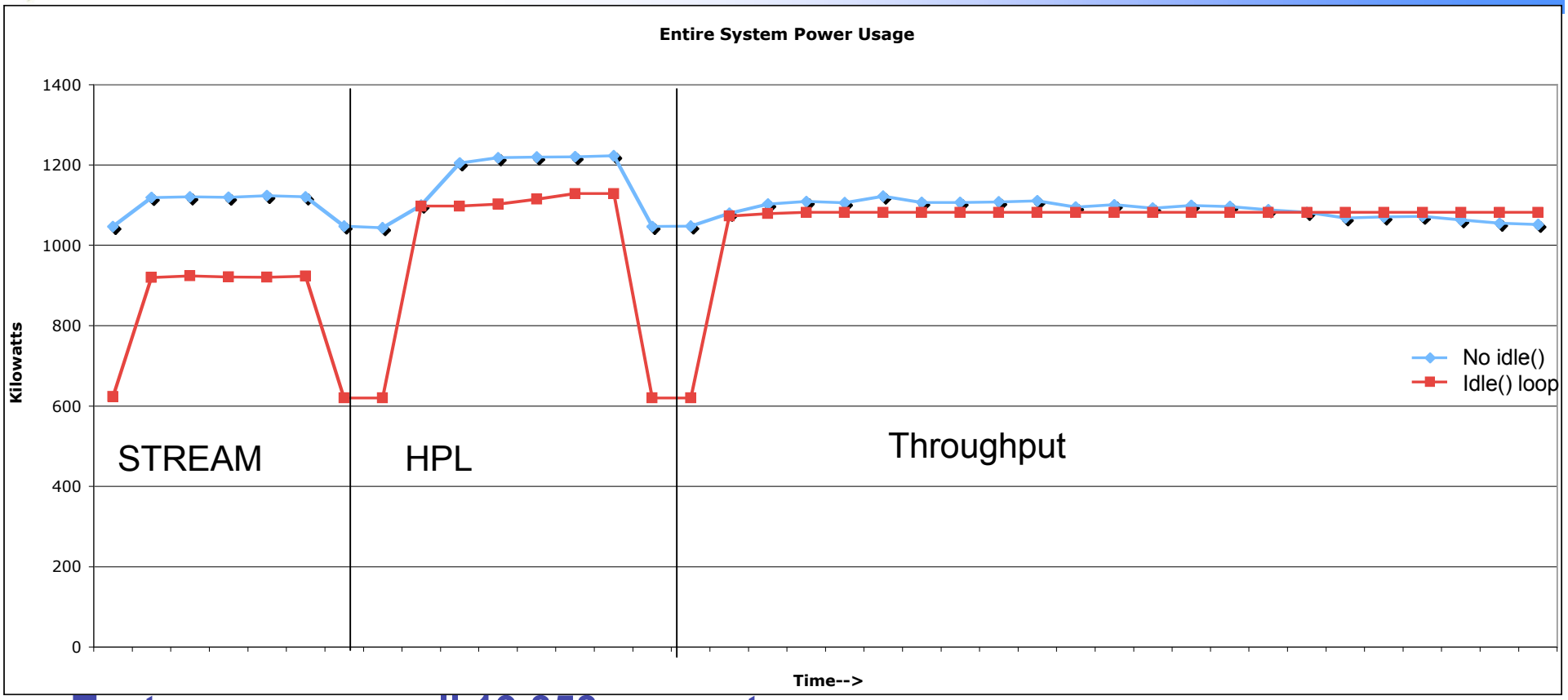


AMD Opteron

IBM PPC970/G5

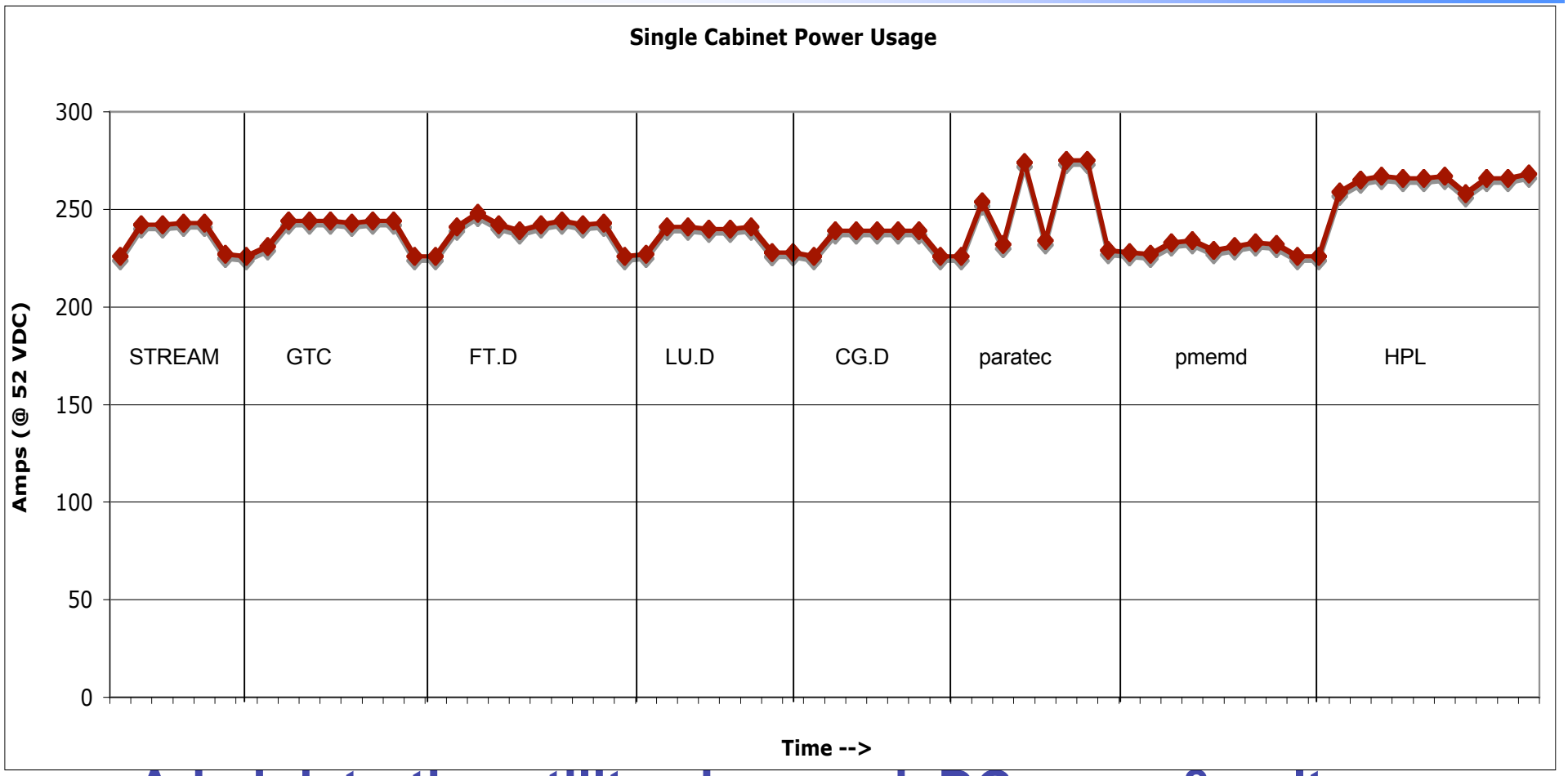
- **Power consumption far less than manufacturer' estimated "nameplate power"**
- **Idle power much lower than active power**
- **Power consumption when running LINPACK is very close to power consumed when running other compute intensive applications**

Full System Test



- 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

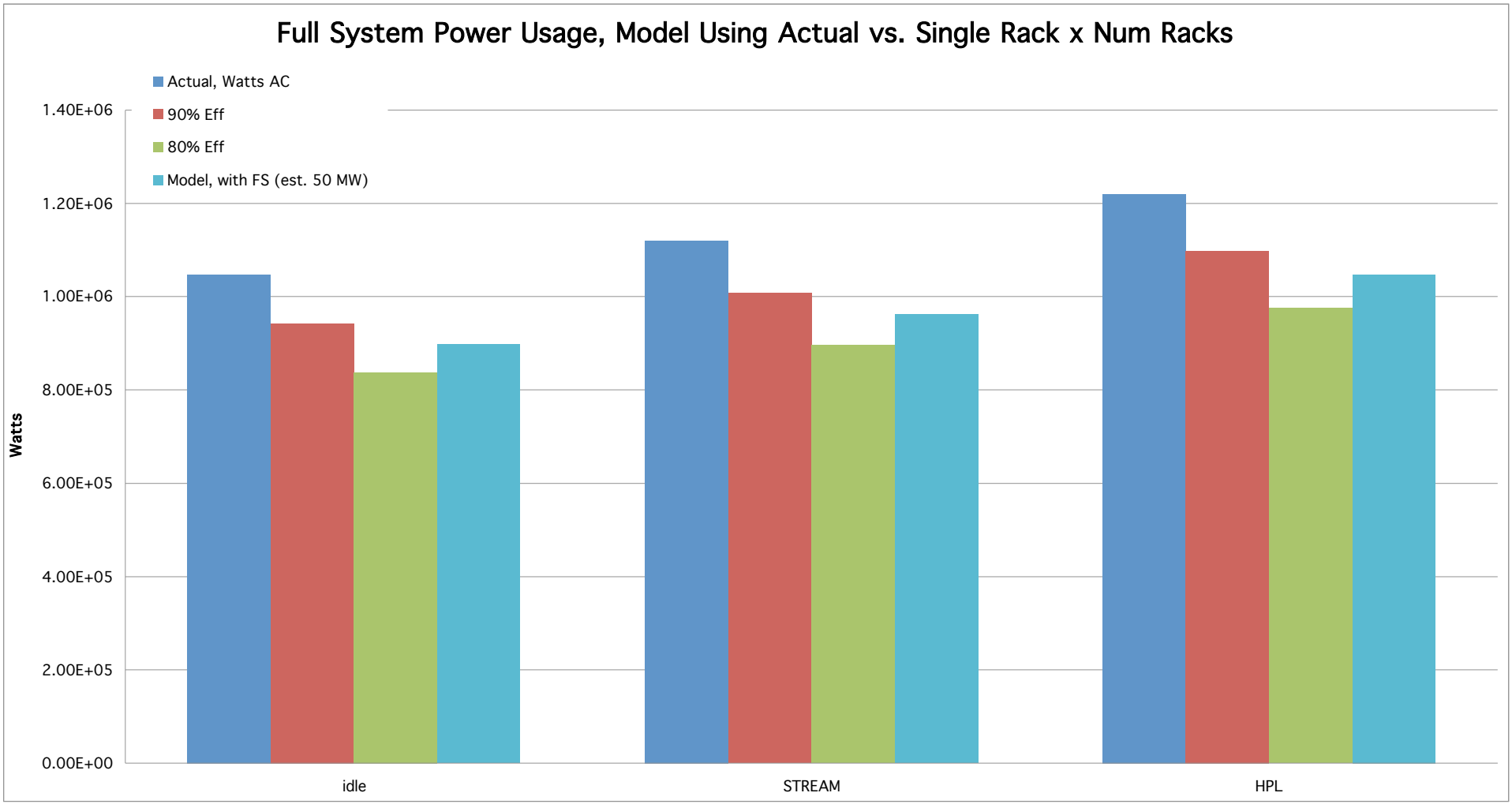


- Administrative utility gives rack DC amps & voltage
- HPL & Paratec are highest power usage

Modeling the Entire System: Disks

- Must take into account disk subsystem
- Drive model *matters*
  - Deskstar 9.6W idle, 13.6W under load
  - Tonka 7.4W idle, 12.6W under load
- Using DDN-provided numbers, estimated power draw for model disk subsystem is 50KW idle, 60KW active
- Observed using PDU panel: ~48KW idle

Modeling the Entire System *(projecting from single cabinet)*



- **Error factor is 0.05 if we assume 90% efficiency**

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 approxmimately 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!
- Disk subsystem is a small fraction of overall power (50-60KW vs 1,200 KW)
  - Disk power dominated by spindles and power supplies
  - Idle power for disks not significantly different from active power