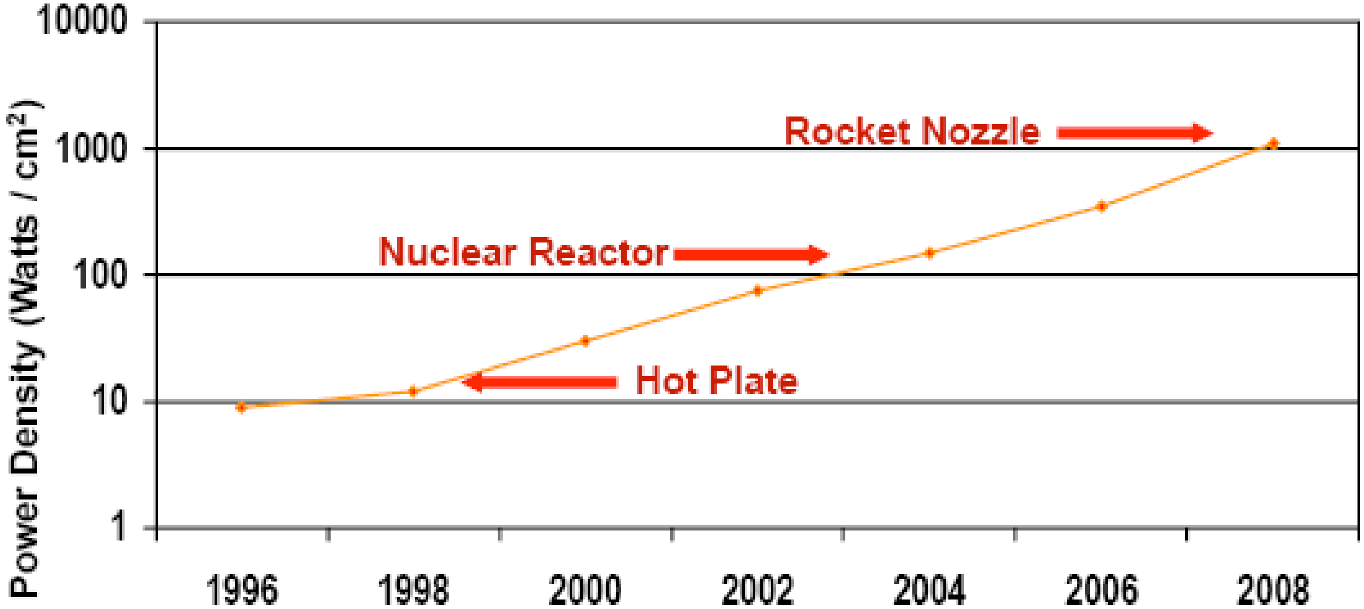
# Power Efficiency Metrics for the Top500

Shoaib Kamil and John Shalf
CRD/NERSC
Lawrence Berkeley National Lab

# Power for Single Processors

## Moore's Law Extrapolation:

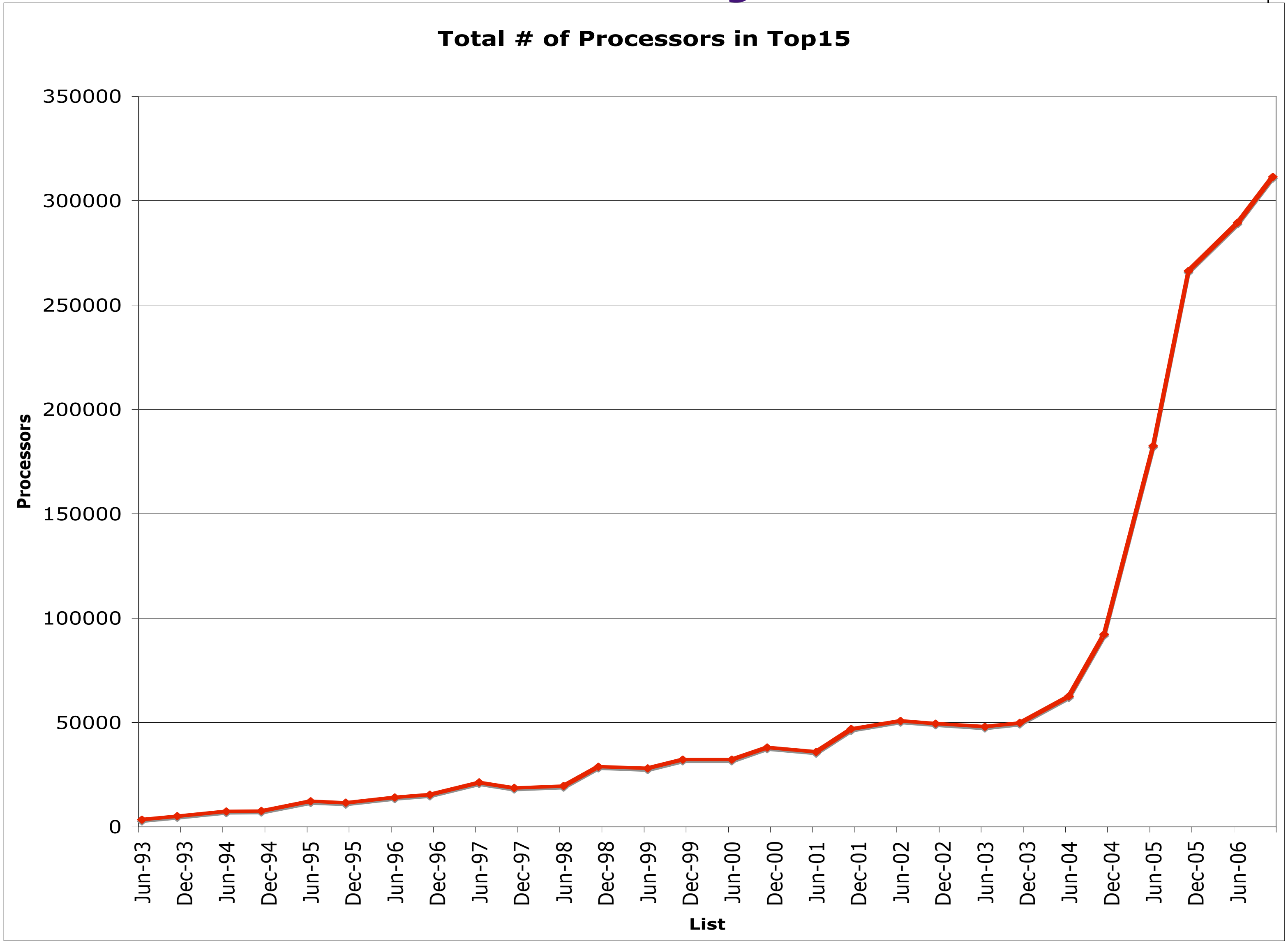
### Power Density for Leading Edge Microprocessors



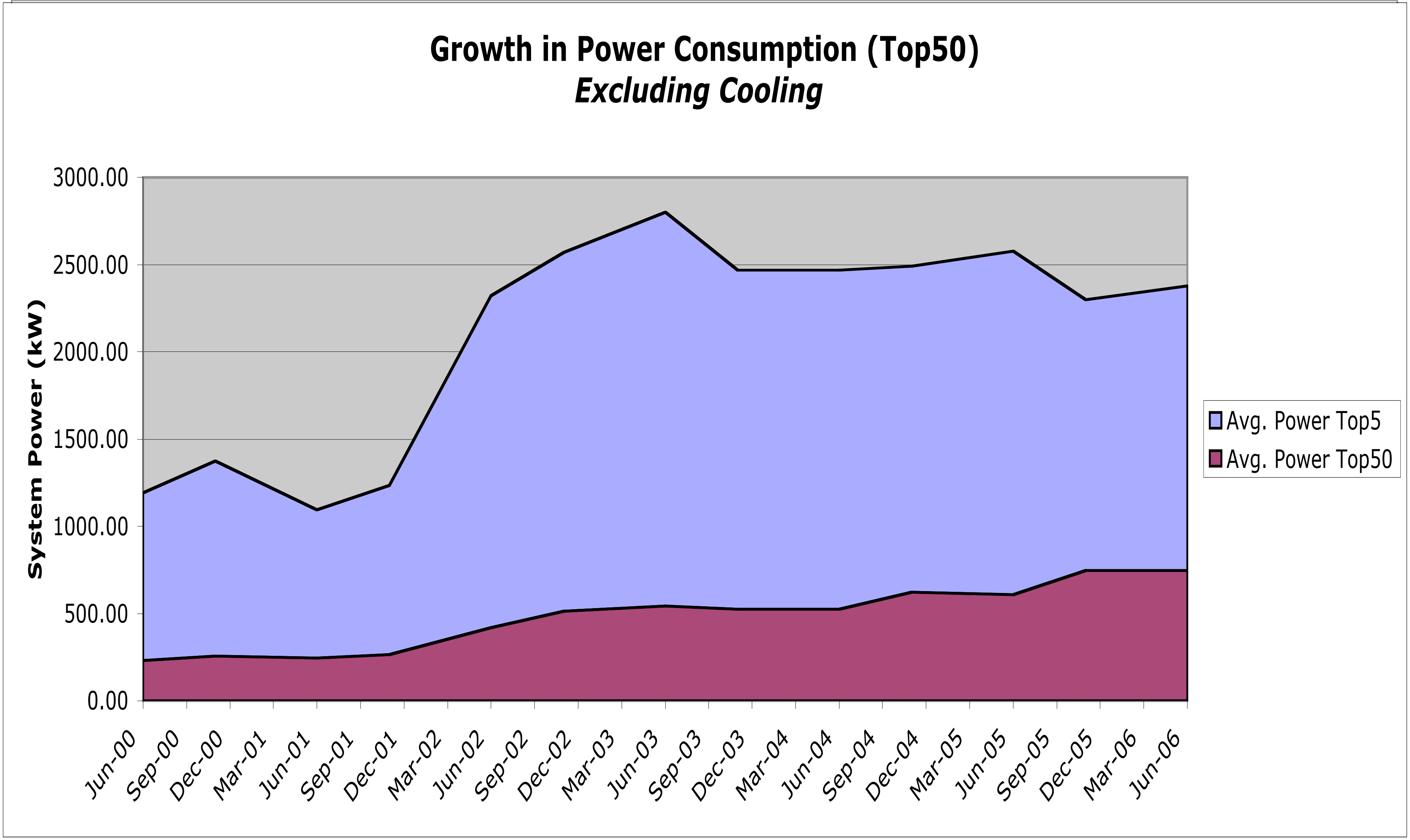
Power Density Becomes Too High to Cool Chips Inexpensively

*Source: Shekhar Borkar, Intel Corp*

# HPC Concurrency on the Rise



# HPC Power Draw on the Rise



# Broad Objective

- 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

# Specific Proposal

- Require all Top500 sites to report system power consumption under a LINPACK workload
  - Establish “rules of engagement” to govern “fair” collection of the data
  - Must establish data collection procedures to make the data collection easy and have little impact on center operations
- We wish to convince you that this can be done with minimal pain
  - If it cannot, we want to hear your input so that the rules can be drafted in a way that will work for ALL Top500 respondents

# What do we mean by “Easy”

- We will try to prove to you that one can measure from a single node or cabinet and project the power consumption for the overall system
- At cabinet and node level,
  - You do not have to take your entire system out of service to measure power under LINPACK
  - sample a few representative pieces running proportionally smaller copies of LINPACK and project it to the full system scale
  - Can use simpler/less-expensive power measurement apparatus

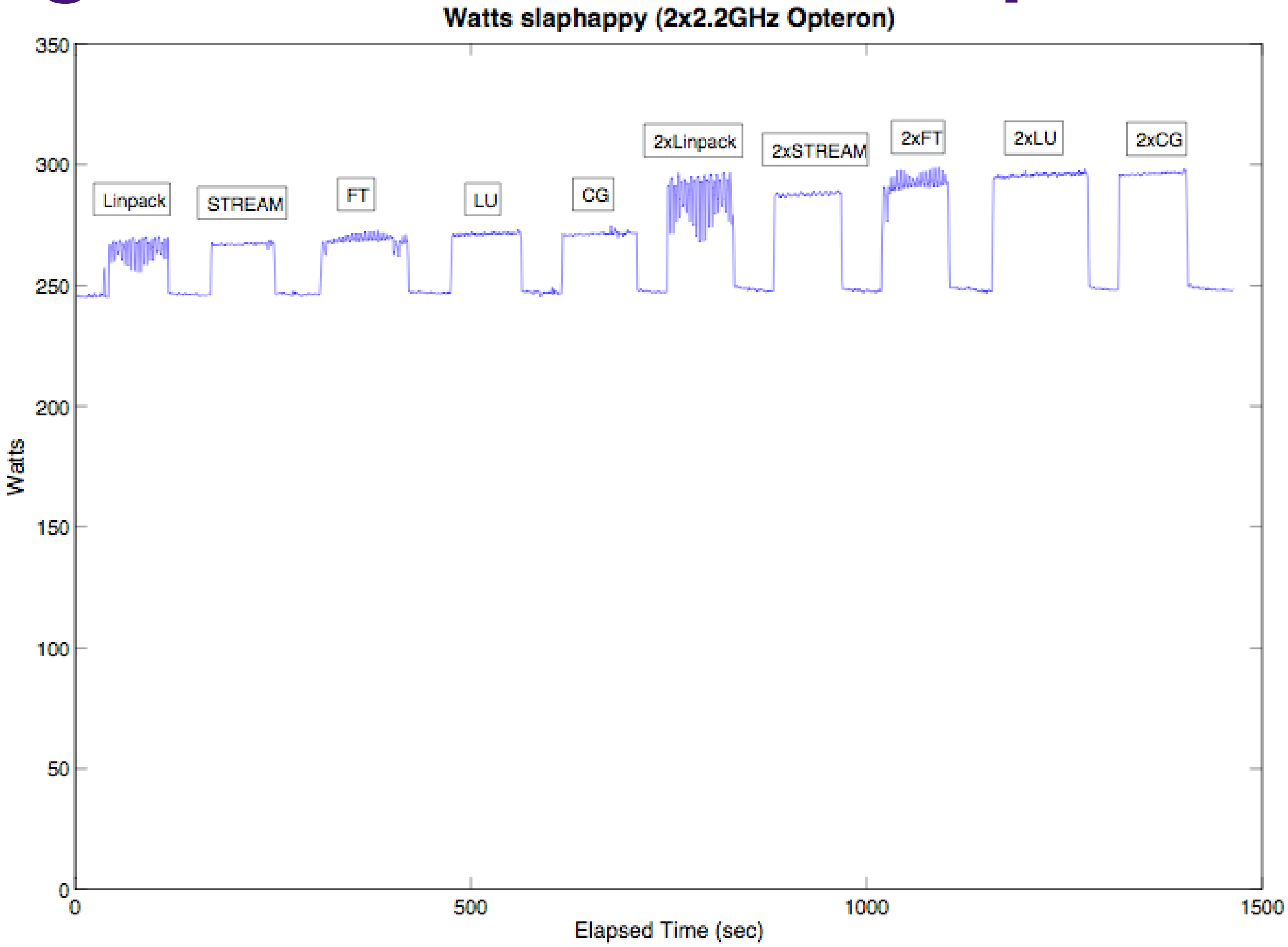
# Many Ways to Measure Power

- 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

# Testing our Methodology

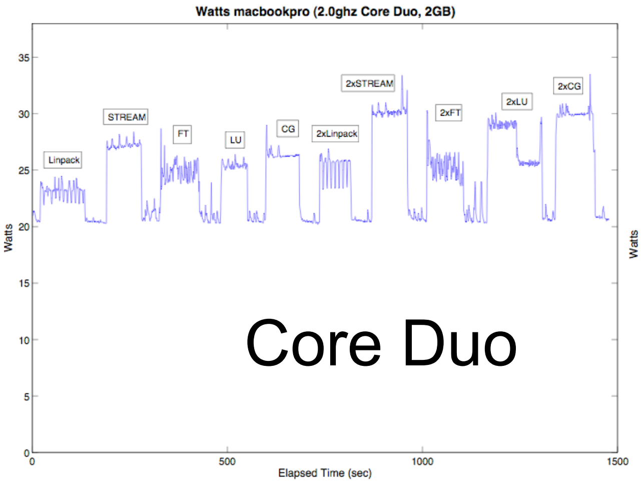
- Look at power usage using variety of synthetic and real benchmarks
  - Memory intensive : STREAM
  - CPU intensive: HPL/Linpack
  - IO intensive: IOZone, MADbench
  - Simulated workloads: NAS PB, NERSC SSP
- Compare single node vs cabinet/cluster vs entire system
  - Is power consumed when running LINPACK similar to that of a real workload?
  - Does power consumed by LINPACK change with concurrency?
  - Can we predict full system power from cabinet/node power?

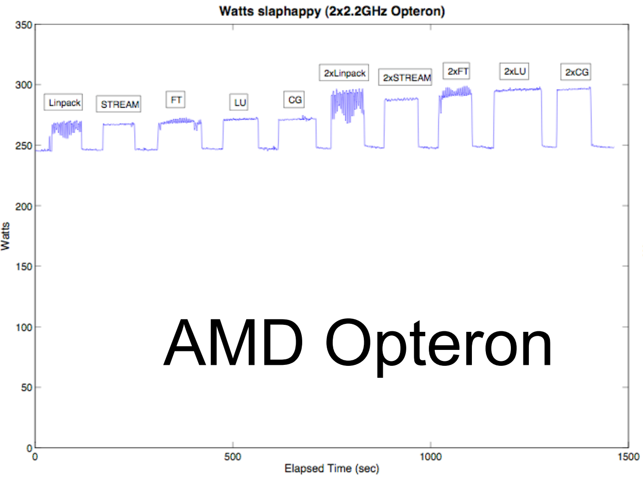
# Single Node Tests: AMD Opteron

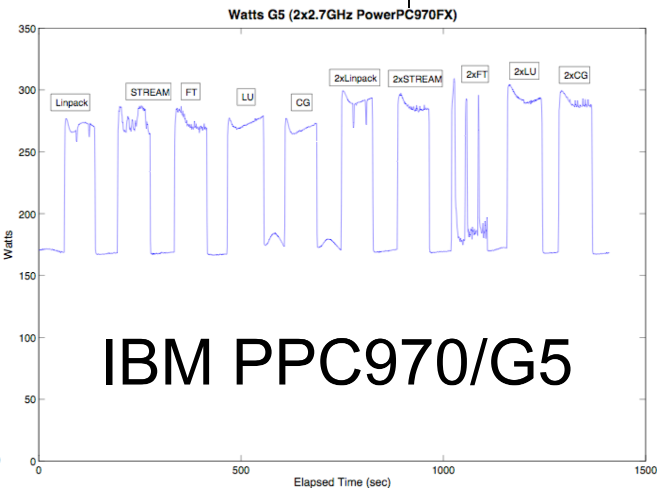


- Highest power usage is 2x NAS FT and LU

# Similar Results when Testing Other CPU Architectures

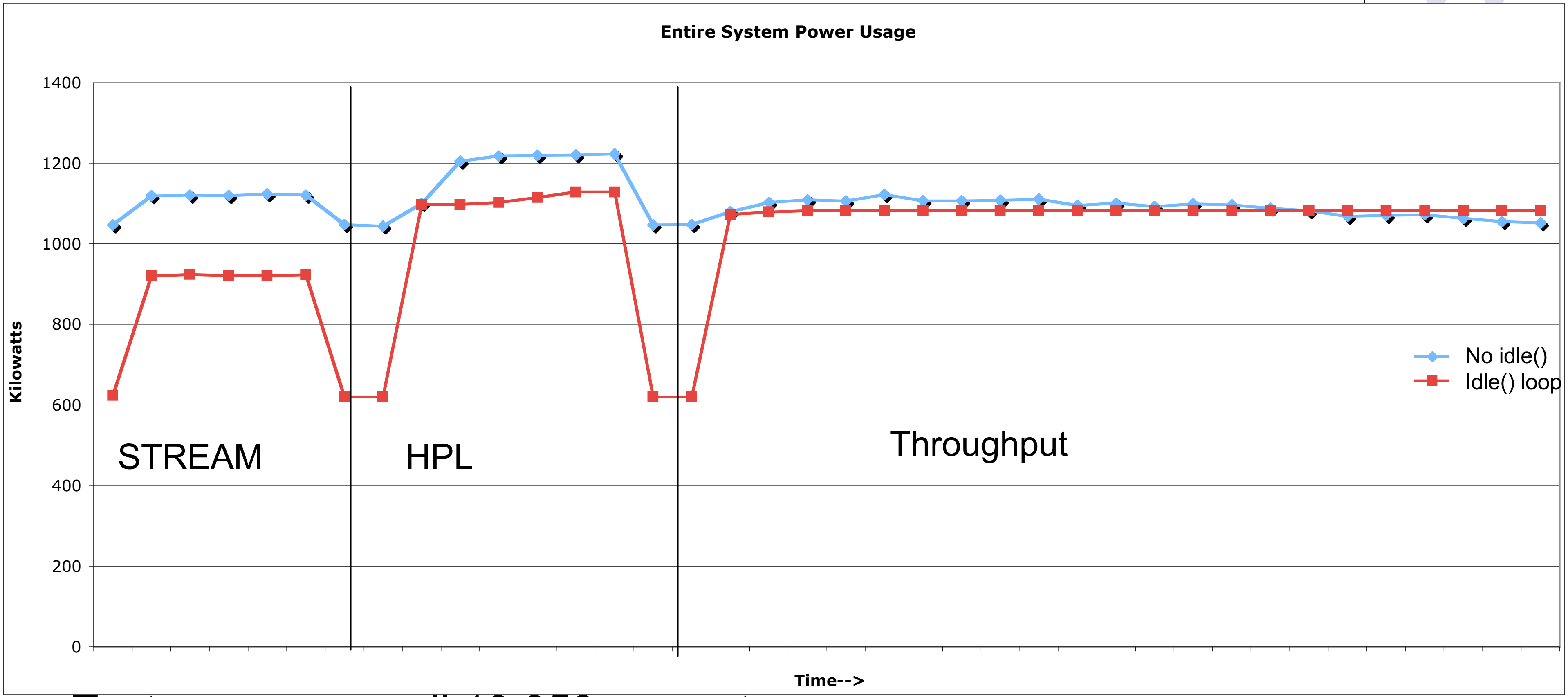






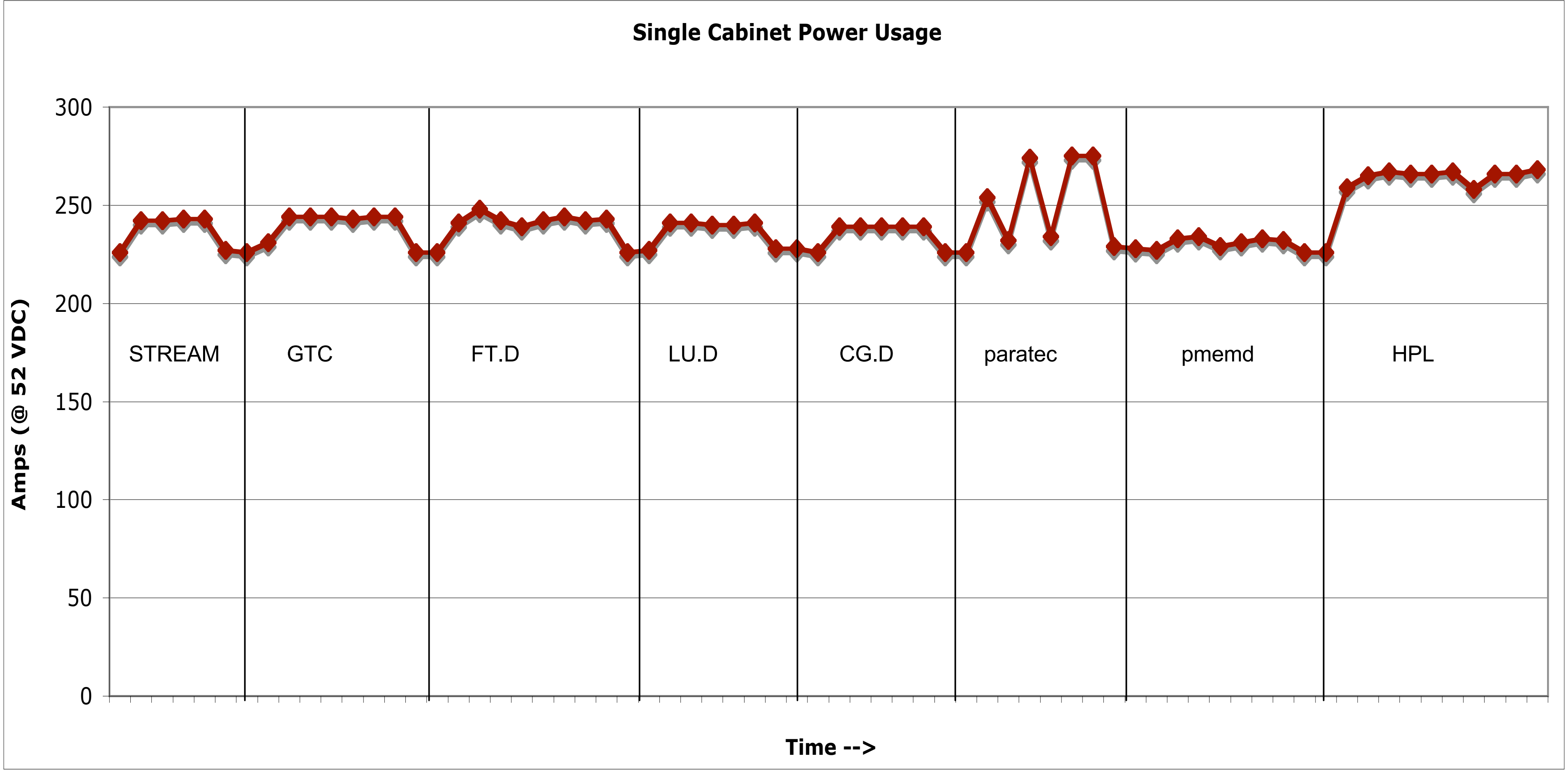
- Power consumption far less than manufacturer' estimated "nameplate power"
- Idle power much lower than active power
- Power consumption when running LINPACK is very close to power consumed when running other compute intensive applications

# Full System Test



- 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

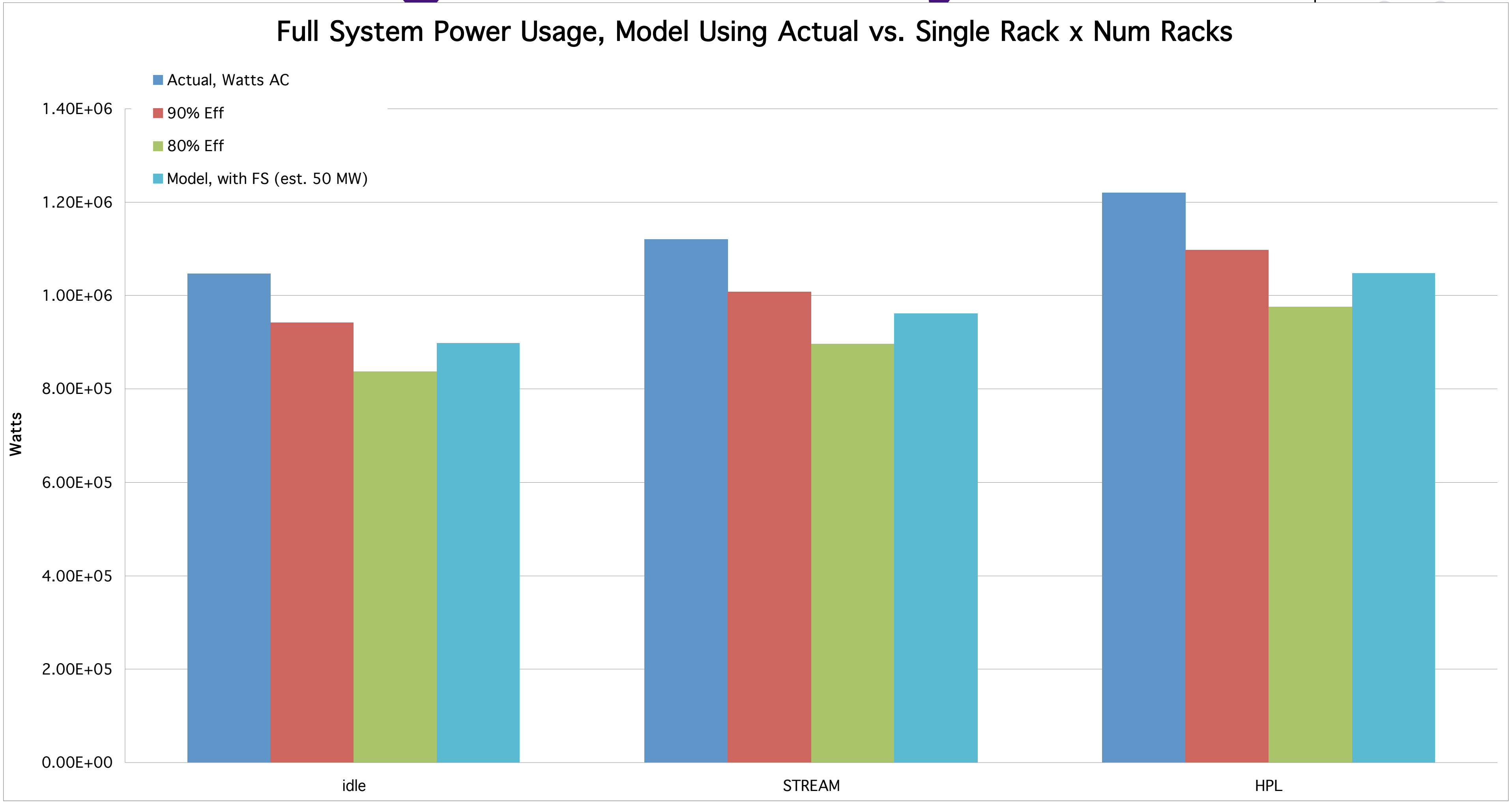


- Administrative utility gives rack DC amps & voltage
- HPL & Paratec are highest power usage

# Modeling the Entire System: AC to DC Conversion

- Commodity desktop machines are ~75% efficient
- Google uses new, 90% efficient power supplies
- Our test system has has ~90% efficient power supplies

# Modeling the Entire System



- 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

# Top500 Power Data Collection

- Measure System Power when running LINPACK
  - If measured on circuit for full system (eg. PDU or Panel), then run LINPACK on full system
  - If cannot differentiate system from other devices sharing same circuit, then isolate components using line meter or inductive clamp meter (can borrow from local Power Co.)
  - If measured from line meter or clamp meter, then just run on one rack, measure representative components comprising system and extrapolate to entire system
  - If measured from integral power supply, account for power supply losses in projections
- Target of projections is RMS AC “wall-socket power” consumed by HPC system
  - Must convert measurements of DC power consumption accordingly

# Top500 Power Data Collection

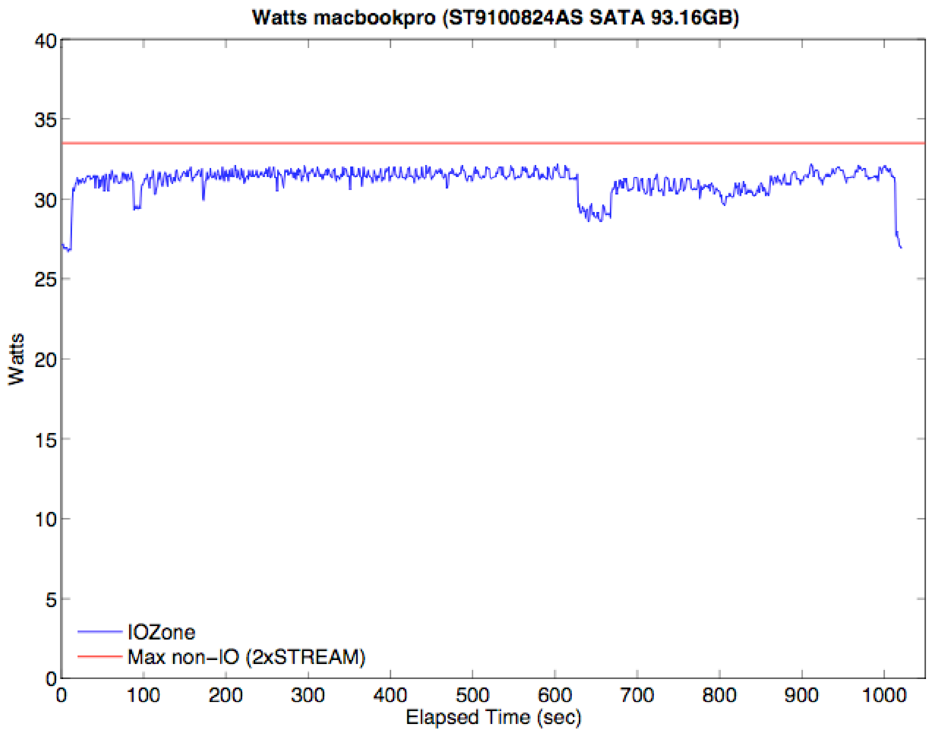
- What to include
  - All components comprising delivered HPC system aside from external disk subsystem (eg. SAN)
  - Can extrapolate from measurement of said components while under a LINPACK load (even if load is local)
- What to exclude
  - Exclude cooling
  - Exclude PDU and other power conversion infrastructure losses that are not part of the deliverd HPC system
  - Exclude disk subsystem (if not integral): *should discuss this further*

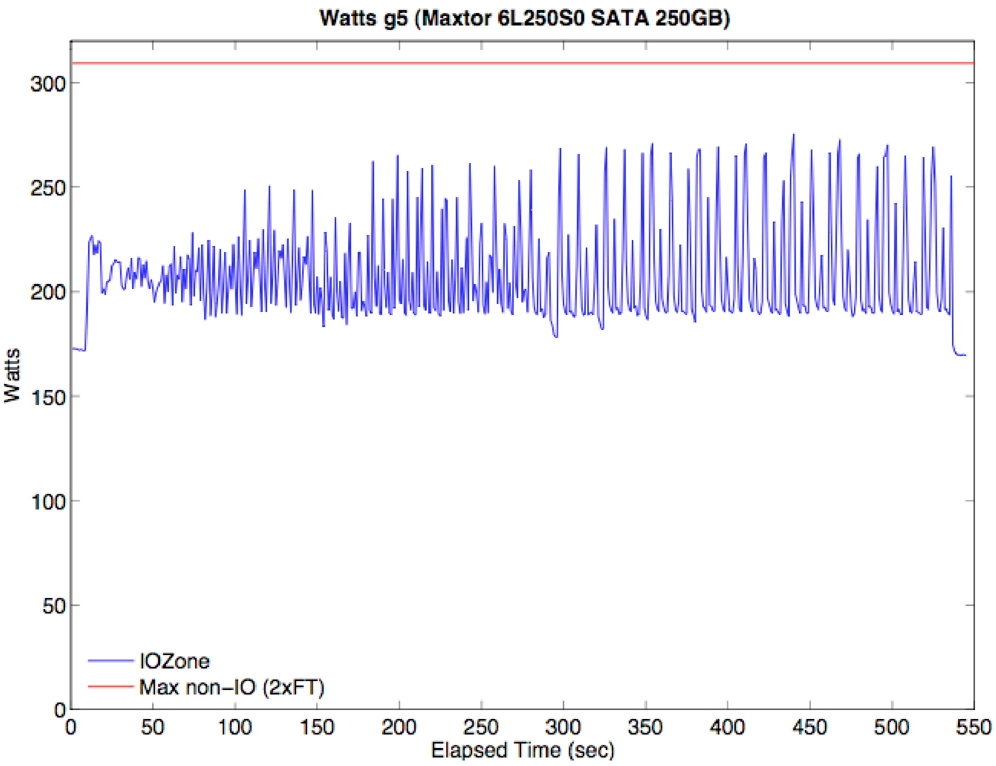
# Complementary Efforts

- Our Effort creates a metric for compute-intensive parallel scientific workloads

- Metrics for I/O intensive workloads: JouleSort by HP Labs
- Metrics for transactional workloads: EPA EnergyStar Server Metrics
- Re-ranking of Top500 for Power Efficiency: http://www.green500.org/

# Single Node Tests: IO





- Highly variable, less than compute-only
- Very difficult to assess power draw for I/O

# 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