



# Benchmarking for Power and Performance

**Heather Hanson (UT-Austin)**  
**Karthick Rajamani (IBM/ARL)**  
**Juan Rubio (IBM/ARL)**  
**Soraya Ghiasi (IBM/ARL)**  
**Freeman Rawson (IBM/ARL)**

# The Future

## UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

WASHINGTON, D.C. 20460

OFFICE OF AIR AND RADIATION

December 28, 2006

Dear Enterprise Server Manufacturer or Other Interested Stakeholder,

*“The purpose of this letter is to inform you that the U.S. Environmental Protection Agency (EPA) is initiating its process to develop an ENERGY STAR specification for enterprise computer servers. In the coming months, EPA will conduct an analysis to determine whether such a specification for servers is viable given current market dynamics, the availability and performance of energy-efficient designs, and the potential energy savings...”*

# Current EPA protocol for power/performance benchmarking

## ■ EPA Server Energy Measurement Protocol

- <http://www.energystar.gov/ia/products/d>
- Recommendation is that system vendors provide curves showing power consumption under different loads
  - Run at maximum load (100%)
  - Repeat with runs at reduced loads, until load reaches 0%
- Expect consumers to use this curve to estimate their own overall energy consumption
  - Multiply average utilization with appropriate point on the curve to get costs

## ■ Considerations

- End-to-end numbers can give the wrong results
- Averaged utilization doesn't necessarily correlate to average power
  - Distribution of utilization?
  - When does utilization peak?
  - Which power management techniques are used?

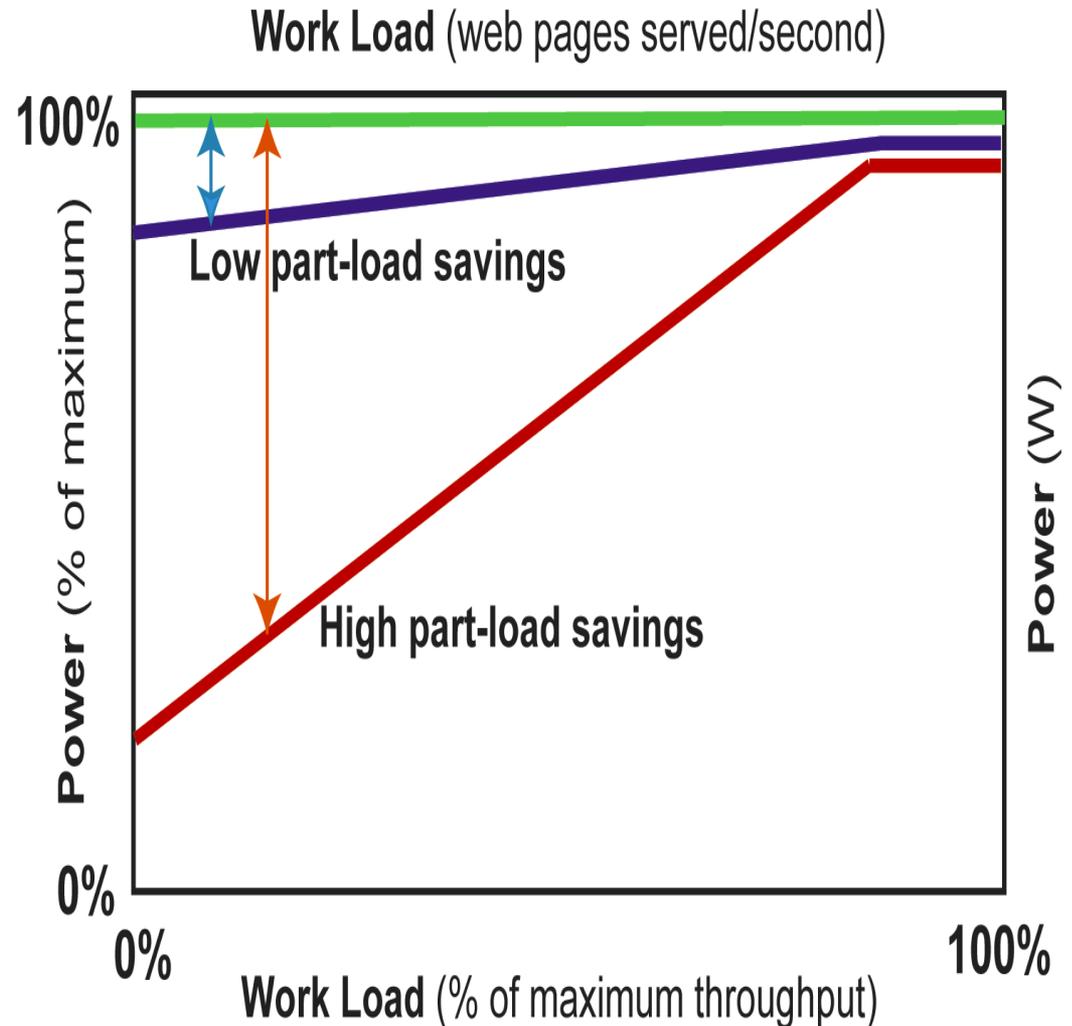


Figure1 from EPA Server Energy Measurement Protocol document

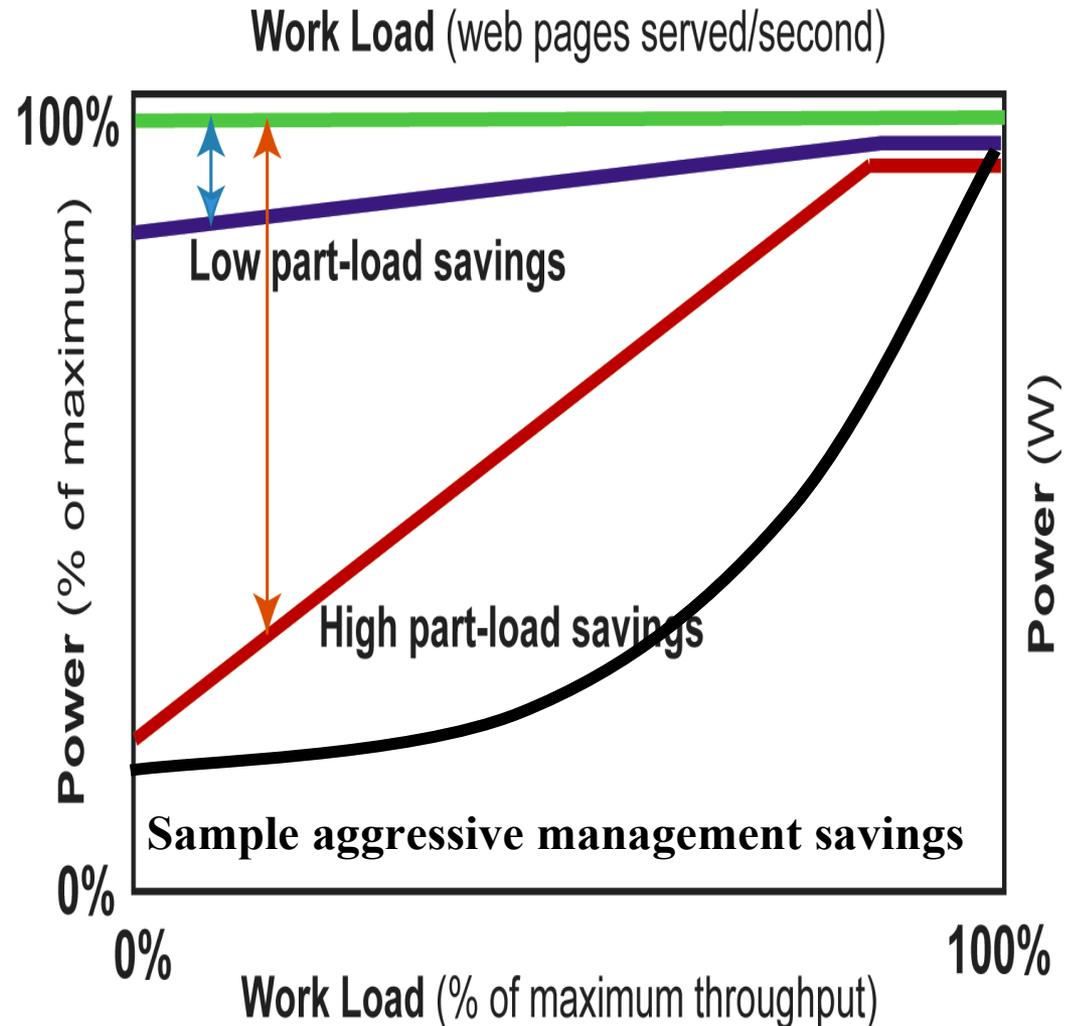
# Current EPA protocol for power/performance benchmarking

## ■ EPA Server Energy Measurement Protocol

- <http://www.energystar.gov/ia/products/d>
- Recommendation is that system vendors provide curves showing power consumption under different loads
  - Run at maximum load (100%)
  - Repeat with runs at reduced loads, until load reaches 0%
- Expect consumers to use this curve to estimate their own overall energy consumption
  - Multiply average utilization with appropriate point on the curve to get costs

## ■ Considerations

- End-to-end numbers can give the wrong results
- Averaged utilization doesn't necessarily correlate to average power
  - Distribution of utilization?
  - When does utilization peak?
  - Which power management techniques are used?



Modified to show the behavior of a power-managed system

# Overview

- **Power and thermal problems in computer systems are becoming more common.**
  - Power and thermal management techniques have significant implications for system performance.
- **Researchers rely on benchmarks to develop models of system behavior and experimentally evaluate new ideas.**
- **Recent EPA announcement and SPEC OSG activities add urgency to resolving the power/performance benchmarking issues.**
- **We present our experiences with adapting performance benchmarks for use in power/performance research.**
- **We focus on two problems:**
  - Variability and its effect on system power management
  - Collecting correlated power and performance data.
- **Benchmarking for combined power and performance analysis has unique features distinct from traditional performance benchmarking.**

# What should power/performance benchmarks expose?

## ■ Intensity Variation

- Workload variation in system utilization
  - Workloads differ from one another
  - A single workload may vary over time

## ■ Nature of activity variation

- Workload variation in program characteristics
  - A single workload may change how it uses system components over time
  - The rate at which this change occurs also varies over time.

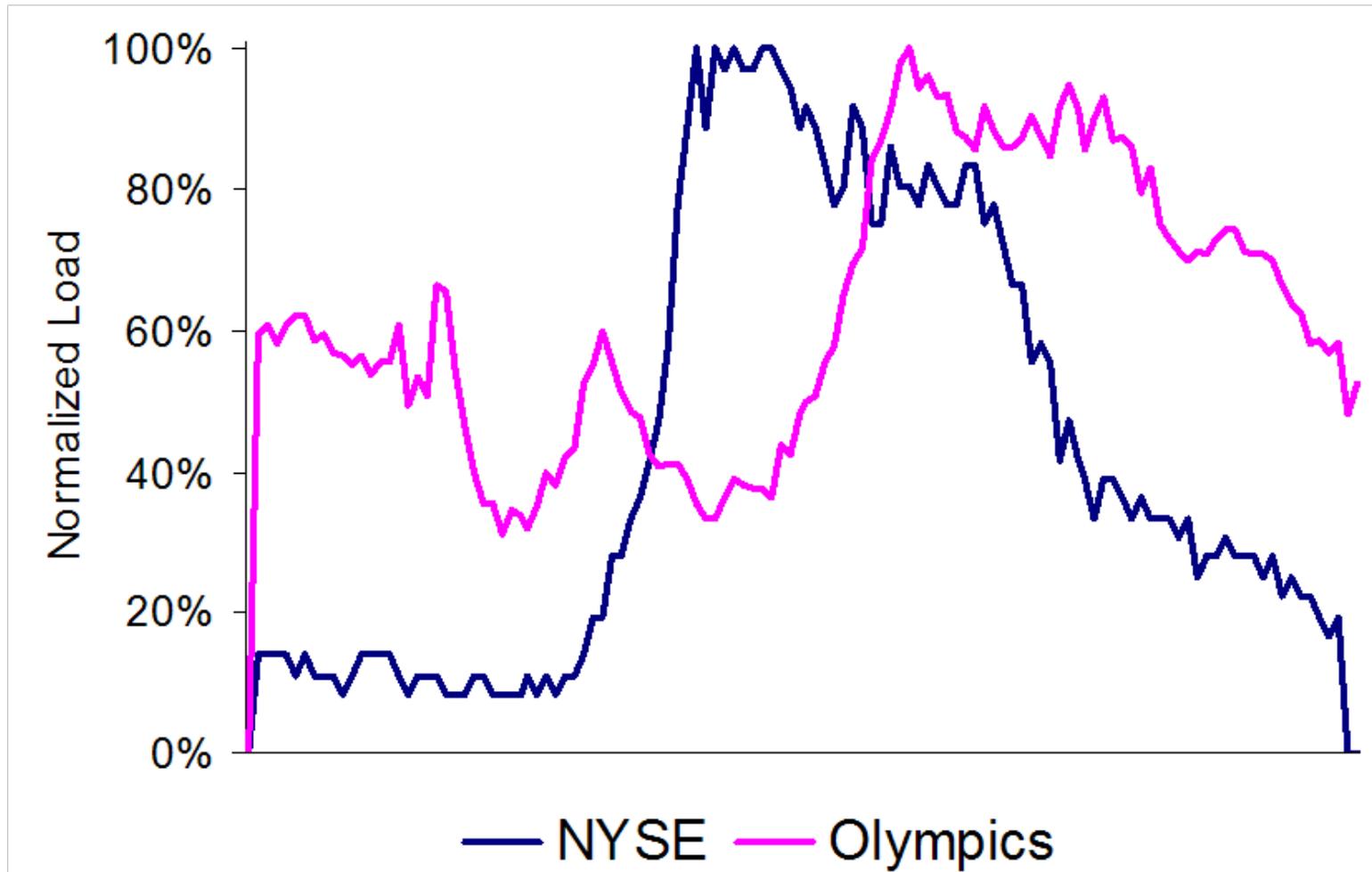
## ■ System response

- How quickly does the system respond to changing characteristics?
- How well does the system respond to changing characteristics?
- Specifically for power-managed systems, how well does the system
  - Meet power constraints
  - Meet performance constraints (response time? throughput?)
  - Manage the power versus performance trade-off
  - Maximize its energy efficiency

## ■ Component-level variation

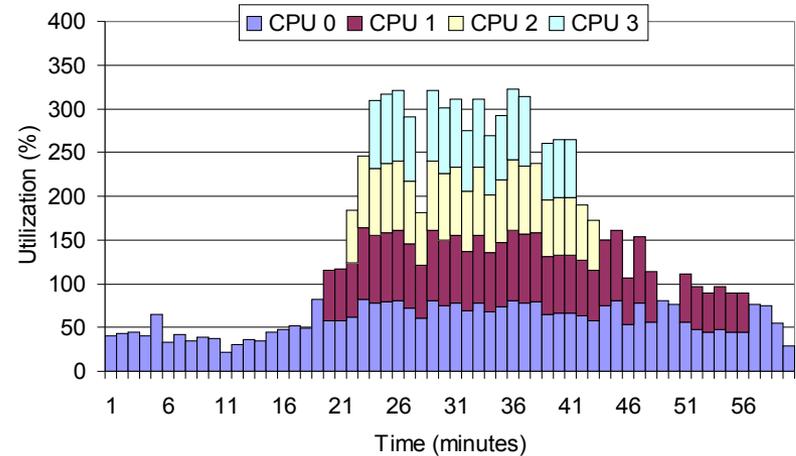
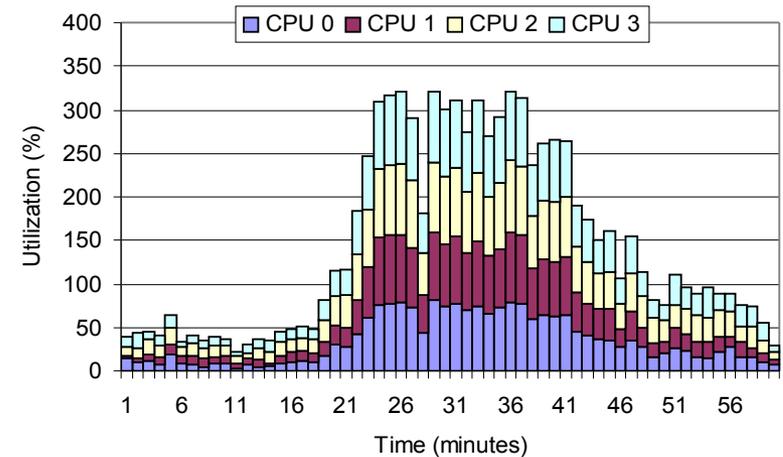
- “Identical” components are not actually identical in power

## Utilization variation



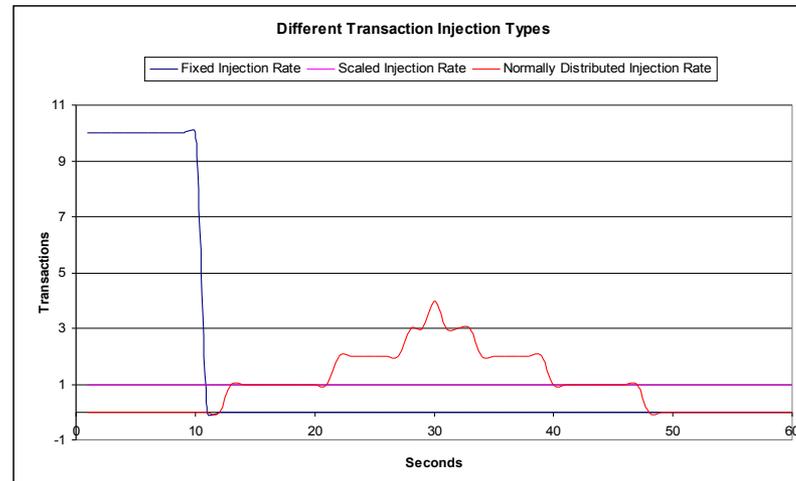
# A closer look at the NYSE trace

- **How can the changes in utilization be exploited to reduce power consumption without harming performance unduly?**
  - DVFS is the most typical solution
    - Lower the frequency and voltage while still meeting the response time criterion
  - Other techniques can be employed instead
    - Throttling
    - CPU Packing with deep power saving
  - Different techniques will give different results depending on the system design and workload
    - DVFS may provide better results for System 1 with Workload A while CPU Packing provides better results for System 2 running Workload B
- **Current EPA protocol does not capture time varying nature of system utilization.**
  - Any new power/performance benchmarks should capture this behavior.
  - **Response to time-varying behavior is a key feature of any power management implementation**



# Cautionary note on scaling transactions for benchmarking

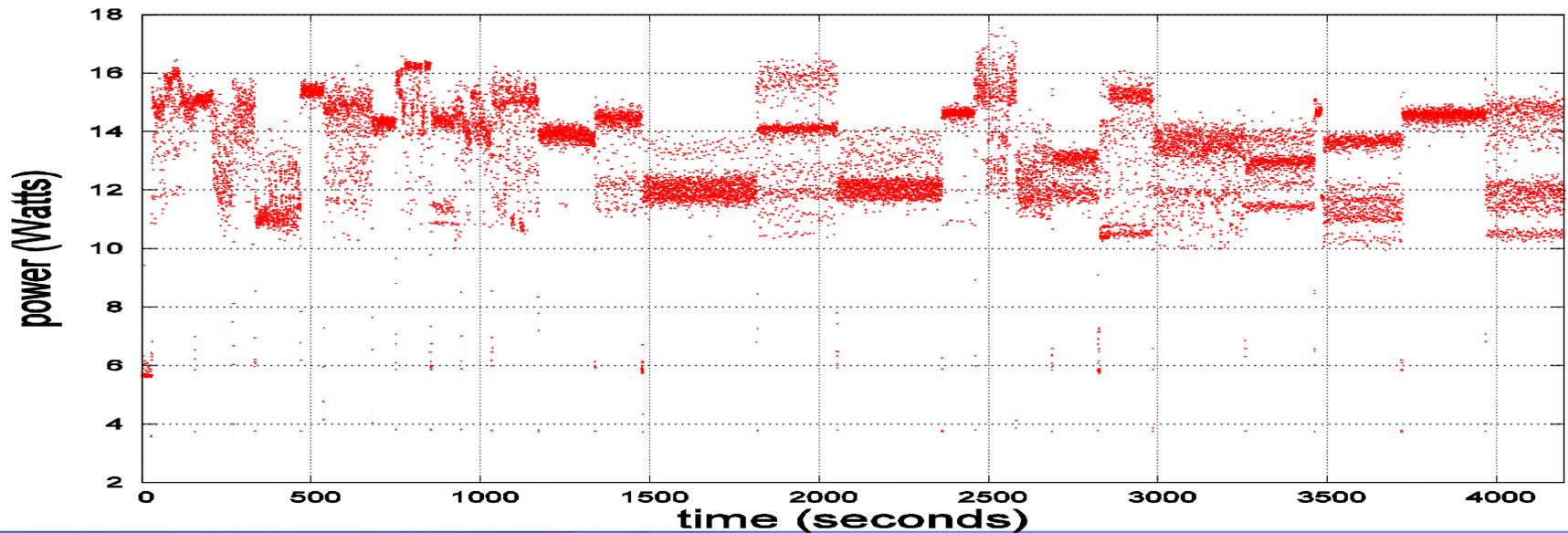
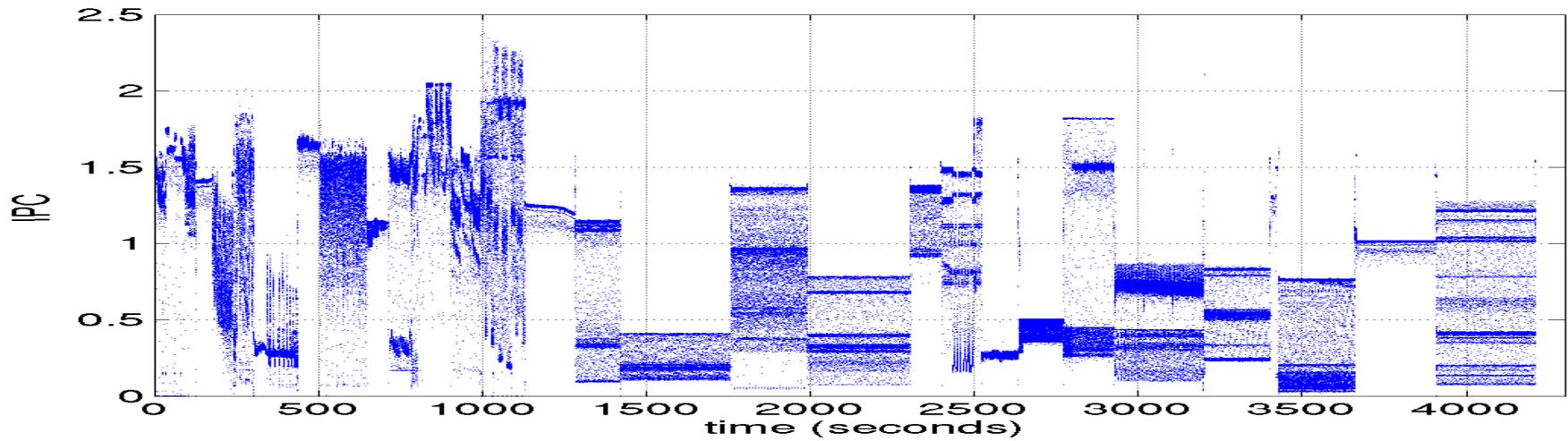
- **Trivial example:**
  - Let max permissible utilization = 85% per processor (340% on previous graph)
    - 600 transactions/minute
    - Approximately 10 transactions a second
  - 10% of the max permissible utilization is 34%
    - Assume this is 60 transactions/minute
    - Many different ways to distribute these over time
- **Different ways of scaling transactions will give different results depending on what the available power management methods are.**
  - Total number of transactions are the same in the 3 cases below.
  - A distribution centered on the average time is probably the most realistic option
- **Different ways of scaling transactions may make it easy to “cheat” on the results.**
- **How the time-varying nature of workloads changes is important because it determines whether or not certain techniques are responsive enough both in the rapidity of response and the range of response.**



## Variation in the nature of a workload's activity

- **What about systems under identical utilization, but running different workloads?**
  
- **How a workload uses systems resources, including the processor, has a significant impact on power consumption.**
  - On the Pentium M, power consumption at the same system utilization can vary by a factor of 2.
  
- **Power variation due to workload will increase**
  - Processors will have more clock gating and will employ other, more aggressive, power savings techniques more extensively.
  - Additional components are adding power reducing techniques such as memory power down, disk idle power reduction and so on

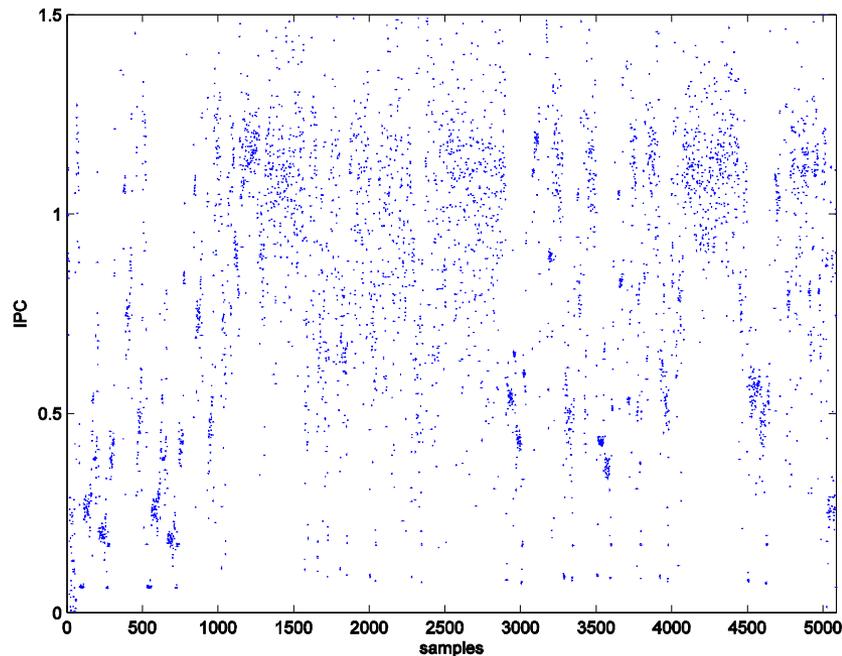
# Variation during SPEC CPU2000 run



# A closer look at gcc and gzip

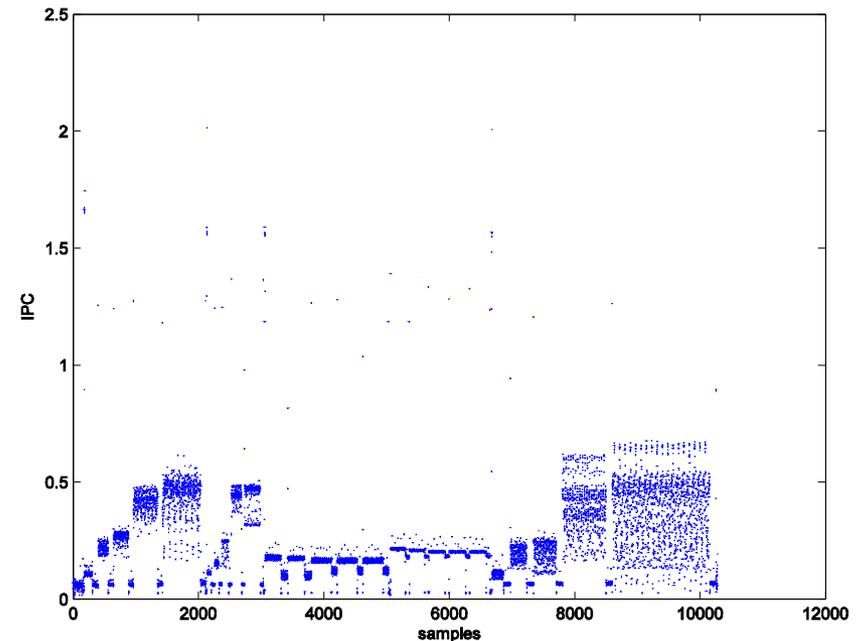
## gcc

- Behavior is nearly chaotic at OS-level time scales
- Power management techniques
  - Must respond on microarchitectural-level time scales



## gzip

- Exhibits a number of different stable phases at OS-level time scales
- Power management techniques
  - Can be slower to respond
  - Can have some associated overhead which gets amortized



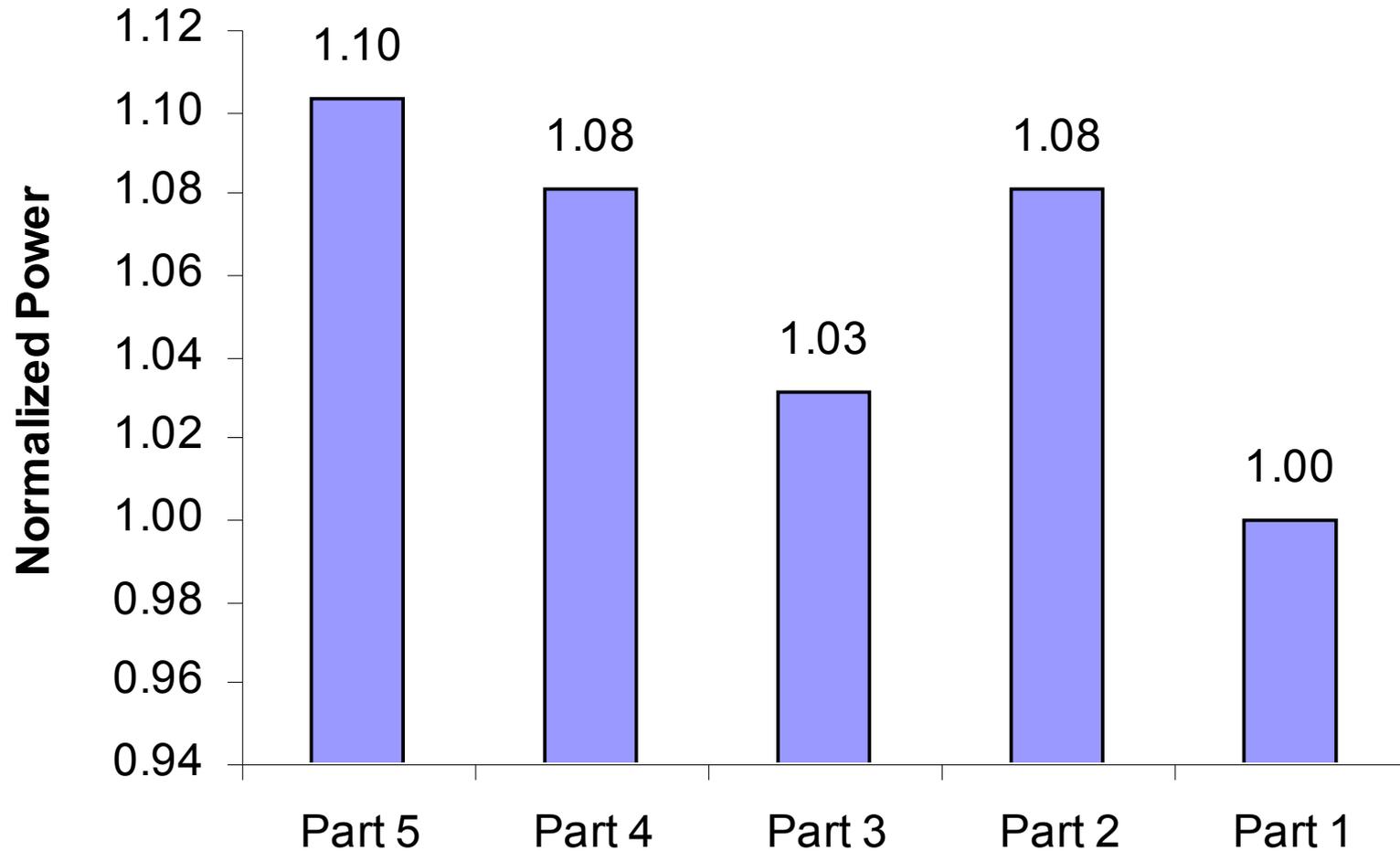
# Why does nature of activity matter?

- **What should be considered here?**
  - What system and processor resources are being used by the workload?
    - Memory bound? CPU-bound with many mispredictions?
  - How rapidly is the behavior of the workload changing?
    - Stable phases? At what time scale?
  - How rapidly can the various power management techniques respond to changing conditions and when?
    - Sometimes “too fast” can be as bad as “too slow”
  
- **Strive to capture realistic temporal variation in workload activity to contrast different power management solutions of different systems**
  
- **Current EPA protocol relies only on level scaling of the throughput of an application relative to its maximum**
  - Ignores differences between and within applications
  - Has no variation in intensity over very long periods of time
  - Can breed dependence on slow-response techniques only
    - DVFS
    - Low power sleep modes

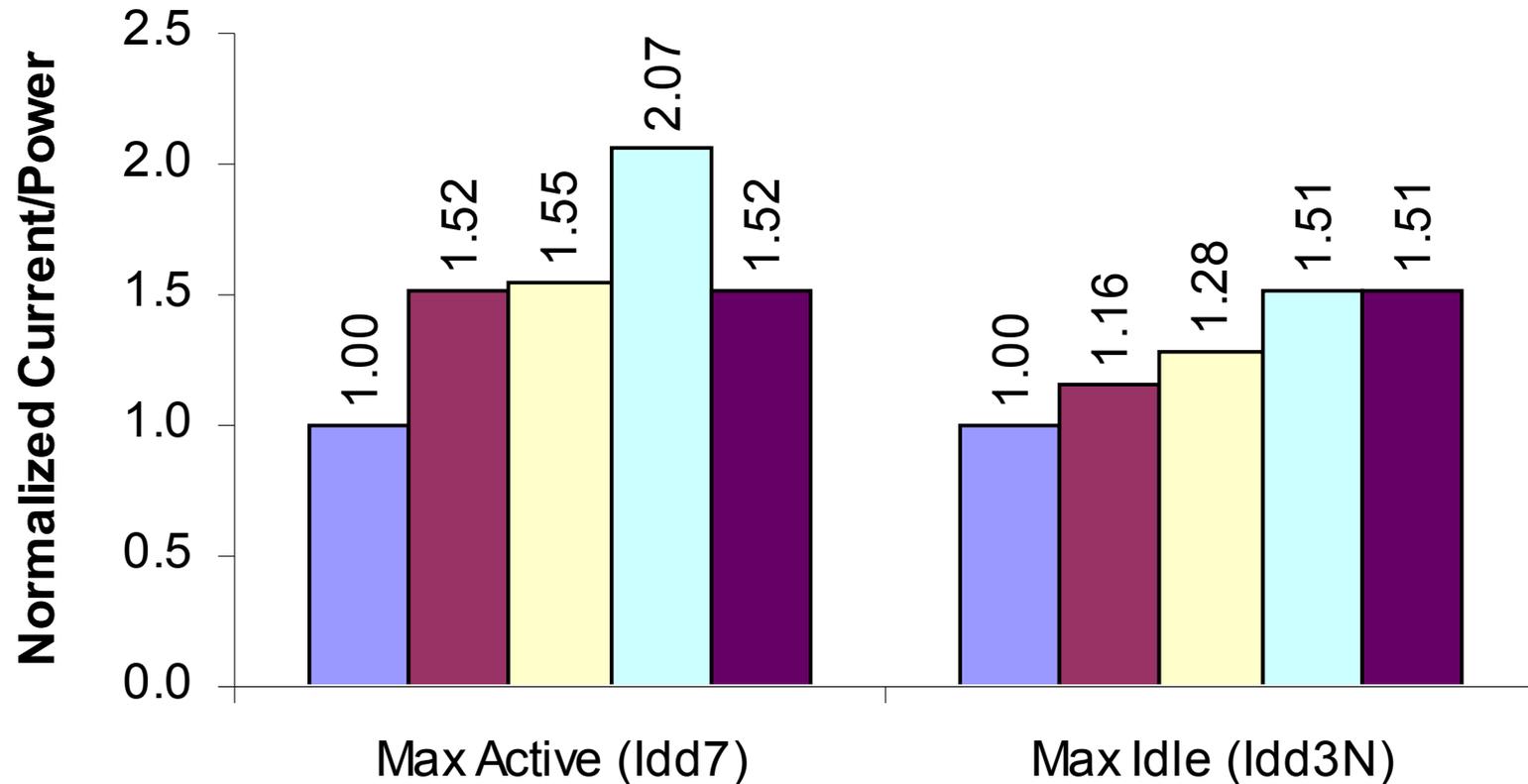
## Other sources can impact power, but are harder to capture

- **“Identical” components from the same manufacturer consume different amounts of power**
  - Process variation, part binning
  
- **“Identical” components from different manufacturers consume different amounts of power**
  - Different design criteria may produce same functional spec, but different implementations
  
- **Power supply efficiency varies**
  - Different loads
  - Different power supplies
  
- **Different environments cause components to consume different amounts of power**
  - Temperature, humidity, type of heat sink, etc
  - Ex: Temperature of the datacenter can cause power consumption to increase as
    - cooling becomes less effective (smaller  $\Delta T$ )
    - leakage power increases (exponential dependence on T).

## Variation across “identical” Pentium M processors



# Variation across “identical” DRAM from different vendors



■ Vendor 1 ■ Vendor 2 ■ Vendor 3 ■ Vendor 4 ■ Vendor 5

## Consequences of additional variation?

**Example: A vendor benchmarks a system with 100 processors of type A fabricated by manufacturer X, 200 DIMMs of memory of type B fabricated by manufacturer Y and placed in an “ideal” datacenter...**

- **The vendor can benchmark with energy-efficient parts that may not match the system purchased by the customer.**
- **What happens when parts start failing and are replaced?**
- **The initial estimate is incorrect as more parts are replaced with “identical” parts.**
- **Power consumption can “drift” as more expensive parts are replaced with cheaper parts that consume more energy.**
- **System may eventually raise the temperature of the datacenter beyond the threshold at which it cools efficiently.**
- **System temperatures go up in response to rising datacenter temperatures and datacenter temperatures go up in response to rising system temperatures.**
- **Initial benchmark measurements may be useless.**

# Conclusions

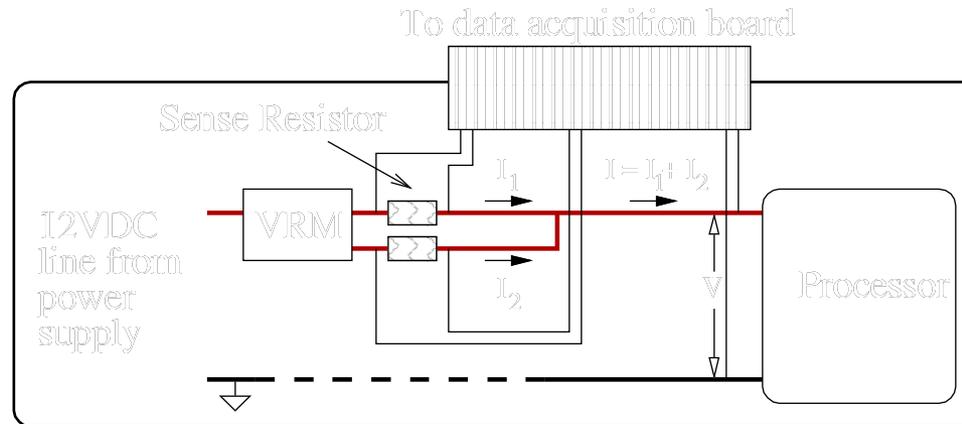
- **Systems exhibit many types of variation that need to be captured in power/performance benchmarking**
  - Intensity Variation
    - Workload variation in system utilization
      - Workloads differ from one another
      - A single workload may vary over time
  - Nature of activity variation
    - Workload variation in program characteristics
      - A single workload may change how it uses system components over time
  - Component-level variation
    - “Identical” components are not actually identical in power or performance
  - System response
    - How quickly does the system respond to changing characteristics?
    - How well does the system respond to changing characteristics?

Questions?

## IBM/ARL – Power Aware Systems Group

- **Focus is on power and thermal management techniques**
- **Can't evaluate techniques without good benchmarks**
- **Developing and adapting benchmarks for power/performance characterizations since 2002.**

# Correlating traces



- Send the power high to mark the beginning of a trace
- Capture both power and performance traces
- Send the power high to mark the end of a trace
  
- Done through various utilities for Windows and Linux.