IBM Research and U. Tennessee, Knoxville

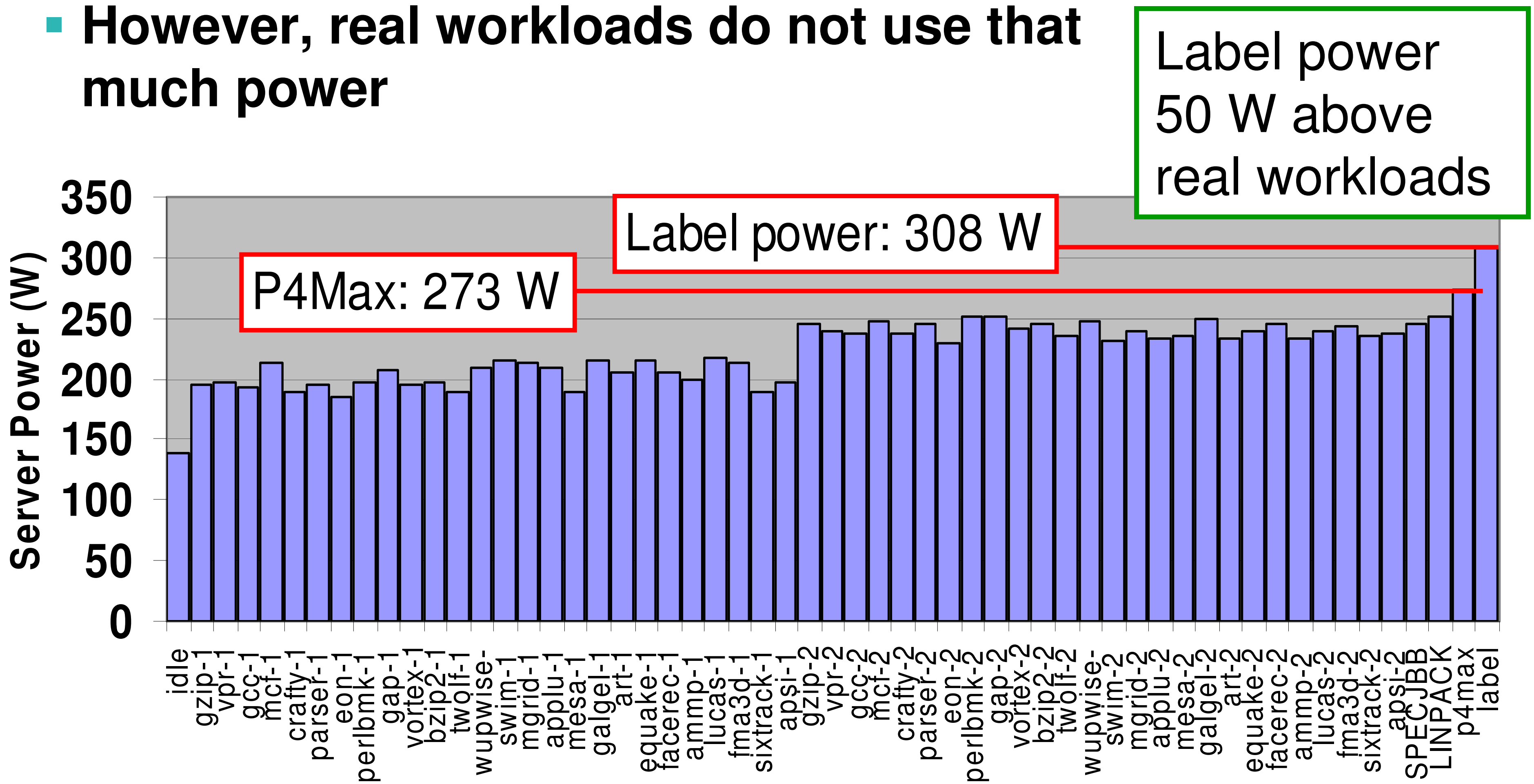
Server-level Power Control

Charles Lefurgy, Xiaorui Wang, Malcolm Ware

The 4th IEEE International Conference on Autonomic Computing June 12, 2007

Server power supplies

- **Datacenter must wire to label power**
- **However, real workloads do not use that much power**



The problem

- **Server power consumption is not well controlled.**
  - System variance (workload, configuration, process, etc.)
  - Design for worst-case power

- **Results:**
  - Power supplies are significantly over-provisioned
  - Therefore, datacenters provision for power that cannot be used
  - High cost, with no benefit in most environments

Our approach

- **Use "better-than-worst-case" design**
  - Example: Intel's Thermal Design Power (TDP)
  - Power, like temperature, can be controlled
- **Reduce design-time power requirements**
  - Run real workloads at full performance
  - Use smaller, cost-effective power supplies
- **Enforce run-time power constraint with feedback control**
  - Slow system when running power virus

Our contributions

- **Control of peak server-level power (to 0.5 W in 1 second)**
- **Derivation and analysis [see paper]**
  - Guaranteed accuracy and stability
- **Verified on real hardware**
- **Better application performance than previous methods**

Caveats

- **Our prototype is a blade server**
  - The results of the study also apply to rack-mount servers.
- **Power controller uses clock throttling, not dynamic voltage and frequency scaling (DVFS)**
  - At the time of the study, only clock throttling was available on our prototype system.
  - DVFS is not available on all processors (lower speed grades)
  - Recently, we have built a prototype using DVFS

Rest of the talk

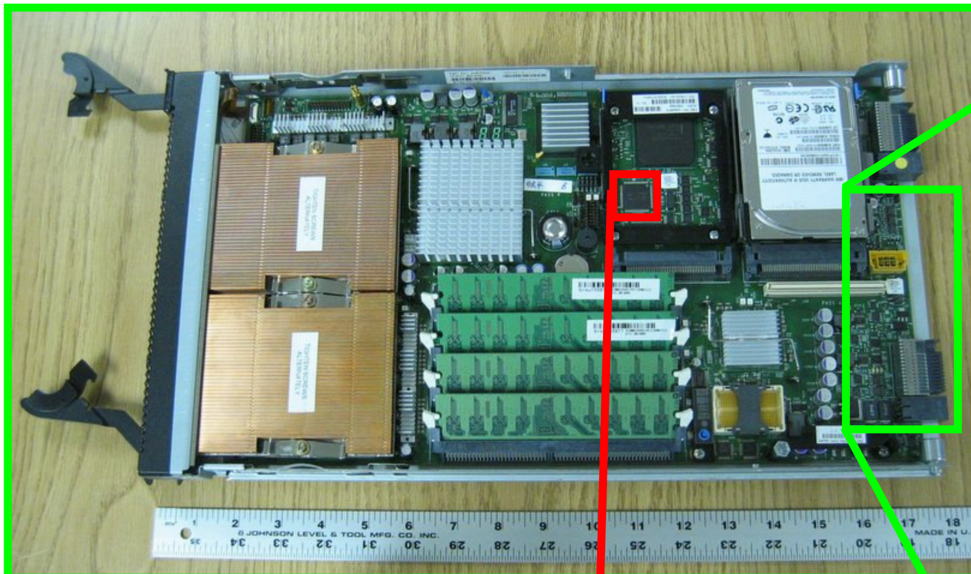
- **Power measurement**
- **Power control**
  - Open loop controller
  - Ad-hoc controller
  - Proportional controller
- **Experimental results**
- **Conclusions**

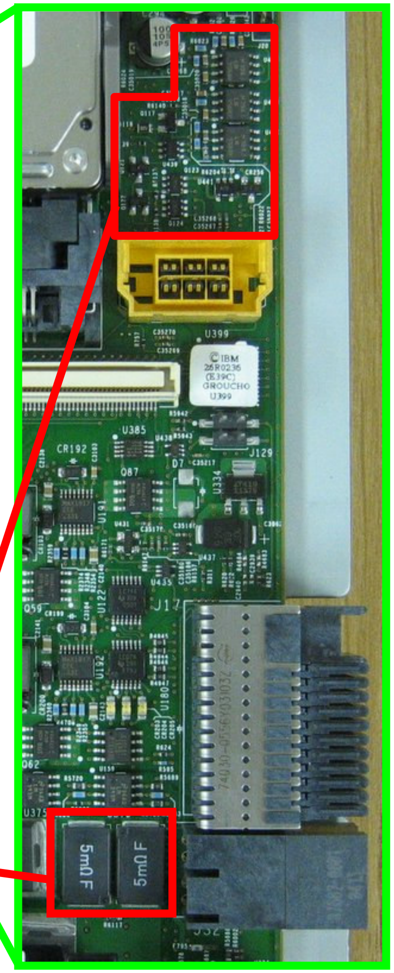
Power measurement

Measure 12V bulk power

0.1 W precision, 2% error

HS20 8843 (Intel Xeon blade)



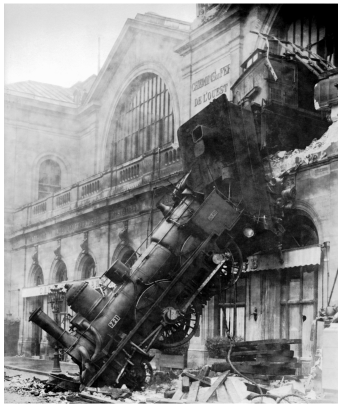


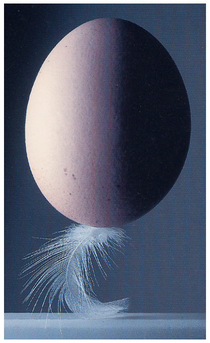
controller firmware on service processor (Renesas H8 2168)

Measurement/calibration circuit

Sense resistors

Options for power control

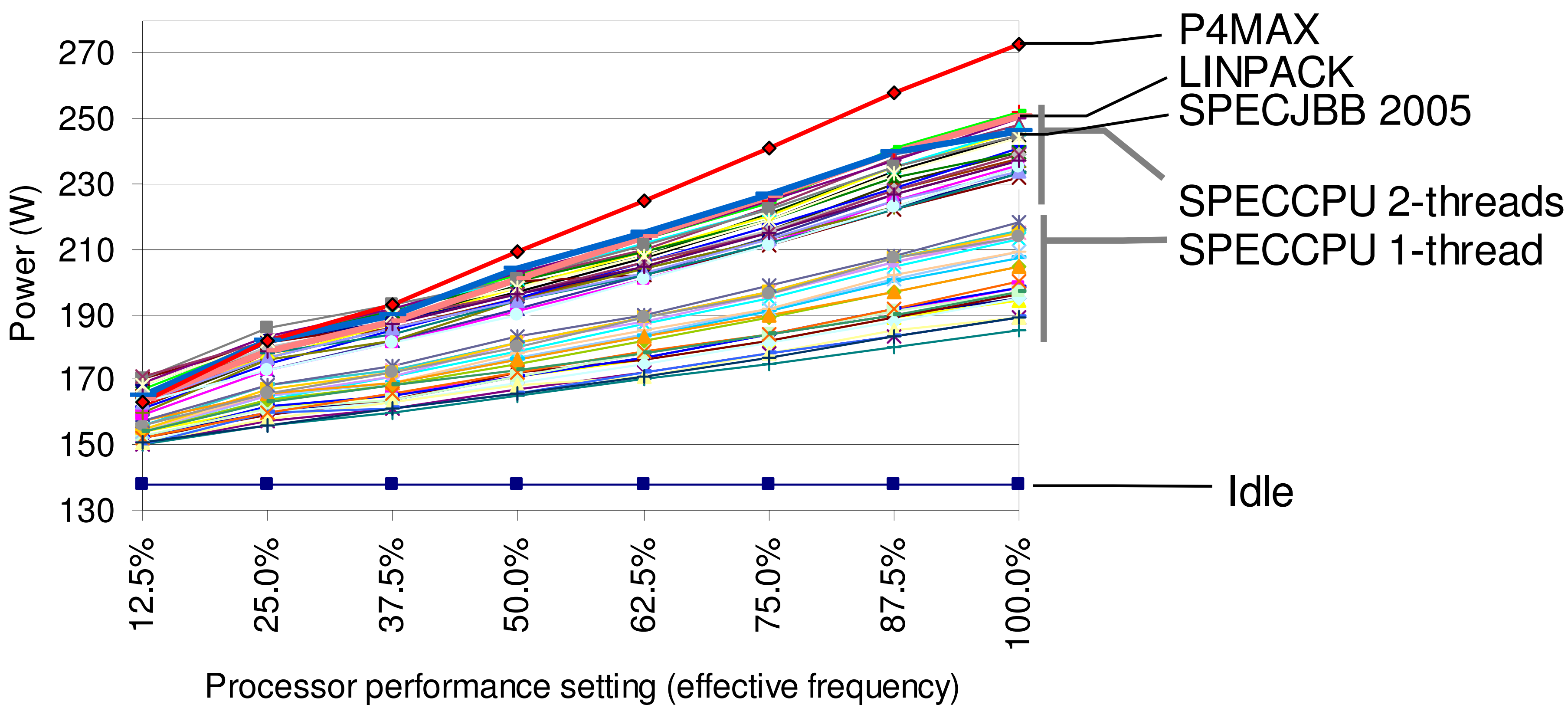




- **Open-loop**
  - No measurement of power
  - Chooses fixed speed for a given power budget
  - Based on most power hungry workload
- **Ad-hoc**
  - Measures power and compares to power budget
  - +1/-1 adjustments to processor clock throttle register
- **Proportional Controller ("P control")**
  - Designed using control theory
  - Guaranteed controller performance

Open loop design

- **P4MAX workload used as basis for open-loop controller**
- **Graph shows maximum 1 second power for workload**



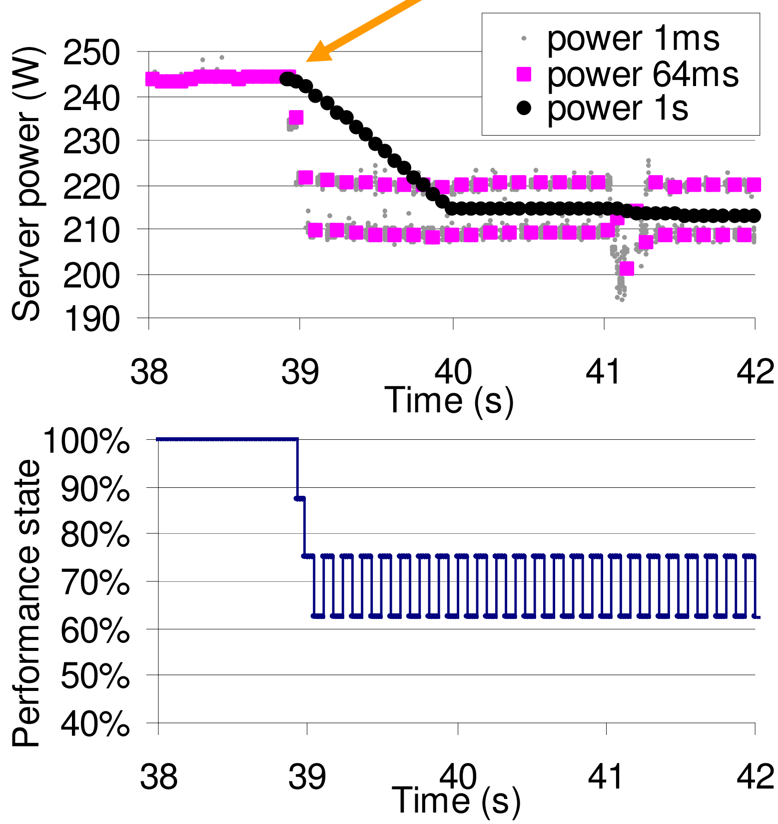
Proportional controller design

- **Settle to within 0.5 W of desired power in 1 second**
  - Based on BladeCenter power supply requirements
- **Every 64 ms**
  - Compare power to target power
  - Use proportional controller to select desired processor speed
    - 12.5% - 100% in units of 0.1%
- **Clock throttling**
  - Intel processor: 8 settings in units of 12.5% (12.5% - 100%)
  - Use delta-sigma modulation to achieve finer resolution

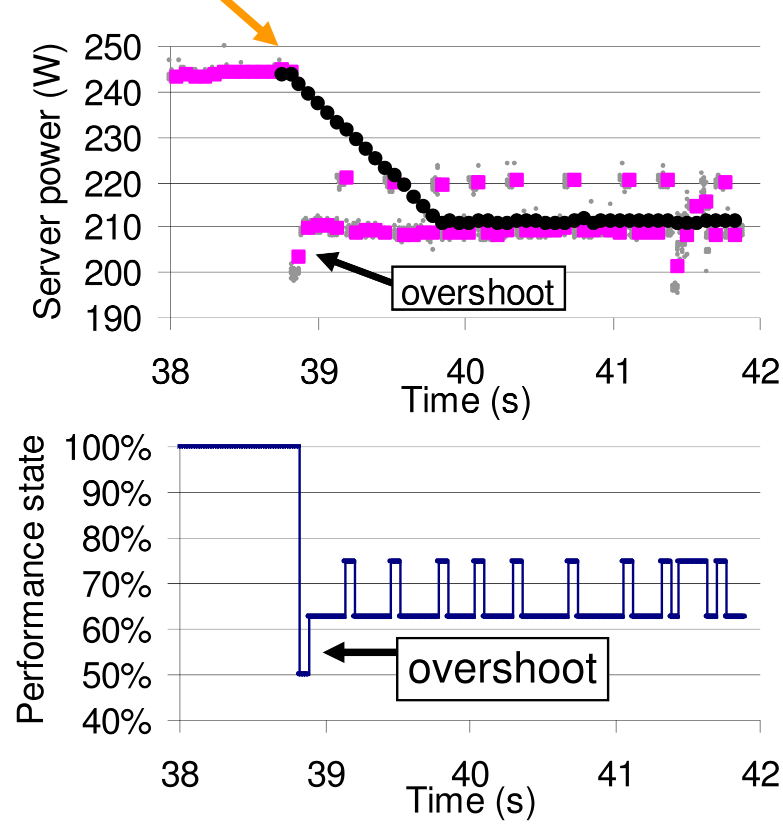
Why not use ad-hoc control?

Set point = 211.0 W

Ad-hoc



P Controller

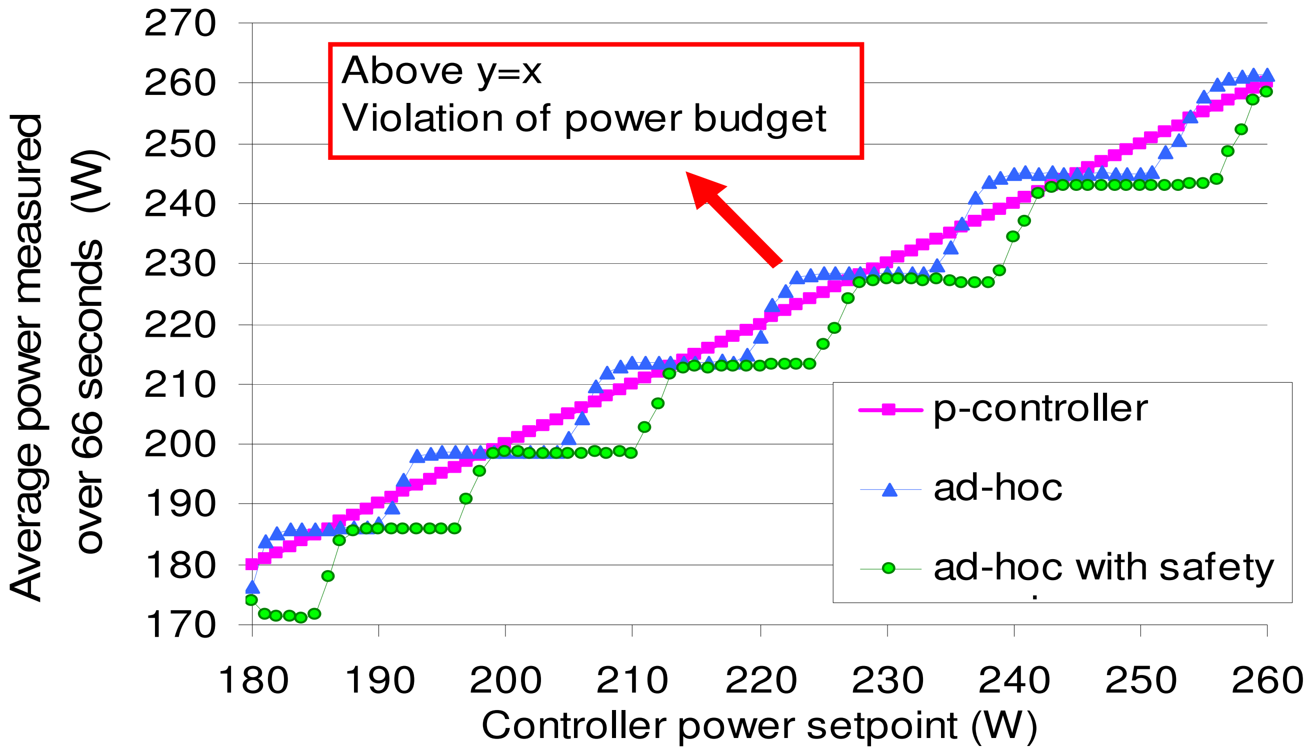


Settles to 216.0 W 5 W Violation
CPU speed: 68.8%

Settles to 211.0 W No violation
CPU speed: 65.8%

Steady-state error

- **P controller has no steady-state error (x=y)**
- **Ad-hoc controller has steady-state error**
  - Add safety margin of 6.1 W to ad-hoc



Comparison of 3 controllers

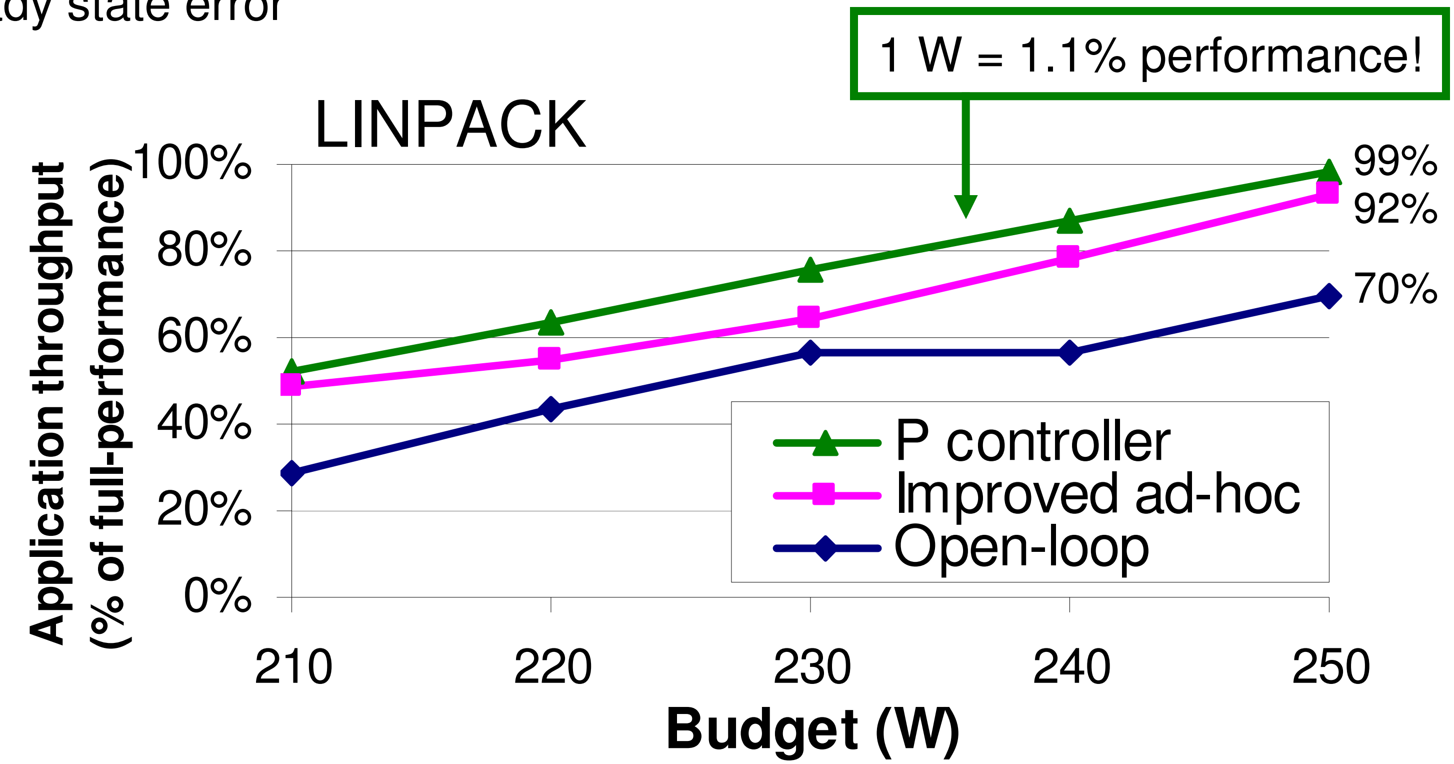
- **Run each controller with 5 power budgets**
- **Compare throughput of workloads**

- **Table shows settings used for each controller**

Power budget	Open-loop processor performance setting	Ad-hoc (with safety margin) set point	P control set point
250 W	75%	238.9 W	245.0 W
240 W	62.5%	229.1 W	235.2 W
230 W	62.5%	219.3 W	225.4 W
220 W	50%	209.5 W	215.6 W
210 W	37.5%	199.7 W	205.8 W

Application performance summary

- **P controller**
  - 31-82% higher performance than open-loop
  - 1-17% higher performance than ad-hoc
    - Quicker settling time
    - Zero steady state error



Power supply reduction

- **308 W: Label power of HS20 blade**
- **260 W: Real workloads run at full performance**
  - A reduction of 15% in supply power.
- **Fit 15% more servers per circuit**

Conclusions

- **Power is a 1st class resource that can be managed.**
  - Power is no longer the accidental result component configuration, manufacturing variation, and workload.
- **Reduce power supply capacity, safely.**
  - Relax design-time constraints, enforce run-time constraints.
  - Install more servers per rack.
- **Power control is a fundamental mechanism for power management in a power-constrained datacenter.**
  - Move power to critical workloads.