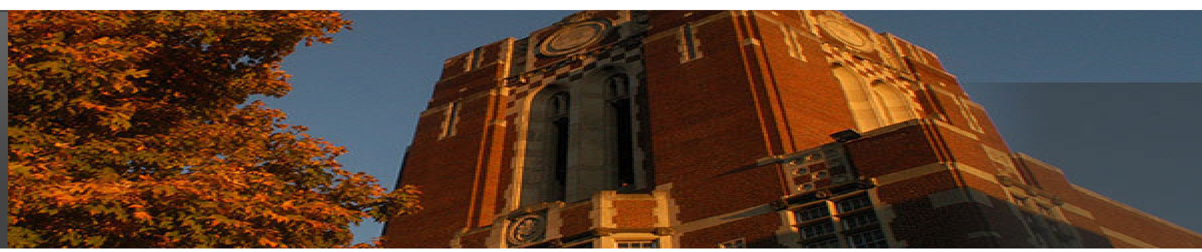
How Much Power Oversubscription is Safe and Allowed in Data Centers?

Xing Fu[1,2], Xiaorui Wang[1,2], Charles Lefurgy[3]

[1]EECS @ University of Tennessee, Knoxville

[2]ECE @ The Ohio State University

[3]IBM Research, Austin



Introduction

- Power: a first-class constraint in data center design
- Power oversubscription by power capping
  - Improves power facility utilization
  - Improves server performance
- Power capping at different levels
  - Servers, racks, and data centers
  - However, they all share a common assumption

Power should never exceed the rated power capacity?

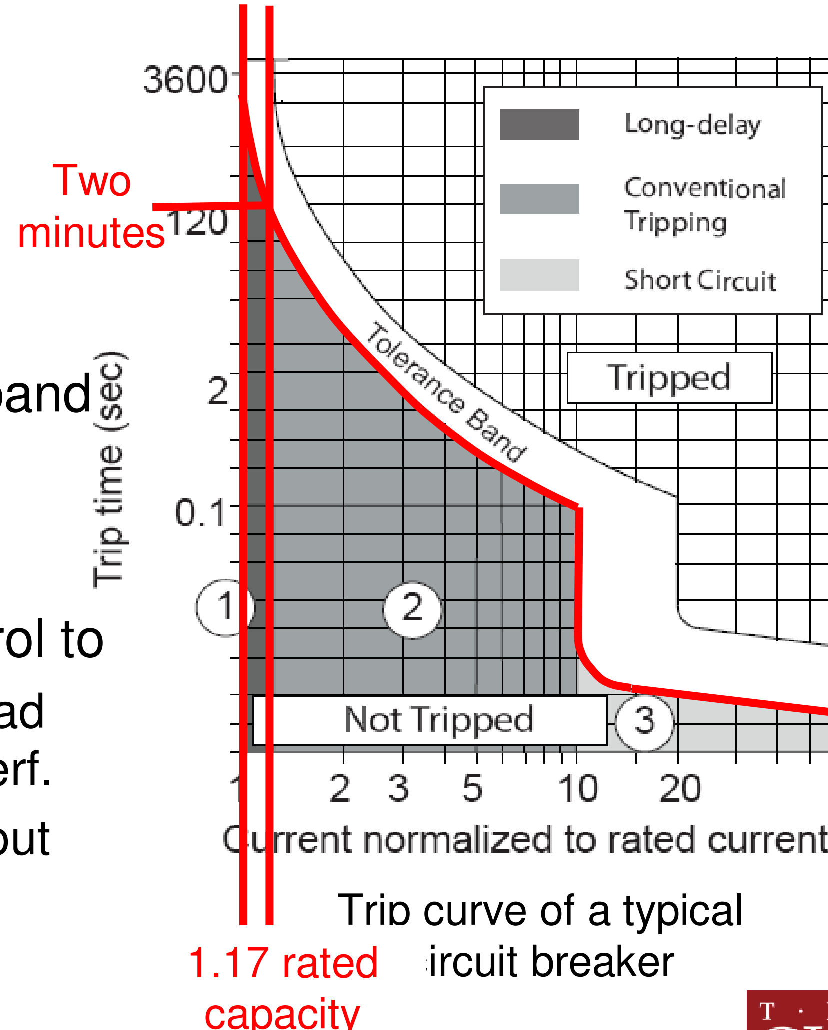
- Otherwise the circuit breaker (CB) would trip?
- Not really! circuit breakers can sustain short overloads.

THE UNIVERSITY of TENNESSEE KNOXVILLE

THE OHIO STATE UNIVERSITY

How Much Power Subscription is Safe?

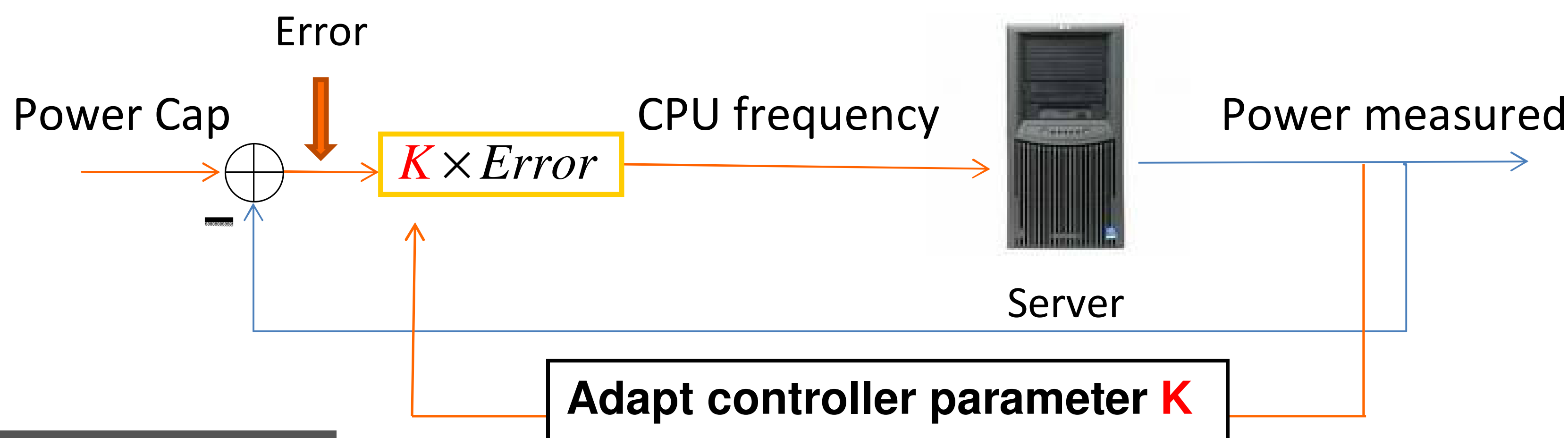
- A CB trips or not depends on
  - Magnitude of the overload
  - Duration of the overload
- Ideal upper bound?
  - Lower bound of the tolerance band
- This paper
  - Investigates CB trip features
  - Proposes adaptive power control to
    1. Fully utilizes the allowed overload interval for maximized server perf.
    2. Safely hosts more servers without upgrading power facilities



Trip curve of a typical circuit breaker

Proposed Solution: CB-Adaptive

- More than just a standalone controller
  - A methodology that adapts the parameters of existing power controllers to engineer their settling times
- Example: adapts a server power controller [Lefurgy ICAC’07]
  1. Obtain the tripping time from the CB tripping curve
  2. The desired settling time should be the tripping time
  3. Adapt controller parameter K to enforce the settling time



CB-Adaptive Design Details

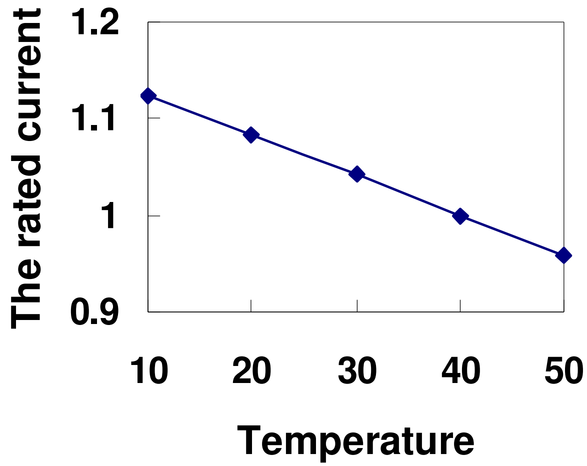
- System model

$$p(k+1) = p(k) + Ad(k)$$

  - *p(k)* is the power of the server
  - *d(k)* is the change to the CPU frequency
  - *A* is a hardware-specific parameter when the server runs LINPACK
- How to adapt the controller parameter?
  - The relationship between the parameter and the settling time
  - The parameter is a function of the measured power, the rated current of CB, and the control period.

Two CB-Adaptive Improvements

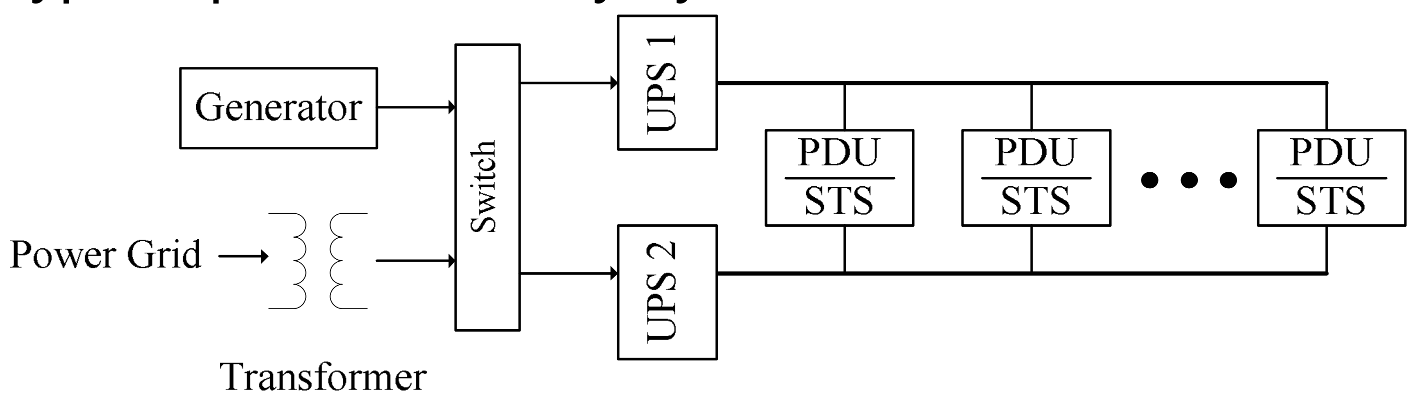
- Temperature-aware CB-Adaptive
  - The CB trip curve is impacted by the ambient temperature.
  - The rated current of CB is a linear function of the temperature.
  - K is also a function of ambient temperature.



- CB-Proactive
  - Delicately increases DVFS level in a proactive way
  - Further improves the server performance
  - When and to what extent the DVFS level is increased?
    - CB enters the long-delay region
    - Increase the frequency to the highest level

Discussion on Power Oversubscription

- Possible applications of CB-Adaptive
  - Hosting additional servers
- Safety issues
- A typical power delivery system



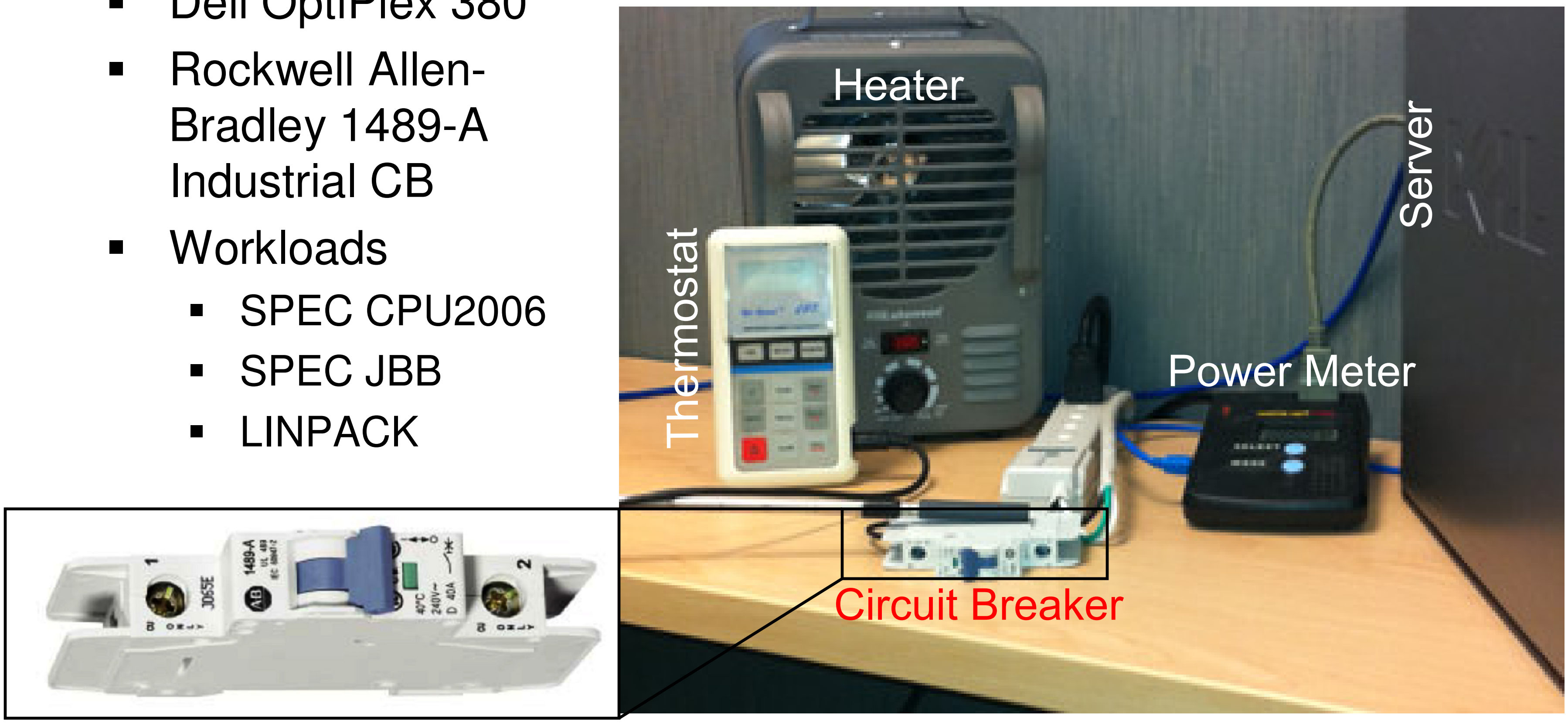
- Every component can tolerate overloads like CBs
  - Overload capacity: power beyond which permanent damage occurs to the component

More Discussion

- Components other than CBs do not experience overloads frequently.
  - It is less likely that many servers reach their peak power simultaneously.
  - Evidenced by a real Google data center [Fan ISCA'07]
- When only a branch circuit is overloaded
  - CB-Adaptive can be applied directly
- When multiple branch circuits are overloaded
  - CB-Adaptive needs to consider the tripping time of components other than CBs.

Hardware Testbed

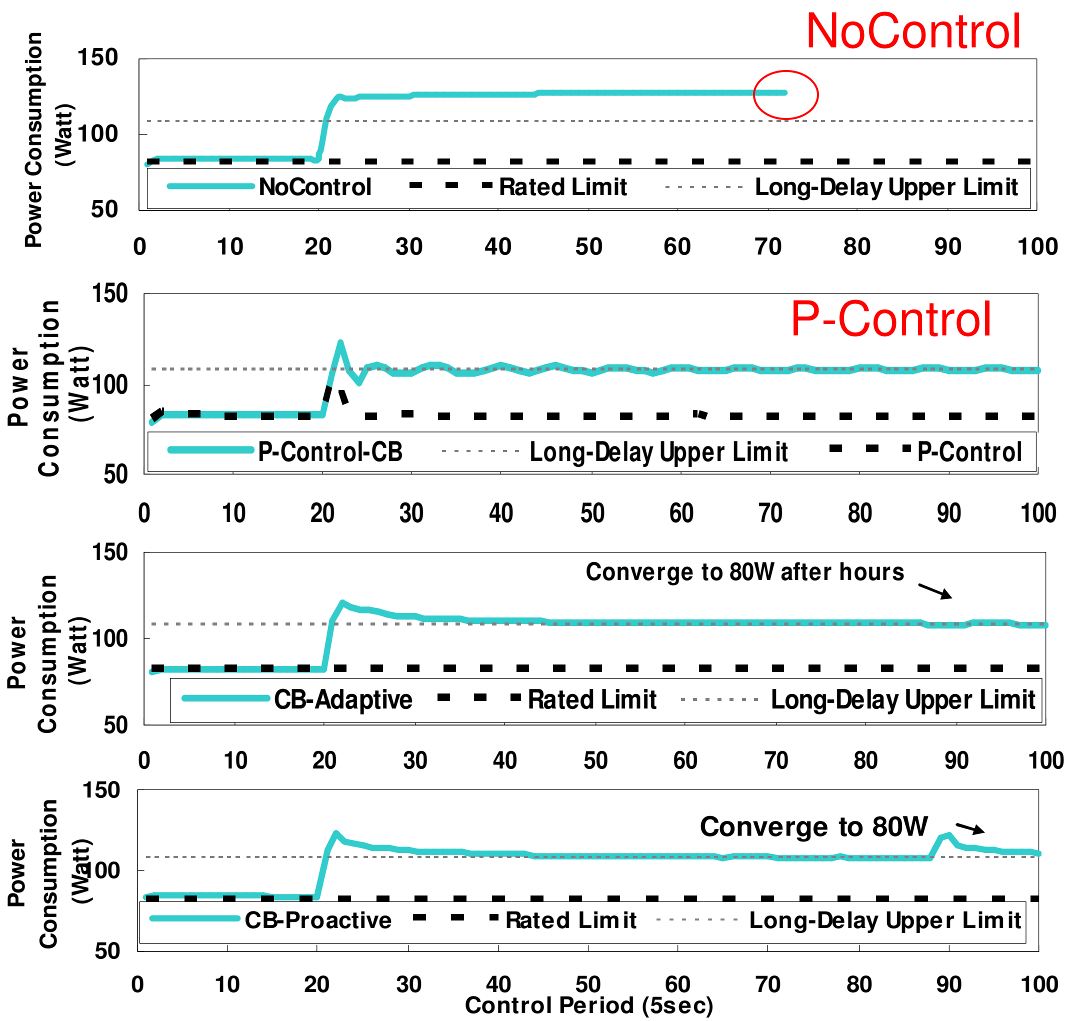
- Dell OptiPlex 380
- Rockwell Allen-Bradley 1489-A Industrial CB
- Workloads
  - SPEC CPU2006
  - SPEC JBB
  - LINPACK



Baselines

- NoControl
  - Estimates the peak power consumption of a server
    - No power caps
    - Unsafe and conservative
- P-Control
  - Measures the power in every control period
  - A non-adaptive proportional controller calculates frequency changes to enforce a power budget.
- P-Control-CB
  - The power budget is different from that of P-Control
    - Upper bound of the long-delay region of the CB

Power Control Comparison



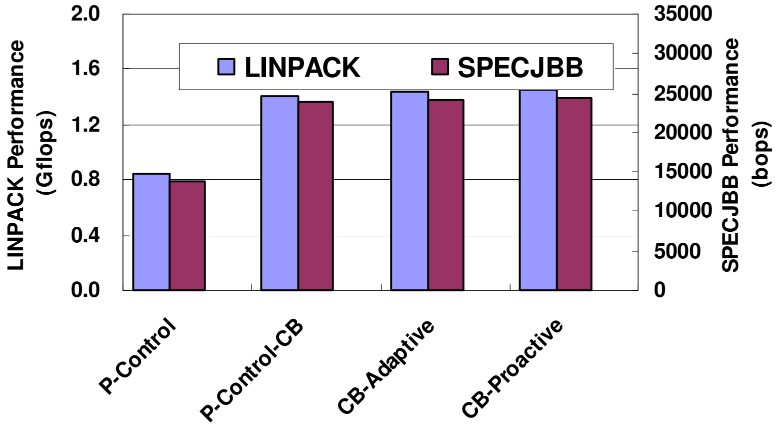
- NoControl causes the CB trips. Unsafe
- P-Control & P-Control-CB Unsafe and conservative
- CB-Adaptive fully utilizes overload intervals of CBs.
- Raise CPU freq for higher performance

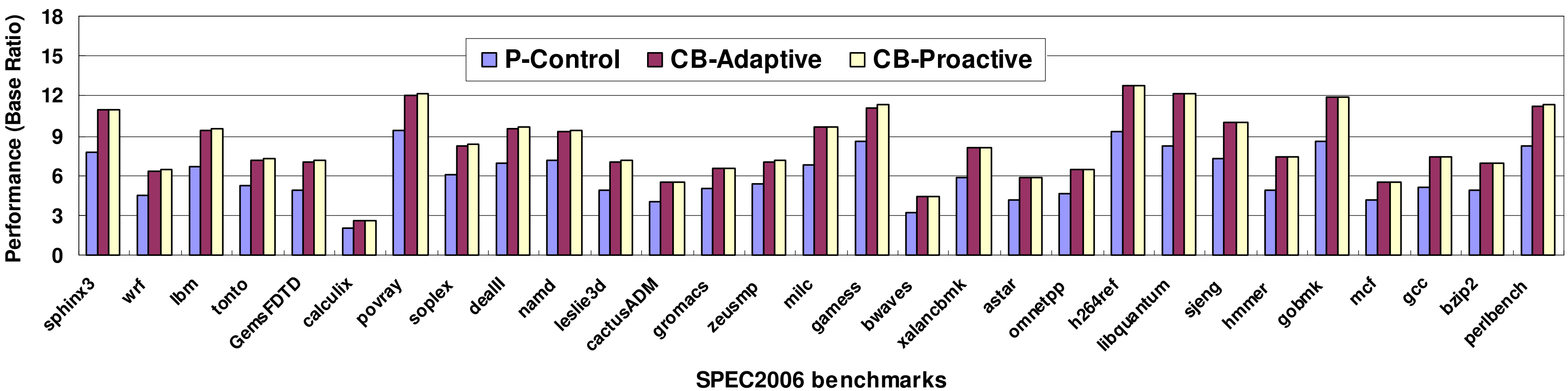
THE UNIVERSITY of TENNESSEE KNOXVILLE

THE OHIO STATE UNIVERSITY

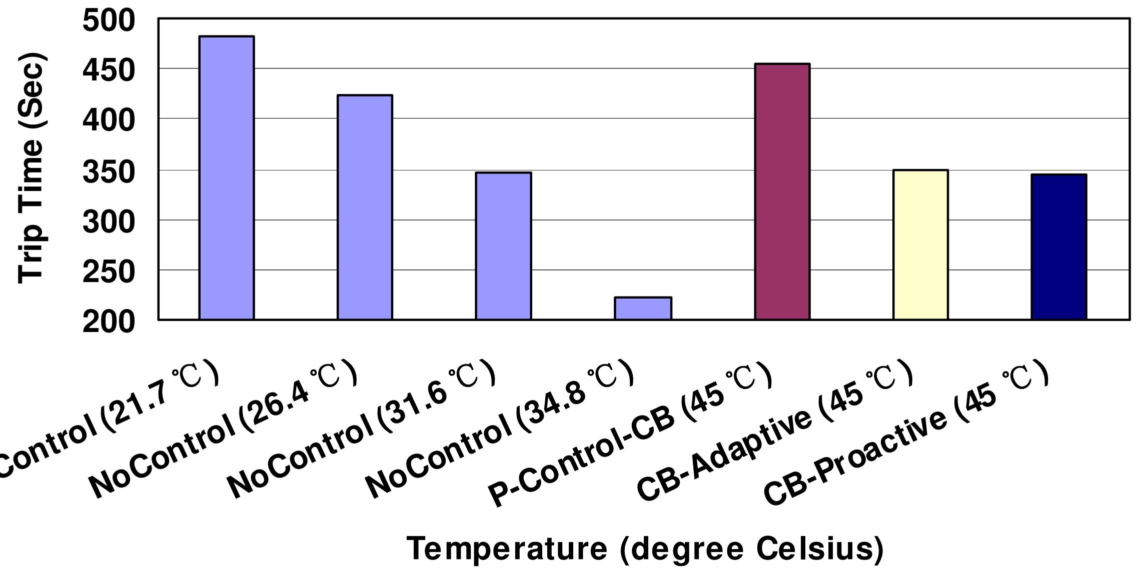
Performance Comparison

- CB-Adaptive outperforms P-Control by
  - 66%, for LINPACK
  - 29 % to 49%, for SPEC CPU 2006
  - 74%, for SPEC JBB



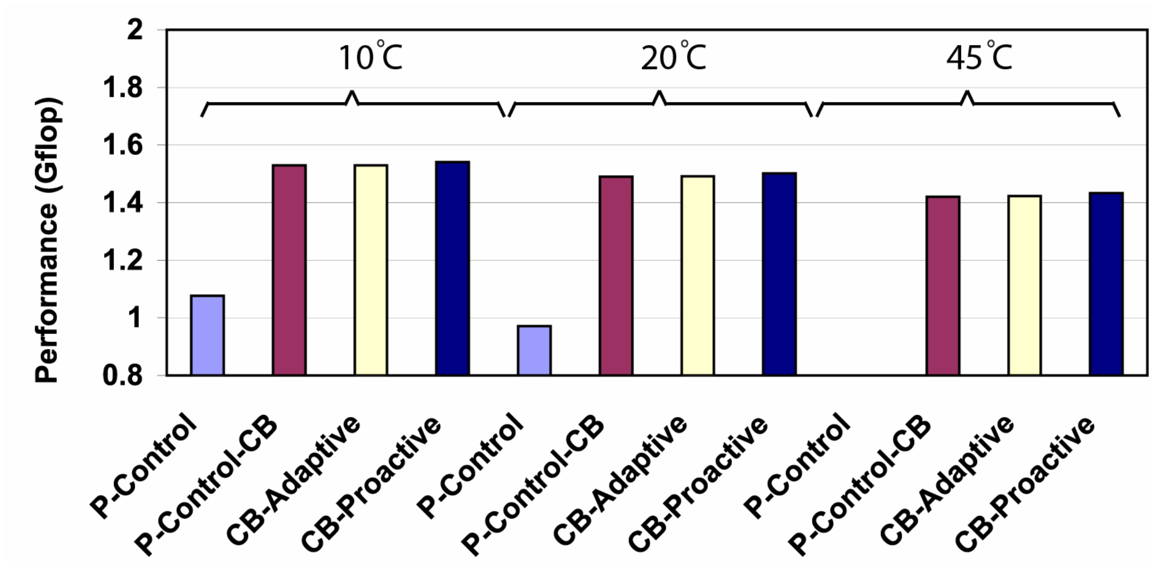


Impact of Temperature



- Temperature impacts the trip time significantly.
- Temperature-blind solutions P-Control-CB, CB-Adaptive and CB-Proactive are not safe.

Temperature-Aware CB-Adaptive



- As the temperature increases, the performance of servers decreases.
- The performance decrease is modest.

Power Provisioning Analysis

- NoControl

$$The\ number\ of\ servers = \frac{Rated\ power\ of\ the\ CB}{estimated\ server\ power}$$

  - The estimation is too conservative
  - 7 servers hosted per branch
- P-Control
  - Enforce a power budget instead of an estimation of power
  - 13 servers hosted per branch
- CB-Adaptive
  - Enforce a higher power budget than P-Control
  - 20 servers hosted per branch

Conclusions

- A common assumption of existing power capping
  - Peak power should never exceed the rated CB capacity
- This paper
  - Systematically studies the CB tripping characteristics
  - Identifies ideal upper bound of safe power oversubscription
  - Proposes two adaptive power control strategies
- Evaluation on safe power oversubscription
  - A single server: 38% performance improvement
  - Circuit branch: host 54% more servers without upgrading power infrastructure

Questions?

- Acknowledgements
  - NSF CAREER Award CNS-0845390
  - NSF CSR Grant CNS-0720663
  - NSF SHF Grant CCF-1017336
  - Prof. Leon Tolbert at the University of Tennessee

Thank you!

BackUp

Control Theoretic Analysis

- How to adapt the controller parameter?

$$K = \frac{1 - \sqrt[m]{0.02}}{A}$$

- Details of the derivation
  - Z transform of the system model
  - Z-domain controller
  - Calculate the close loop transfer function
  - Reverse Z transform

Power Provisioning Analysis

Table 2: Overload capacities of power delivery components.

Components	Overload capacity normalized to the rating	Trip time (minutes)
Static Transfer Switch	125%	60
Various cables	125%	3.5 to 110
UPS	125%	0.5
Generator	110%	60
Transformer	150%	30

- UPS cannot tolerate overloads
  - Not a problem because each UPS run at 50% its capacity
- Factors limiting overload capacities