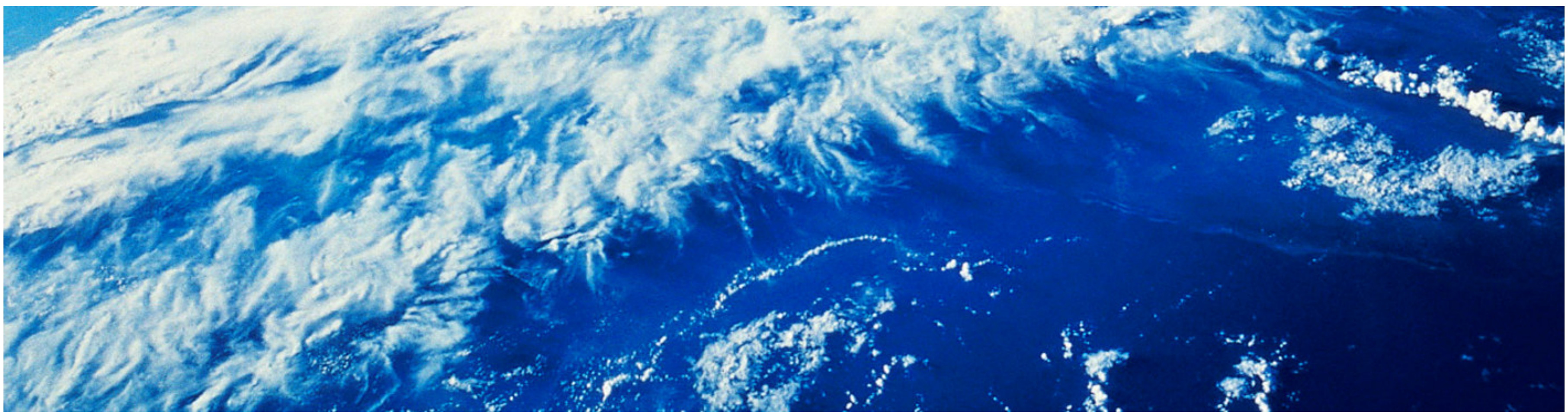
TAPO: Thermal-Aware Power Optimization Techniques for Servers and Data Centers



Wei Huang

IBM Research

Overview

- Team work at IBM Research:
  - Austin: Malcolm Allen-Ware, John Carter, Mootaz Elnozahy, Tom Keller, Charles Lefurgy, Jian Li, Karthick Rajamani, Juan Rubio
  - T. J. Watson: Hendrik Hamann
- **Objective: Power Optimization of an entire system (e.g., server, DC), with explicit consideration of Cooling Power**
- Hierarchical Techniques:
  - Server-level power (TAPO-server):
    - Fan power vs. leakage power
    - Goal: minimize aggregate **fan+leakage** power
    - Prototyped on a POWER 750 Express server (POWER7-based).
  - Datacenter-level power (TAPO-dc):
    - HVAC power vs. server fan power
    - Goal: minimize aggregate **HVAC+server** power
    - Analysis based on realistic models

Background

- Thermal setpoints are fixed
  - Server temperature setpoint, e.g. 70C for POWER7 processors
  - Data Center (DC) HVAC chiller setpoint (cooled water), e.g. 10C
  - **System dynamics are not considered, can be power inefficient → overcooling and wasting cooling power.**

- Cooling-related power components
  - DC HVAC power (chiller, blower, etc)
    - Comparable to IT power
    - Characteristics: warmer environment, higher chiller setpoint, lower chiller power
  - Server fan power:
    - Has been part of IT power, but really should be considered separately
      - PUE is not an accurate indicator
    - Strong superlinear (~ quadratic or cubic) relationship to fan speed
  - Server (processor) leakage power:
    - Strongly temperature dependent
    - To reduce leakage, want more server fan power to cool chips down

Overview

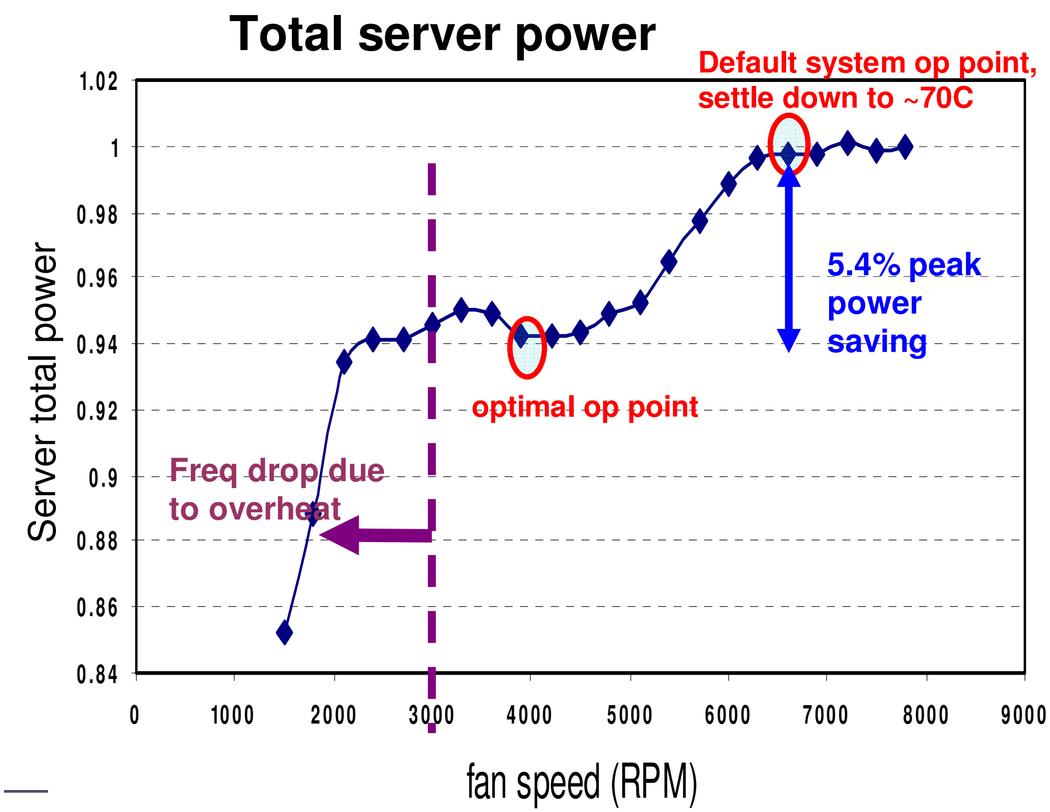
- Team work:
  - Austin: Wei Huang, Malcolm Allen-Ware, John Carter, Mootaz Elnozahy, Tom Keller, Charles Lefurgy, Jian Li, Karthick Rajamani, Juan Rubio
  - Watson: Hendrik Hamann

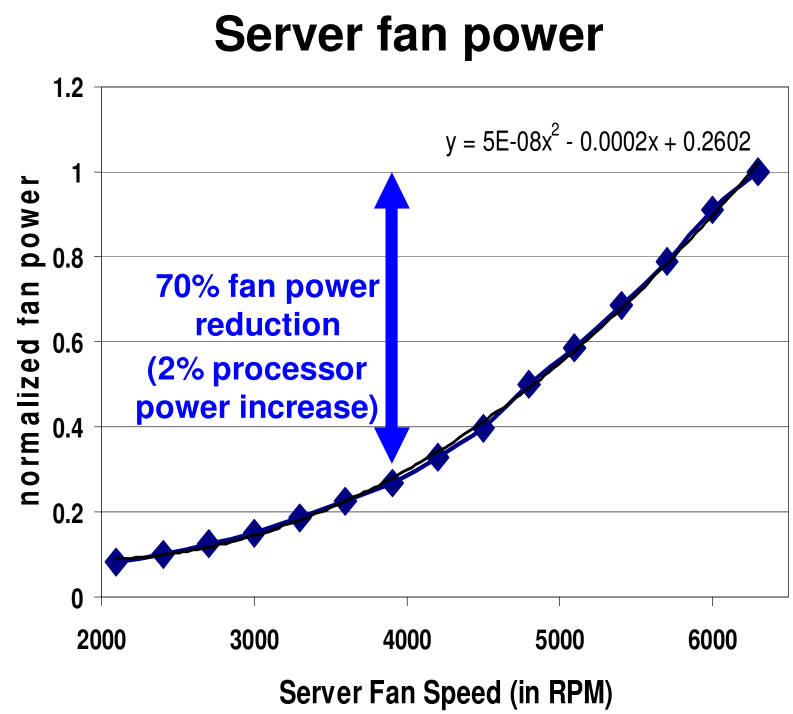
- **Objective**: Optimize power and/or performance of an entire system (e.g., server, DC), with explicit consideration of **cooling power**

- Hierarchical Techniques:
  - **Server-level power (TAPO-server):**
    - Fan power vs. leakage power
    - Goal: minimize aggregate **fan+leakage** power
    - Prototyped on a POWER 750 Express server.
  - Datacenter-level power (TAPO-dc):
    - HVAC power vs. server fan power
    - Goal: minimize aggregate **HVAC+server** power
    - Analysis based on realistic models

TAPO-server

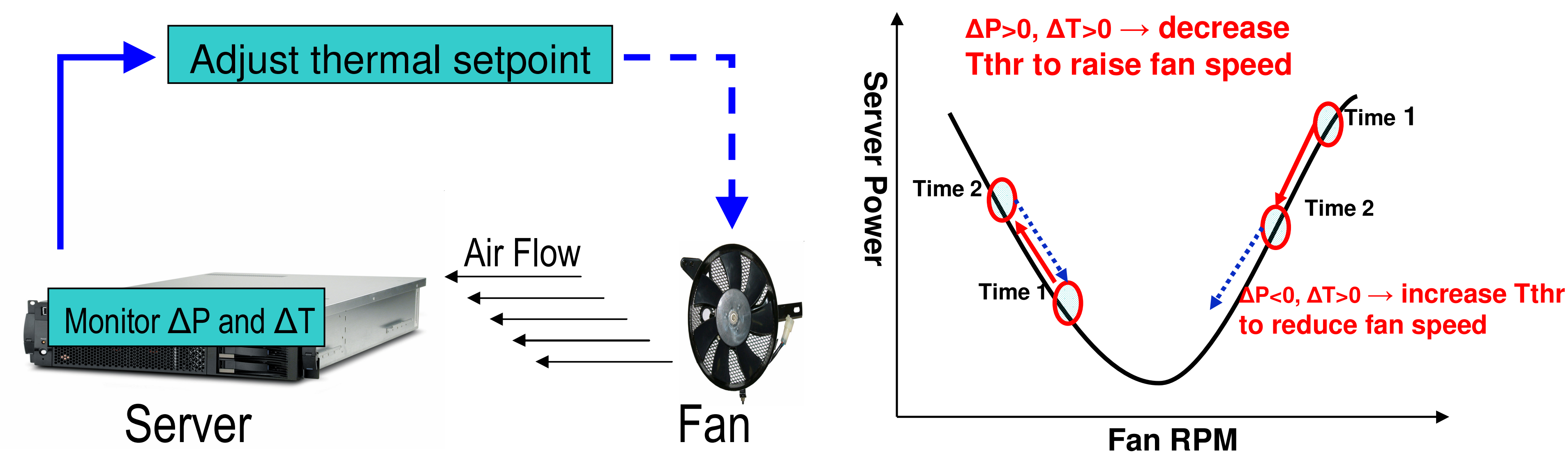
- Optimize server fan + processor leakage power, what is the power saving potential?
  - Manual characterization:
    - POWER7-based server
    - Turbo frequency (3.864GHz), CPU-intensive workload, L2 resident, 32 SMT4 cores





Search for optimal thermal setpoint in TAPO-server

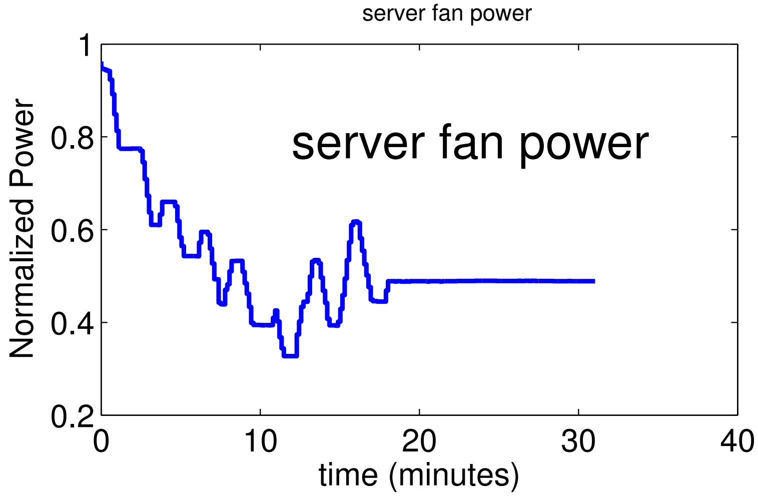
- Change processor thermal setpoint
  - Indirectly change fan speed
- On the curve:
  - Left: fan speed low, more thermal-induced leakage power
  - Right: system is cool, but more fan power

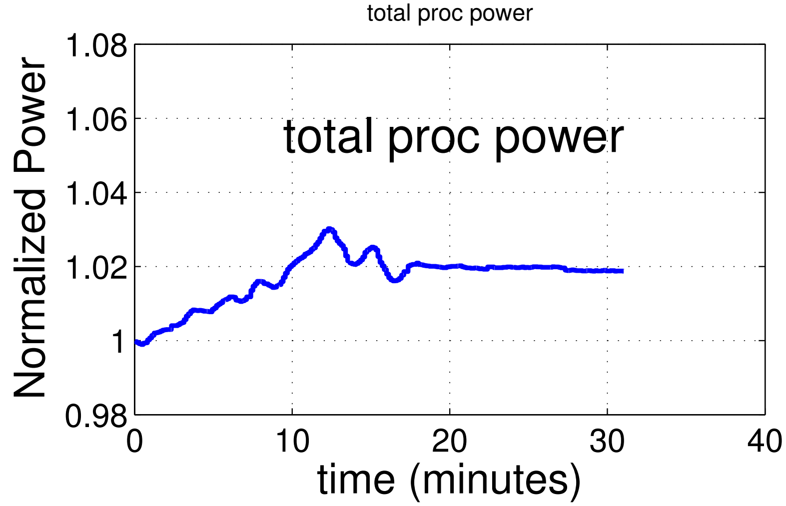


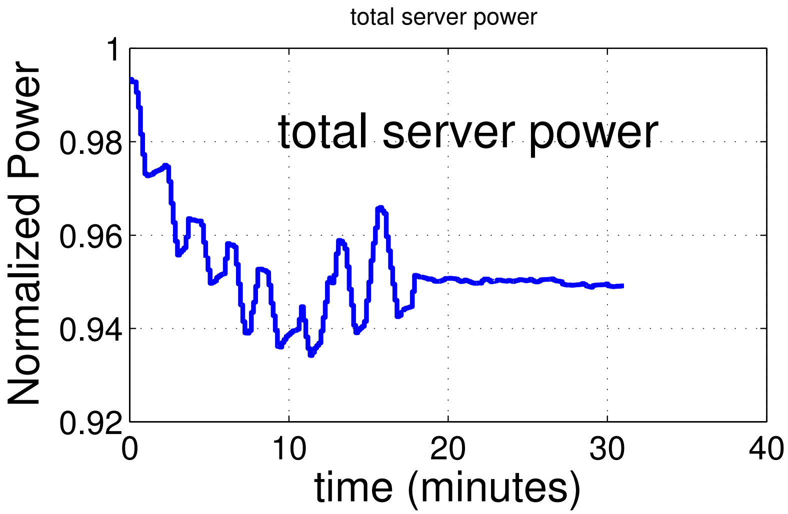
TAPO-server discussions

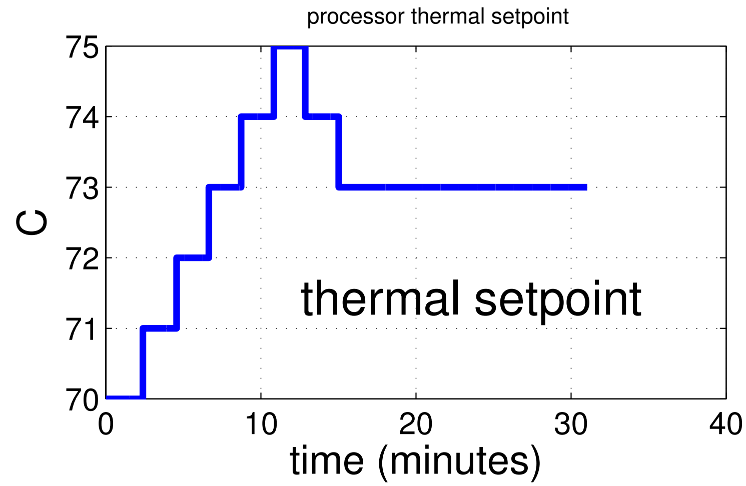
- Power convergence threshold: 5 Watts.
- Sampled every 32ms.
- Entirely depends on measurements, no models involved.
- reduce peak power at peak performance.
- Save ~5% peak power, a perfect solution would have been 5.4%
- No observed performance loss (frequency and voltage are fixed).
- Regardless of workload, chip variations and environment, TAPO-server should adaptively find the optimal point.
- Slow convergence: wait long enough (30 seconds to 2 minutes) for temperature to settle down after fan speed changes.
- For safety, there is an upper limit on thermal threshold (if exceeded, use DVFS to prevent thermal emergency).

TAPO-server results







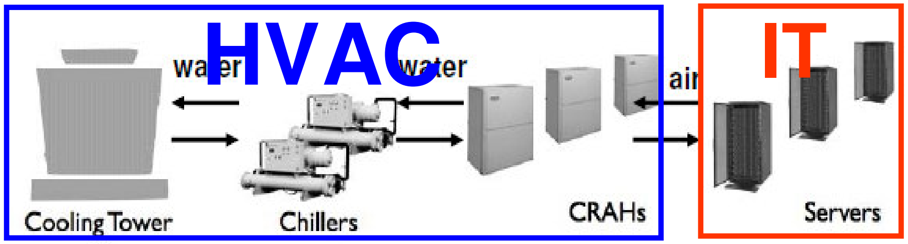


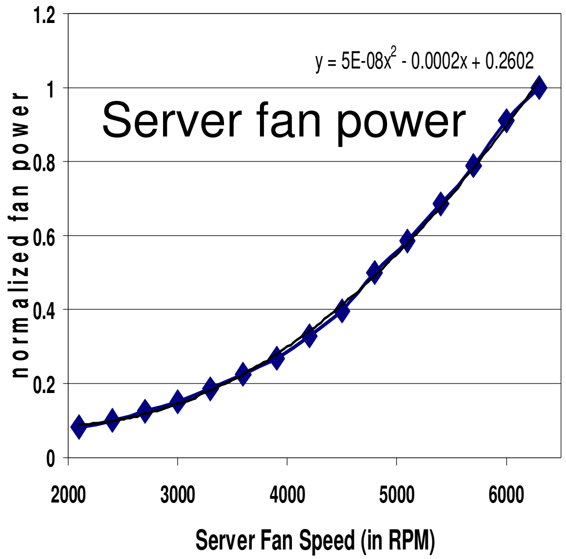
Prototyped new model-based control method reduces convergence time to ~1 minute

Overview

- Team work:
  - Austin: Wei Huang, Malcolm Allen-Ware, John Carter, Mootaz Elnozahy, Tom Keller, Charles Lefurgy, Jian Li, Karthick Rajamani, Juan Rubio
  - Watson: Hendrik Hamann
- **Objective**: Optimize power and/or performance of an entire system (e.g., server, DC), with explicit consideration of **cooling power**
- Hierarchical Techniques:
  - Server-level power (TAPO-server):
    - Fan power vs. leakage power
    - Goal: minimize aggregate **fan+leakage** power
    - Prototyped on a P7 HV32 server.
  - **Datacenter-level power (TAPO-dc):**
    - HVAC power vs. server fan power
    - Goal: minimize aggregate **HVAC+server** power
    - Analysis based on realistic models

TAPO-dc





- Tradeoff between HVAC power and server fan power
- Use chilled water setpoint to adjust HVAC power
  - Based on published component power models
  - Two chiller designs (COP 3.0-6.0 and 4.1-5.5)
  - T_inlet = T_chiller + 10C
  - Server inlet temperature range 20C ~ 40C

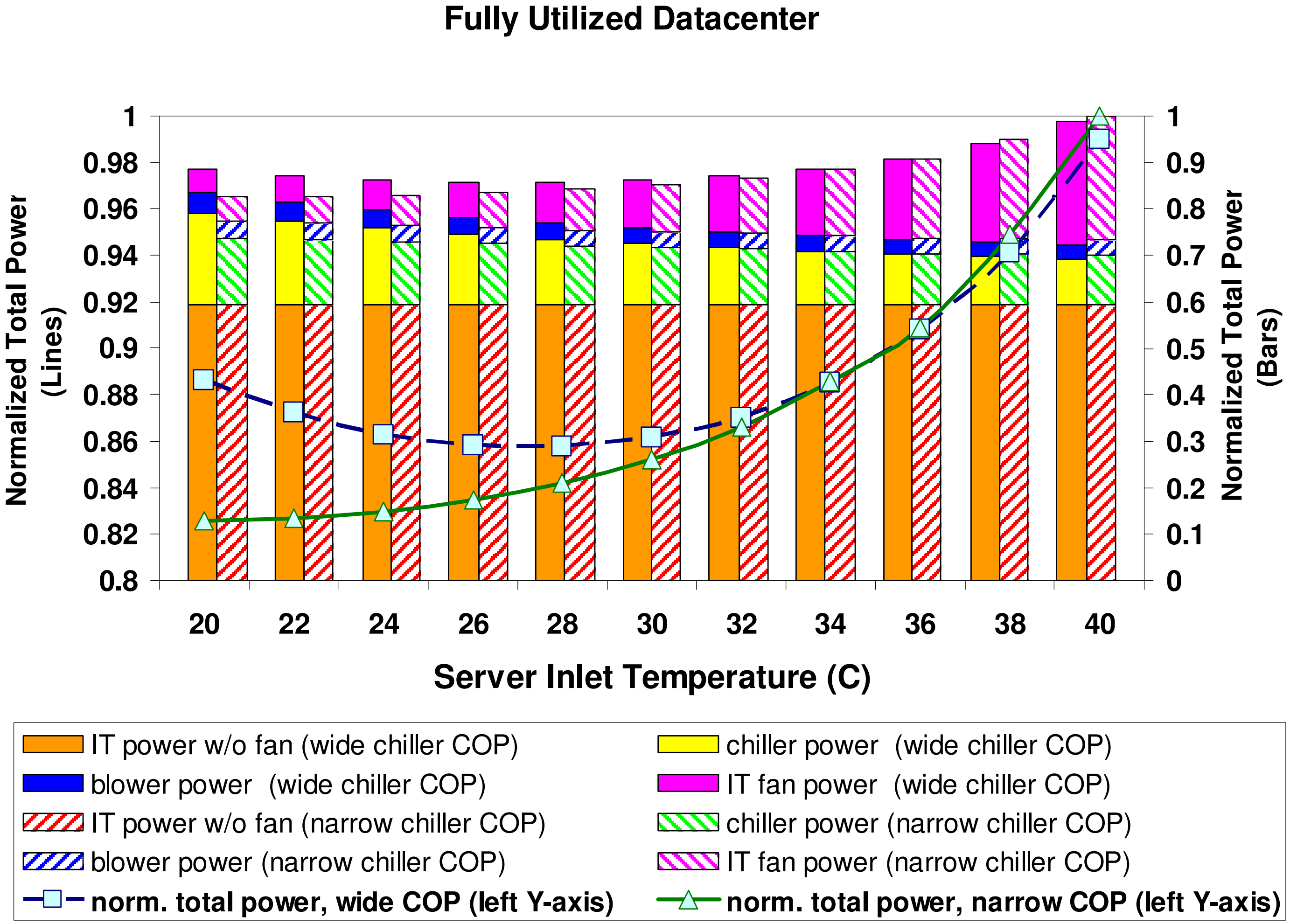
$$P_{chiller} = \frac{1}{a(1+b\cdot(T_{s\_chiller}-T_0))}P_{DC\_total} + 0.05P_{DC\_total} = \frac{P_{DC\_total}}{COP_{chiller}}$$

$$P_{blower} = P_{blower\_max}\cdot\left(\frac{P_{IT}+P_{chiller}}{100kW}\cdot\frac{10kCFM}{maxCFM}\right)^{\alpha}$$

$$P_{IT\_without\_fan} = P_{idle} + \mu\cdot(P_{full} - P_{idle})$$

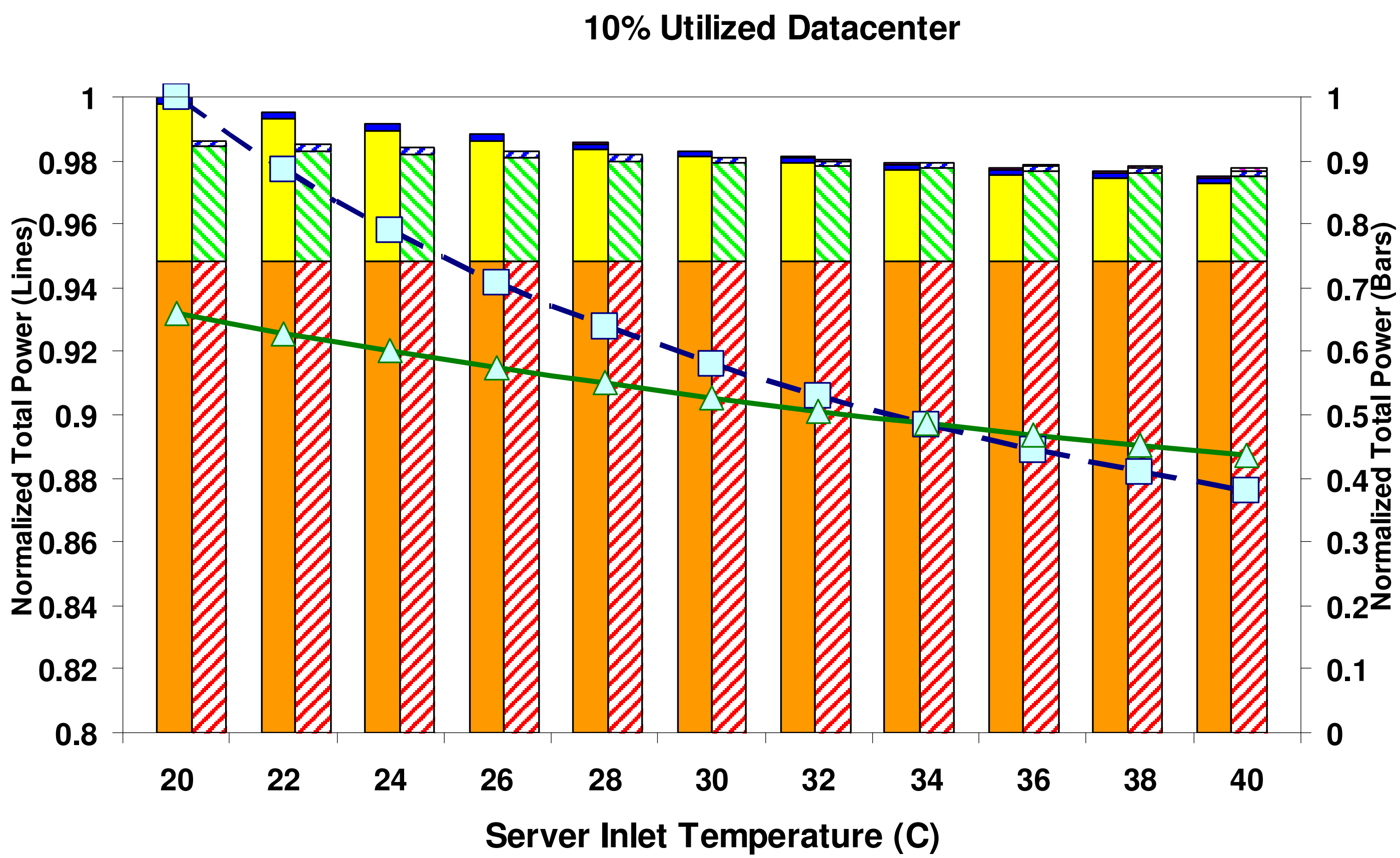
TAPO-dc results

- Assuming a rack of ten POWER 750 Express servers
- Fully utilized DC cooling zone



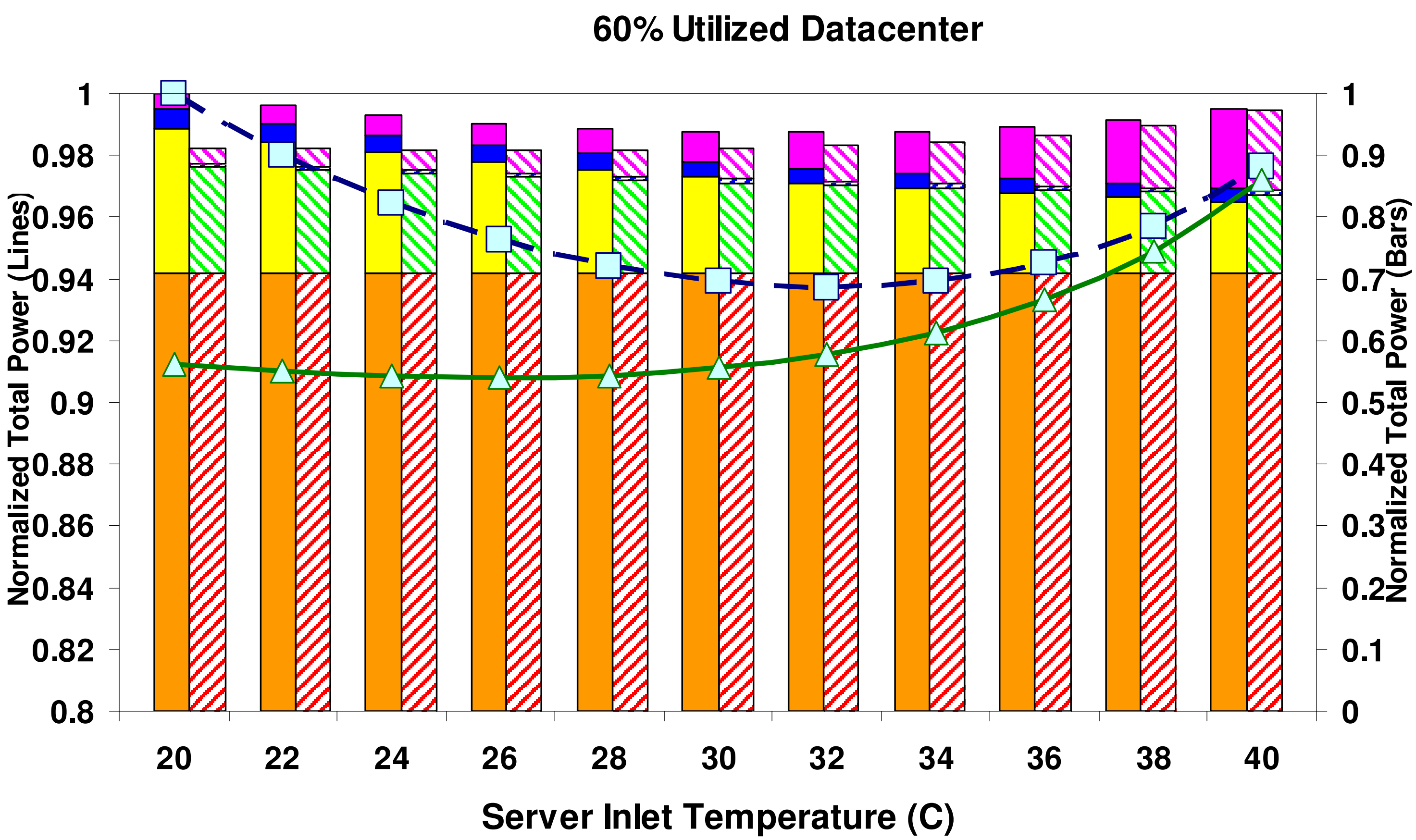
TAPO-dc results (cont'd)

- 10% utilized DC cooling zone



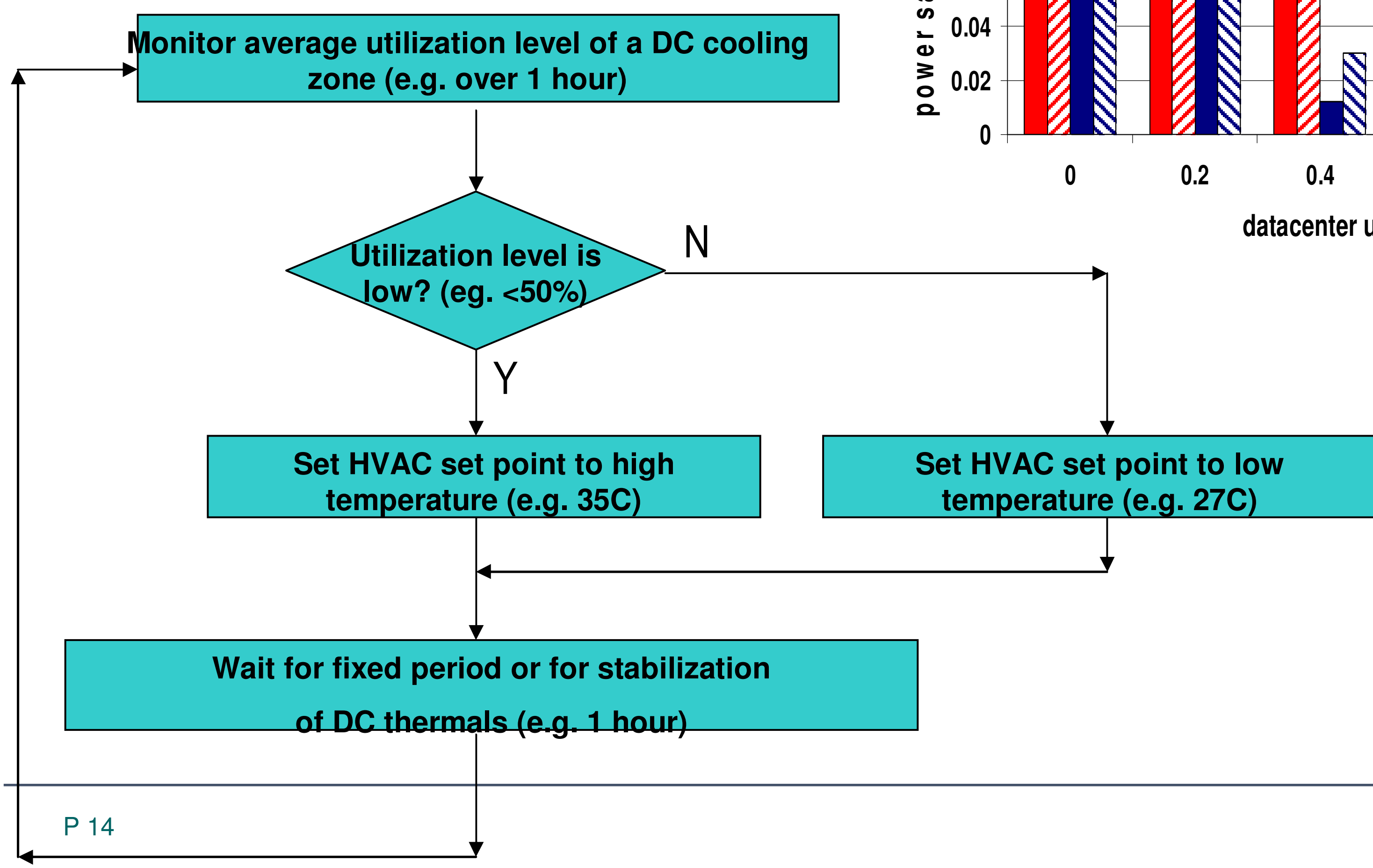
TAPO-dc results (cont'd)

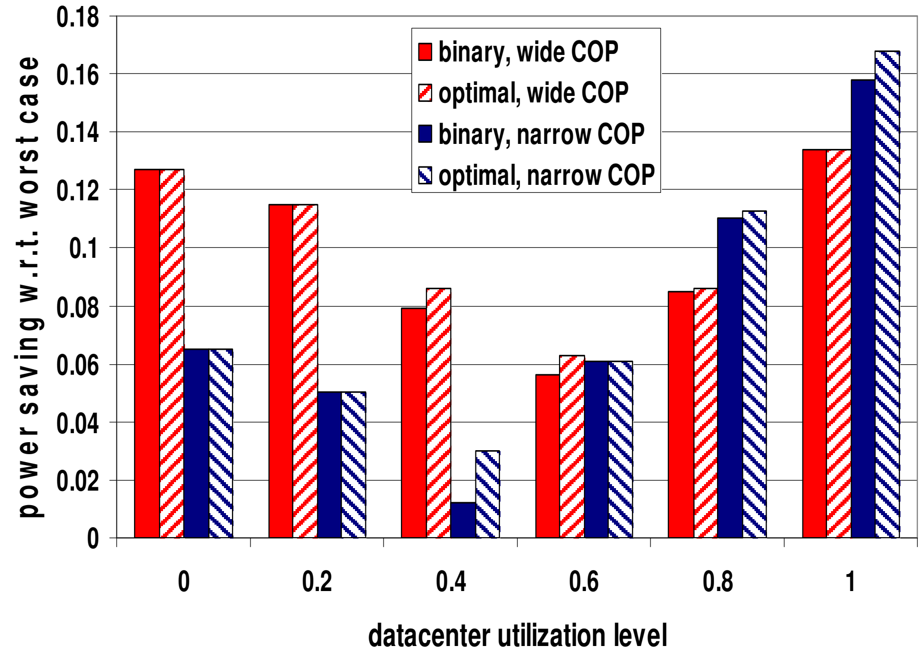
- 60% utilized DC cooling zone



TAPO-dc control method

- No single thermal setpoint is optimal
- Dynamically searching for optimal point is not tractable
  - Thermal mass, HVAC complexity
- **Binary control, based on utilization level**





Conclusions and Ongoing work

- Finding the right thermal setpoint helps save total system power, <u>without performance hit</u>
  - TAPO-server and TAPO-dc
- Ongoing work
  - Prototype TAPO-dc in a real data center
  - Make TAPO-server converge faster
  - Understand the delicate interactions among the two techniques
    - Warmer ambient from TAPO-dc makes TAPO-server more valuable
    - TAPO-server lowers server fan power, favoring TAPO-dc with warmer chiller setpoint to reduce HVAC power.
    - Reliability concerns of server components running at slightly hotter temperatures

Thank you. Questions?

More materials…

Overview

- It is a team work:
  - Austin: Wei Huang, Malcolm Allen-Ware, John Carter, Mootaz Elnozahy, Tom Keller, Charles Lefurgy, Jian Li, Karthick Rajamani, Juan Rubio
  - Watson: Hendrik Hamann
- **Objective**: Optimize power and/or performance of an entire system (e.g., server, DC), with explicit consideration of **cooling power**
- Hierarchical Techniques:
  - Server-level power (TAPO-server):
    - Fan power vs. leakage power
    - Goal: minimize aggregate fan+leakage power
    - Prototyped on a P7 HV32 server.
  - Datacenter-level power (TAPO-dc):
    - HVAC power vs. server fan power
    - Goal: minimize aggregate HVAC+server power
  - **Server-level performance (TAPO-shift):**
    - Load imbalance in different cooling zones
    - State of the art can't fully exploit power shifting from an idle zone to an active zone, due to thermal limitations
    - Goal: maximize active zone performance, within power and thermal budgets

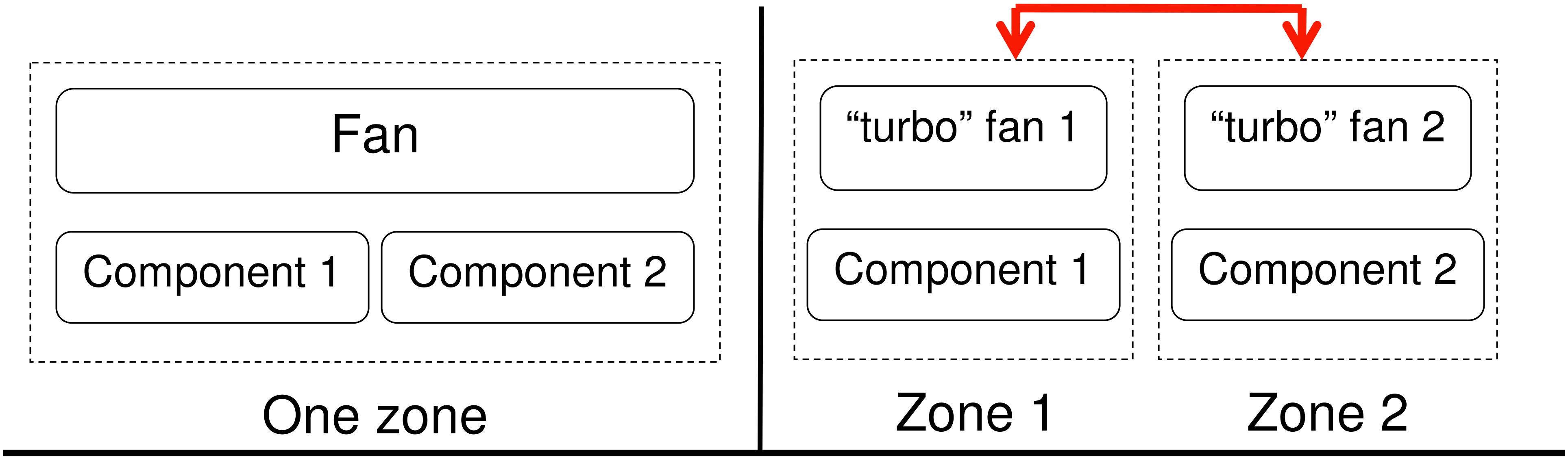
TAPPO-shift

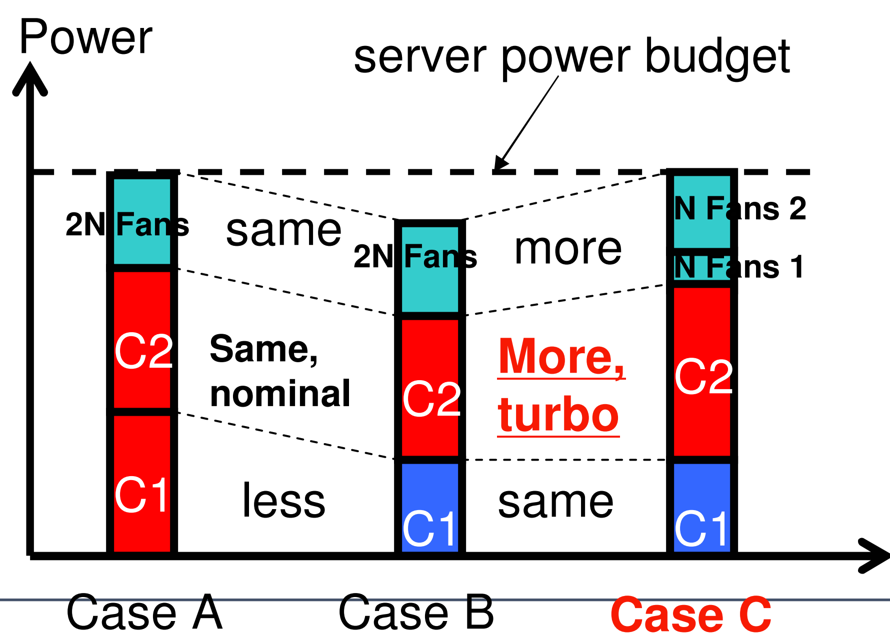
- Power shifting:
  - Idea: Shift unused power budget in underutilized parts to boost performance of highly utilized parts
  - Total power constraint, thermal constraint
  - Shifting among cooling zones. Example: socket to socket, server to server, rack to rack, DC zone to DC zone, etc
- Limitations:
  - Each cooling zone is design independently, without cooling capability for significantly more power
  - On the other hand, server processors can be overclocked by ~25% above nominal – hard to achieve in reality due to thermal limits

TAPPO-shift

- Solution: over-provisioned cooling capacity (by a large margin) in each cooling zone
- Cost is small: better/more fans
- Benefit: higher performance (e.g. processor can run at much higher frequency with shifted power)
- Within the same overall power budget across cooling zones, no thermal violation

TAPPO-shift: an illustrative example





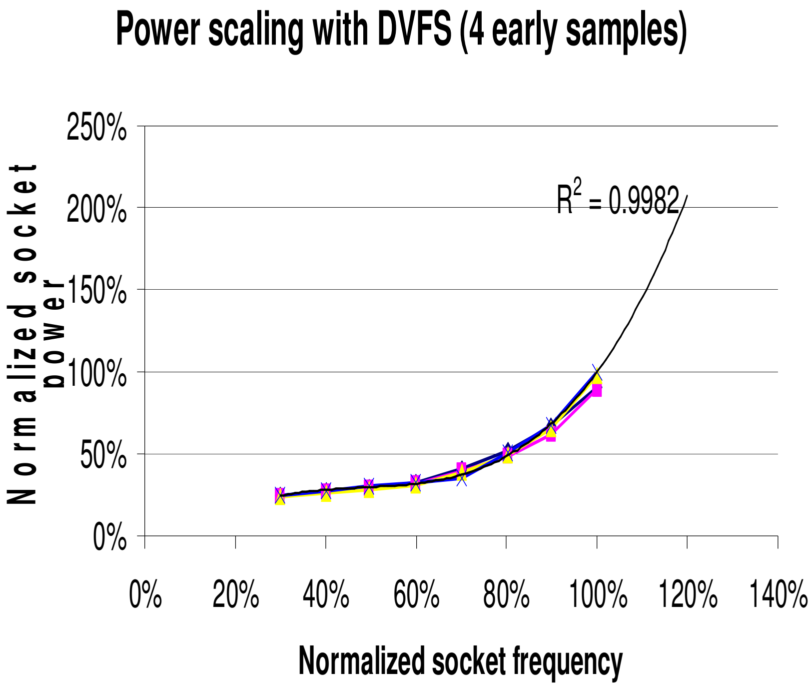
Case A: one zone, balanced, fully loaded

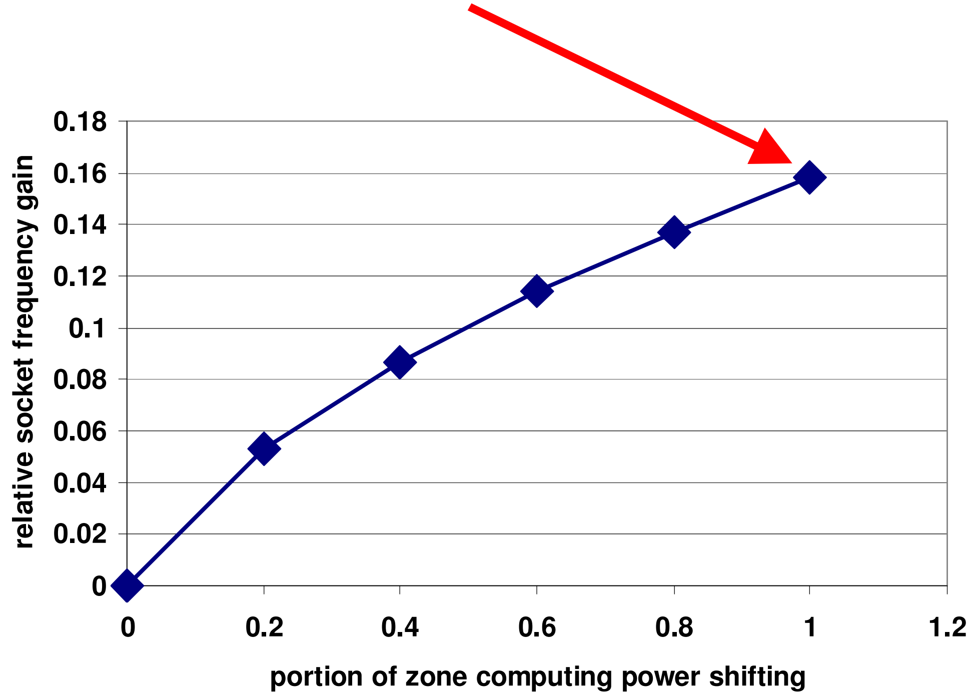
Case B: one zone, unbalanced

Case C: two zones, unbalanced, shifting

TAPPO-shift results

- Use P7 power-frequency relationship (cubic)
- Use P7 HV32 system power and fan power (almost cubical to rpm)
- 4 sockets divided into two cooling zones (each has separate fan control and better fans)
- Potentially 16% higher than P7 Turbo frequency





Combined TAPPO techniques – qualitative example

- Two DC cooling zones, Zone1 is 80% utilized, Zone2 is 10% utilized
- Workload migration to make Zone2 idle
- Observations:
  - Migration itself does not save power, but turning off idle zone does!
  - TAPPO-dc and -server can save about 9% power in this example
  - Combined with TAPPO-shift, can boost active zone utilization by 10% with about the same power