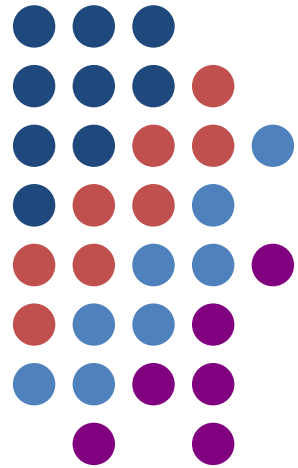
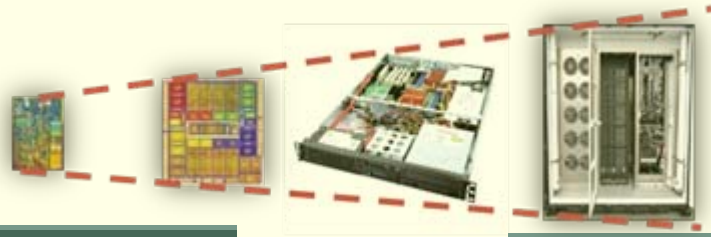
# *Cool and Save*: Cooling Aware Dynamic Workload Scheduling in Multi-socket CPU Systems

**Raid Ayoub  Tajana Simunic Rosing**

**Computer Science and Engineering Department**
**UC San Diego**



# Thermal stress challenges



Market demand for high integration

Market demand for high performance

**High power density**

**High temp**

☹ **Reliability**

- (10-15) °C ↑ → 1/2 MTTF (MTTF reduction is exponential)
- Skew problem: (15-20) °C spatial variations

☹ **Leakage power**

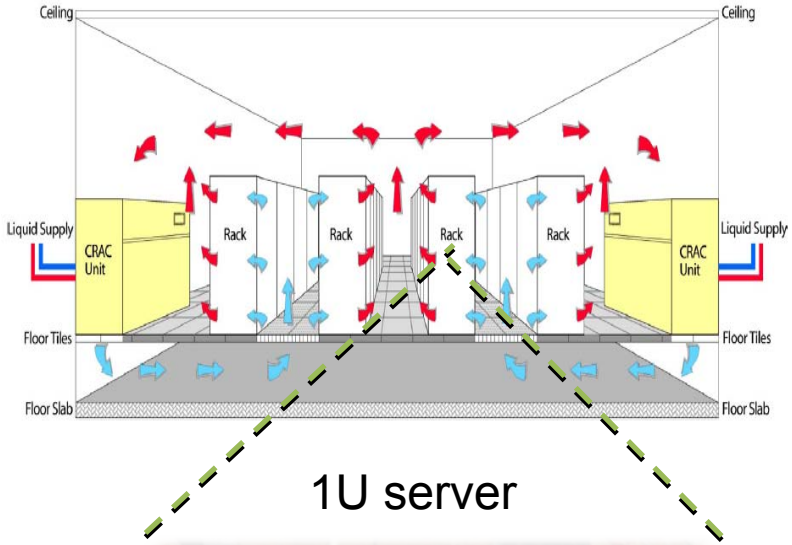
- Increases exponentially with temperature
- 30-40% of total power (65nm)

☹ **Performance**

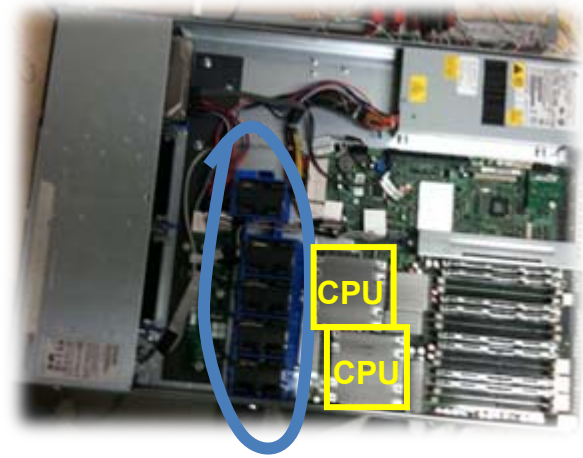
- Wire delay: 20°C ↑ → 6% delay

# Cooling subsystem challenges

- Use cooling systems to maintain reliability
- ***Challenges:***
  - Cooling must withstand the max temperature
- Cooling subsystem consumes 44% of total data centers energy [Scheihing 08]
  - Servers fan subsystem consumes large energy (up to **50%**) of total server power [Lefurgy 03]
  - We focus on a single machine




1U server

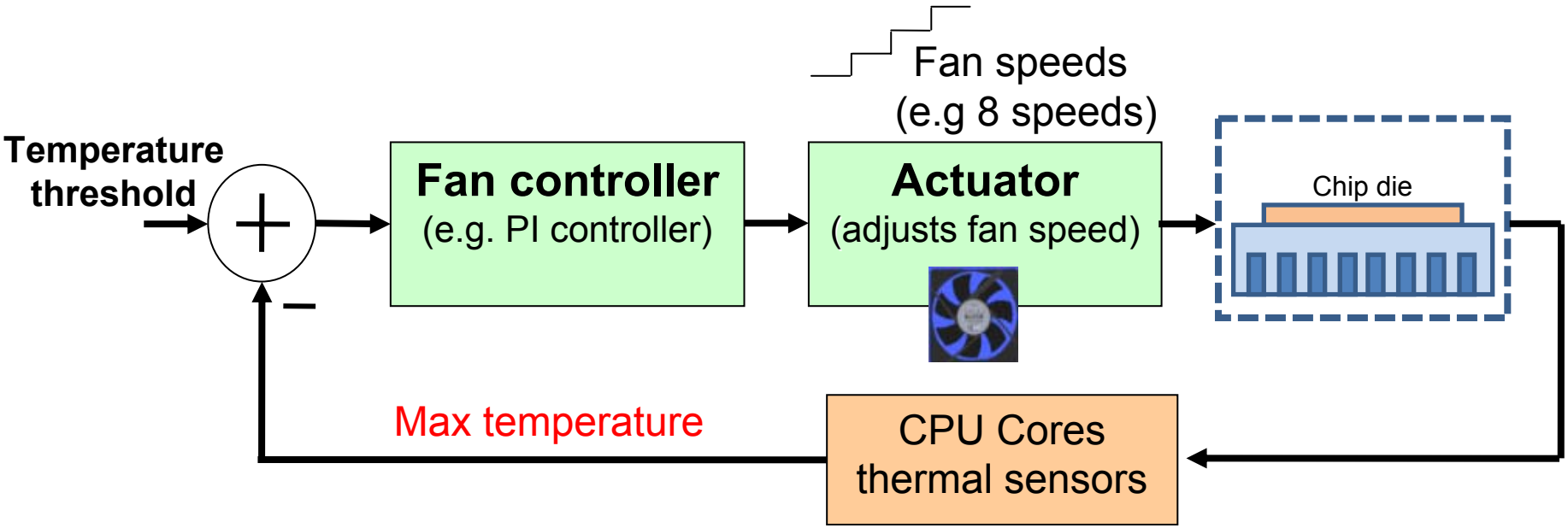



**Fan subsystem**

**P. Scheihing .** U.S. Department of Energy Energy Efficiency and Renewable Energy, 2008

**C. Lefurgy** et al. Energy management for commercial servers *IEEE Computer, pages 39–48, 2003.*

# Fan controller




- Efficient fan controller can be constructed based on *control theory*

☹ Traditionally, cooling optimizations focus only on the fan controller without including workload management

# Cooling aware workload scheduling

☹ State of the art ***load balancing*** algorithms in operating systems do not consider cooling costs

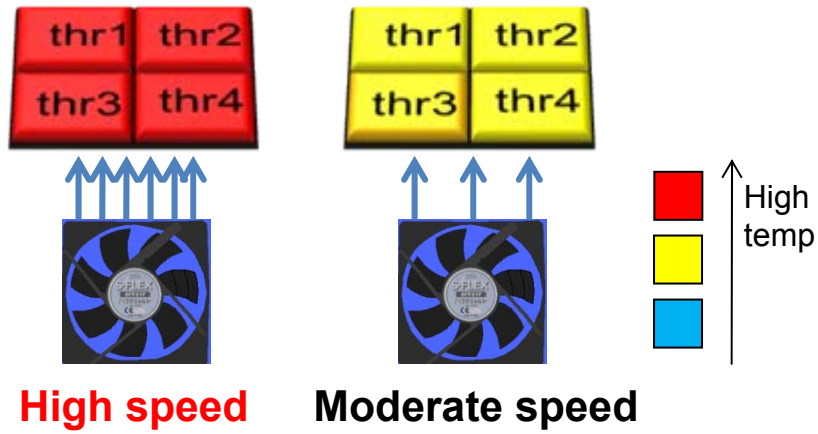
Assume a case of two sockets where one runs intensive workload while the other executes moderately active workload

😐 The workload is balanced from the performance point of view

- ☹ **One fan runs at high speed while the other at moderate speed**
- ➢ **Sources of inefficiency in cooling costs?**

**fan power ~ (fan speed)$^3$** **Nonlinear relation**

**How and when to schedule the workload to minimize the cooling costs?**

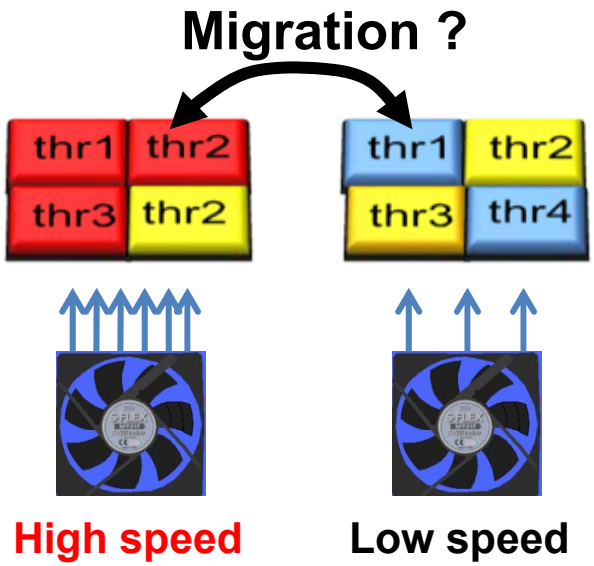


# Handling cooling inefficiency

- **Migrate computations to minimize the cooling cost**

☺ **Migration overhead is acceptable:**

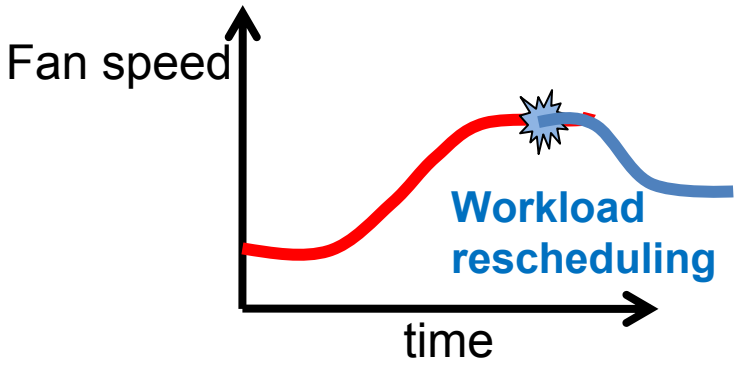
➢Heat sink *temperature time constant* (10*s* of *seconds*) **>>** *migration latency* ((10s – 100s) *of μs* )

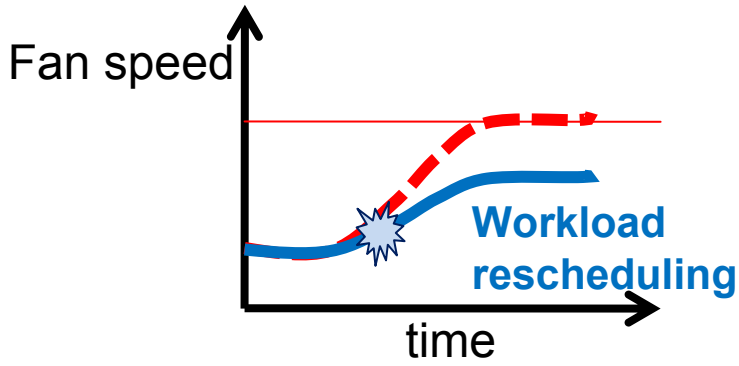



**Challenge:**
**Which threads to migrate?**

**Thermal and cooling model is needed to estimate the benefits of workload reassignment**

# Triggering workload rescheduling




**Reactive approach**
- ☹ **Lowers cooling savings**
- ☹ **Cannot minimize the noise level**
- ☹ **Impacts stability**

**Predictive approach:**
- ☺ **Improves energy**
- ☺ **Lowers noise level**
- ☺ **Provides better stability**

**Challenge:** Design of efficient proactive dynamic cooling aware workload management technique

# Related work and our contribution

❑ **CPU Fan control [Chiueh 00, Wang 2009]**

➢Techniques have been focused on improving the fan controller

☹ No cooling aware workload management

❑ **Improving cooling efficiency in data centers [Moore 05, Tang 2007]**

➢Workload scheduling is used to lower cooling costs by reducing air recirculation problems across the data center racks

☹ Applicable only at the rack level

**Our contribution:**

**A cooling aware workload management strategy**

**capable of 73% energy savings on average**


H. Chiueh, et al. A Novel Fully Integrated Fan Controller for Advanced Computer Systems. *SSMSD, 2000.*

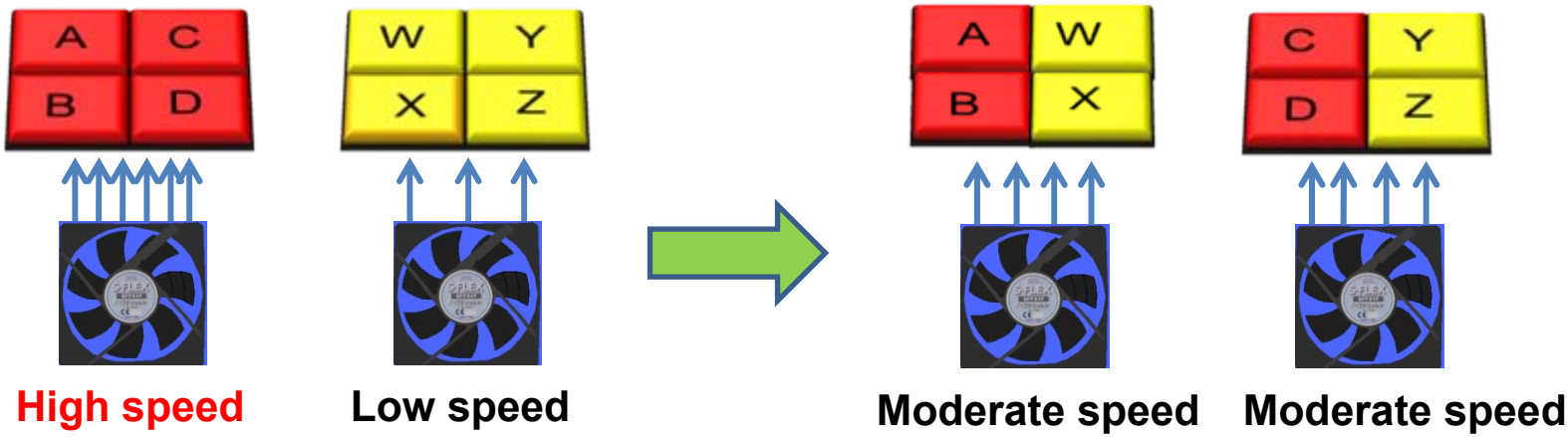
Z. Wang, et al. Optimal fan speed control for thermal management of servers. *Proc. IPAC, 2009.*

J. Moore et al. Making scheduling ”cool”: temperature-aware workload placement in data centers. *USENIX, 2005.*

Q. Tang et al. Thermal-aware task scheduling for data centers through minimizing heat recirculation *ICCC, 2007.*

# Cooling savings: spreading jobs

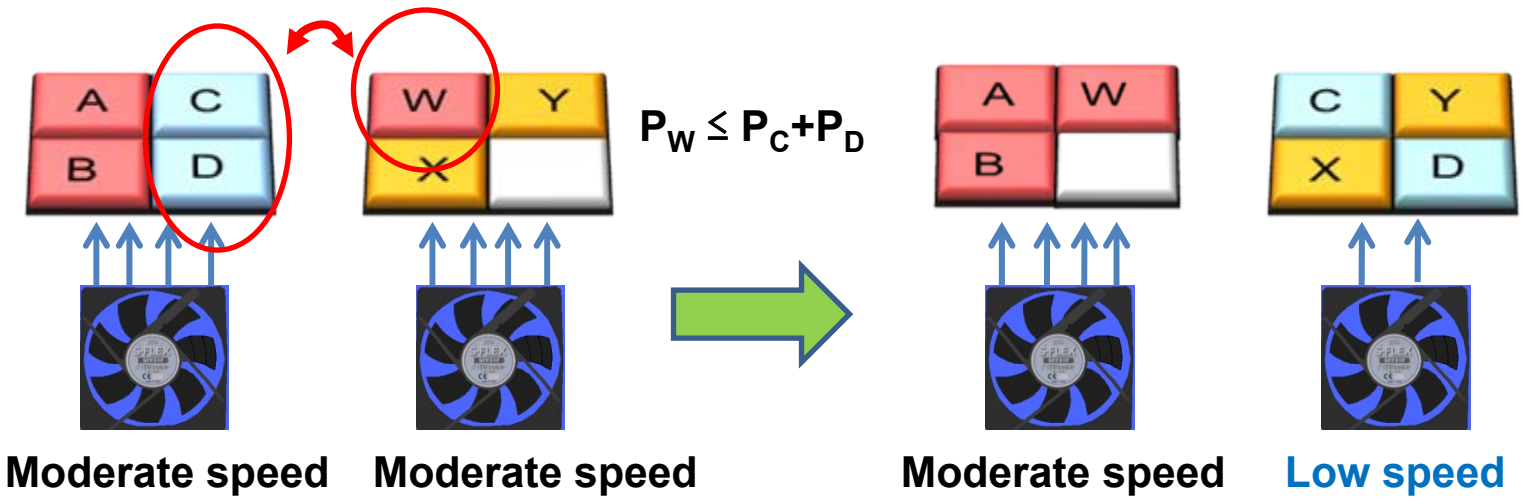
- Fan speed can be reduced by **creating a better temperature distribution**
  - Migrate some of the active threads from the sockets with high fan speed to sockets with lower speed
  - Swap some of the hot threads from sockets with high fan speed with colder threads from sockets with lower speed.



# Cooling savings: consolidating jobs

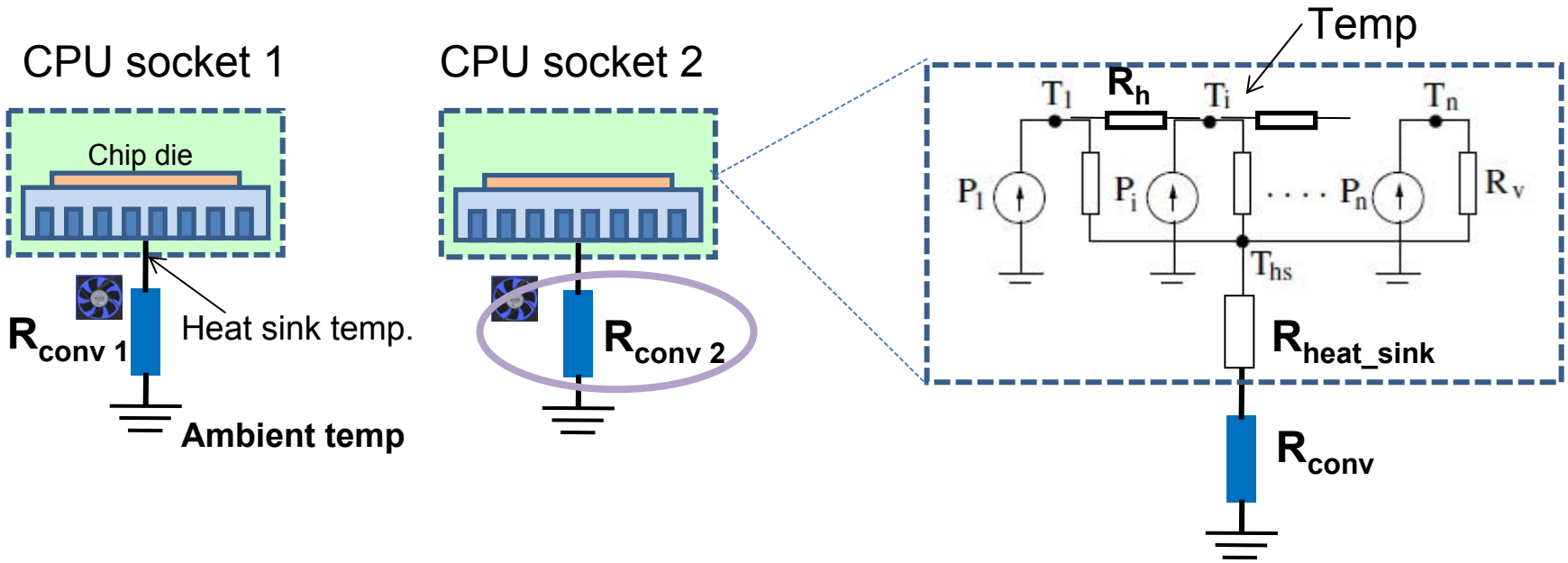
Concentrate the hot threads into fewer sockets while keeping the fan at the same speed:

- Migrate hot threads to the socket with fan that is spinning higher than is required
- If *Fan speed*$_M \geq$ *Fan speed*$_N$, we can swap the hot thread from socket ***N*** with colder threads from socket ***M***



# On-line cooling & thermal model

- Needed to evaluate cooling aware job reassignments




- Thermal model of fan uses single resistance $R_{conv}$ per socket
- $R_h$ (horizontal resistor) can be neglected since it is *much larger than* $R_v$ (vertical resistor)

**☺ Very low overhead**

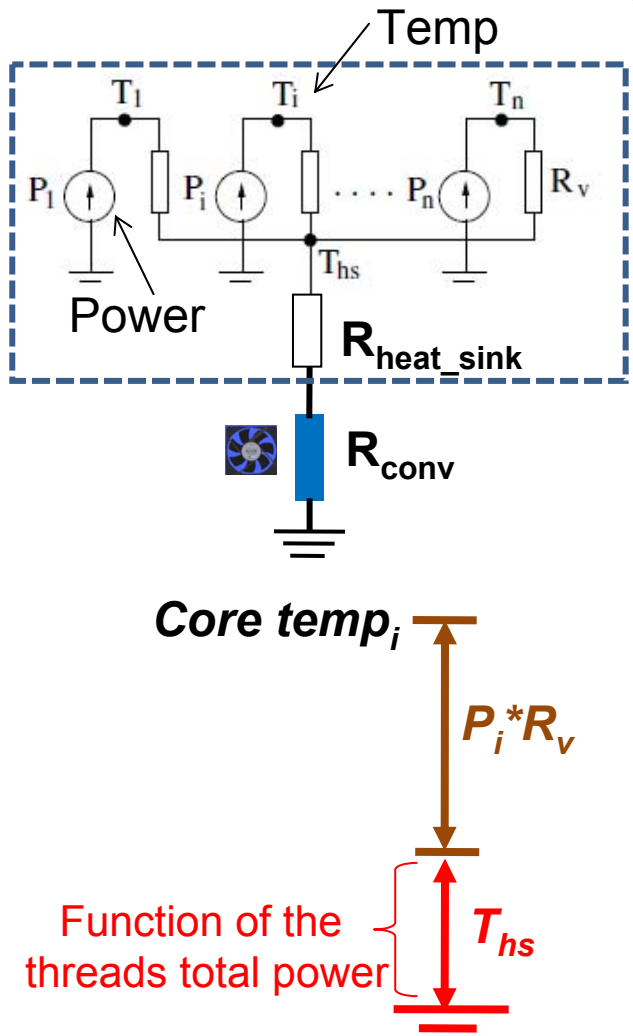
# Impact of co-located threads on cooling cost

- **Core temp$_i$ = $P_i$*$R_v$ + $T_{hs}$**
- To calculate $T_{hs}$ *we use superposition theory*:

$$\Delta T_i = P_i (R_{hs} + R_{conv})$$

$$T_{hs} = \sum \Delta T_i$$

Each thread increases the $T_{hs}$ *by* $\Delta T$ that is proportional to its power consumption



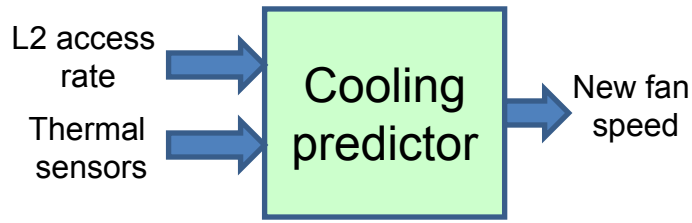
# Cooling cost estimation

- Model to predict the steady-state cooling cost of rescheduling

$$Fan\ speed^{new} \sim max(\{core\ temp_1^{new}, core\ temp_2^{new}, \ldots\})$$

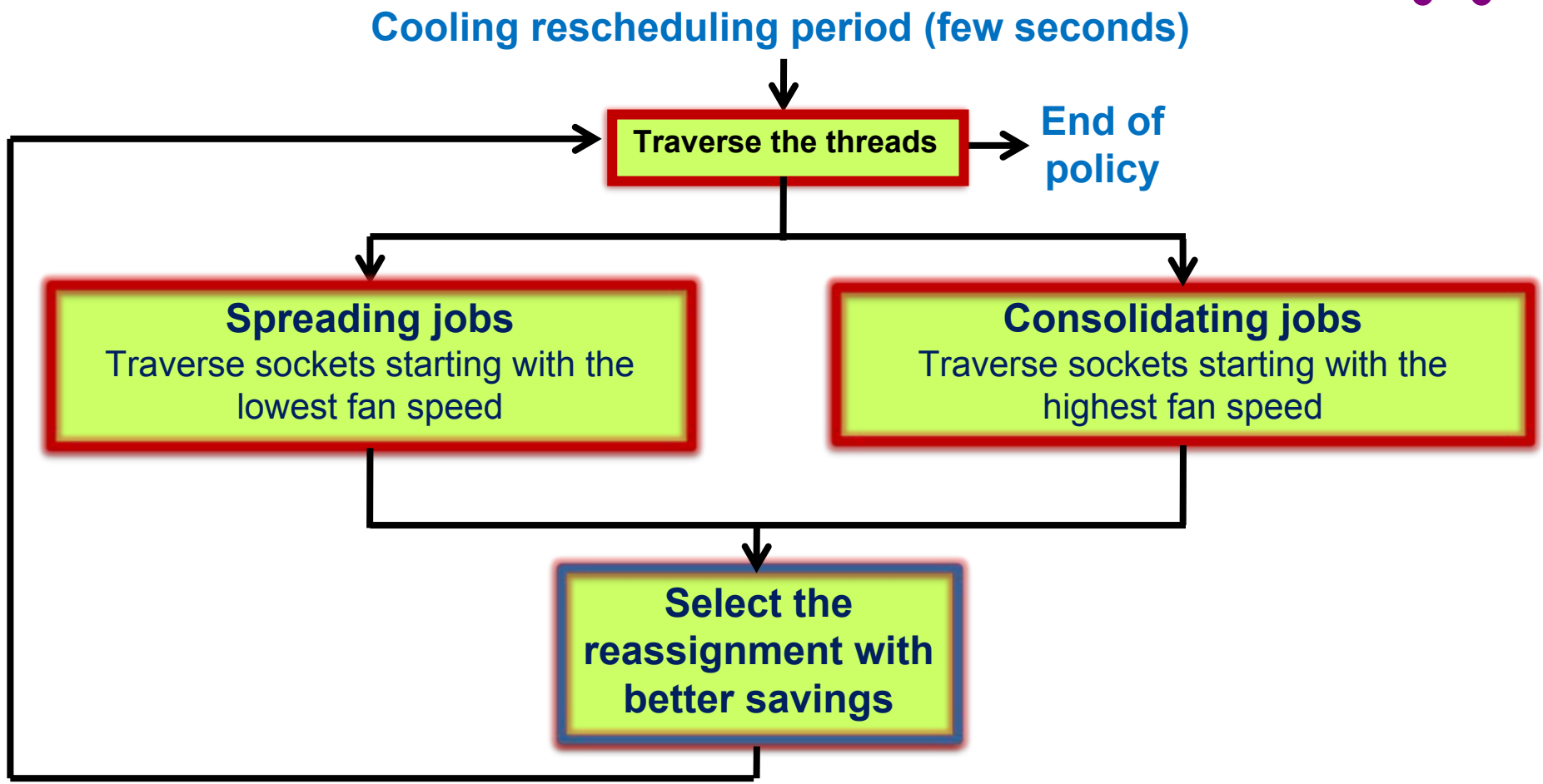
$$Core\ temp^{new} = P_i * R_v + T_{hs}^{new}$$

$$T_{hs}^{new} = f(\ workload\ induced\ power^{new}\ )$$




**Thread power consumption:**
**Core power**: Convert thermal sensor values into power estimates using our model
**L2 cache**: Performance counters to estimate power

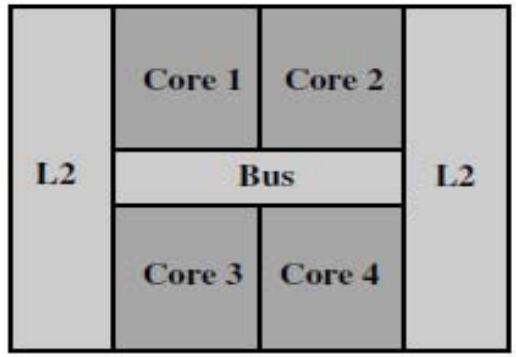
# Scheduling policy



# Experimental setup

**Platform:** Two sockets, each has 4 cores similar to EV6

**Simulators:** M5 Microarchitectural simulator [Binkert 06]
Wattch: Power simulator [Brooks 00]
HotSpot: Thermal simulator [Skadron 03]




**Technology and CPU packaging**

| | |
|---|---|
| Technology | 65nm |
| Critical temperature | 85°C |
| Processor clock | 2GHz |
| Core area | 3.3mm x 3.3mm |
| Die thickness | 0.2mm |
| Heat spreader | 20mmx20mmx1mm |

**Thermal packaging and fan**

| | |
|---|---|
| Heat sink (R_conv) | 0.18 °C/W at 23 CFM |
| Fan air flow | 23 CFM/socket |
| Fan control | Simple closed loop control |


**N. Binkert**, et al. The m5 simulator: Modeling networked systems. IEEE Micro, 26(4), 2006.
**D. Brooks**, et al. Wattch: A framework for architectural-level power analysis and optimisations, ISCA 2000.
**K. Skadron**, et al. Temperature-aware microarchitecture, ISCA 2003.

# Experimental setup

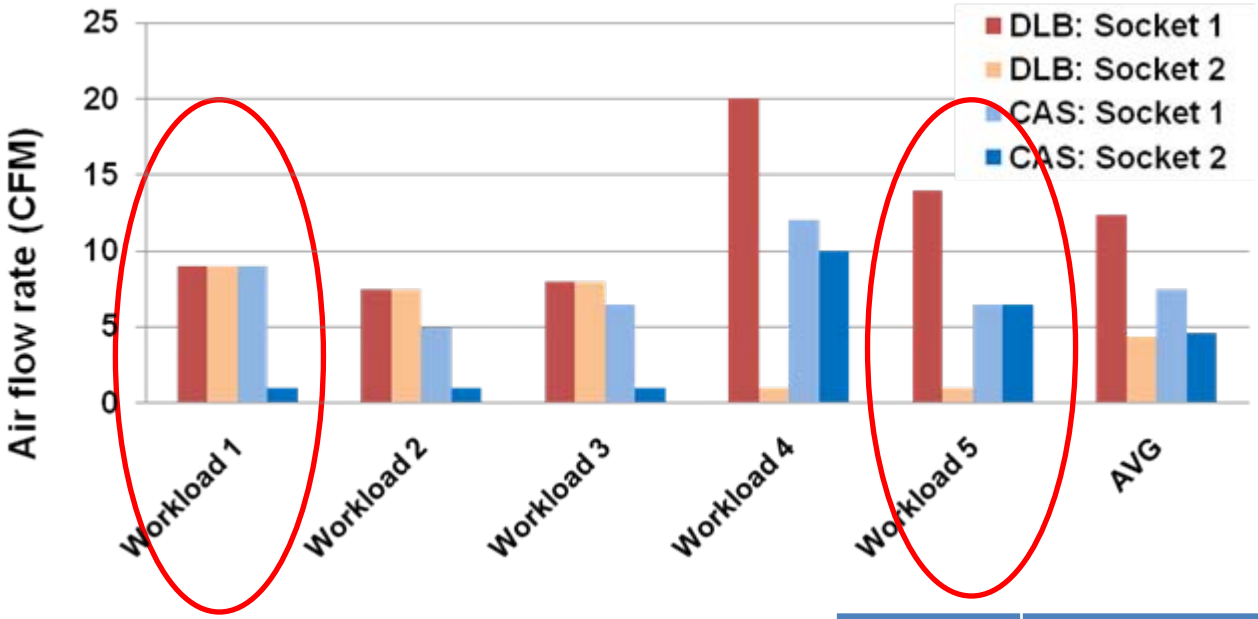
- Benchmark combinations from **SPEC 2000**
  - Longer traces obtained by replicating 5sec traces

| Benchmark | Avg. power | Standard deviation |
|---|---|---|
| **gcc** | 4.12 | 1.8 |
| **gzip** | 5.51 | 1.65 |
| **swim** | 6.35 | 1.77 |
| **bzip2** | 7.9 | 2.25 |
| **crafty** | 9.11 | 0.66 |

| Workload | Benchmarks (socket 1/socket 2) |
|---|---|
| 1 | {crafty+gzip+gcc}/{crafty+gzip+gcc} |
| 2 | {bzip2+gzip+swim+swim}/{crafty+gzip+gcc} |
| 3 | {bzip2+swim+gcc+gcc}/{bzip2+swim+gzip+gzip} |
| 4 | {crafty+crafty+swim}/{gzip+swim+swim} |
| 5 | {bzip2+bzip2+bzip2+bzip2}/{gcc+gcc+gcc+gcc} |

**Polices:**
**CAS (Cooling aware scheduling):** Proposed policy
**DLB (Dynamic load balancing):** Minimizes the differences in task queue lengths

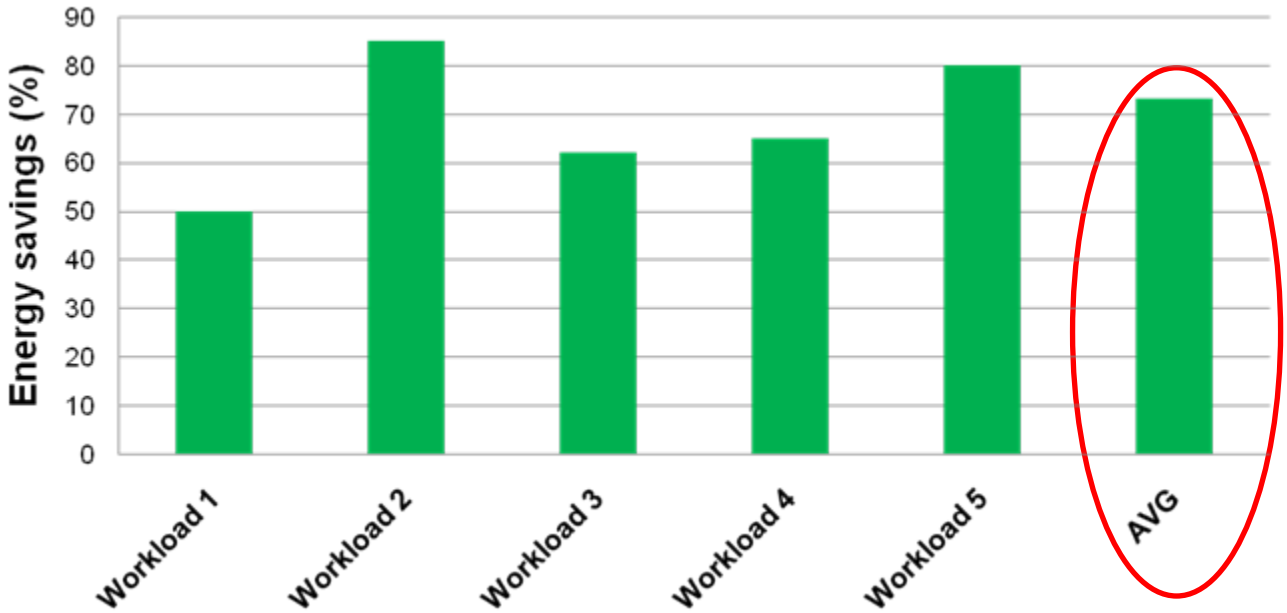
# Lowering air flow (CFM)




| Workload | Benchmarks (socket 1/socket 2) |
|---|---|
| 1 | {crafty+gzip+gcc}/{crafty+gzip+gcc} |
| 2 | {bzip2+gzip+swim+swim}/{crafty+gzip+gcc} |
| 3 | {bzip2+swim+gcc+gcc}/{bzip2+swim+gzip+gzip} |
| 4 | {crafty+crafty+swim}/{gzip+swim+swim} |
| 5 | {bzip2+bzip2+bzip2+bzip2}/{gcc+gcc+gcc+gcc} |

- **Average reduction in max fan speed is 35%**

# Energy savings




| Workload | Benchmarks (socket 1/socket 2) |
|---|---|
| 1 | {crafty+gzip+gcc}/{crafty+gzip+gcc} |
| 2 | {bzip2+gzip+swim+swim}/{crafty+gzip+gcc} |
| 3 | {bzip2+swim+gcc+gcc}/{bzip2+swim+gzip+gzip} |
| 4 | {crafty+crafty+swim}/{gzip+swim+swim} |
| 5 | {bzip2+bzip2+bzip2+bzip2}/{gcc+gcc+gcc+gcc} |

- **Average cooling energy savings of 73%**

# Summary

- Cooling aware workload scheduling is an efficient way to lower the cooling costs
- Our approach is scalable and not limited to a specific platform
- Rescheduling overhead is below 1%

- Average cooling energy savings is **73%**
- With average reduction in max fan speed of **35%**

# Backup slides

# Cooling cost non-linearity

- $R_{conv} \approx \alpha (FS)^{-1}$ FS: Fan speed
- $FP_2/FP_1 \approx (R_{conv1}/R_{conv2})^3$ FP: Fan power
- Reduction in max temp. is proportional to the reduction in $R_{conv2}$
  - **Cubic increase in cooling cost**
- Existing of such Non-linearity open opportunities for cooling savings