# PCFS: A Power Credit based Fair Scheduler under DVFS for Multi-core Virtualization Platform

Chengjian Wen, Jun He, Jiong Zhang, Xiang Long

School of Computer Science and Technology
Beihang University
Beijing 100191, P.R. China

2010 IEEE/ACM International Conference on Green Computing and Communications
2010 IEEE/ACM International Conference on Cyber, Physical and Social Computing

# Outline

- Introduction
- Related Work
- Design Goals for Power Efficient Scheduler
- Design and Implementation
- Evaluation
- Conclusions

# Introduction

- The necessity of power efficiency management on multicore virtualization platform has become increasingly evident.

- How to make further use of multicore virtualization platform for better performance per watt becomes a focus of green computing.

- However, most current approaches on energy management are developed for standard, legacy OSes and not suitable for virtual machine monitor (VMM).

- In order to integrate DVFS policy to a hypervisor scheduler, we propose the approach of Power Credit based Fair scheduler (PCFS for short).

# Related Work

1. Power management
   - DVFS
   - Linux CPU Governor Policy
   - OpenSolaris Project Tesla: CPU Power Management

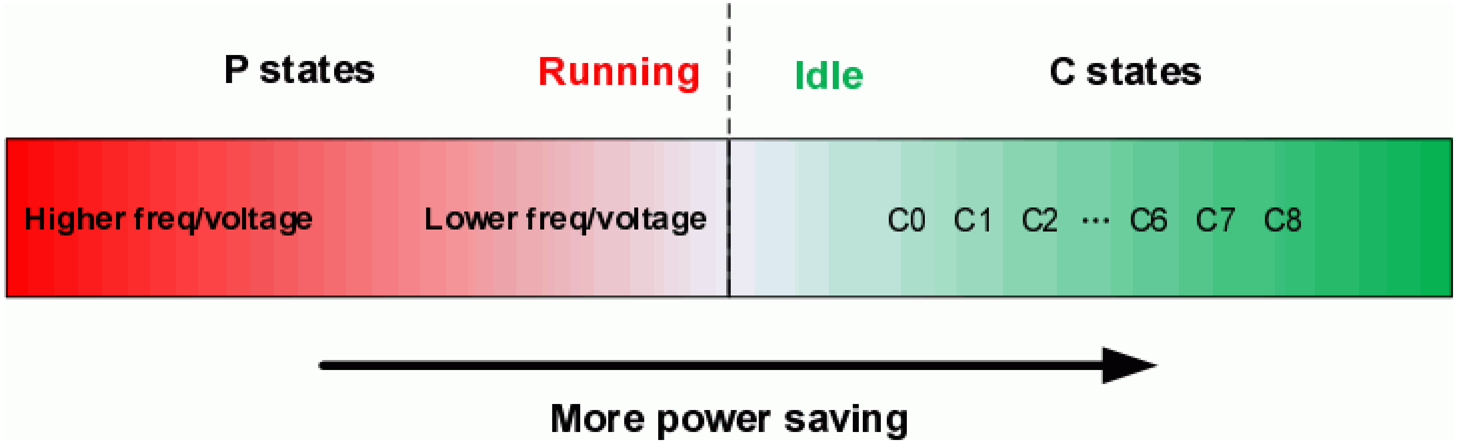
2. Online performance prediction and energy model

3. Combine the virtualization and power management
   - Intel Open Source Technology Center: Xenpm

# Design Goals for Power Efficient Scheduler (1/5)

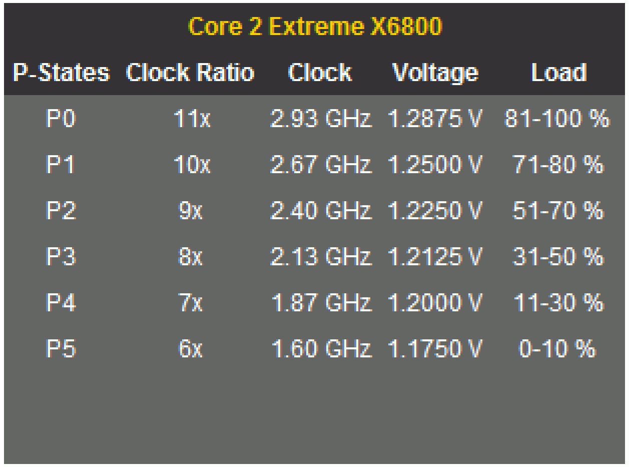
- In the virtualization environment, the basic unit of scheduling is VCPU. The role of scheduler is to assign physical processor cores to each VCPU.
- General processors have multiple P-states and C-states. The higher the frequency and voltage, the more energy will be consumed.

$$P \propto fV^2$$

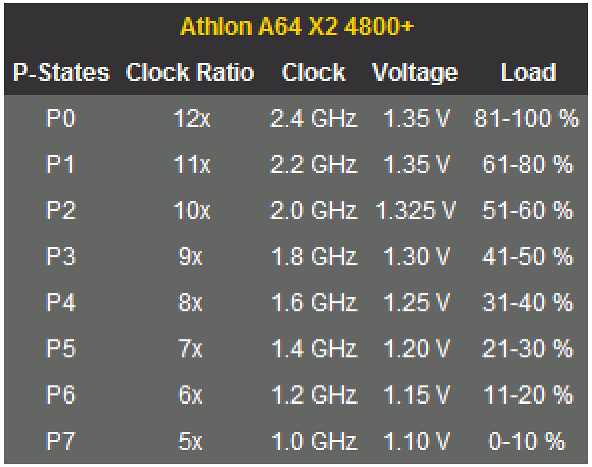


# Design Goals for Power Efficient Scheduler (2/5)

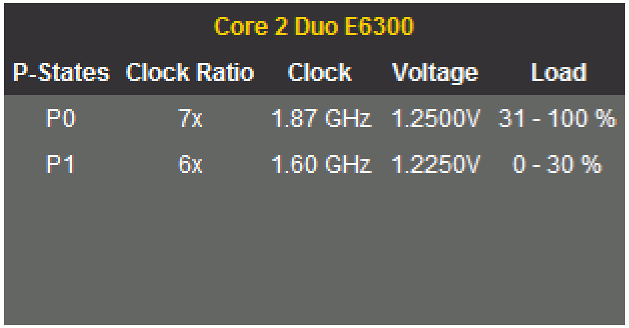
- P-states and C-states Examples



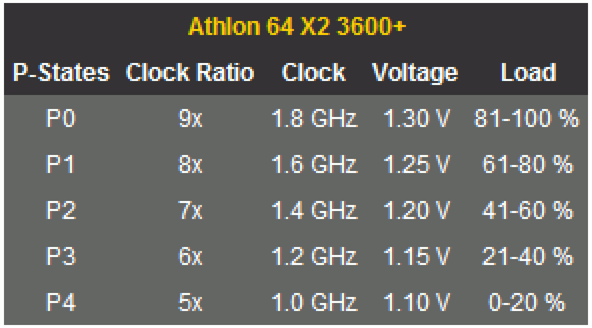
| Core 2 Extreme X6800 | | | | |
|---|---|---|---|---|
| P-States | Clock Ratio | Clock | Voltage | Load |
| P0 | 11x | 2.93 GHz | 1.2875 V | 81-100 % |
| P1 | 10x | 2.67 GHz | 1.2500 V | 71-80 % |
| P2 | 9x | 2.40 GHz | 1.2250 V | 51-70 % |
| P3 | 8x | 2.13 GHz | 1.2125 V | 31-50 % |
| P4 | 7x | 1.87 GHz | 1.2000 V | 11-30 % |
| P5 | 6x | 1.60 GHz | 1.1750 V | 0-10 % |



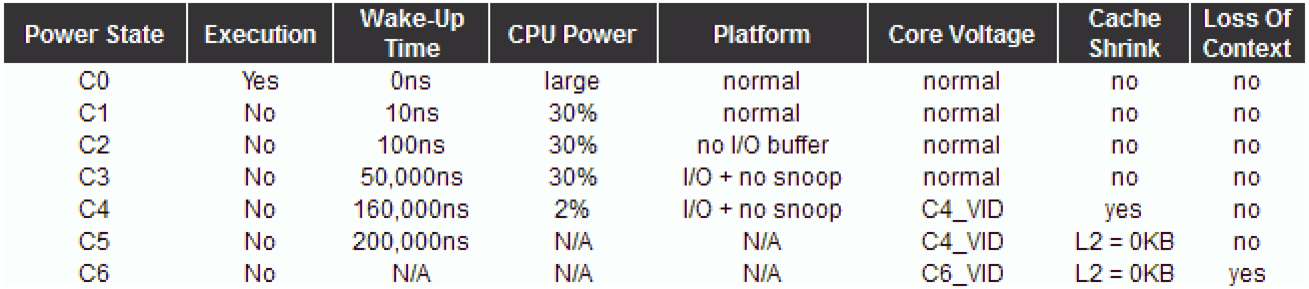
| Athlon A64 X2 4800+ | | | | |
|---|---|---|---|---|
| P-States | Clock Ratio | Clock | Voltage | Load |
| P0 | 12x | 2.4 GHz | 1.35 V | 81-100 % |
| P1 | 11x | 2.2 GHz | 1.35 V | 61-80 % |
| P2 | 10x | 2.0 GHz | 1.325 V | 51-60 % |
| P3 | 9x | 1.8 GHz | 1.30 V | 41-50 % |
| P4 | 8x | 1.6 GHz | 1.25 V | 31-40 % |
| P5 | 7x | 1.4 GHz | 1.20 V | 21-30 % |
| P6 | 6x | 1.2 GHz | 1.15 V | 11-20 % |
| P7 | 5x | 1.0 GHz | 1.10 V | 0-10 % |



| Core 2 Duo E6300 | | | | |
|---|---|---|---|---|
| P-States | Clock Ratio | Clock | Voltage | Load |
| P0 | 7x | 1.87 GHz | 1.2500V | 31 - 100 % |
| P1 | 6x | 1.60 GHz | 1.2250V | 0 - 30 % |



| Athlon 64 X2 3600+ | | | | |
|---|---|---|---|---|
| P-States | Clock Ratio | Clock | Voltage | Load |
| P0 | 9x | 1.8 GHz | 1.30 V | 81-100 % |
| P1 | 8x | 1.6 GHz | 1.25 V | 61-80 % |
| P2 | 7x | 1.4 GHz | 1.20 V | 41-60 % |
| P3 | 6x | 1.2 GHz | 1.15 V | 21-40 % |
| P4 | 5x | 1.0 GHz | 1.10 V | 0-20 % |



| Power State | Execution | Wake-Up Time | CPU Power | Platform | Core Voltage | Cache Shrink | Loss Of Context |
|---|---|---|---|---|---|---|---|
| C0 | Yes | 0ns | large | normal | normal | no | no |
| C1 | No | 10ns | 30% | normal | normal | no | no |
| C2 | No | 100ns | 30% | no I/O buffer | normal | no | no |
| C3 | No | 50,000ns | 30% | I/O + no snoop | normal | no | no |
| C4 | No | 160,000ns | 2% | I/O + no snoop | C4_VID | yes | no |
| C5 | No | 200,000ns | N/A | N/A | C4_VID | L2 = 0KB | no |
| C6 | No | N/A | N/A | N/A | C6_VID | L2 = 0KB | yes |

Source: http://www.techarp.com/showarticle.aspx?artno=420

# Design Goals for Power Efficient Scheduler (3/5)

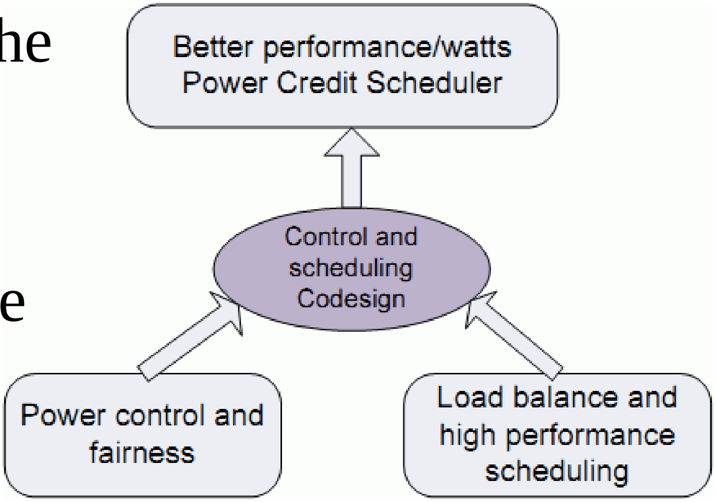
- Xen Credit Scheduler
  - 2 VCPU priority: OVER and UNDER fair share.
  - Period (30ms)
  - Every VM has a weight and cap.
  - Each core can provide 300 credit and these credit will be shared by all active VCPUs.
  - If a running VCPU does not have any runnable task, it will be blocked and leave the run queue.
  - VCPUs in the UNDER state are always run before those run in the OVER state.

# Design Goals for Power Efficient Scheduler (4/5)

- Xen Credit Scheduler
  - Shortage
    - Its design is oriented to whole system performance fairness between different VMs and load balance between processors.
    - Thus, the processor cores get little chance to enter a more power-saving state.
    - An efficient power-aware scheduler needs to know the full information about the topology of system. But Xen Credit Scheduler does not support it very well.

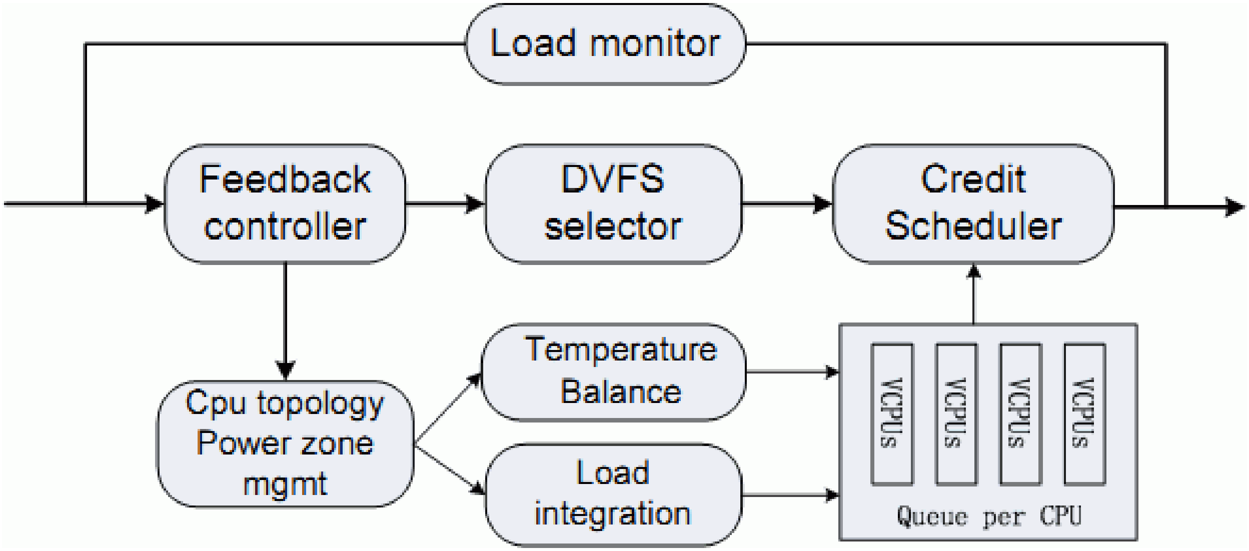
# Design Goals for Power Efficient Scheduler (5/5)

- Design principles:
  1. Take energy as a kind of system resource, and distinguish statistics of VCPU running time at different frequencies.
  2. Scheduler monitors the load of physical cores and select the frequency according to historical trends not only to the static mapping relations between load and frequency.
  3. Fairly schedule the guest OSes on cores at the higher frequency to meet the fairness requirements of guest OSes.
  4. Fairly schedule frequency levels on the same core to satisfy the physical core cooling limitations.


Better performance/watts
Power Credit Scheduler
Control and scheduling Codesign
Power control and fairness
Load balance and high performance scheduling

# Design and Implementation (1/5)

- The PCFS framework



  - Feedback controller
  - DVFS selector
  - CPU topology & Power zone management
  - Credit Scheduler

# Design and Implementation (2/5)

- Feedback controller
  - It is the entry of the entire scheduler.
  - It is in charge of choosing one action from DVFS selector and power zone management.
  - Judge whether one core has the load sample under 20% for 8 times
    - If yes, turn to power zone management.
      Migrate the load of this core to the other core running at highest frequency. Then set this core to a more power saving state.
    - If no, turn to DVFS selector.
      Choose the frequency which can keep CPU load on 90%.

# Design and Implementation (3/5)

- DVFS selector
  - Define the up-threshold: 95%
  - If the load is decreased under the threshold, we choose the frequency which can keep CPU load on 90%.

# Design and Implementation (4/5)

- CPU topology & Power zone management
  - Power zone is a set of processor cores which run with the same frequency and their workloads would be kept balanced by the scheduler of hypervisor.
  - Only when a power zone is idle, we’ll take further power saving measurement such as offline a core or whole physical package according to the topology.
  - When a temperature threshold at one core is reached, the associated power zone will search for lower frequency and substitute the overheated core.
  - When one core has the load sample under 20% for 8 times, it will migrate the load of this core to the other core running at highest frequency.

# Design and Implementation (5/5)

- Credit Scheduler
  - Power credit
    - Every CPU has P-states: $P_0, P_1, P_{\ldots}, P_{n-1}$ ($P_0$ is the highest frequency)
    - Frequency of core number i is denoted by $f_i$.
    - $Credit_{total} = \sum_{i=1}^{m} \frac{f_i}{p_0}$
  - Consumed by VCPU
    - $\lambda = \left(\frac{P_i}{P_0}\right)^2, i = 1,2,\ldots,m$
    - $Credit_{consumed} = \lambda * 100$
    - $Credit_{left} = Credit_p - Credit_{consumed}$

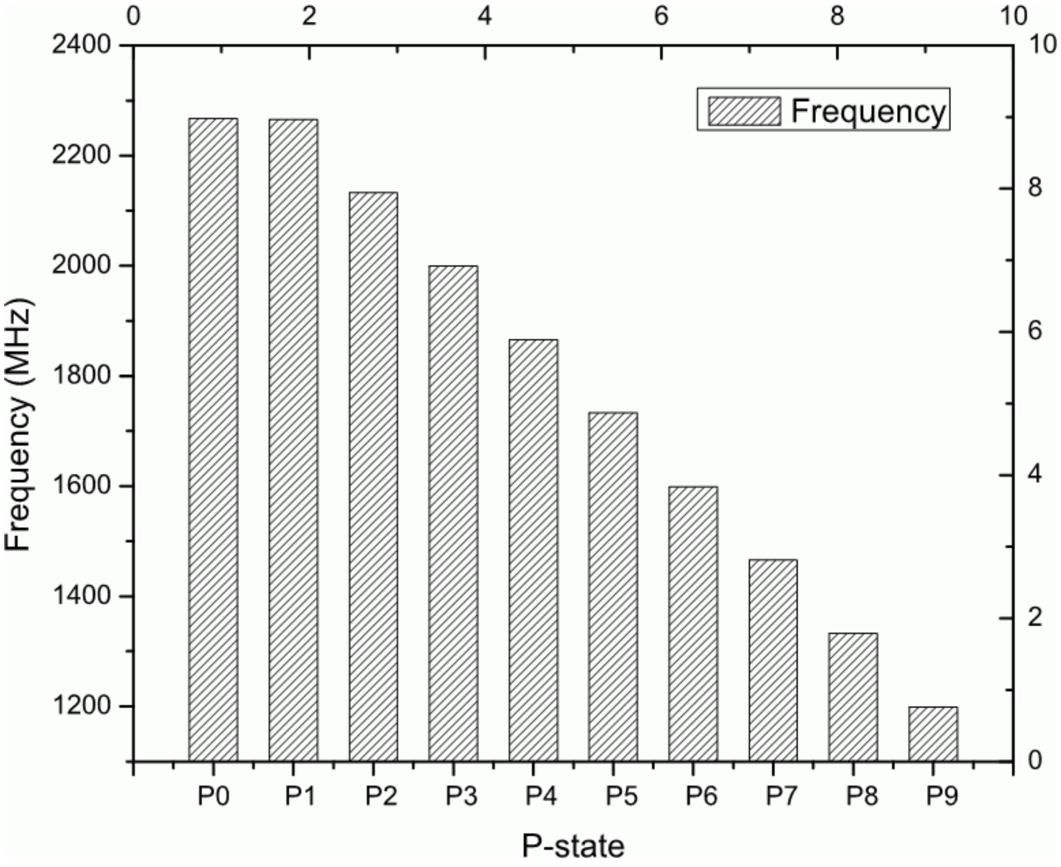
# Evaluation (1/7)

- Evaluation objectives:
  1. How effective PCFS can use DVFS as power saving measurement.
  2. Whether PCFS utilizes the variety of power saving measurements in the context of performance conservation.
  3. Whether PCFS can improve equal sharing capability of the high frequency core?
  4. Whether PCFS can schedule high frequency on cores to avoid some core become over heated?

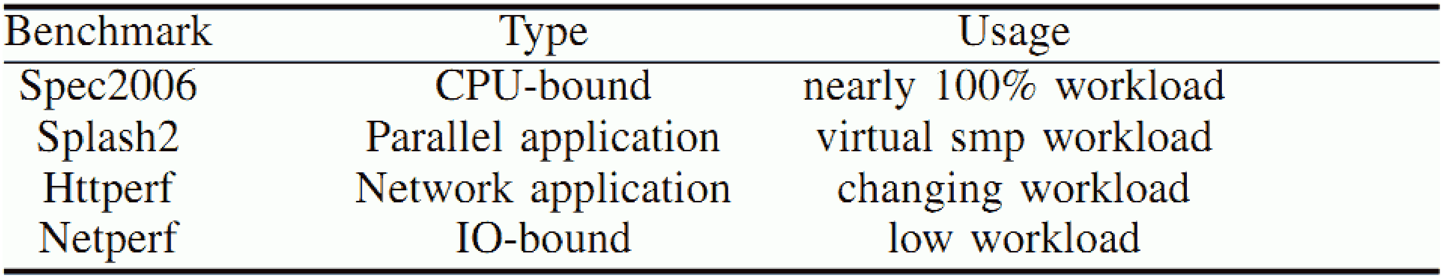
# Evaluation (2/7)

- CPU: Intel Core i5 Quad-core
  RAM: 4GB

  Each core supports 10 P-state and 4 C-state. The maximum frequency is 2.267Ghz, and the frequency step is 133MHz.

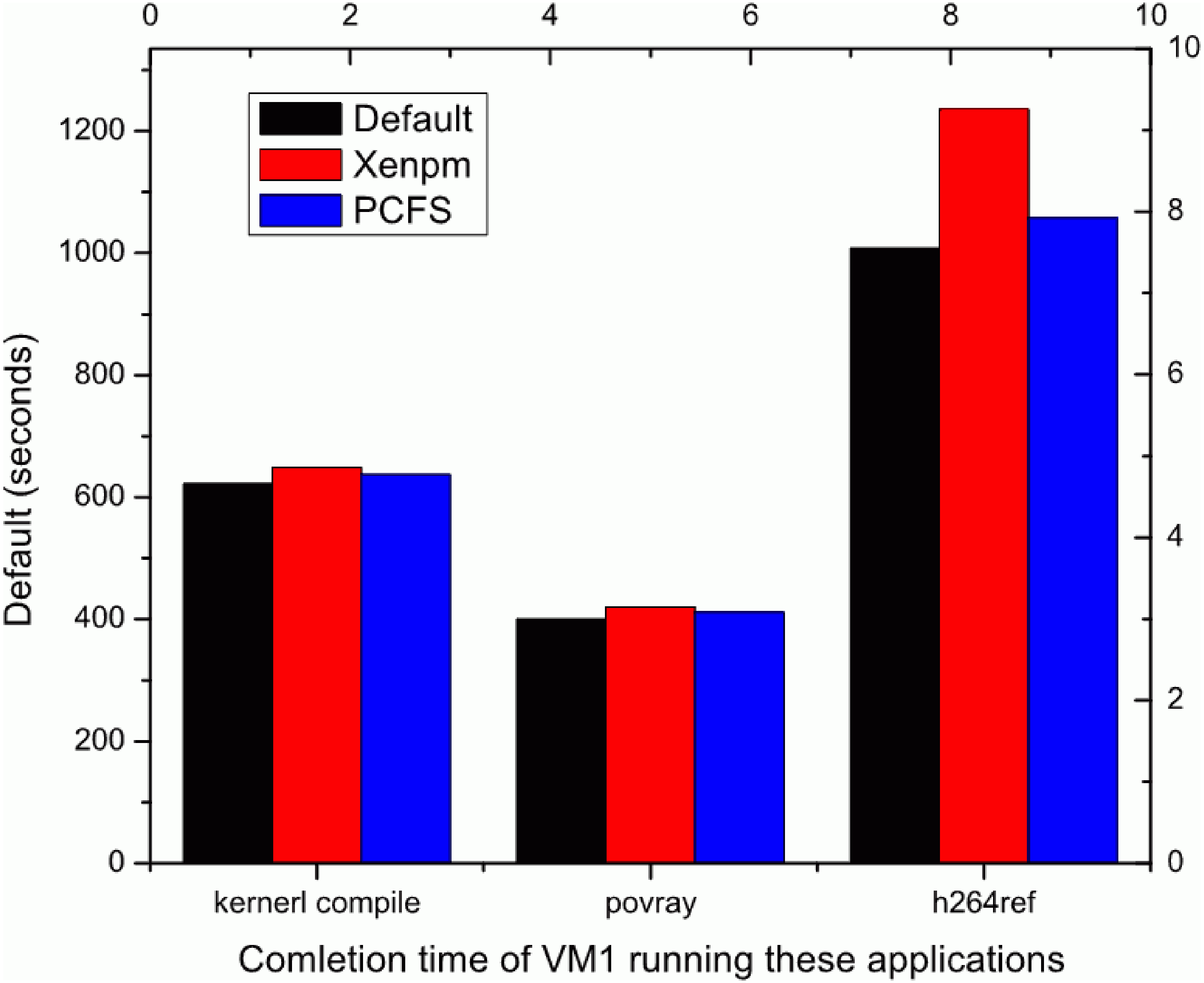


- The version of Xen hypervisor is 3.4.3 while domain0 and domainU both use Linux 2.6.32-5 kernel.
- The virtualized guest operating system is Debian 5.0 squeeze.
- Three kinds of applications are selected as benchmarks workload which are listed below.

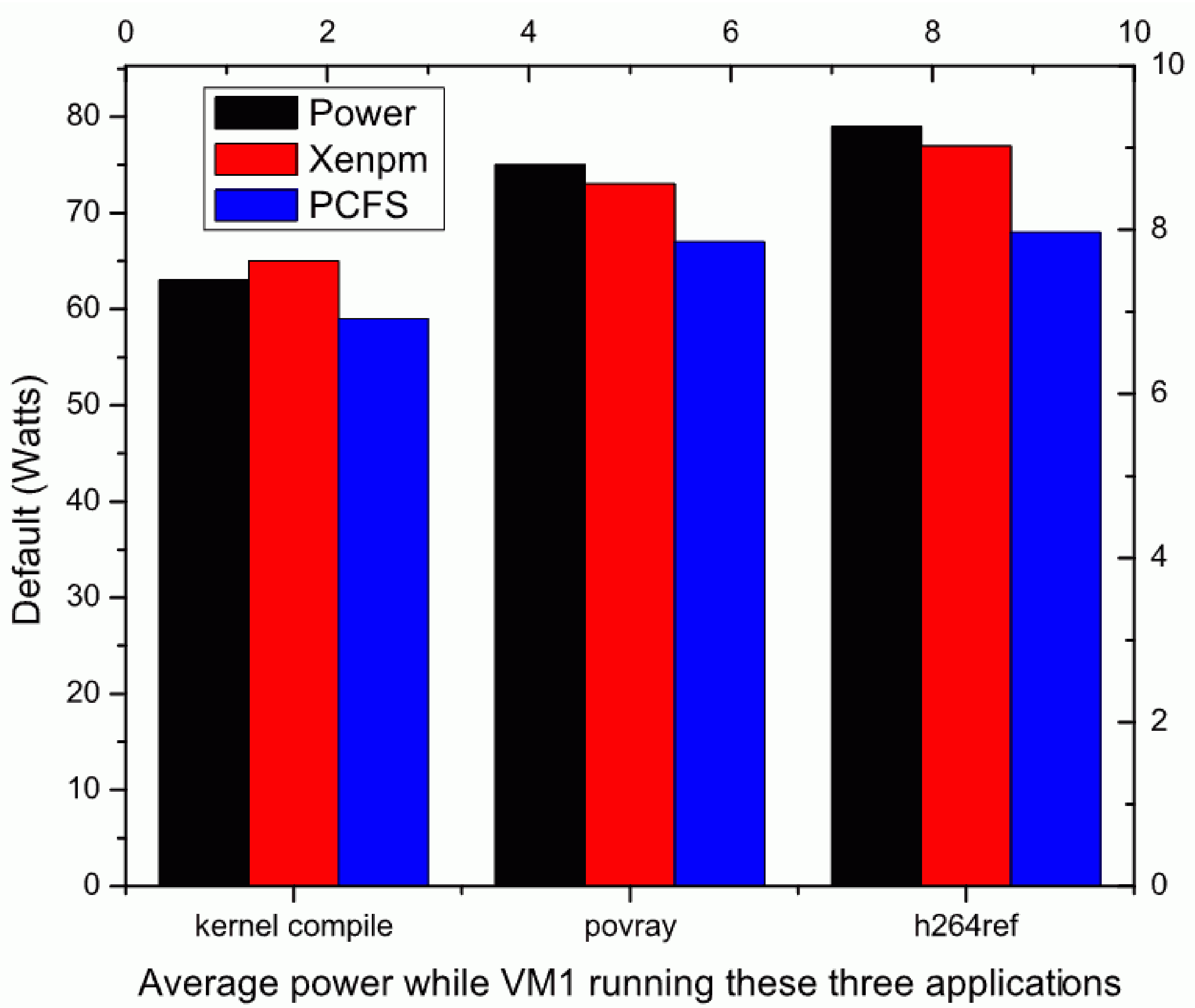


| Benchmark | Type | Usage |
|---|---|---|
| Spec2006 | CPU-bound | nearly 100% workload |
| Splash2 | Parallel application | virtual smp workload |
| Httperf | Network application | changing workload |
| Netperf | IO-bound | low workload |

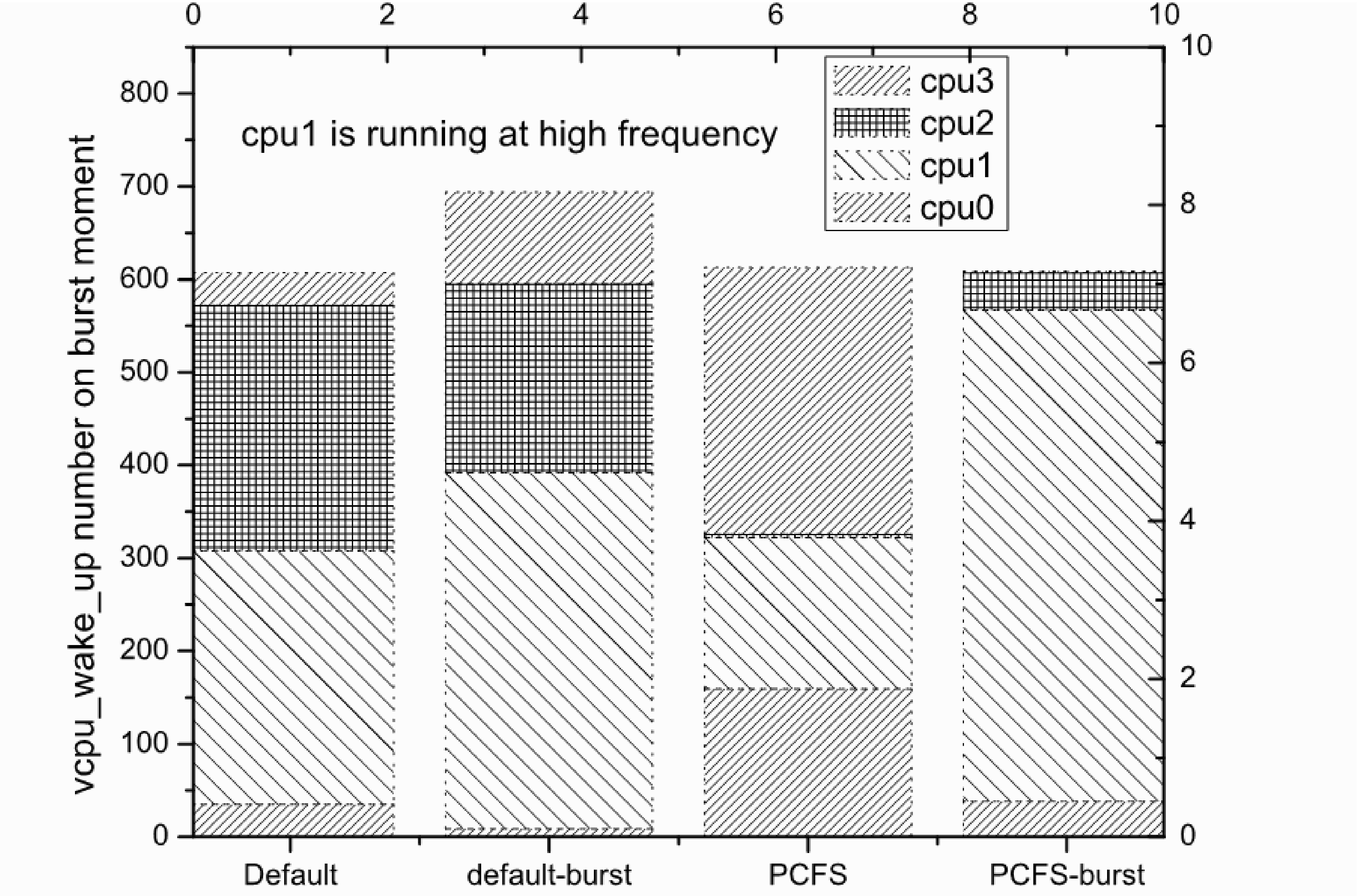
# Evaluation (3/7)



# Evaluation (4/7)

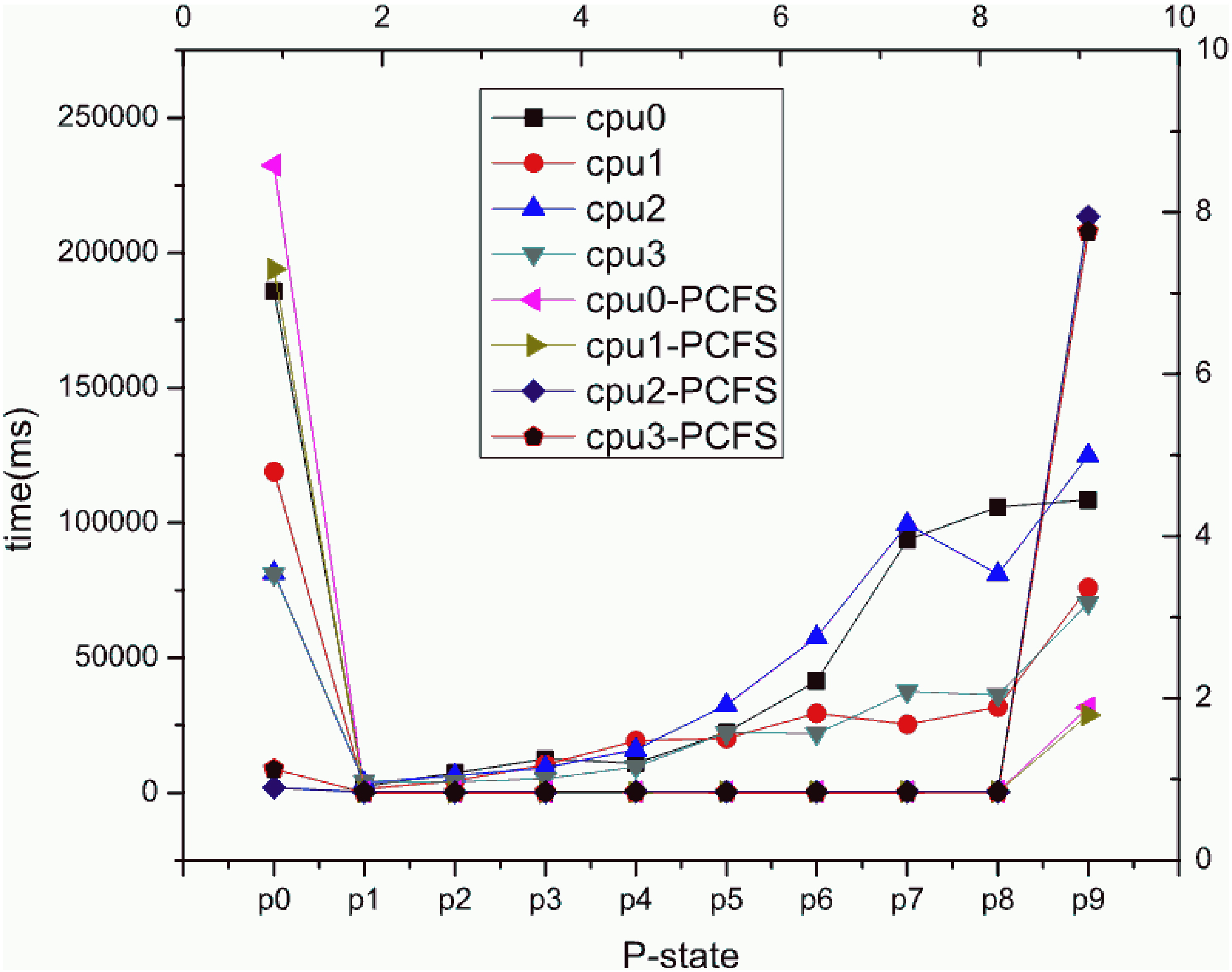


# Evaluation (5/7)



when VM 2 has burst load cpu1 is running at high frequency and has more run time.

# Evaluation (6/7)



# Evaluation (7/7)

- Temperature balance between physical cores
  - 3 VMs (Each one has 1 VCPU)
    - VM1 runs SPEC2006
    - VM2 runs netperf (Average 20% workload)
    - VM3 runs httperf (Average 80% load)
  - If the CPU exceeded 73°C, an interrupt is triggered.
  - The scheduler will firstly search the low frequency power zone and migrate the VCPU on that heated core to one core of that power zone.
  - Finally it sets the according frequency in the target power zone.

# Conclusions

- PCFS could keep some cores running at higher frequency and the other at lower frequency as long as possible.

- Experiments show that PCFS can make good use of these power saving measures.

- In the future, we will improve the PCFS to adapt to the heterogeneous multicore system and virtual SMP guest operating system.