Power Management Challenges in Virtualization Environments

Congfeng Jiang, Jian Wan, Xianghua Xu,
Yunfa Li, Xindong You
Grid and Service Computing Technology Lab,
Hangzhou Dianzi University, Hangzhou ,
310037, China
Sep., 2009
cjiang@hdu.edu.cn

Outline

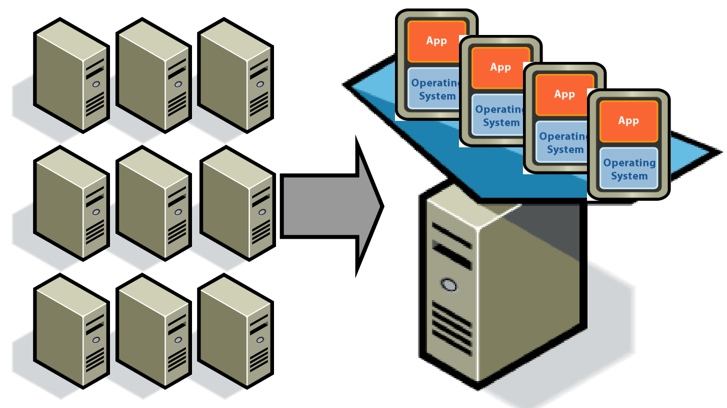
- Introduction
- Implications of Virtualization
- Power management challenges
- Discussions
- Acknowledgments
- Q&A

1 Introduction

- Power has been a critical resource for :
  - Battery-powered devices
  - PCs
  - Large scale server systems
  - Data centers

Example: Data centers

- Server consolidations and virtualizations in data centers
  - Higher power densities → higher power consumptions
  - Expensive cooling → Total Cost of Ownership (TCO)
- Thermal emergencies
  - Failed fans or air conditioners
  - Poor cooling or air distribution
  - Hot spots
  - Brownouts
- Component reliability decreases
  - Unpredictable behaviors or failures
  - Can impact system performance and availability



Why Power Management(PM)?

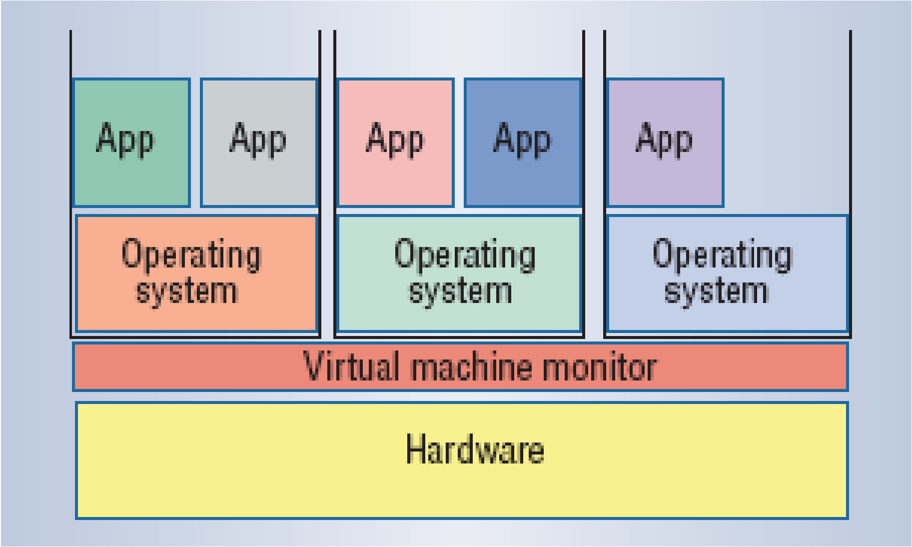
- Use less electricity
  - E.g.,half of power used to power PCs is wasted
- Reducing cooling loads and costs
- Reducing peak load demand charges

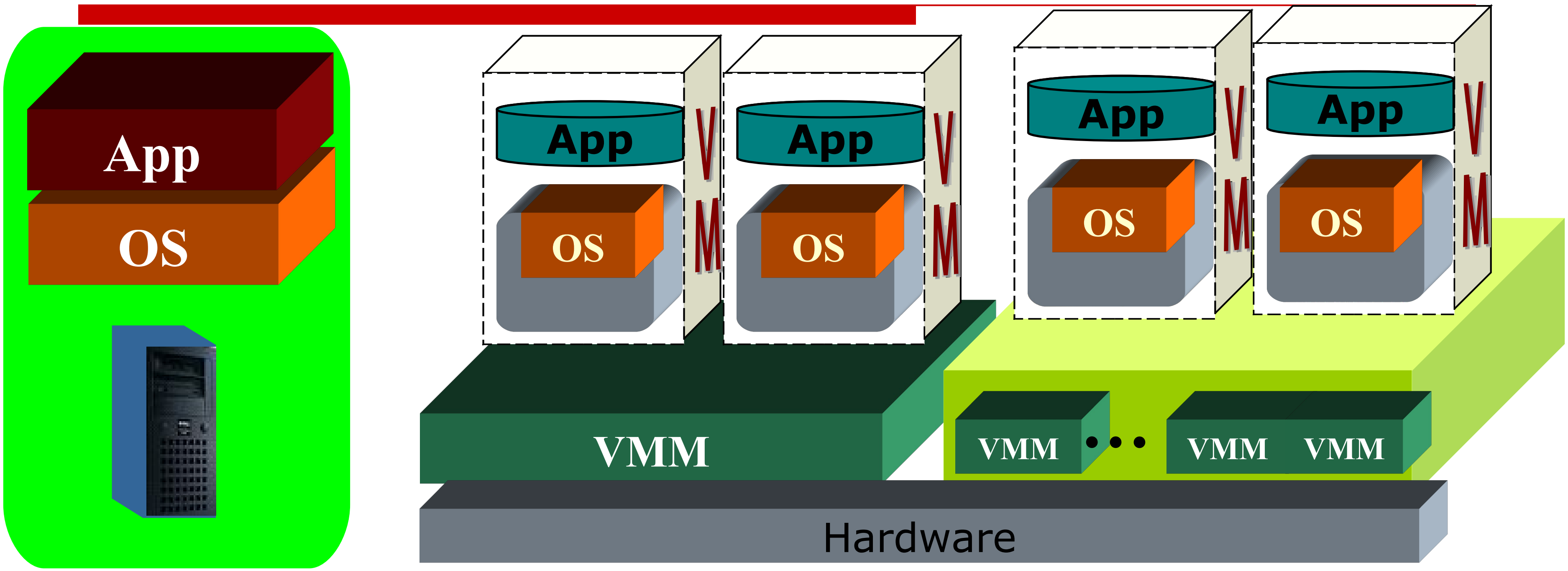
2. Implications of Virtualization

- Characteristics
- Implications

Characteristics → Issues

- Transparency → VMs know nothing about hardware power consumption
- Isolation → PM coordination among VMs





- **Conventional Computing Systems:** OS with full knowledge of and full control over the underlying hardware
- **Virtualization Environments:** multi-layered, PM coordination among VMs

Implications

- Conventional power management methods are not applicable to virtualization environments without modifications

Bad news!

- Soft-level fine grained power management can save more power in virtualization environment through live VM migration, job scheduling, power hotspots elimination

Good news!

3. Power management challenges

- Power consumption accounting and estimation of VMs
- Power management Coordination among VMs

3.1 Power consumption accounting and estimation

- Non-Virtualized environments
- Virtualized environments

Non-Virtualized environments

- Code profiling
- Hardware Performance Counters
- Power-driven statistical sampling

Power estimation

Application and thread level PM

Virtualized environments

- Devices are shared among multiple VMs
- Hardware heterogeneity

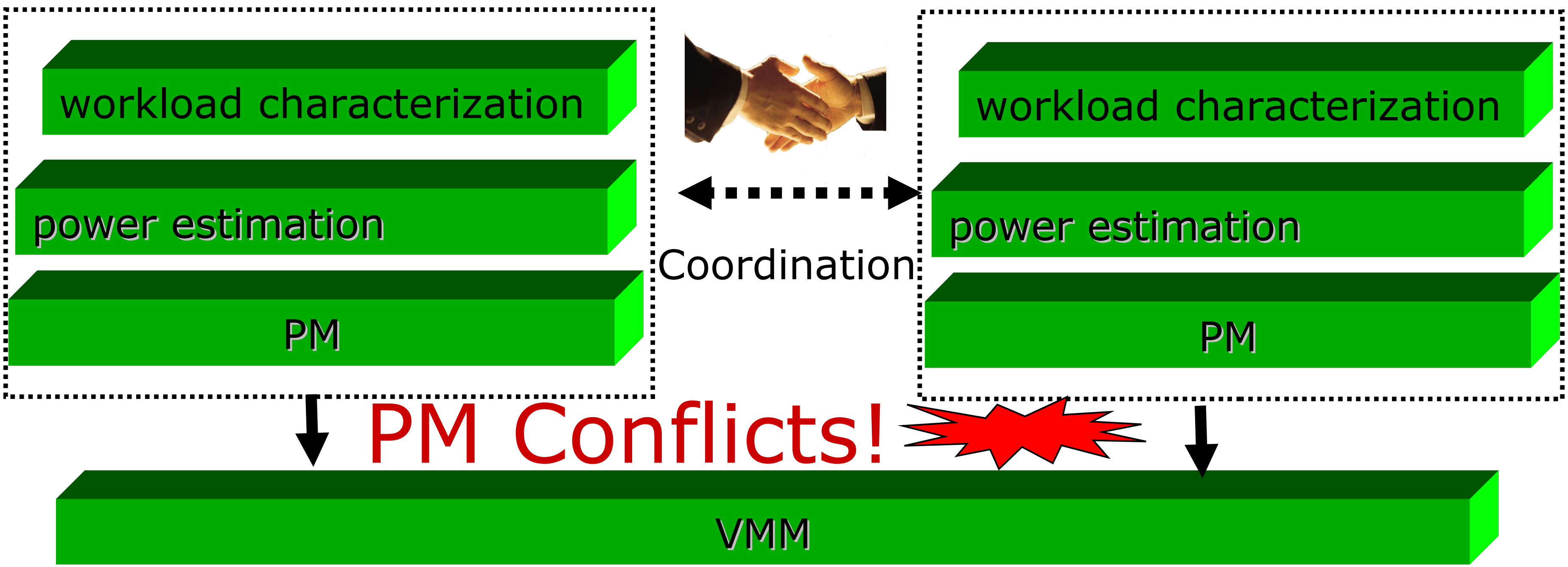
↓ Power estimation

VM level PM

Considerations

- Overheads
- Prediction accuracy
- Highly expensive workload characterization in large scale data centers

3.2 Power Coordination among VMs



3.3 Comparison of Existing Solutions

- Conventional techniques
- Virtualization environments
  - VirtualPower Management (VPM)
  - Magnet
  - ClientVisor
  - Stoess et al Framework

3.3.1 Conventional techniques

- Hardware level
  - *Micro-architectural design(VLSI& CMOS)*
  - Per-component adaptations
  - Multi-components adaptations
- Software level
  - OS
  - Scheduling
  - Virtualization

System-wide PM

- Reduce power consumption and maximize hosting revenue
  - power estimation and profiles
  - workload characterization
  - OS support power-aware algorithms

Per-component adaptations

- CPU
- Memory
- Hard drives
- Network Interface Cards(NICs)
- Display devices

slowing down the devices or switching the devices to low-power modes

CPU

- ACPI (Advanced Configuration and Power Interface) specifications
  - C0, C1, C2, C3, . . . , Cn.
- DVS (Dynamic Voltage Scaling)
- DFS (Dynamic Frequency Scaling)
- UDFS (User-Driven Frequency Scaling)
- PDVS (Process-Driven Voltage Scaling)
- Per-core DVS/DFS

Memory

- DRAM power consumptions is significant
  - 45% of total system power (Lefurgy et al., IEEE Computer 2003)
- Opportunity:DRAM is usually installed in an over-provisioned style to avoid swapping between memory and hard disks

Memory

- Decide to power down which memory units and into which low-power state to transition
- Queue-Aware Power-Down Mechanism
- Power/Performance-Aware Scheduling
- Adaptive Memory Throttling
  - Power Shifting: (Felter et al., ICS 2005),Dynamically assign power budgets to CPU and DRAM

Reducing DRAM power consumptions

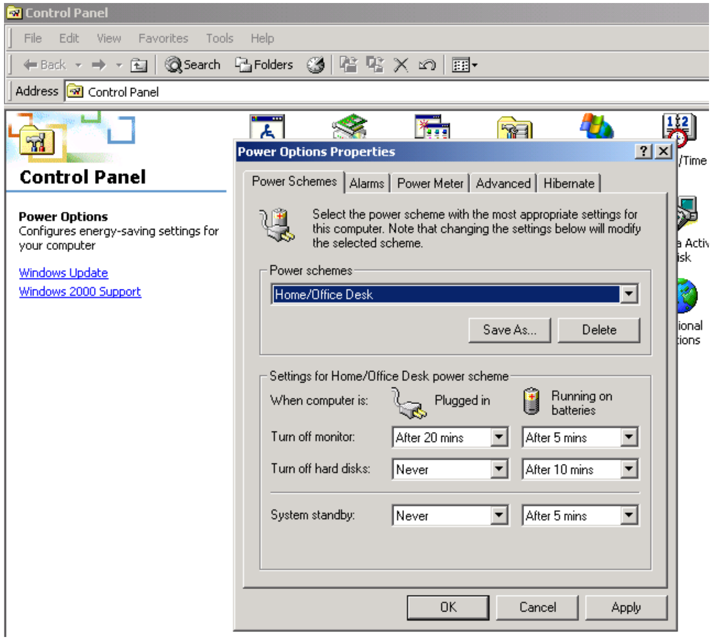
- Put certain ranks of DRAM into low-power mode
  - Entering and exiting has overhead
  - Ranks must remain in low-power mode for some minimum number of cycles
- How to enter and exit low-power mode?
  - Enter and exit too frequently → increased DRAM latency
  - Enter and exit too infrequently → less power savings

Hard drives

- Slowing down
- Switching to low-power modes
  - Hibernate

Monitor PM

- Monitor power management (MPM) places ***monitors*** into low power sleep mode after period of inactivity
- System standby and hibernate place the ***computer*** (CPU, hard drive, etc.) into sleep mode
- Built into Windows 95,98, ME, 2000, XP and Vista
- Settings simply need to be activated



Tradeoffs

- Hardware-level PM
  - Disregards high-level information
  - E.g. CPU shutdown mechanism
  - Unnecessary performance loss
- Software policies
  - More sophisticated reactions to emergencies
  - E.g. reduce load on “hot server” in a datacenter
  - Example: Freon for Internet services [ASPLOS’06]

- Tradeoffs between power reductions and performance degradions
- New trends
  - Multi-component joint-adaptations
  - Hardware-software joint-adaptations

3.3.2 PM in XEN

- what CPU load is suitable for reduction in speed and at what level do we increase the CPU speed
- Tricks: Switching to low power mode when all VMs are idle
  - An event channel to tell the Domain 0 guest to perform PM actions
  - Transitions between PM states
- XEN 3.1 No good ACPI&PM support

- XEN 3.3: ACPI C/P States support
  - The idle governor is triggered when the CPU is fully idle, and then the governor chooses the appropriate low power state based on the power budget and latency tolerance accordingly.
  - The deeper C-state is, less power is consumed with longer entry/exit latency.
  - governor monitors CPU utilization (using a call into Xen).
- No PM estimation and coordination features

- XEN 3.4
  - A new algorithms to better manage the processor including schedulers and timers optimized for peak power savings.
- No PM estimation and coordination features

3.3.3 Comparison of existing PM methods for Virtualization environments

Metrics\schemes	VirtualPower	Magnet	ClientVisor	Ref.[20]
Testbed configuration	Multiple PCs machines with Intel Dual Core Pentium 4 processors	A 64-hosts cluster with AMD Athlon 3500+processors	Desktop virtualization environment with Intel Core2 Duo T9400 processor	A machine with Intel Pentium D processor
Hardware Heterogeneity	Identical +Heterogeneous	Homogeneous	Homogeneous	Homogeneous
VMM	Xen	Xen	Xen	L4 micro-kernel
Using DVS/DFS	Yes	N/A	N/A	N/A
Number of VMs	>=4	N/A	3	N/A
Online/Offline	online	online	online	online
Power consumption estimation	measured 'at the wall'	N/A	measured 'at the wall'	external high performance data acquisition (DAQ) system
Power management coordination	(i)system-level abstractions including VPM states, channels, mechanisms, and rules (ii)VM-level 'soft' power scaling mapping to real power states and scaling	centric, concentric non-overlapping rings with heartbeats exchange	coordinate only "at the key points"	budget allotment
Max. Power savings	34%	74.8%	22%	N/A
Overheads	Little performance penalties	Adjustably acceptable	Degration about 2%~3%.	N/A
With QoS/SLA guarantees	Yes	Yes	N/A	N/A
VM migration	Yes	Yes	N/A	N/A
Workload	RUBiS	bit-r, m-sort, m-m, t-sim, metis, r-sphere, and r-wing	SPECpower_ssj	DAQ/bzip2 application

4 Discussions

- Two possible goals of PM
  - Reduce power consumptions with minimal performance degradation
  - Stay within some given power budget while degrading performance as little as possible

- Fine-grained VM-level PM is necessary for virtualization
- Conventional power estimation techniques are designed only for monolithic kernels
- Negotiations among VMs
- SLAs and QoS guarantees

Future work

- Possible DMTF standards for power efficiency specifications, evaluation ,benchmarking & metrics
- Power Management interoperability among different virtualized devices

5 Acknowledgments

- State Key Development Program of Basic Research of China ("973"Project,Grant No. 2007CB310900)
- Natural Science Fund of China (NSFC) (Grant No. 60873023)

Thank you
&
Any question ?