Towards More Power Friendly Xen

June 2008

Yu Ke <ke.yu@intel.com>

Tian Kevin <kevin.tian@intel.com>

Wei Gang <gang.wei@intel.com>

Liu Jinsong <jinsong.liu@intel.com>

Open Source Technology Center

Software and Solutions Group

Agenda

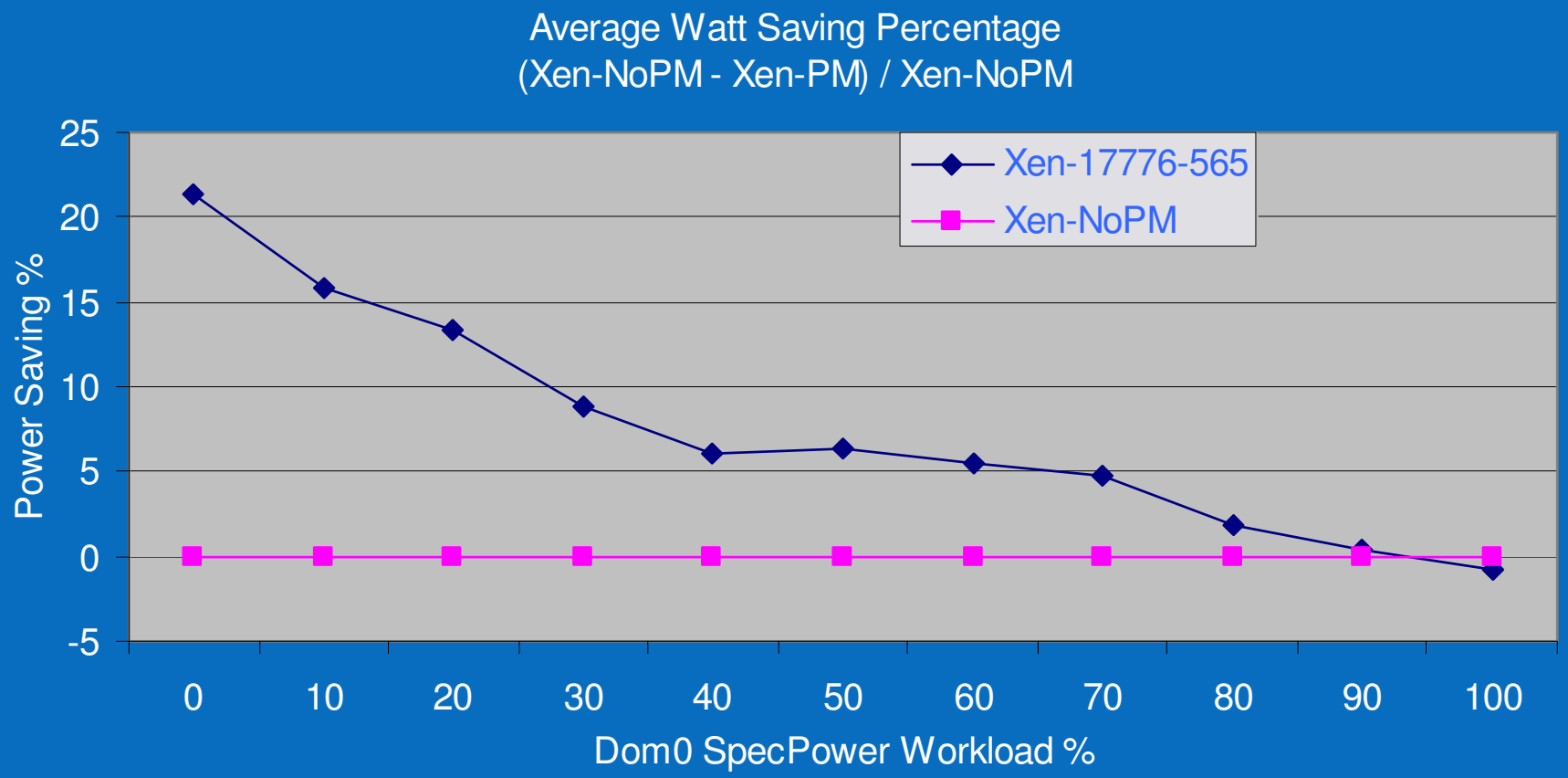
- Xen power management current status
- Power management tuning
  - Deep C State
  - Dynamic Tick
  - Deferrable Timer
  - Power Aware Scheduler

Xen Power Management Status

- P state support has been done
  - Hypervisor based P state Control
  - On-demand policy
- C state support has been done
  - Support ACPI C1 & C2
- Dynamic Tick Support has been done

Xen Power Management for server has been in place

Xen Power Management Experimental Results



Legal Disclaim: experimental result only, not guaranteed result from Intel

Xen Power Management Status (Cont.)

- Significant improvement compared to Xen without PM support
- Still has gap between Xen Dom0 and Native Linux
  - When idle, dom0 consumes 16% more power than native 2.6.25 (tickless idle + device PM support)
- Need continue to tune Xen power management

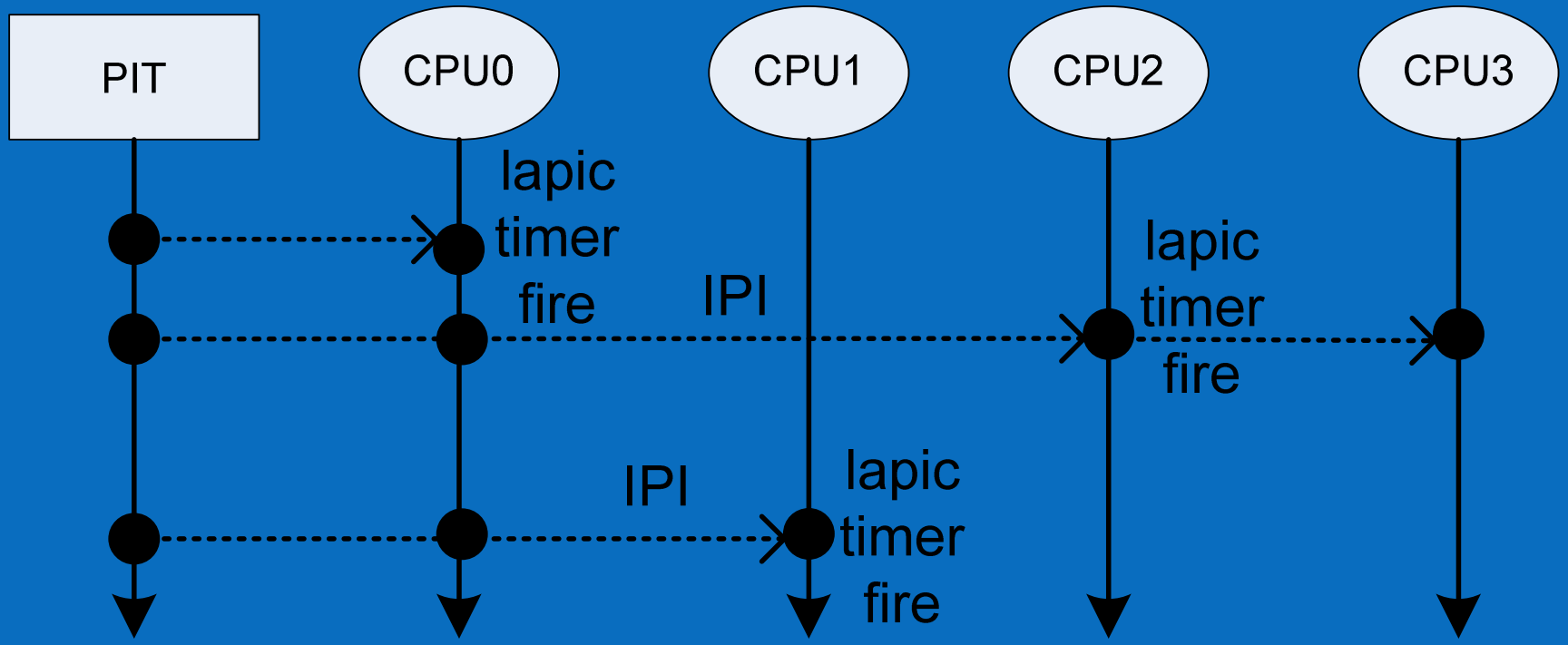
Power Management Tuning

- Enable power management feature
  - Support Deep C states
- Optimize C state residency
  - Key point is to reduce breaks events (e.g. external interrupts)
  - Dynamic tick mode
  - Deferrable AC timer

Deep C states support

- Issue with Deep C states (C3 and deeper)
  - TSC stops during C state today
  - Local APIC timer stops during C state
- Solution for TSC stop issue
  - Save TSC during C state entry and re-sync TSC after C state exit
- Solution for local APIC timer issue
  - Broadcast platform timer

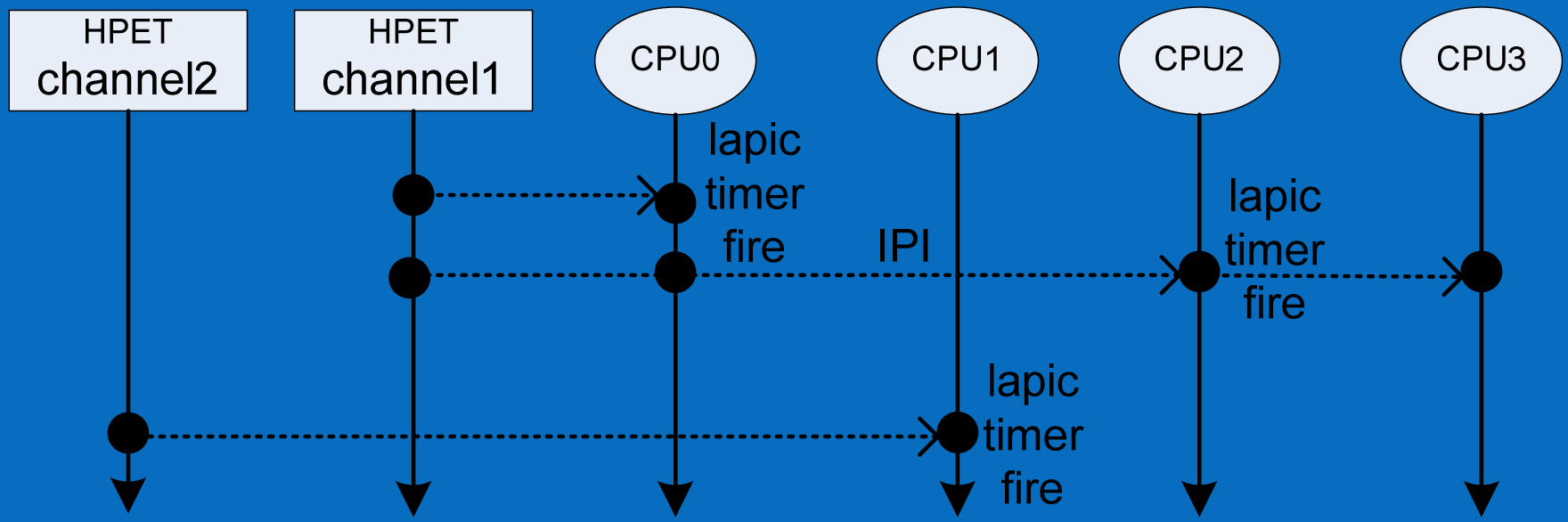
Broadcast Timer: PIT



PIT as platform timer

- periodic 100HZ timer
- send IPI to wakeup target CPU

Broadcast Timer: HPET



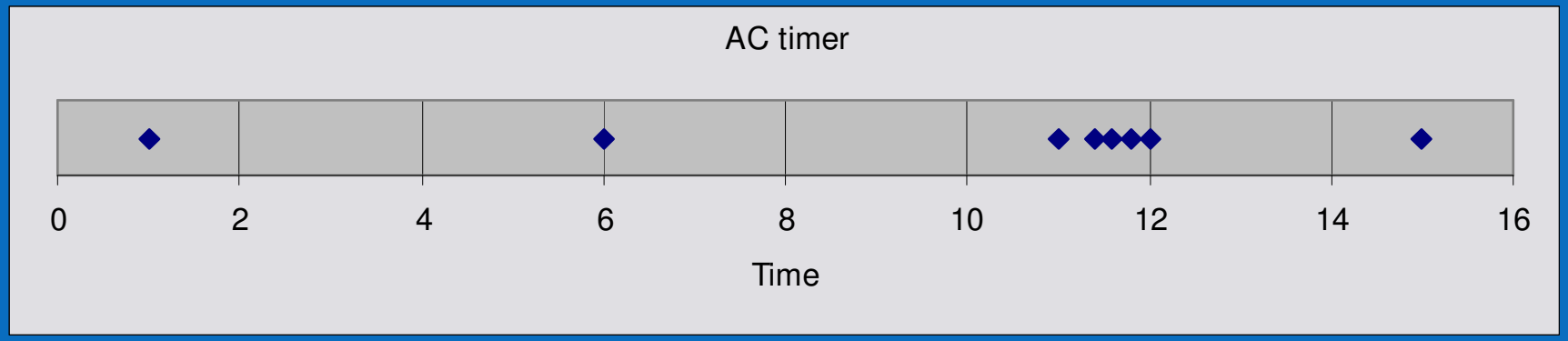
HPET as platform timer

- One-shot timer
- send IPI to broadcast or use dedicated channel

Dynamic Tick

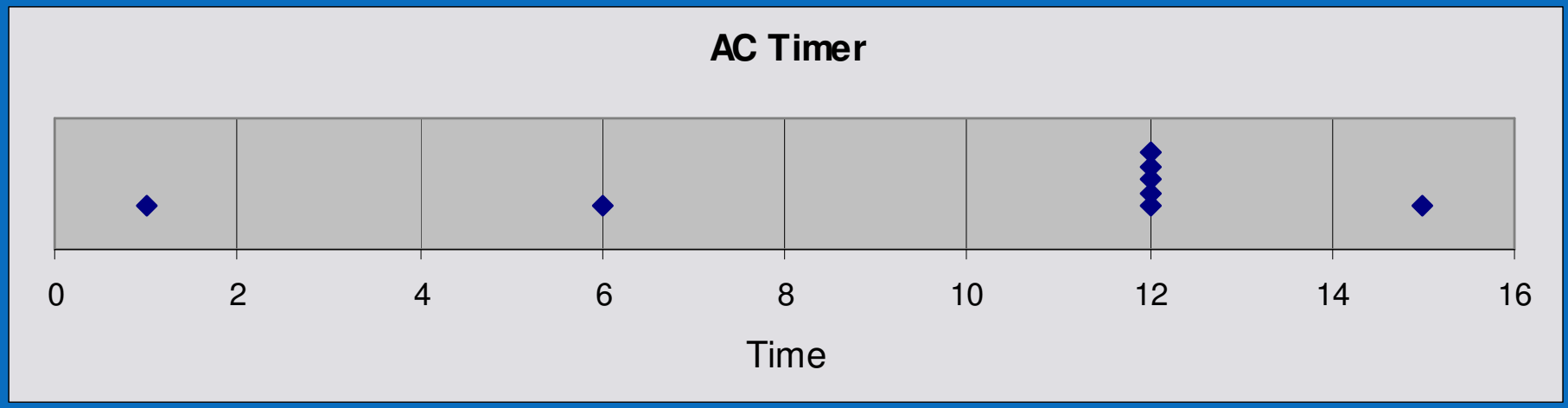
- Tuning point: Xen original has 100HZ PIT timer, mostly for jiffies tick, which put 10ms limitation to C state residency
- Optimization:
  - Jiffies is obsolete in Xen, so remove jiffies usage
  - Disable PIT timer if applicable
- Status:
  - Has been done in Xen

Deferrable Timer



- Tuning point: as AC timer number increase, timers will aggregates (see timers near 12), this will break C state frequently

Deferrable Timer (Cont.)

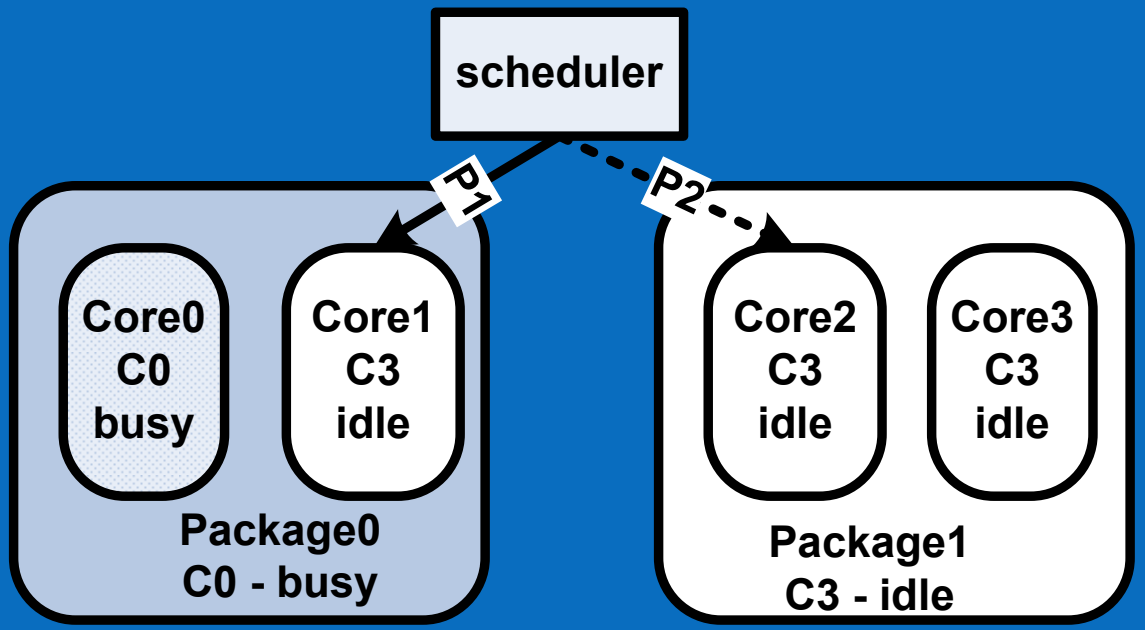


- Optimization:
  - If the timer is not time critical, defer the timer, and fire them one time
  - New API: set_timer_deferable (struct timer *timer, s_time_t expires, s_time_t period), expire @ [expire, expire+period]
  - User use the deferrable timer API if it does not require precise firing point

Deferrable Timer Current Status

- Patch prototype patch is done
- Preliminary evaluation
  - Px on-demand governor has 20ms periodically dbs_timer to update P state, which does not require exact timer firing point
  - After changing the dbs_timer to deferrable timer, the timer interrupt frequency reduce 10%
- Next step
  - Evaluate more code path to use deferrable timer

Power Aware Scheduler



- Package C state
  - Package C state is decided by the C state of the most busy core
  - Package C state also save power
- Scheduler: pick the idle core whose sibling is already busy
  - P1 better than more power-saving that P2

Summary

- Xen power management for server has been in place
  - Significant power saving compared to Xen without PM support
- We are optimizing Xen power management to save more power
  - Enabling new PM feature, e.g. deep C states
  - Reduce breaking event for longer C state residency
  - Make Xen sub-system power aware, e.g. Scheduler

Legal Information

- INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY RELATING TO SALE AND/OR USE OF INTEL PRODUCTS, INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT, OR OTHER INTELLECTUAL PROPERTY RIGHT.
- Intel may make changes to specifications, product descriptions, and plans at any time, without notice.
- All dates provided are subject to change without notice.
- Intel is a trademark of Intel Corporation in the U.S. and other countries.
- *Other names and brands may be claimed as the property of others.
-


intel
Leap ahead™