Power Management Challenges in Linux*

Rafael J. Wysocki

Intel Open Source Technology Center

September 15, 2017
What Power Management Is About
Linux Power Management Overview

- **Sleeping**
  - Hibernate
  - Suspend (Deep, Shallow)
  - Suspend-to-Idle

- **Working**
  - Runtime Idle
  - Runtime Active

Rafael J. Wysocki (Intel OTC)
System-Wide PM (Sleep States) and Working-State PM
System Suspend Control Flows

Full Suspend

- **Suspend**
  - Call Notifiers
  - Freeze Tasks
    - Device Suspend
      - `.prepare()`
      - `.suspend()`
      - `.suspend泐late()`
      - `.suspend_noirq()`
    - Nonboot CPU Offline
    - System Core Offline
    - Platform Offline
  - Wait For a Wakeup Event

- **Resume**
  - Call Notifiers
  - Thaw Tasks
    - Device Resume
      - `.complete()`
      - `.resume()`
      - `.resume泐early()`
      - `.resume_noirq()`
    - Nonboot CPU Online
    - System Core Online
    - Platform Online

Suspend to Idle

- **Suspend**
  - Call Notifiers
  - Freeze Tasks
    - Device Suspend
      - `.prepare()`
      - `.suspend()`
      - `.suspend泐late()`
      - `.suspend_noirq()`
    - Platform Offline
  - Wait For a Wakeup Interrupt

- **Resume**
  - Call Notifiers
  - Thaw Tasks
    - Device Resume
      - `.complete()`
      - `.resume()`
      - `.resume泐early()`
      - `.resume_noirq()`
    - Platform Online
Suspend-to-Idle Implementation Challenges

- **Suspend**
  - Call Notifiers
  - Freeze Tasks
  - Device Suspend
    - prepare()
    - suspend()
    - suspend_late()
    - suspend_noirq()
  - Wait For a Wakeup Interrupt
- **Resume**
  - Call Notifiers
  - Thaw Tasks
  - Device Resume
    - complete()
    - resume()
    - resume_early()
    - resume_noirq()

**Must be absolutely reliable**

Wakeup via in-band interrupts

- Rafael J. Wysocki (Intel OTC)
Example: EC-Based Power Button Events Signaling

- Power Button
- Embedded Controller
- Battery
- Thermal
- ACPI SCI
- PCI PME
- Device Wakeup
- I/O APIC
- CPU

Pressing the power button triggers a series of events that eventually lead to the system being suspended to idle. The process involves signaling through the embedded controller, bus/device check, EC GPE, ACPI SCI, and other components as described in the diagram.
Suspend-to-Idle With Detection of Spurious Wakeups

- **Suspend**:
  - Call Notifiers
  - Freeze Tasks
  - Device Suspend
    - prepare()
    - suspend()
    - suspend_late()
    - suspend_noirq()
  - Wait For a Wakeup Interrupt

- **Resume**:
  - Call Notifiers
  - Thaw Tasks
  - Device Resume
    - complete()
    - resume()
    - resume_early()
    - resume_noirq()

Wakeup?

- YES
- NO
Challenges Related to Hibernation

- Hibernation
  - Freeze Transition
    - Device Freeze
      - System Core Online
        - Nonboot CPU Online
          - Device Thaw
            - System Core Offline
          - Support in drivers
        - Support in drivers
      - Call Notifiers
    - Support in drivers
  - Freeze Tasks
  - Save Image
  - Create Image
  - Encryption

- Hibernation
  - Power Off Transition
    - Device Power Off
      - System Core Offline
    - Restore Transition
      - Nonboot CPU Offline
      - System Core Offline
      - Device Restore
        - Restore Transition
        - Call Notifiers

- Secure Boot
  - Support
  - Load Image

- Restore
  - Call Notifiers
  - Freeze Tasks
  - Support in drivers
  - Support in drivers
  - Support in drivers
  - JUMP
  - Restore Memory
  - Restore Kernel
  - Image Kernel

- Rafael J. Wysocki (Intel OTC)
- Linux PM Challenges
- September 15, 2017
CPU Performance Scaling

- **intel_pstate**
  - HWP
  - CPU scheduler
  - Governor callback
  - Estimate target performance
  - Adjust performance
  - P-state selection

- **schedutil governor (Example)**
  - CPU scheduler
  - Governor callback
  - Utilization metric
  - Compute frequency
  - Fast switching supported?
    - NO: Queue up work
    - YES: Adjust performance (Driver)

Rafael J. Wysocki (Intel OTC)
CPU Idle Time Management

1. CPU scheduler
2. CPU is idle
3. Estimate idle time
4. Select state
5. Enter state
6. Wake up

- Not deterministic
- Latency constraints

Rafael J. Wysocki (Intel OTC)
Task Placement and CPU PM

- Task placement
- Available CPU capacity
  - CPU utilization
    - Performance scaling
      - CPU idle time
        - Idle states selection
          - Performance
            - Energy usage

Rafael J. Wysocki (Intel OTC)
Device Runtime PM Framework

- **Active**
- **Suspended**
- **Resume**
- **Suspend**
- **Is Idle?**
  - YES
  - NO

How to trigger?

Rafael J. Wysocki (Intel OTC)
Driver Core and Device PM Operations
Constraints and Requirements

Thermal

Power budget

Throughput

Latency
Questions?
References


Disclaimer

Intel technologies’ features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at www.intel.com.

Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries. *Other names and brands may be claimed as the property of others.

© Intel Corporation