Power Management in Linux* - State of The Art

Rafael J. Wysocki

Intel System Software Engineering

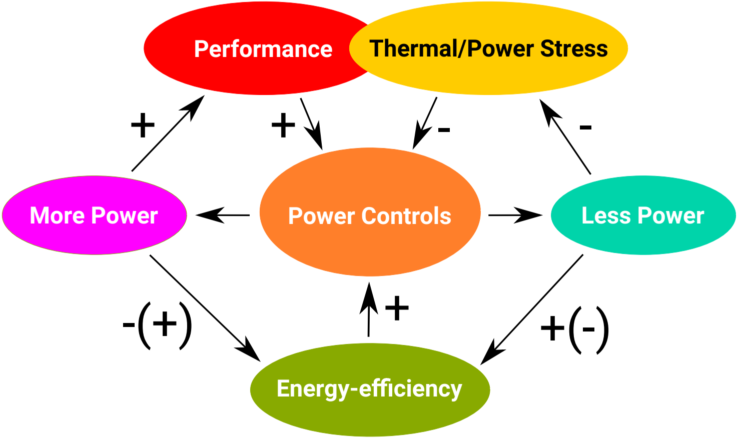
August 10, 2020

* Other names and brands may be claimed as the property of others

What Power Management Is About

Intel OpenSource TECHNOLOGY CENTER

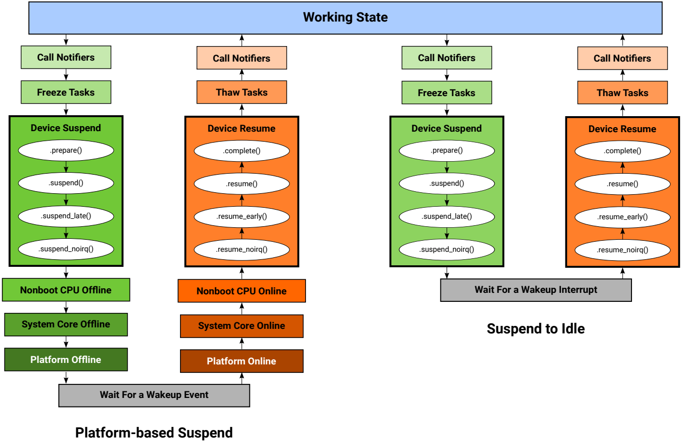
General Overview Of Power Management



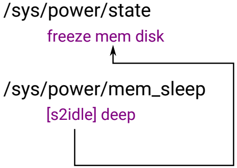
Intel OpenSource TECHNOLOGY CENTER

Two Different Ways To Get There

System Suspend



System Suspend/Hibernation Interface In sysfs



/sys/power/state

freeze mem disk

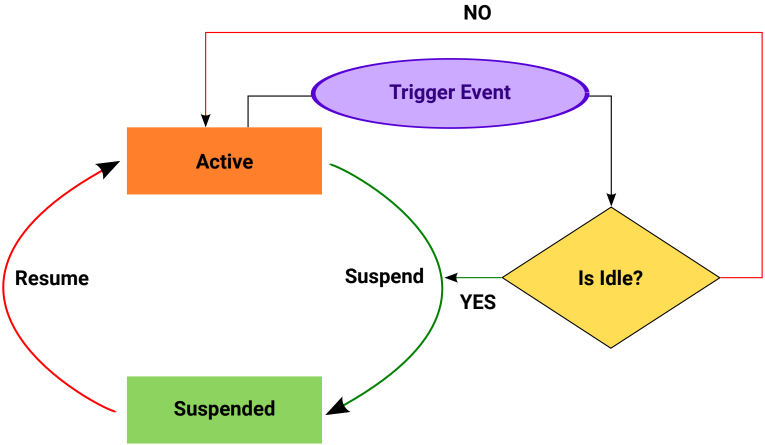
/sys/power/mem_sleep

[s2idle] deep

/sys/devices/.../power/wakeup

enabled
disabled

Overview Of PM-runtime



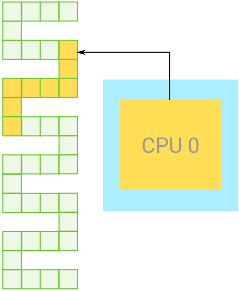
PM-runtime Control Through sysfs

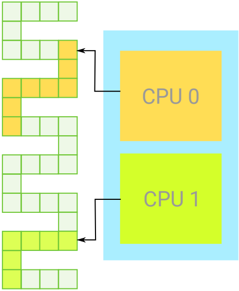
/sys/devices/.../power/control

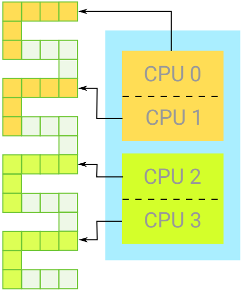
- auto
- on

/sys/devices/.../power/autosuspend_delay_ms

CPUs Are Logical Entities





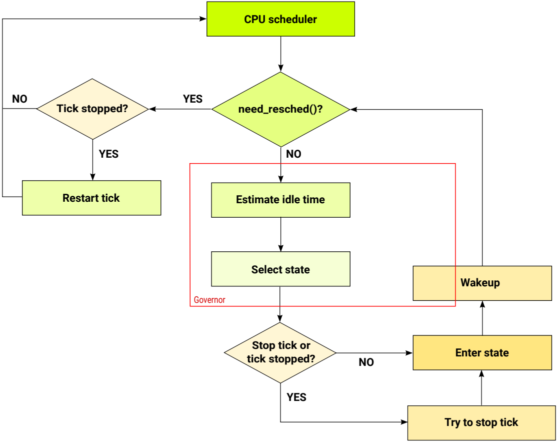


Intel OpenSource TECHNOLOGY CENTER

CPUs: Busy Vs Idle

Intel OpenSource TECHNOLOGY CENTER

CPU Idle Loop (Linux* 4.17 And Later)



* Other names and brands may be claimed as the property of others

CPU Idle Time Management Control

Kernel command line

idle=**poll**
halt
nomwait

cpuidle.off=1

intel_idle.max_cstate=0
1 ... 9

processor.max_cstate=1 ... 9
0

Special device

/dev/cpu_dma_latency

sysfs

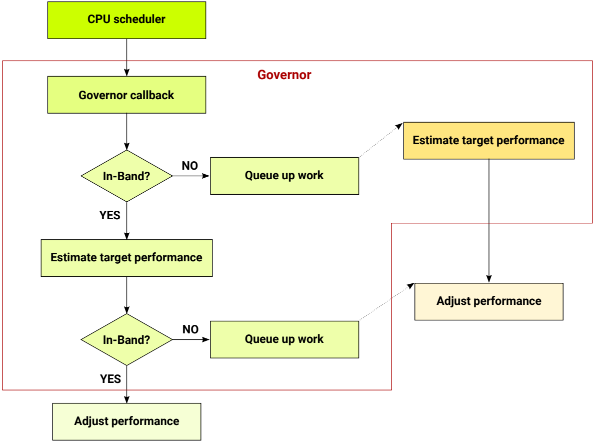
/sys/devices/system/cpu/cpuidle/
current_driver : intel_Idle
current_governor_ro : menu

/sys/devices/system/cpu/cpu<nr>/cpuidle/state<nr>/
desc : MWAIT 0x00
disable : **0**
latency : 2
name : C1
power : 0
residency : 2
time : 438396006
usage : 4637114

/sys/devices/system/cpu/cpu<nr>/power/pm_qos_resume_latency_us

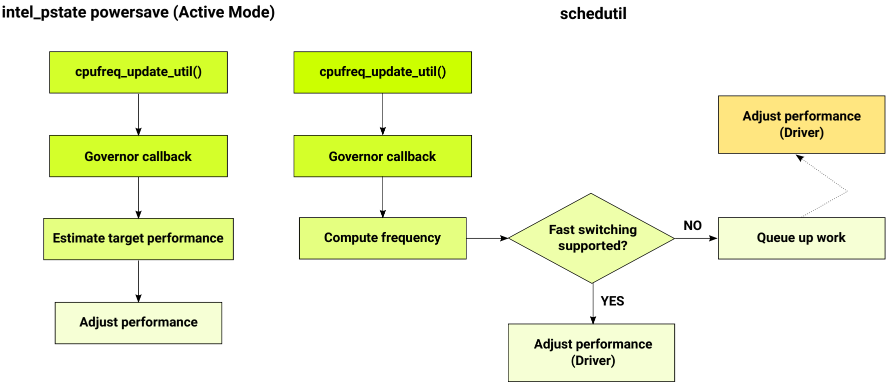
CPUs: Performance And Utilization

Overview Of CPU Performance Scaling



Intel OpenSource TECHNOLOGY CENTER

In-Band Scaling Governors


intel_pstate powersave (Active Mode)
cpufreq_update_util()
Governor callback
Estimate target performance
Adjust performance
schedutil
cpufreq_update_util()
Governor callback
Compute frequency
Fast switching supported?
NO
Queue up work
Adjust performance (Driver)
YES
Adjust performance (Driver)

CPU Performance Scaling Control

sysfs

/sys/devices/system/cpu/cpu<nr>/cpufreq/

affected_cpus : 0

cpuinfo_max_freq : 4300000

cpuinfo_min_freq : 800000

cpuinfo_transition_latency : 0

energy_performance_available_preferences : default performance balance_performance balance_power power

energy_performance_preference : **balance_performance**

related_cpus : 0

scaling_available_governors : performance powersave

scaling_cur_freq : 3039632

scaling_driver : intel_pstate

scaling_governor : **powersave**

scaling_max_freq : **4300000**

scaling_min_freq : **800000**

/sys/devices/system/cpu/intel_pstate/

max_perf_pct : **100**

min_perf_pct : **18**

no_turbo : **0**

num_pstates : 36

status: **active**

turbo_pct : 20

Kernel command line

cpufreq.off=1

intel_pstate=**active**
disable
force
hwp_only
no_hwp
passive
per_cpu_perf_limits
support_acpi_ppc

Intel OpenSource TECHNOLOGY CENTER

PCI Express Active State Power Management (ASPM)

PCIe ASPM Control

Module parameter

```
pcie_aspm.policy=default
                 performance
                 powersave
                 powersupersave
```

sysfs

```
/sys/devices/pci0000:00/.../link/
    clkpm : 1
    l0s_aspm : 1
    l1_1_aspm : 1
    l1_1_pcipm : 1
    l1_2_aspm : 1
    l1_2_pcipm : 1
    l1_aspm : 1
```

Kernel command line

```
pcie_aspm= off
           force
```

Intel OpenSource TECHNOLOGY CENTER

Intel Performance and Energy Bias Hint (EPB)

Intel EPB Control Through sysfs

/sys/devices/system/cpu/cpu<nr>/power/energy_perf_bias

0...15
performance
balance-performance
normal
balance-power
power

Intel OpenSource TECHNOLOGY CENTER

Is Energy-efficiency Always At Odds With Performance?

Connection Between Energy-efficiency And Performance

For individual hardware components

Performance depends on the capacity and latency.

Without power budget sharing

Better energy-efficiency means more latency and/or less capacity.

However

Improving energy-efficiency of one component may change the distribution of power.

PM Features May Depend On One Another

Questions? Comments? Concerns?

References

- The Linux kernel user's and administrator's guide, *Power Management* (https://www.kernel.org/doc/html/latest/admin-guide/pm/index.html).
- Rafael J. Wysocki, *Energy-efficiency and Linux* (https://static.sched.com/hosted_files/osseu19/d1/energy-efficiency_and_Linux.pdf).
- Rafael J. Wysocki, *Advances in CPU Idle Time Management* (http://events19.linuxfoundation.org/wp-content/uploads/2017/11/Advances-in-CPU-Idle-Time-Management-Rafael-Wysocki-Intel.pdf).
- Rafael J. Wysocki, *Power Management Challenges in Linux* (https://www.linuxplumbersconf.org/2017/ocw//system/presentations/4652/original/linux_pm_challenges.pdf).
- Rafael J. Wysocki, *Advances in CPU Performance Scaling* (http://schd.ws/hosted_files/ossna2017/39/advances_in_cpu_perf_scaling.pdf).
- Rafael J. Wysocki, *PM Infrastructure in the Linux Kernel – Current Status and Future* (https://events.linuxfoundation.org/sites/events/files/slides/kernel_PM_infra_0.pdf).

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 OpenSource TECHNOLOGY CENTER