ITMO Advanced OS 2024

Aleksei Romanovskii, PhD,
SPb Research Center (CBG OS Lab)

Lesson 2024.12.18

What triggers a CPU core to enter Cx-state

- C0 is entered at boot, when an interrupt happens or a write event occurs on the address monitored by an MWAIT (WFI, WFE in ARM, use DSB to sync data) instruction.

- C1is entered with HLT or MWAIT instructions.

- C3 is entered with MWAIT instruction, then L1 and L2 caches are flushed to LLC and all core clocks are stopped.
  - However, the core maintains its state since it is still powered.

- C6 is entered with MWAIT instruction, then the processor state is saved into a dedicated SRAM and the voltage to core is reduced to zero.
  - At this state, core has no power.
  - When exiting C6, the processor state is restored back from the SRAM.

- C7 and C8 are same as C6 for the core.

OS-controlled P-states

- OS is aware of the P-states, a specific P-state is requested by the OS.
- This means selecting an operating frequency, whereas the voltage is automatically selected by the processor depending on the frequency and other factors.
- After a P-state is requested by writing it to a model specific register, the voltage is transitioned to an automatically calculated value and the clock generator (PLL) locks to the frequency requested.
- All cores in cluster (big.LITTLE) share the same P-state, so it is not possible to set this individually for a core.
- The current P-state / operating point can be read from another model specific register

What is not controlled (by OS, C/P/G/S-states)?

- Intel AMT and ME (there are similar, less known, technologies in AMD and ARM)

- Intel Active Management Technology (AMT): "hardware and firmware technology for remote out-of-band management of personal computers". AMT has many features which includes power management.

- Intel Management Engine (ME) is the hardware part: an isolated and protected coprocessor, embedded as a non-optional part in all Intel chipsets since 2008.
  - The coprocessor is a special 32-bit ARC microprocessor (RISC architecture) that's physically located inside the PCH chipset (or MCH on older chipsets).

- So, Intel x86s hide another CPU that can take over your machine (you can't audit it)
  - Vulnerabilities, backdoors (FBI, CIA, etc)
  - Can be switched off as discovered by Positive Technologies (Russian based company)

CPU performance scaling in Linux 1 of 2

- OS estimates the required CPU capacity, and decides which P-states to put the CPUs into.
- The activity (is NOT scheduling) by which this happens is referred to as CPU performance scaling or CPU frequency scaling (CPUFreq) - because it involves adjusting the CPU clock frequency
- The Linux kernel supports CPU performance scaling by means of the CPUFreq subsystem that consists of three layers of code:
  - core layer;
  - scaling governors;
  - scaling drivers.
- The CPUFreq core provides the common code infrastructure and user space interfaces for all platforms that support CPU performance scaling.
- Scaling governors implement algorithms to estimate the required CPU capacity. As a rule, each governor implements one, possibly parametrized, scaling algorithm.
- Scaling drivers talk to the hardware. They provide scaling governors with information on the available P-states (or P-state ranges in some cases) and access platform-specific hardware interfaces to change CPU P-states as requested by scaling governors.

CPU performance scaling in Linux 2 of 2

- Generally all available scaling governors can be used with every scaling driver.
- That design is based on the observation: information used by performance scaling algorithms for P-state selection can be represented in a platform-independent form in the majority of cases.
- So it should be possible to use the same performance scaling algorithm implemented in exactly the same way regardless of which scaling driver is used.
- Consequently, the same set of scaling governors should be suitable for every supported platform.
- However, that observation may not hold for performance scaling algorithms based on information provided by the hardware itself, for example through feedback registers.
- That information is typically specific to the hardware interface it comes from and may not be easily represented in an abstract, platform-independent way.
- For this reason, CPUFreq allows scaling drivers to bypass the governor layer and implement their own performance scaling algorithms.

CPUFreq policy objects

- In some cases the hardware interface for P-state control is shared by multiple CPUs.
- For example, the same register (or set of registers) is used to control the P-state of multiple CPUs at the same time and writing to it affects all of those CPUs simultaneously.
- Sets of CPUs sharing hardware P-state control interfaces are represented by CPUFreq as ***struct cpufreq_policy*** objects.
- For consistency, ***struct cpufreq_policy*** is also used when there is only one CPU in the given set.
- The CPUFreq core layer maintains a pointer to a ***struct cpufreq_policy*** object for every CPU in the system, including CPUs that are currently offline.
- If multiple CPUs share the same hardware P-state control interface, all of the pointers corresponding to them point to the same ***struct cpufreq_policy*** object.
- CPUFreq uses struct cpufreq_policy as its basic data type and the design of its user space interface is based on the policy concept.

CPUFreq policy CLI interface

- During the initialization of the kernel, the CPUFreq core layer creates a sysfs directory (kobject) called cpufreq under /sys/devices/system/cpu/.
- That directory contains a policyX subdirectory (where X represents an integer number) for every policy object maintained by the CPUFreq core.
- Each policyX directory is pointed to by cpufreq symbolic links under /sys/devices/system/cpu/cpuY/ (where Y represents an integer that may be different from the one represented by X) for all of the CPUs associated with (or belonging to) the given policy.
- The policyX directories in /sys/devices/system/cpu/cpufreq each contain policy-specific attributes (files) to control CPUFreq behavior for the corresponding policy objects (that is, for all of the CPUs associated with them).

CPUFreq scaling governors 1 of 2

- CPUFreq provides generic scaling governors that can be used with all scaling drivers. Each governor implements a single, possibly parametrized, performance scaling algorithm.

- Scaling governors are attached to policy objects and different policy objects can be handled by different scaling governors at the same time (although that may lead to suboptimal results in some cases).

- The scaling governor for a given policy object can be changed at any time with the help of the scaling_governor policy attribute in sysfs.

- Some governors expose sysfs attributes to control or fine-tune the scaling algorithms implemented by them.
- Those attributes, referred to as governor tunables, can be either global (system-wide) or per-policy, depending on the scaling driver in use.
- If the driver requires governor tunables to be per-policy, they are located in a subdirectory of each policy directory.
- Otherwise, they are located in a subdirectory under /sys/devices/system/cpu/cpufreq/. In either case the name of the subdirectory containing the governor tunables is the name of the governor providing them.

CPUFreq scaling governors 2 of 2

- performance
  - when attached to a policy object, this governor causes the highest frequency, within the scaling_max_freq policy limit, to be requested for that policy.
  - the request is made once at that time the governor for the policy is set to performance and whenever the scaling_max_freq or scaling_min_freq policy limits change after that.

- powersave
  - when attached to a policy object, this governor causes the lowest frequency, within the scaling_min_freq policy limit, to be requested for that policy.
  - the request is made once at that time the governor for the policy is set to powersave and whenever the scaling_max_freq or scaling_min_freq policy limits change after that.

- userspace
  - does not do anything by itself. Instead, it allows user space to set the CPU frequency for the policy it is attached to by writing to the scaling_setspeed attribute of that policy.

- ondemand
  - uses CPU load as a CPU frequency selection metric.
  - In order to estimate the current CPU load, it measures the time elapsed between consecutive invocations of its worker routine and computes the fraction of that time in which the given CPU was not idle.
  - The ratio of the non-idle (active) time to the total CPU time is taken as an estimate of the load.

- schedutil
  - uses CPU utilization data available from the CPU scheduler. It generally is regarded as a part of the CPU scheduler, so it can access the scheduler’s internal data structures directly.

CPUFreq scaling governor schedutil

- It runs entirely in scheduler context, although in some cases it may need to invoke the scaling driver asynchronously when it decides that the CPU frequency should be changed for a given policy
  - that depends on whether or not the driver is capable of changing the CPU frequency from scheduler context
- The actions of this governor for a particular CPU depend on the scheduling class invoking its utilization update callback for that CPU.
- if it is invoked by the CFS/EAS, the governor will use the PELT/ WALT metric for the root control group of the given CPU as the CPU utilization estimate.
- Then, the new CPU frequency to apply is computed in accordance with the formula
  - f = 1.25 * f_0 * util / max
  - where util is the PELT/WALT number, and
  - max is the theoretical maximum of util, and f_0 is either the maximum possible CPU frequency for the given policy, or the current CPU frequency (otherwise).
- This governor also employs a mechanism allowing it to temporarily bump up the CPU frequency for tasks that have been waiting on I/O most recently, called “IO-wait boosting”.
- That happens when the SCHED_CPUFREQ_IOWAIT flag is passed by the scheduler to the governor callback which causes the frequency to go up to the allowed maximum immediately and then draw back to the value returned by the above formula over time.

CPU idle time management 1 of 2

- In CPU idle states the execution of a program is suspended and instructions belonging to it are not fetched from memory or executed
- CPU idle time management operates on CPUs as seen by the *CPU scheduler* (that is the part of the kernel responsible for the distribution of computational work in the system).
- In its view, CPUs are *logical*. That is, they need not be separate physical entities and may just be interfaces appearing to software as individual single-core processors.
- In other words, a CPU is an entity which appears to be fetching instructions that belong to one program from memory and executing them, but it need not work this way physically.

CPU idle time management 2 of 2

Generally, three different cases of logical CPUs:

1. If the whole processor can only follow one program at a time, it is a CPU.
   a) In that case, if the hardware is asked to enter an idle state, that applies to the processor as a whole.
2. If the processor is multi-core, each core in it is able to follow at least one program at a time.
   a) The cores need not be entirely independent of each other (they may share caches), but still most of the time they work physically in parallel with each other, so if each of them executes only one program, those programs run mostly independently of each other at the same time.
   b) The entire cores are CPUs in that case and if the hardware is asked to enter an idle state, that applies to the core that asked for it in the first place, but it also may apply to a cluster that the core belongs to
3. Each core in a multi-core processor may be able to follow more than one program in the same time frame
   a) Each core may be able to fetch instructions from multiple locations in memory and execute them in the same time frame, but not necessarily entirely in parallel with each other.
   b) In that case the cores present themselves to software as “bundles” each consisting of multiple individual single-core “processors”, referred to as hardware threads, that each can follow one sequence of instructions.
   c) Then, the hardware threads are CPUs from the CPU idle time management perspective and if the processor is asked to enter an idle state by one of them, the hardware thread (or CPU) that asked for it is stopped, but nothing more happens, unless all of the other hardware threads within the same core also have asked the processor to enter an idle state.
   d) In that situation, the core may be put into an idle state individually or a cluster containing it may be put into an idle state as a whole (if the other cores within the cluster are in idle states already).

The Idle loop

- CPU idle time management subsystem called CPUIdle
- The idle loop code takes two major steps in every iteration of it.
  - First, it calls into a code module referred to as **the governor** that belongs to the CPUIdle to select an idle state for the CPU to ask the hardware to enter.
  - Second, it invokes another code module from the CPUIdle subsystem, called **the driver**, to actually ask the processor hardware to enter the idle state selected by the governor.
- The role of the governor is to find an idle state most suitable for the conditions at hand.
- For this purpose, idle states that the hardware can be asked to enter by logical CPUs are represented in an abstract way independent of the platform or the processor architecture and organized in a linear array.
- That array has to be prepared and supplied by the CPUIdle driver matching the platform the kernel is running on at the initialization time.
- This allows CPUIdle governors to be independent of the underlying hardware and to work with any platforms that the Linux kernel can run on.

The Idle states

- Each idle state present in the array is characterized by two parameters to be taken into account by the governor, the target residency and the (worst-case) exit latency.
  - The target residency is the minimum time the hardware must spend in the given state, including the time needed to enter it (which may be substantial), in order to save more energy than it would save by entering one of the shallower idle states instead.
    - The “depth” of an idle state roughly corresponds to the power drawn by the processor in that state.
  - The exit latency, in turn, is the maximum time it will take a CPU asking the processor hardware to enter an idle state to start executing the first instruction after a wakeup from that state.
- In general the exit latency also must cover the time needed to enter the given state in case the wakeup occurs when the hardware is entering it and it must be entered completely to be exited in an ordered manner.

Idle CPUs and scheduler ticks

- The scheduler tick is a timer that triggers periodically in order to implement the time sharing strategy of the CPU scheduler.
- The currently running task may not want to give the CPU away voluntarily, and the scheduler tick is there to make the switch happen regardless.

- If the tick is allowed to trigger on idle CPUs, it will not make sense for them to ask the hardware to enter idle states with target residencies above the tick period length.
  - Since the time of an idle CPU need not be shared between multiple runnable tasks, the primary reason for using the tick goes away if the given CPU is idle.
  - Consequently, it is possible to stop the scheduler tick entirely on idle CPUs in principle, even though that may not always be worth the effort.

- If the tick is stopped and the wakeup does not occur any time soon, the hardware may spend indefinite amount of time in the shallow idle state selected by the governor, which will be a waste of energy.

- Hence, if the governor is expecting a wakeup of any kind within the tick range, it is better to allow the tick trigger

- The systems that run kernels configured to allow the scheduler tick to be stopped on idle CPUs are referred to as tickless systems and they are generally regarded as more energy-efficient than the systems running kernels in which the tick cannot be stopped.

- If the given system is tickless, it will use the menu governor by default and if it is not tickless, the default CPUIdle governor on it will be ladder.

CPUIdle governor decisions

- There are two types of information that can influence the governor's decisions.
  - The governor knows exactly the time until the closest timer event, because
    - the kernel programs timers and it knows exactly when they will trigger, and
    - it is the maximum time the hardware that the given CPU depends on can spend in an idle state, including the time necessary to enter and exit it.
  - However, the CPU may be woken up by a non-timer event at any time (in particular, before the closest timer triggers) and it generally is not known when that may happen.
  - The governor can only see how much time the CPU actually was idle after it has been woken up (that time will be referred to as the idle duration from now on) and it can use that information somehow along with the time until the closest timer to estimate the idle duration in future.
  - How the governor uses that information depends on what algorithm is implemented by it and that is the primary reason for having more than one governor in the CPUIdle subsystem.
  - There are four CPUIdle governors available, menu, TEO, ladder and haltpoll.

CPUIdle governors

- menu: is the default CPUIdle governor for tickless systems. When invoked to select an idle state for a CPU (i.e. an idle state that the CPU will ask the processor hardware to enter), it attempts to predict the idle duration and uses the predicted value for idle state selection.

- TEO (timer events oriented): is an alternative CPUIdle governor for tickless systems

- ladder and haltpoll: for systems with ticks enabled for CPU idle (not interesting for us)