Hardware-Feedback Interface for scheduling on Intel Hardware

Overview

Intel has described the Hardware Feedback Interface (HFI) in the Intel 64 and IA-32 Architectures Software Developer's Manual (Intel SDM) Volume 3 Section 14.6 [1].

The HFI gives the operating system a performance and energy efficiency capability data for each CPU in the system. Linux can use the information from the HFI to influence task placement decisions.

The Hardware Feedback Interface

The Hardware Feedback Interface provides to the operating system information about the performance and energy efficiency of each CPU in the system. Each capability is given as a unit-less quantity in the range [0-255]. Higher values indicate higher capability. Energy efficiency and performance are reported in separate capabilities. Even though on some systems these two metrics may be related, they are specified as independent capabilities in the Intel SDM.

These capabilities may change at runtime as a result of changes in the operating conditions of the system or the action of external factors. The rate at which these capabilities are updated is specific to each processor model. On some models, capabilities are set at boot time and never change. On others, capabilities may change every tens of milliseconds. For instance, a remote mechanism may be used to lower Thermal Design Power. Such change can be reflected in the HFI. Likewise, if the system needs to be throttled due to excessive heat, the HFI may reflect reduced performance on specific CPUs.

The kernel or a userspace policy daemon can use these capabilities to modify task placement decisions. For instance, if either the performance or energy capabilities of a given logical processor becomes zero, it is an indication that the hardware recommends to the operating system to not schedule any tasks on that processor for performance or energy efficiency reasons, respectively.

Implementation details for Linux

The infrastructure to handle thermal event interrupts has two parts. In the Local Vector Table of a CPU's local APIC, there exists a register for the Thermal Monitor Register. This register controls how interrupts are delivered to a CPU when the thermal monitor generates and interrupt. Further details can be found in the Intel SDM Vol. 3 Section 10.5 [1].

The thermal monitor may generate interrupts per CPU or per package. The HFI generates package-level interrupts. This monitor is configured and initialized via a set of machine-specific registers. Specifically, the HFI interrupt and status are controlled via designated bits in the IA32_PACKAGE_THERM_INTERRUPT and IA32_PACKAGE_THERM_STATUS registers, respectively. There exists one HFI table per package. Further details can be found in the Intel SDM Vol. 3 Section 14.9 [1].

The hardware issues an HFI interrupt after updating the HFI table and is ready for the operating system to consume it. CPUs receive such interrupt via the thermal entry in the Local APIC's Local Vector Table.

When servicing such interrupt, the HFI driver parses the updated table and relays the update to userspace using the thermal notification framework. Given that there may be many HFI updates every second, the updates relayed to userspace are throttled at a rate of CONFIG_HZ jiffies.

References

[1] (*1*,*2*,*3*) https://www.intel.com/sdm