`amd-pstate` CPU Performance Scaling Driver

Copyright: © 2021 Advanced Micro Devices, Inc.
Author: Huang Rui <ray.huang@amd.com>

Introduction

`amd-pstate` is the AMD CPU performance scaling driver that introduces a new CPU frequency control mechanism on modern AMD APU and CPU series in Linux kernel. The new mechanism is based on Collaborative Processor Performance Control (CPPC) which provides finer grain frequency management than legacy ACPI hardware P-States. Current AMD CPU/APU platforms are using the ACPI P-states driver to manage CPU frequency and clocks with switching only in 3 P-states. CPPC replaces the ACPI P-states controls and allows a flexible, low-latency interface for the Linux kernel to directly communicate the performance hints to hardware.

`amd-pstate` leverages the Linux kernel governors such as `schedutil`, `ondemand`, etc. to manage the performance hints which are provided by CPPC hardware functionality that internally follows the hardware specification (for details refer to AMD64 Architecture Programmer's Manual Volume 2: System Programming [1]). Currently, `amd-pstate` supports basic frequency control function according to kernel governors on some of the Zen2 and Zen3 processors, and we will implement more AMD specific functions in future after we verify them on the hardware and SBIOS.

AMD CPPC Overview

Collaborative Processor Performance Control (CPPC) interface enumerates a continuous, abstract, and unit-less performance value in a scale that is not tied to a specific performance state / frequency. This is an ACPI standard [2] which software can specify application performance goals and hints as a relative target to the infrastructure limits. AMD processors provide the low latency register model (MSR) instead of an AML code interpreter for performance adjustments. `amd-pstate` will initialize a `struct cpufreq_driver` instance, `amd_pstate_driver`, with the callbacks to manage each performance update behavior.

```
 Highest Perf ------>+-----------------------+                         +-----------------------+
                     |                       |                         |                       |
                     |                       |                         |                       |
                     |                       |          Max Perf  ---->|                       |
                     |                       |                         |                       |
                     |                       |                         |                       |
 Nominal Perf ------>+-----------------------+                         +-----------------------+
                     |                       |                         |                       |
                     |                       |                         |                       |
                     |                       |                         |                       |
                     |                       |                         |                       |
                     |                       |                         |                       |
                     |                       |                         |                       |
                     |                       |      Desired Perf  ---->|                       |
                     |                       |                         |                       |
                     |                       |                         |                       |
                     |                       |                         |                       |
                     |                       |                         |                       |
                     |                       |                         |                       |
                     |                       |                         |                       |
                     |                       |                         |                       |
                     |                       |                         |                       |
                     |                       |                         |                       |
  Lowest non-        |                       |                         |                       |
  linear perf ------>+-----------------------+                         +-----------------------+
                     |                       |                         |                       |
                     |                       |       Lowest perf  ---->|                       |
                     |                       |                         |                       |
  Lowest perf ------>+-----------------------+                         +-----------------------+
                     |                       |                         |                       |
                     |                       |                         |                       |
                     |                       |                         |                       |
          0   ------>+-----------------------+                         +-----------------------+

                                     AMD P-States Performance Scale
```

AMD CPPC Performance Capability

Highest Performance (RO)

This is the absolute maximum performance an individual processor may reach, assuming ideal conditions. This performance level may not be sustainable for long durations and may only be achievable if other platform components are in a specific state; for example, it may require other processors to be in an idle state. This would be equivalent to the highest frequencies supported by the processor.

Nominal (Guaranteed) Performance (RO)

This is the maximum sustained performance level of the processor, assuming ideal operating conditions. In the absence of an external constraint (power, thermal, etc.), this is the performance level the processor is expected to be able to maintain continuously. All cores/processors are expected to be able to sustain their nominal performance state simultaneously.

Lowest non-linear Performance (RO)

This is the lowest performance level at which nonlinear power savings are achieved, for example, due to the combined effects of voltage and frequency scaling. Above this threshold, lower performance levels should be generally more energy efficient than higher performance levels. This register effectively conveys the most efficient performance level to `amd-pstate`.

Lowest Performance (RO)

This is the absolute lowest performance level of the processor. Selecting a performance level lower than the lowest nonlinear performance level may cause an efficiency penalty but should reduce the instantaneous power consumption of the processor.

AMD CPPC Performance Control

`amd-pstate` passes performance goals through these registers. The register drives the behavior of the desired performance target.

Minimum requested performance (RW)

`amd-pstate` specifies the minimum allowed performance level.

Maximum requested performance (RW)

`amd-pstate` specifies a limit the maximum performance that is expected to be supplied by the hardware.

Desired performance target (RW)

`amd-pstate` specifies a desired target in the CPPC performance scale as a relative number. This can be expressed as percentage of nominal performance (infrastructure max). Below the nominal sustained performance level, desired performance expresses the average performance level of the processor subject to hardware. Above the nominal performance level, the processor must provide at least nominal performance requested and go higher if current operating conditions allow.

Energy Performance Preference (EPP) (RW)

This attribute provides a hint to the hardware if software wants to bias toward performance (0x0) or energy efficiency (0xff).

Key Governors Support

`amd-pstate` can be used with all the (generic) scaling governors listed by the `scaling_available_governors` policy attribute in sysfs. Then, it is responsible for the configuration of policy objects corresponding to CPUs and provides the CPUFreq core (and the scaling governors attached to the policy objects) with accurate information on the maximum and minimum operating frequencies supported by the hardware. Users can check the scaling_cur_freq information comes from the CPUFreq core.

amd-pstate mainly supports schedutil and ondemand for dynamic frequency control. It is to fine tune the processor configuration on amd-pstate to the schedutil with CPU CFS scheduler. amd-pstate registers the adjust_perf callback to implement performance update behavior similar to CPPC. It is initialized by sugov_start and then populates the CPU's update_util_data pointer to assign sugov_update_single_perf as the utilization update callback function in the CPU scheduler. The CPU scheduler will call cpufreq_update_util and assigns the target performance according to the struct sugov_cpu that the utilization update belongs to. Then, amd-pstate updates the desired performance according to the CPU scheduler assigned.

Processor Support

The amd-pstate initialization will fail if the _CPC entry in the ACPI SBIOS does not exist in the detected processor. It uses acpi_cpc_valid to check the existence of _CPC. All Zen based processors support the legacy ACPI hardware P-States function, so when amd-pstate fails initialization, the kernel will fall back to initialize the acpi-cpufreq driver.

There are two types of hardware implementations for amd-pstate: one is Full MSR Support and another is Shared Memory Support. It can use the :c:macro:`X86_FEATURE_CPPC` feature flag to indicate the different types. (For details, refer to the Processor Programming Reference (PPR) for AMD Family 19h Model 51h, Revision A1 Processors [3].) amd-pstate is to register different static_call instances for different hardware implementations.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\admin-guide\pm\ (linux-master) (Documentation) (admin-guide) (pm) amd-pstate.rst, line 195); *backlink*

Unknown interpreted text role "c:macro".

Currently, some of the Zen2 and Zen3 processors support amd-pstate. In the future, it will be supported on more and more AMD processors.

Full MSR Support

Some new Zen3 processors such as Cezanne provide the MSR registers directly while the :c:macro:`X86_FEATURE_CPPC` CPU feature flag is set. amd-pstate can handle the MSR register to implement the fast switch function in CPUFreq that can reduce the latency of frequency control in interrupt context. The functions with a pstate_xxx prefix represent the operations on MSR registers.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\admin-guide\pm\ (linux-master) (Documentation) (admin-guide) (pm) amd-pstate.rst, line 209); *backlink*

Unknown interpreted text role "c:macro".

Shared Memory Support

If the :c:macro:`X86_FEATURE_CPPC` CPU feature flag is not set, the processor supports the shared memory solution. In this case, amd-pstate uses the cppc_acpi helper methods to implement the callback functions that are defined on static_call. The functions with the cppc_xxx prefix represent the operations of ACPI CPPC helpers for the shared memory solution.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\admin-guide\pm\ (linux-master) (Documentation) (admin-guide) (pm) amd-pstate.rst, line 219); *backlink*

Unknown interpreted text role "c:macro".

AMD P-States and ACPI hardware P-States always can be supported in one processor. But AMD P-States has the higher priority and if it is enabled with :c:macro:`MSR_AMD_CPPC_ENABLE` or cppc_set_enable, it will respond to the request from AMD P-States.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\admin-guide\pm\ (linux-master) (Documentation) (admin-guide) (pm) amd-pstate.rst, line 226); *backlink*

Unknown interpreted text role "c:macro".

User Space Interface in sysfs

amd-pstate exposes several global attributes (files) in sysfs to control its functionality at the system level. They are located in the /sys/devices/system/cpu/cpufreq/policyX/ directory and affect all CPUs.

```
root@hr-test1:/home/ray# ls /sys/devices/system/cpu/cpufreq/policy0/*amd*
/sys/devices/system/cpu/cpufreq/policy0/amd_pstate_highest_perf
/sys/devices/system/cpu/cpufreq/policy0/amd_pstate_lowest_nonlinear_freq
/sys/devices/system/cpu/cpufreq/policy0/amd_pstate_max_freq
```

amd_pstate_highest_perf / amd_pstate_max_freq

Maximum CPPC performance and CPU frequency that the driver is allowed to set, in percent of the maximum supported CPPC performance level (the highest performance supported in AMD CPPC Performance Capability). In some ASICs, the highest CPPC performance is not the one in the _CPC table, so we need to expose it to sysfs. If boost is not active, but still supported, this maximum frequency will be larger than the one in cpuinfo. This attribute is read-only.

amd_pstate_lowest_nonlinear_freq

The lowest non-linear CPPC CPU frequency that the driver is allowed to set, in percent of the maximum supported CPPC performance level. (Please see the lowest non-linear performance in AMD CPPC Performance Capability.) This attribute is read-only.

Other performance and frequency values can be read back from /sys/devices/system/cpu/cpuX/acpi_cppc/, see :ref:`cppc_sysfs`.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\linux-master\Documentation\admin-guide\pm\ (linux-master) (Documentation) (admin-guide) (pm) amd-pstate.rst, line 264); *backlink*

Unknown interpreted text role "ref".

amd-pstate vs acpi-cpufreq

On the majority of AMD platforms supported by acpi-cpufreq, the ACPI tables provided by the platform firmware are used for CPU performance scaling, but only provide 3 P-states on AMD processors. However, on modern AMD APU and CPU series, hardware provides the Collaborative Processor Performance Control according to the ACPI protocol and customizes this for AMD platforms. That is, fine-grained and continuous frequency ranges instead of the legacy hardware P-states. amd-pstate is the kernel module which supports the new AMD P-States mechanism on most of the future AMD platforms. The AMD P-States mechanism is the more performance and energy efficiency frequency management method on AMD processors.

Kernel Module Options for amd-pstate

shared_mem Use a module param (shared_mem) to enable related processors manually with **amd_pstate.shared_mem=1**. Due to the performance issue on the processors with Shared Memory Support, we disable it presently and will re-enable this by default once we address performance issue with this solution.

To check whether the current processor is using Full MSR Support or Shared Memory Support :

```
ray@hr-test1:~$ lscpu | grep cppc
Flags:                           fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush mmx fxsr sse sse2 ht sys
```

If the CPU flags have `cppc`, then this processor supports Full MSR Support. Otherwise, it supports Shared Memory Support.

`cpupower` tool support for `amd-pstate`

`amd-pstate` is supported by the `cpupower` tool, which can be used to dump frequency information. Development is in progress to support more and more operations for the new `amd-pstate` module with this tool.

```
root@hr-test1:/home/ray# cpupower frequency-info
analyzing CPU 0:
  driver: amd-pstate
  CPUs which run at the same hardware frequency: 0
  CPUs which need to have their frequency coordinated by software: 0
  maximum transition latency: 131 us
  hardware limits: 400 MHz - 4.68 GHz
  available cpufreq governors: ondemand conservative powersave userspace performance schedutil
  current policy: frequency should be within 400 MHz and 4.68 GHz.
                  The governor "schedutil" may decide which speed to use
                  within this range.
  current CPU frequency: Unable to call hardware
  current CPU frequency: 4.02 GHz (asserted by call to kernel)
  boost state support:
    Supported: yes
    Active: yes
    AMD PSTATE Highest Performance: 166. Maximum Frequency: 4.68 GHz.
    AMD PSTATE Nominal Performance: 117. Nominal Frequency: 3.30 GHz.
    AMD PSTATE Lowest Non-linear Performance: 39. Lowest Non-linear Frequency: 1.10 GHz.
    AMD PSTATE Lowest Performance: 15. Lowest Frequency: 400 MHz.
```

Diagnostics and Tuning

Trace Events

There are two static trace events that can be used for `amd-pstate` diagnostics. One of them is the `cpu_frequency` trace event generally used by `CPUFreq`, and the other one is the `amd_pstate_perf` trace event specific to `amd-pstate`. The following sequence of shell commands can be used to enable them and see their output (if the kernel is configured to support event tracing).

```
root@hr-test1:/home/ray# cd /sys/kernel/tracing/
root@hr-test1:/sys/kernel/tracing# echo 1 > events/amd_cpu/enable
root@hr-test1:/sys/kernel/tracing# cat trace
# tracer: nop
#
# entries-in-buffer/entries-written: 47827/42233061   #P:2
#
#                                _-----=> irqs-off
#                               / _----=> need-resched
#                              | / _---=> hardirq/softirq
#                              || / _--=> preempt-depth
#                              ||| /     delay
#           TASK-PID     CPU#  ||||   TIMESTAMP  FUNCTION
#              | |         |   ||||      |         |
         <idle>-0       [015] dN...  4995.979886: amd_pstate_perf: amd_min_perf=85 amd_des_perf=85 amd_max_perf=166 cpu_id=15 changed=fal
         <idle>-0       [007] d.h..  4995.979893: amd_pstate_perf: amd_min_perf=85 amd_des_perf=85 amd_max_perf=166 cpu_id=7 changed=fals
            cat-2161    [000] d....  4995.980841: amd_pstate_perf: amd_min_perf=85 amd_des_perf=85 amd_max_perf=166 cpu_id=0 changed=fals
           sshd-2125    [004] d.s..  4995.980968: amd_pstate_perf: amd_min_perf=85 amd_des_perf=85 amd_max_perf=166 cpu_id=4 changed=fals
         <idle>-0       [007] d.s..  4995.980968: amd_pstate_perf: amd_min_perf=85 amd_des_perf=85 amd_max_perf=166 cpu_id=7 changed=fals
         <idle>-0       [003] d.s..  4995.980971: amd_pstate_perf: amd_min_perf=85 amd_des_perf=85 amd_max_perf=166 cpu_id=3 changed=fals
         <idle>-0       [011] d.s..  4995.980996: amd_pstate_perf: amd_min_perf=85 amd_des_perf=85 amd_max_perf=166 cpu_id=11 changed=fal
```

The `cpu_frequency` trace event will be triggered either by the `schedutil` scaling governor (for the policies it is attached to), or by the `CPUFreq` core (for the policies with other scaling governors).

Tracer Tool

`amd_pstate_tracer.py` can record and parse `amd-pstate` trace log, then generate performance plots. This utility can be used to debug and tune the performance of `amd-pstate` driver. The tracer tool needs to import intel pstate tracer.

Tracer tool located in `linux/tools/power/x86/amd_pstate_tracer`. It can be used in two ways. If trace file is available, then directly parse the file with command

```
./amd_pstate_trace.py [-c cpus] -t <trace_file> -n <test_name>
```

Or generate trace file with root privilege, then parse and plot with command

```
sudo ./amd_pstate_trace.py [-c cpus] -n <test_name> -i <interval> [-m kbytes]
```

The test result can be found in `results/test_name`. Following is the example about part of the output.

```
common_cpu  common_secs  common_usecs  min_perf  des_perf  max_perf  freq    mperf   apef    tsc       load   duration_ms  sample_num  el
CPU_005     712          116384        39        49        166       0.7565  9645075 2214891 38431470  25.1   11.646       469         2
CPU_006     712          116408        39        49        166       0.6769  8950227 1839034 37192089  24.06  11.272       470         2
```

Reference

[1] AMD64 Architecture Programmer's Manual Volume 2: System Programming,
https://www.amd.com/system/files/TechDocs/24593.pdf

[2] Advanced Configuration and Power Interface Specification,
https://uefi.org/sites/default/files/resources/ACPI_Spec_6_4_Jan22.pdf

[3] Processor Programming Reference (PPR) for AMD Family 19h Model 51h, Revision A1 Processors
https://www.amd.com/system/files/TechDocs/56569-A1-PUB.zip