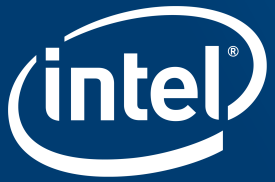


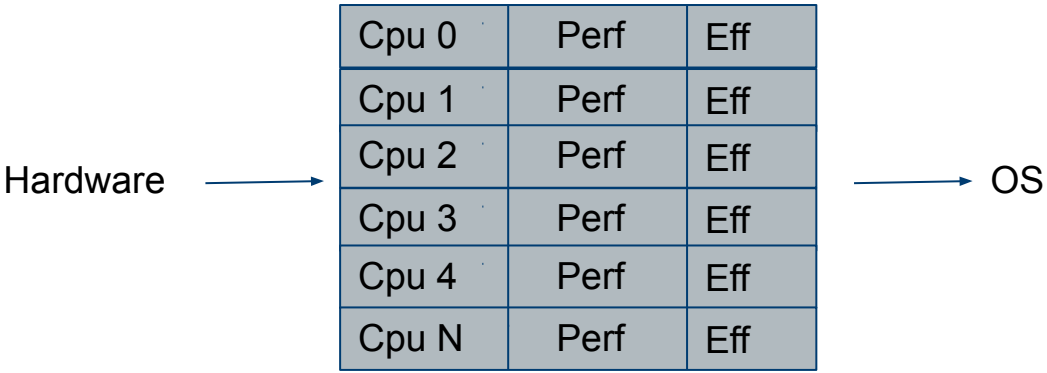
CPU offline using Hardware Feedback Interface

Linux Plumbers 2020

Background

- Intel® Speed Select Technology-Performance Profile
  - Dynamically allow to increase TDP by keeping some specific cores in C6 (+95% residency)
  - Current deployment model is driven by user space (two step process)
    - User gives command to switch
    - Then based on the profile, offline/online CPUs
      - Trust user to identify correct CPUs for offline/online
    - Exploring using hardware feedback Interface to avoid user error

Intel® Hardware Feedback Interface



Special : Perf == 0 to indicate, don’t schedule anything on that CPU.
This needs something more than setting arch_scale_cpu_capacity().

Implementation choice for perf == 0

- CPU online/offline
- Idle injection
- Use arch_scale_cpu_capacity()

Trigger CPU online/offline from OS

- Use add_cpu() and remove_cpu()
  - Standard interface from include/linux/cpu.h
  - Used at other places in the kernel
  - CPU0 online/offline is conditional
    - CONFIG option or kernel command line
    - Conditional, may fail on some platforms

Idle Injection

- Natural choice, as we want to force some CPUs to idle
  - But not for thermal urgency but for long term
- Can’t set idle_inject_set_duration: run_time to 0
- Any low run time to meet +95% C6, causes issues
  - For example on a live system:

INFO: task kworker/x: blocked for more than xx seconds.

NOHZ: local_softirq_pending …

Marking TSC unstable due to clocksource watchdog

NOHZ: local_softirq_pending 20a

NETDEV WATCHDOG: enp1s0 (igb): transmit queue 1 timed out

Use arch_scale_cpu_capacity()

- We will have this support for other platforms to notify asym performance
- On a busy (overloaded)system, every CPU will be used irrespective of any capacity

Capacity == 0 or 1 on CPU 2 and 6

```
  1  [||||||||||||||||||||||||||||||||||||||||||||||||100.0%]   5  [||||||||||||||||||||||||||||||||||||||||||||||||100.0%]
  2  [||||||||||||||||||||||||||||||||||||||||||||||||100.0%]   6  [||||||||||||||||||||||||||||||||||||||||||||||||100.0%]
  3  [||||||||||||||||||||||||||||||||||||||||||||||||100.0%]   7  [||||||||||||||||||||||||||||||||||||||||||||||||100.0%]
  4  [||||||||||||||||||||||||||||||||||||||||||||||||100.0%]   8  [||||||||||||||||||||||||||||||||||||||||||||||||100.0%]
  Mem[||||||||                                    1.28G/15.4G]   Tasks: 129, 248 thr; 8 running
  Swp[                                              0K/2.00G]   Load average: 6.02 4.69 2.46
                                                                Uptime: 00:28:10
```

PID	USER	PRI	NI	VIRT	RES	SHR	S	CPU%	MEM%	TIME+	Command
810	labuser	20	0	111M	2920	28	S	0.0	0.0	0:00.00	(sd-pam)
1606	labuser	20	0	29792	4904	3228	S	0.0	0.0	0:00.03	-bash
1683	labuser	20	0	29920	5400	3588	S	0.0	0.0	0:00.06	-bash
1759	labuser	20	0	29792	4992	3392	S	0.0	0.0	0:00.04	-bash
1835	labuser	20	0	29792	4832	3156	S	0.0	0.0	0:00.03	-bash
1931	**root**	20	0	29812	5144	3492	S	0.0	0.0	0:00.05	/bin/bash
809	labuser	20	0	76944	8220	6752	S	0.0	0.1	0:00.04	/lib/systemd/systemd --user
328	**root**	19	-1	100M	20292	19328	S	0.0	0.1	0:00.29	/lib/systemd/systemd-journald
625	**root**	20	0	70700	6100	5320	S	0.0	0.0	0:00.05	/lib/systemd/systemd-logind
551	**systemd-r**	20	0	70628	5392	4828	S	0.0	0.0	0:00.05	/lib/systemd/systemd-resolved
356	**root**	20	0	44780	5532	3140	S	0.0	0.0	0:00.29	/lib/systemd/systemd-udevd
794	**root**	20	0	23068	2260	2124	S	0.0	0.0	0:00.00	/sbin/agetty -o -p -- \u --keep-baud 115200,38400,9600

This can be addressed:

```
  1  [||||||||||||||||||||||||||||||||||||||||||||||||100.0%]   5  [||||||||||||||||||||||||||||||||||||||||||||||||100.0%]
  2  [                                                  0.0%]   6  [                                                  0.0%]
  3  [||||||||||||||||||||||||||||||||||||||||||||||||100.0%]   7  [||||||||||||||||||||||||||||||||||||||||||||||||100.0%]
  4  [||||||||||||||||||||||||||||||||||||||||||||||||100.0%]   8  [||||||||||||||||||||||||||||||||||||||||||||||||100.0%]
  Mem[||||||||                                     1.27G/15.4G]   Tasks: 127, 248 thr; 7 running
  Swp[                                               0K/2.00G]   Load average: 5.90 4.13 2.00
                                                                 Uptime: 00:26:09

    PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
    810 labuser    20   0  111M  2920    28 S  0.0  0.0  0:00.00 (sd-pam)
   1606 labuser    20   0 29792  4904  3228 S  0.0  0.0  0:00.03 -bash
   1683 labuser    20   0 29920  5400  3588 S  0.0  0.0  0:00.06 -bash
   1759 labuser    20   0 29792  4992  3392 S  0.0  0.0  0:00.04 -bash
   1835 labuser    20   0 29792  4832  3156 S  0.0  0.0  0:00.03 -bash
   1931 root       20   0 29812  5144  3492 S  0.0  0.0  0:00.04 /bin/bash
    809 labuser    20   0 76944  8220  6752 S  0.0  0.1  0:00.04 /lib/systemd/systemd --user
    328 root       19  -1  100M 20236 19272 S  0.0  0.1  0:00.28 /lib/systemd/systemd-journald
    625 root       20   0 70700  6100  5320 S  0.0  0.0  0:00.05 /lib/systemd/systemd-logind
    551 systemd-r  20   0 70628  5392  4828 S  0.0  0.0  0:00.04 /lib/systemd/systemd-resolved
    356 root       20   0 44780  5532  3140 S  0.0  0.0  0:00.29 /lib/systemd/systemd-udevd
    794 root       20   0 23068  2260  2124 S  0.0  0.0  0:00.00 /sbin/agetty -o -p -- \u --keep-baud 115200,38400,9600
    830 root       20   0 25996  6288  5012 S  0.0  0.0  0:00.01 /sbin/dhclient -d -q -sf /usr/lib/NetworkManager/nm-dhc
      1 root       20   0  220M  9464  6692 S  0.0  0.1  0:04.59 /sbin/init _nomodeset
    633 root       20   0 45228  5228  4692 S  0.0  0.0  0:00.00 /sbin/wpa_supplicant -u -s -O /run/wpa_supplicant
   1040 labuser    20   0 49928  4068  3596 S  0.0  0.0  0:00.00 /usr/bin/dbus-daemon --config-file=/usr/share/defaults/
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
```

Summary of changes

- Treat capacity_orig_of(cpu) == 0 special
  - Wakeup_path (fast, slow); Don't select this CPU
  - Busy load balance, same as !cpu_active(target_cpu)
  - Idle load balance, same as !cpu_active(target_cpu)

Some differences with CPU online/offline

- CPUs are still part of CPU set, so user can still use taskset
- RT tasks and interrupts are still scheduled on those CPUs
- CPU offline moves the affenitized tasks

Questions

- Is there any issue is doing special scheduling for capacity_orig_of() == 0?
  - May be it already means something on other platforms.
- What type of tests to execute to check side effect?
- What other issues, I have to look for?