Balancing Power and Performance in the Linux Kernel

Kristen Accardi

intel Software

Intel OpenSource TECHNOLOGY CENTER

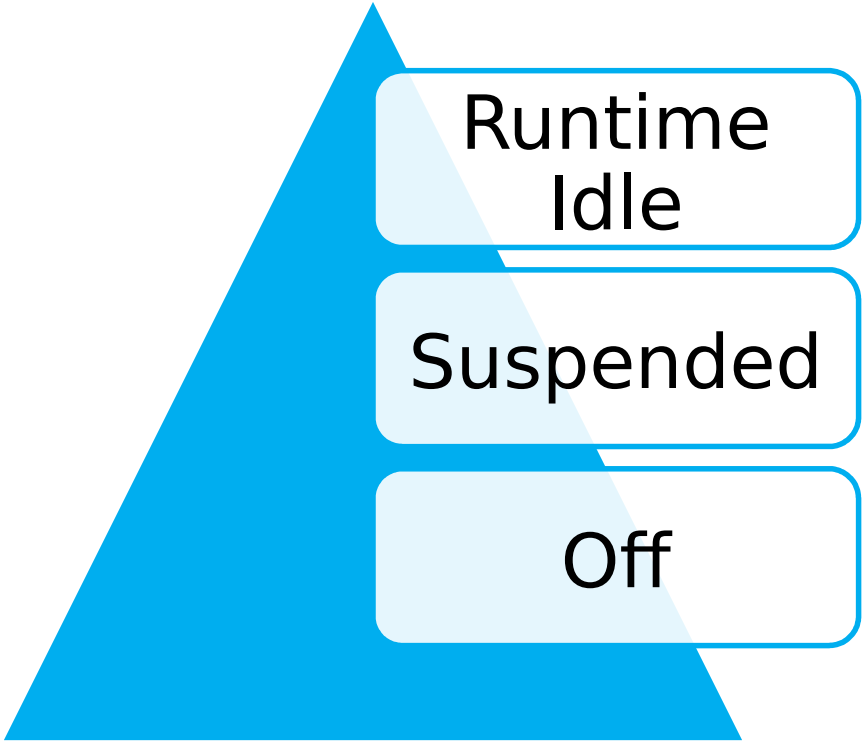
Power Management basics

P-states on Intel platforms

Limitations of OS p-state selection

Hardware P-states (HWP)

Degrees of Idleness



Suspend

- User Initiated
- User tasks frozen
- Devices forced idle
  - d-states
- S-states or Idle
- S3/S0iX/MWAIT

Runtime Idle

- Opportunistic
- User tasks scheduled
- Opportunistic device idle
  - d-states
- CPU C-states (S0iX)

Active Power Management

Active Power Management in Linux – Dim the Lights

- CPU active power management (cpufreq)
  - Aka P-states or DVFS
- Device Active Power management
  - Some device support such as PCIe ASPM
  - Device specific how to activate
- GPU does it's own thing

P-States != Frequency

P-States != Power

How do we pick the right p-state?

- Governors reflect user policy decision
  - intel_pstate supports only “powersave” and “performance” policies. Other drivers support more policies.
  - intel_pstate is actually a governor and hw driver all in one, whereas traditionally the governor is separate from the hw driver.
- intel_pstate “performance” policy always picks the highest p-state
  - Race to Halt – or just don't care about energy
- intel_pstate “powersave” policy attempts to balance performance with energy savings
- The driver looks at utilization and capacity to determine whether to increase or decrease the p-state. This is similar to many other governors.

P-state basics

P-states on Intel platforms

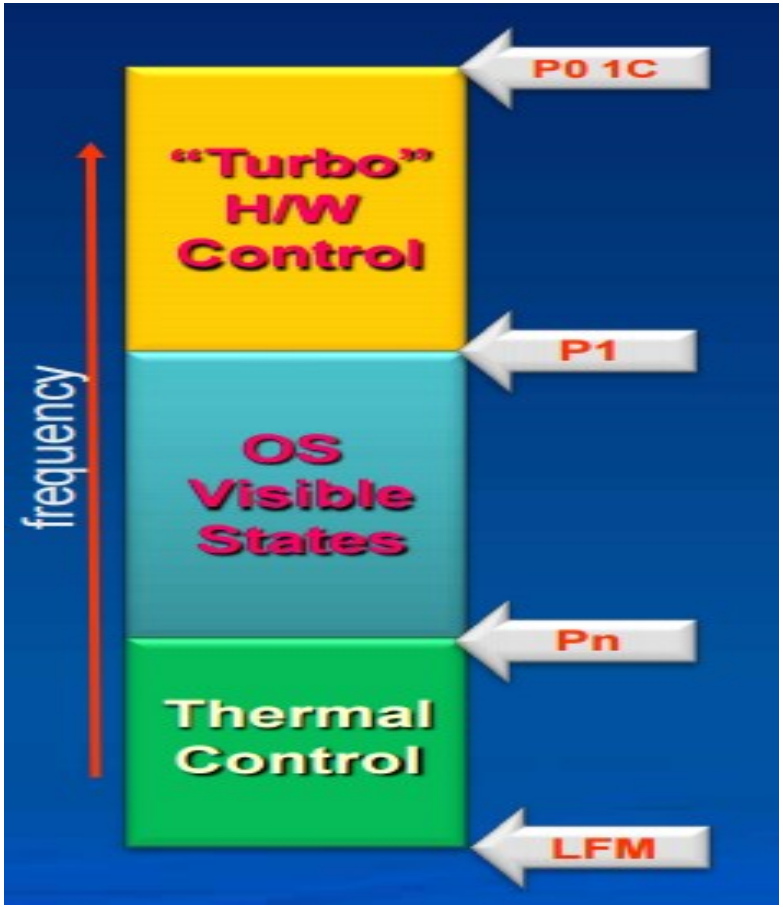
Limitations of OS p-state selection

Hardware P-states (HWP)

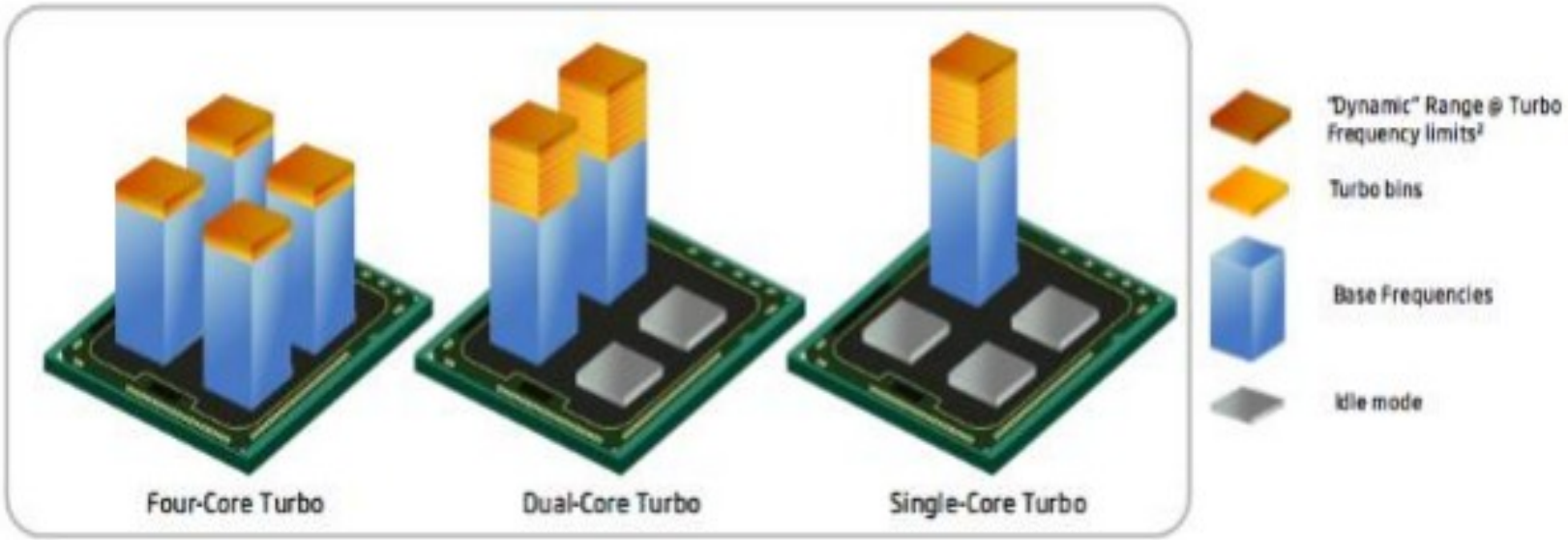
P1 - Pn is Guaranteed

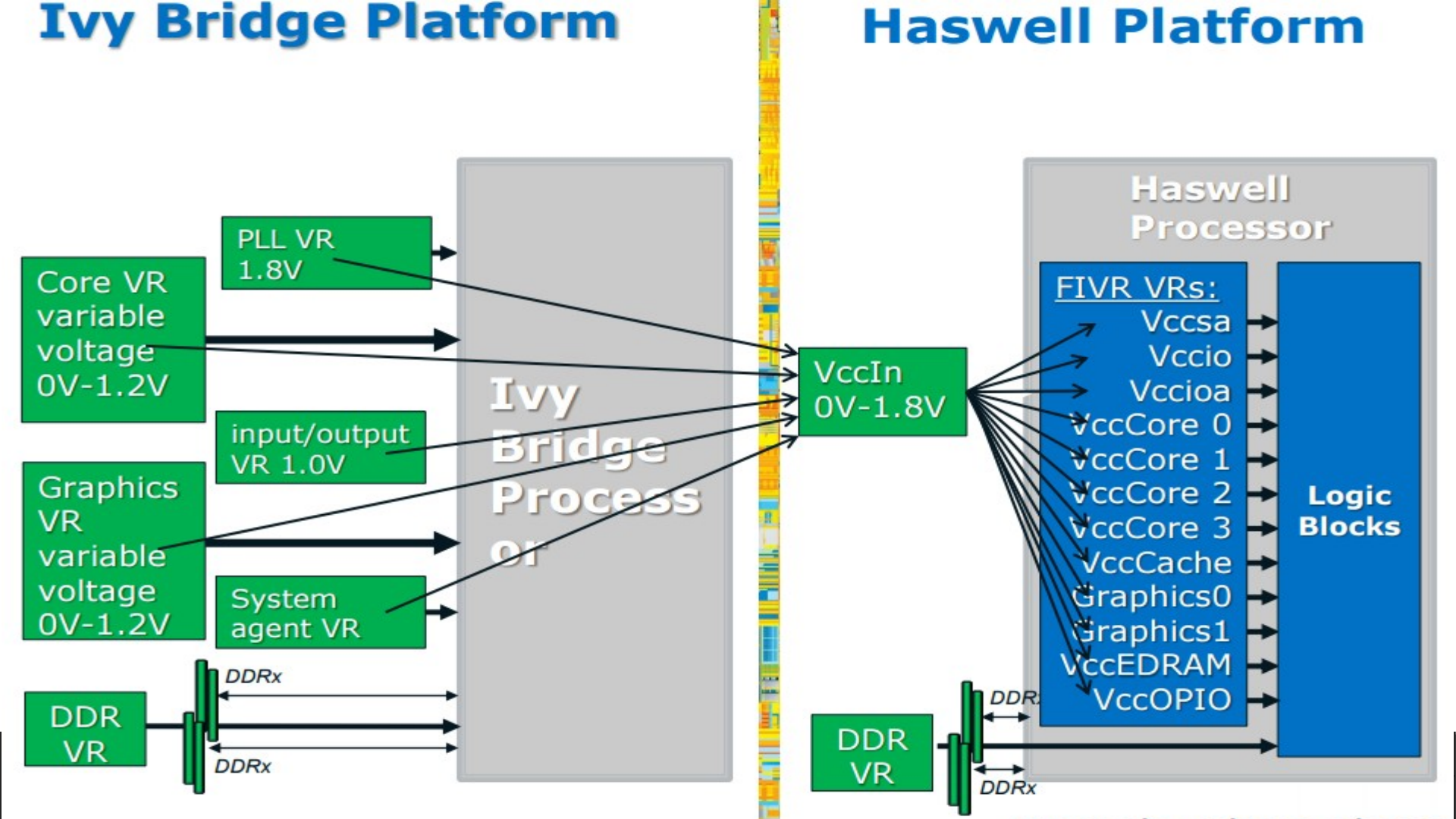
P0 - P1 is turbo

Pn - LFM for Thermal



How Turbo works (Intel Speed Step®)




Ivy Bridge Platform
Haswell Platform
PLL VR
1.8V
Core VR
variable
voltage
0V-1.2V
input/output
VR 1.0V
Graphics
VR
variable
voltage
0V-1.2V
System
agent VR
DDR
VR
DDRx
DDRx
Ivy
Bridge
Process
or
VccIn
0V-1.8V
Haswell
Processor
FIVR VRs:
Vccsa
Vccio
Vccioa
VccCore 0
VccCore 1
VccCore 2
VccCore 3
VccCache
Graphics0
Graphics1
VccEDRAM
VccOPIO
Logic
Blocks
DDR
VR
DDRx
DDRx

Hardware Coordination of P-states

- Cores which share the same voltage domain vote for a p-state

Hardware Coordination of P-states

- Cores which share the same voltage domain vote for a p-state
- The highest p-state for each core wins

Hardware Coordination of P-states

- Cores which share the same voltage domain vote for a p-state
- The highest p-state for each core wins
- APERF/MPERF must be used to see what p-state was granted

Hardware Coordination of P-states

- Cores which share the same voltage domain vote for a p-state
- The highest p-state for each core wins
- APERF/MPERF must be used to see what p-state was granted
- acpi_cpufreq lies!!!!

P-state basics

P-states on Intel platforms

Limitations of OS p-state selection

Hardware P-states (HWP)

Limitations of OS P-state selection

- Capacity/Utilization is insufficient for determining whether to scale

Limitations of OS P-state selection

- Capacity/Utilization is insufficient for determining whether to scale
- Sample rate may cause incorrect utilization calculation

Limitations of OS P-state selection

- Capacity/Utilization is insufficient for determining whether to scale
- Sample rate may cause incorrect utilization calculation
- Scaling benefits unclear

P-state basics

P-states on Intel platforms

Limitations of OS p-state selection

Hardware P-states (HWP)

Intel® Speed Shift Technology (HWP)

- Most Efficient Frequency is Calculated at Runtime (Pe)
  - Depends on system & workload
- EPP is Energy Performance Preference – will dictate how aggressive the algorithm (Pa)
  - Depends on system, workload, OS
- Algorithm will operate between Pa and Pe

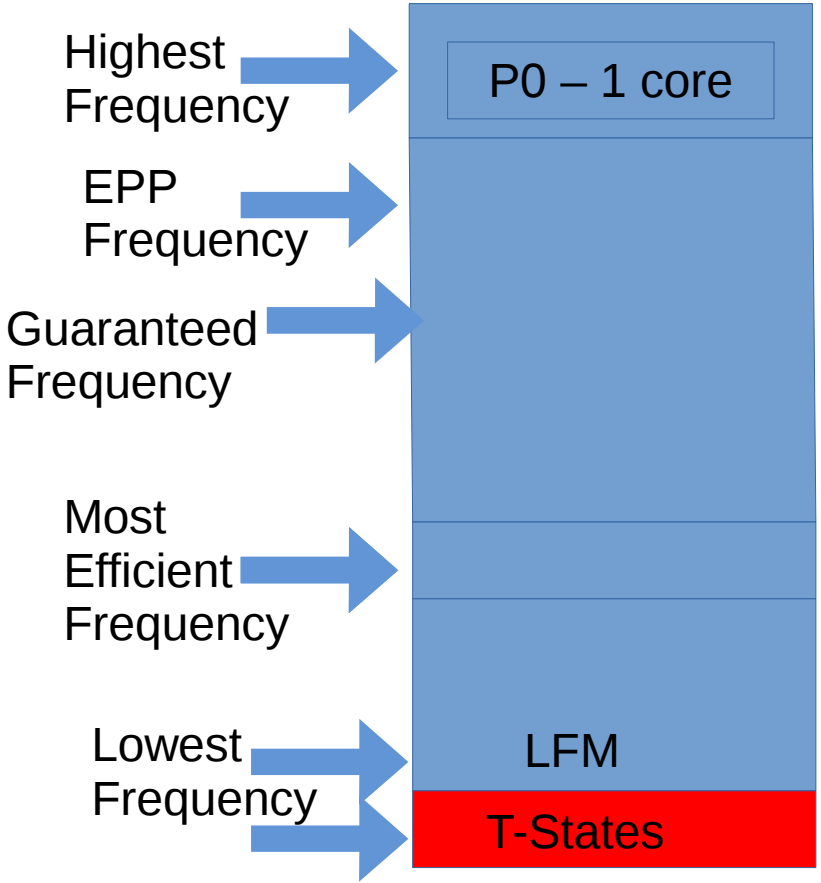


Table 14-1. Architectural and Non-Architectural MSRs Related to HWP

Address	Architectural	Register Name	Description
770H	Y	IA32_PM_ENABLE	Enable/Disable HWP.
771H	Y	IA32_HWP_CAPABILITIES	Enumerates the HWP performance range (static and dynamic).
772H	Y	IA32_HWP_REQUEST_PKG	Conveys OSPM's control hints (Min, Max, Activity Window, Energy Performance Preference, Desired) for all logical processor in the physical package.
773H	Y	IA32_HWP_INTERRUPT	Controls HWP native interrupt generation (Guaranteed Performance changes, excursions).
774H	Y	IA32_HWP_REQUEST	Conveys OSPM's control hints (Min, Max, Activity Window, Energy Performance Preference, Desired) for a single logical processor.
777H	Y	IA32_HWP_STATUS	Status bits indicating changes to Guaranteed Performance and excursions to Minimum Performance.
19CH	Y	IA32_THERM_STATUS[bits 15:12]	Conveys reasons for performance excursions
64EH	N	MSR_PPERF	Productive Performance Count.

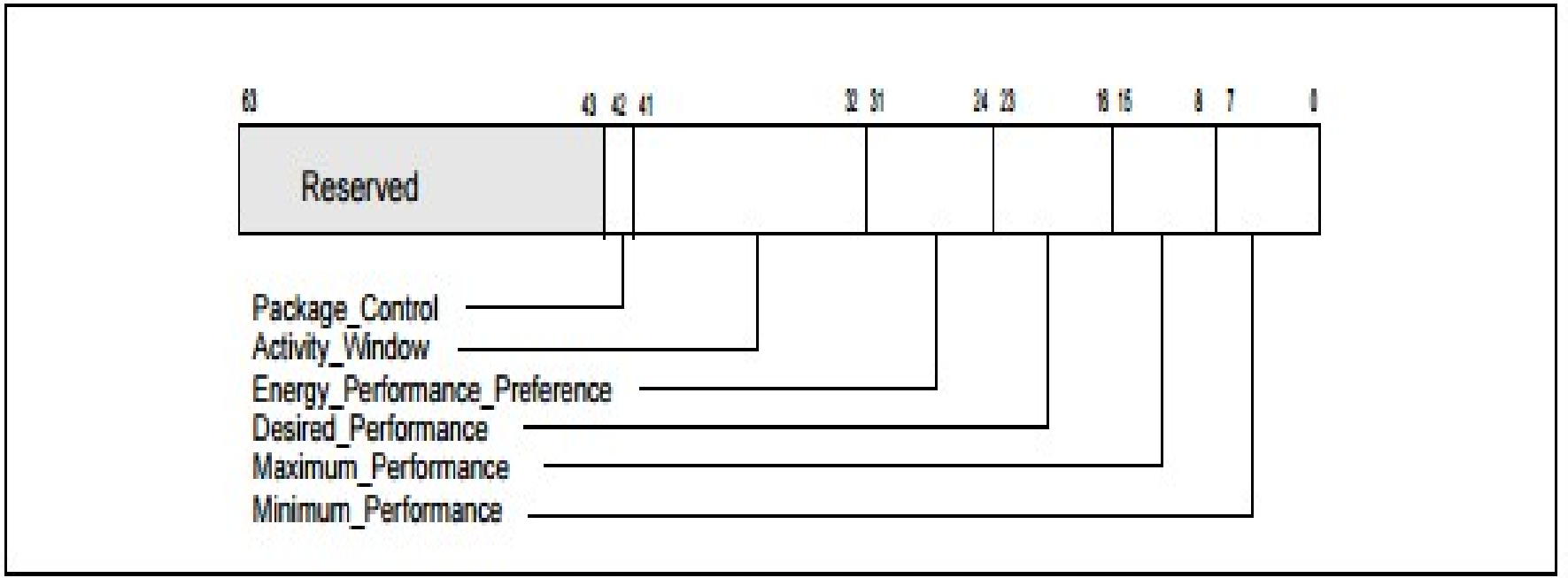


Figure 14-7. IA32_HWP_REQUEST Register

Linux Implementation

- intel_pstate driver checks cpuflags for support
- Enabled by default for whitelisted CPUs
- Autonomous mode only
- No EPP exposed today
- Min and Max pstate can be requested via min and max perf_pct sysfs files

References

1. http://www.hotchips.org/wp-content/uploads/hc_archives/hc23/HC23.19.9-Desktop-CPUs/HC23.19.921.SandyBridge_Power_10-Rotem-Intel.pdf
2. http://events.linuxfoundation.org/sites/events/files/slides/LinuxCon_Japan_2015_idle_injection1_0.pdf
3. http://www.hotchips.org/wp-content/uploads/hc_archives/hc25/HC25.80-Processors2-epub/HC25.27.820-Haswell-Hammarlund-Intel.pdf
4. http://www.intel.com/content/dam/www/public/us/en/documents/manuals/64-ia-32-architectures-software-developer-manual-325462.pdf

Q & A