Dynamic voltage and frequency scaling in modern Intel® TCC equipped architectures

Matthias Hahn, Markus Schweikhardt

Notices & Disclaimers

- Performance varies by use, configuration and other factors. Learn more on the Performance Index site.
- Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.
- Your costs and results may vary.
- Intel technologies may require enabled hardware, software or service activation.
- Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.
- The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.
- Altering clock frequency or voltage may void any product warranties and reduce stability, security, performance, and life of the processor and other components. Check with system and component manufacturers for details.
- Customer is responsible for safety of the overall system, including compliance with applicable safety-related requirements or standards.
-

Agenda

- Introduction
- Dynamic Voltage & Frequency Scaling
- Hybrid Intel Architectures
- Measurements
- Next Steps, Q&A

Presenters

Matthias Hahn
PhD CS

Markus Schweikhardt
MS El Eng

- SW Application Engineers
- Intel NEX NESG CEED CASE EMEA
- Focus: Industrial Real-Time Applications

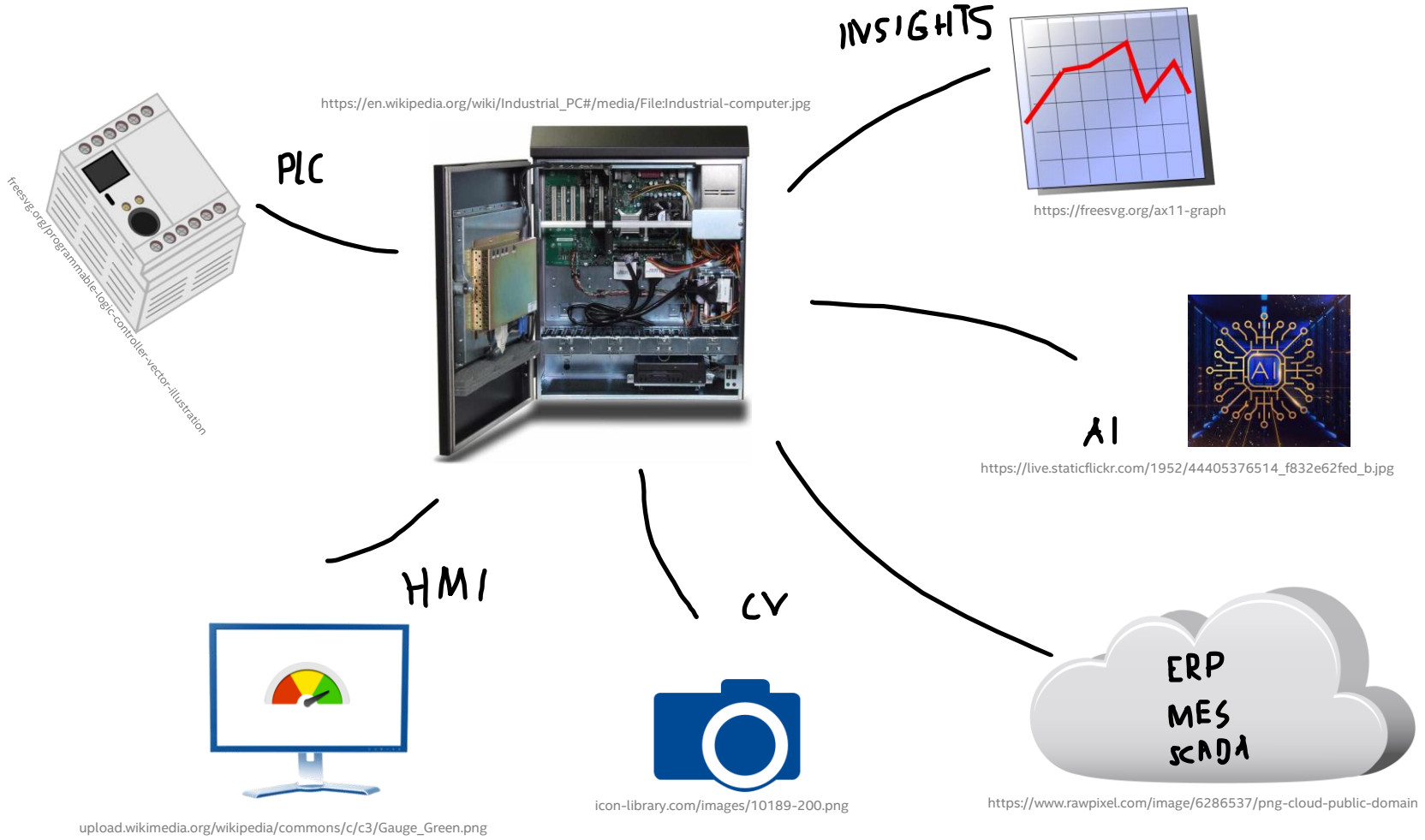
Mixed Criticality Workload Consolidation

Real-time Applications: vPLC, Motion etc..

+

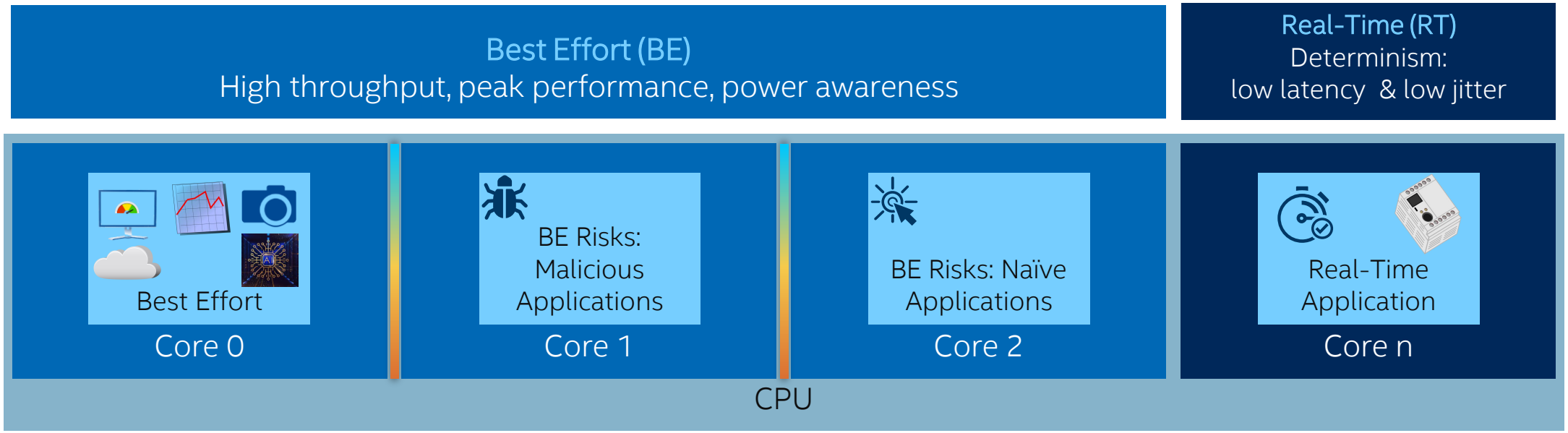
Best Effort Applications: HMI, AI, CV, etc..

Needs:
- QoS on HW Platform
- Real-time capable SW stack (BIOS, Hypervisor, OS, etc)



Illustrations provided for representative purposes only

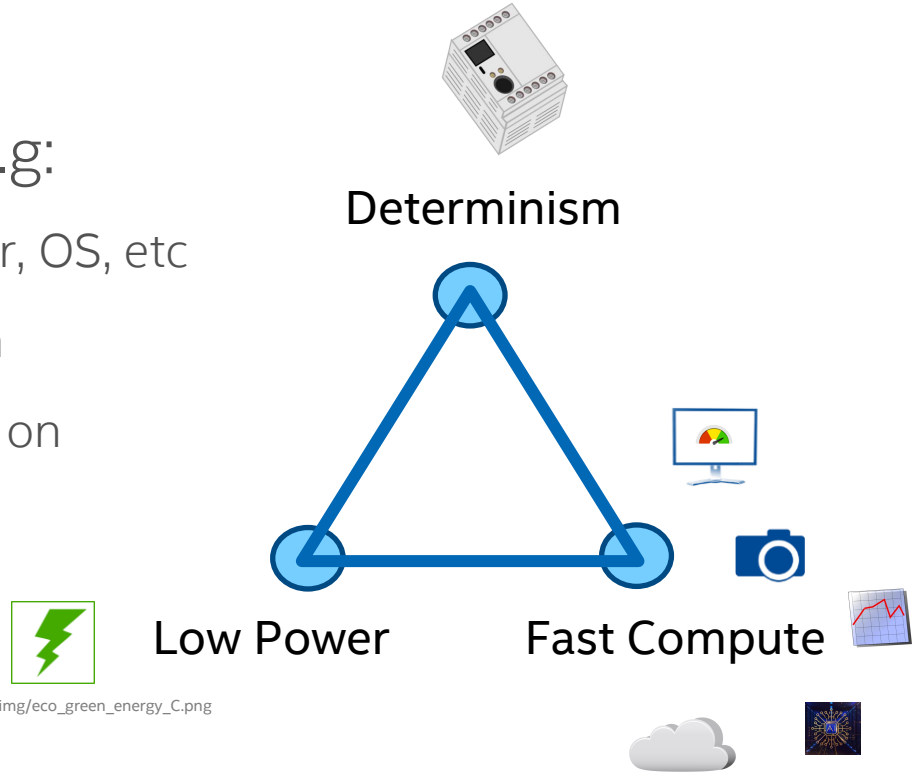
Mixed-Criticality Goal: Temporal Isolation



Concurrent Mixed-Criticality Workloads

Concurrent mixed criticality workloads may:

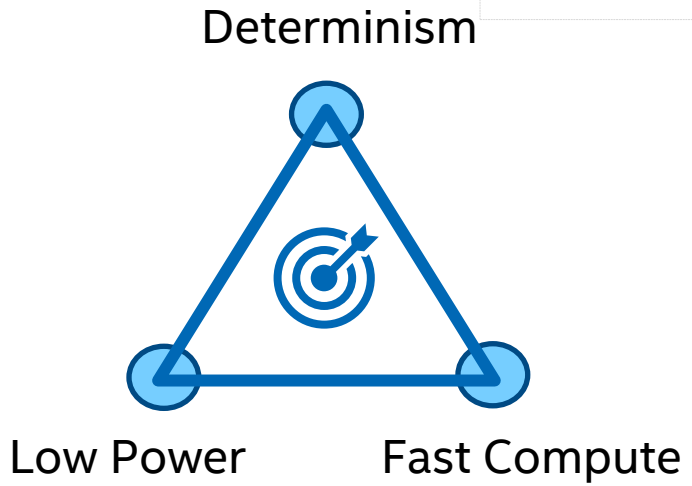
- Compete on shared resources, like cache, I/O buffers, etc ➔ Intel® TCC
- Impact each other, or even whole platform, e.g:
  - MSR, SMI, cache flush, split locks, … ➔ BIOS, hypervisor, OS, etc
  - Power, current, or thermal stress ➔ Cooling solution
  - Power savings, dynamic freqs ➔ Addressed later on
- Compete on constraints

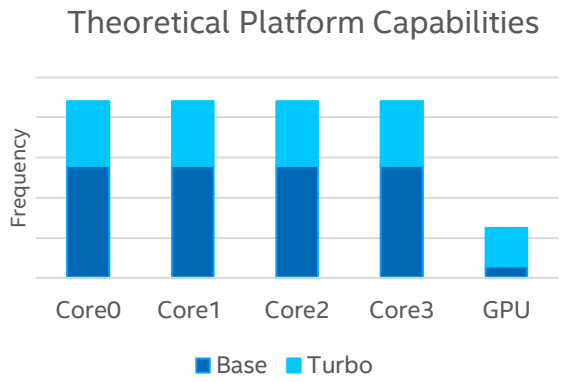


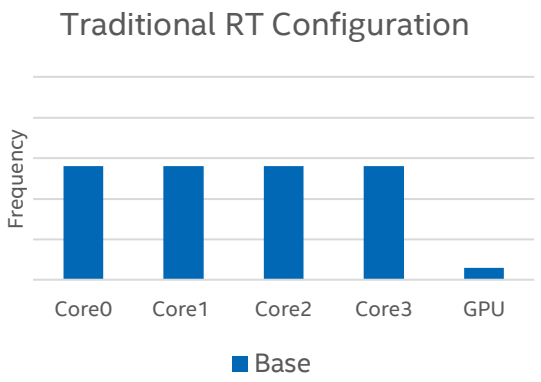
https://freesvg.org/img/eco_green_energy_C.png

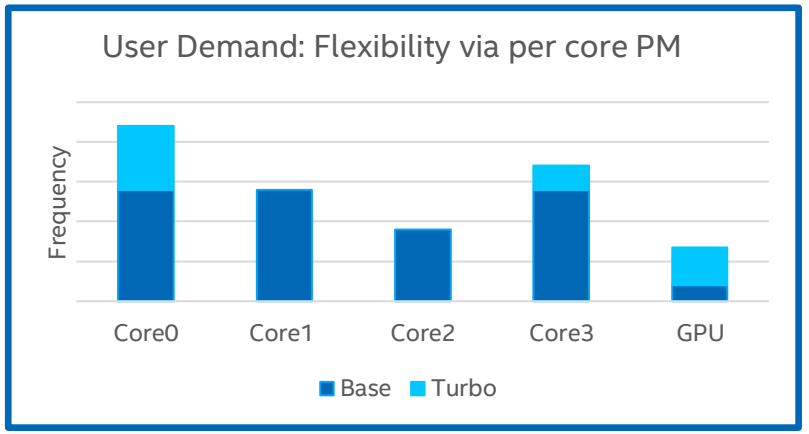
Dynamic Frequency and Voltage Scaling (DVFS)

- May impact all cores
- May yield jitter
  - Depending on Power Management (PM) subsystem
  - Dynamic frequency adjustments on an active core may impact compute latencies of all active cores ⇒ increased jitter (e.g. additional >10 µs)





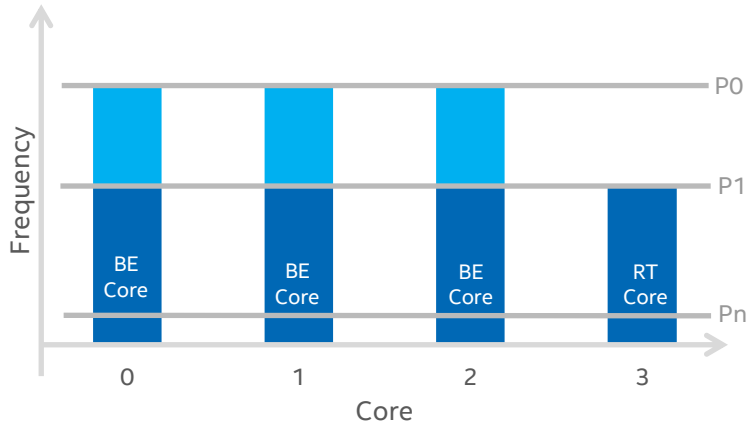




Illustrations provided for representative purposes only

DVFS Use Cases

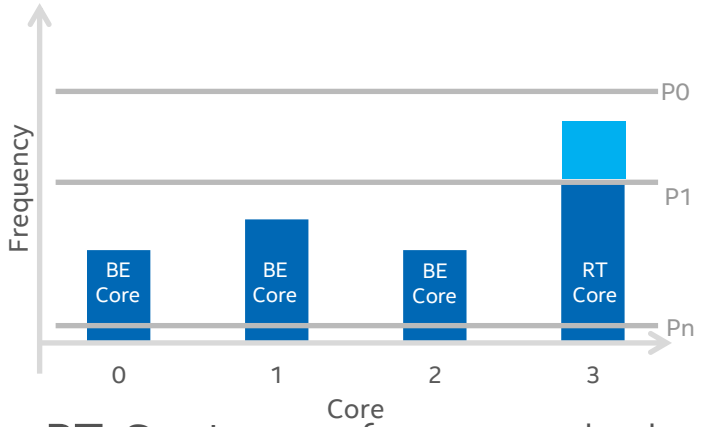
Standard – BE high peak Perf



RT: Static core frequency locked to base frequency

BE : Dynamic core frequency

High RT Perf



RT: Static core frequency locked to Turbo frequency

BE: Max frequency capped to stay within a given TDP target.

→ PM subsystem with optimized DVFS to provide more flexibility starting with **11th Gen Intel® Core™ Processors**

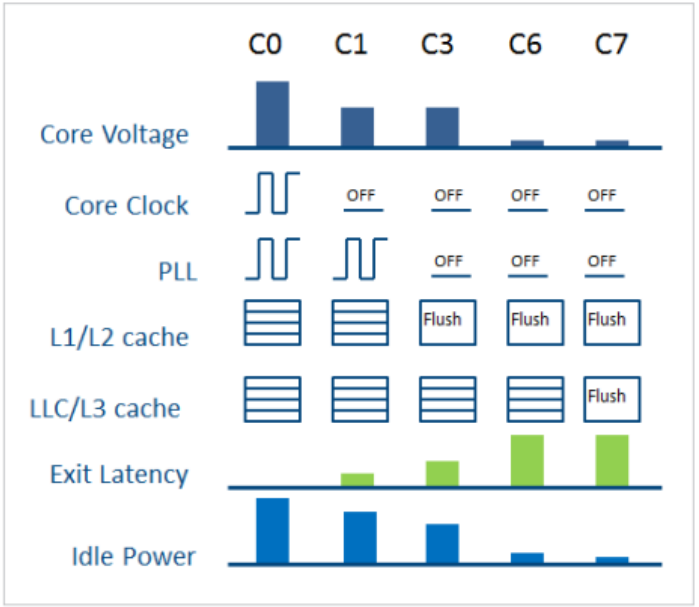
Illustrations provided for representative purposes only

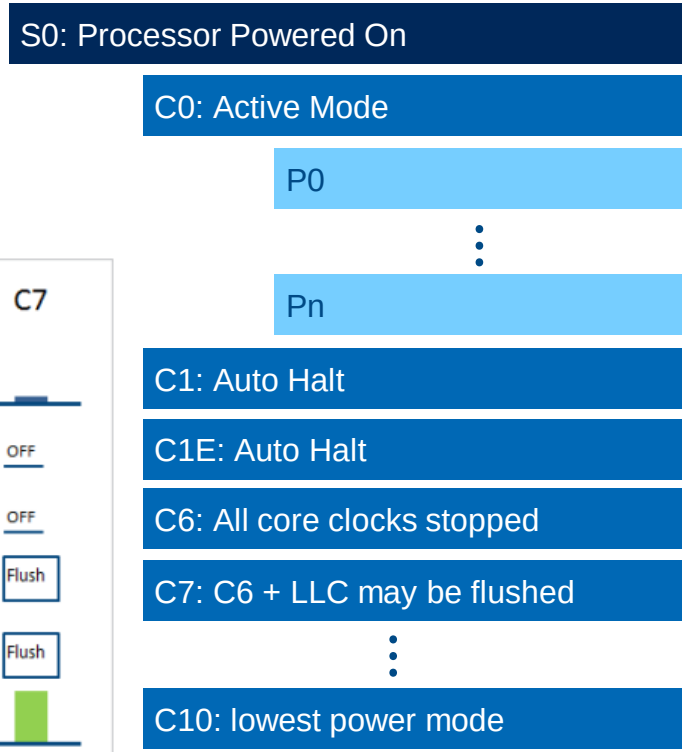
Core Frequency: P0 = Turbo; P1 = Base; Pn = most efficient

Processer Power Management in a Nutshell

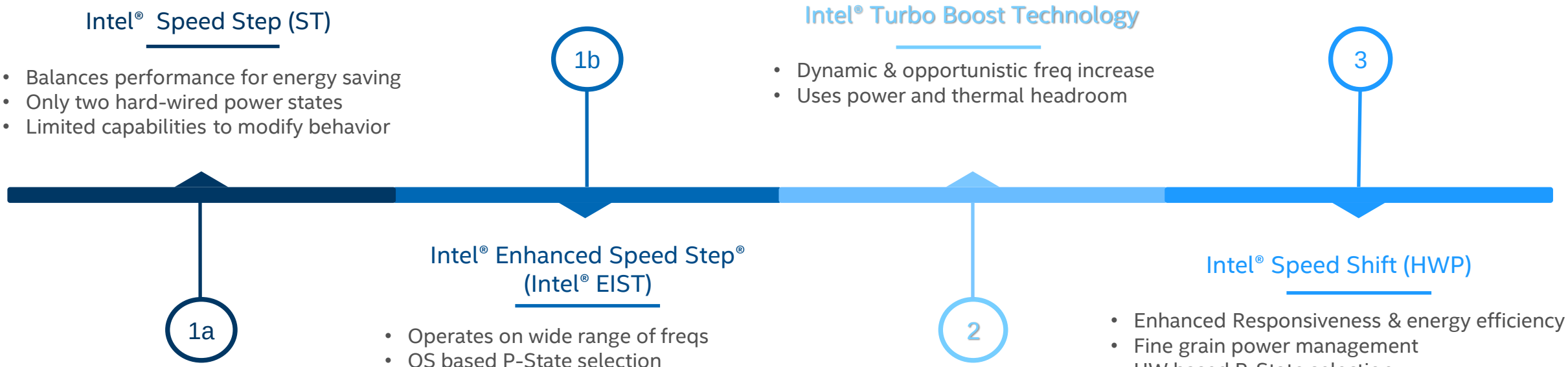
Processor Power Management

- Ways to reduce power consumption, e.g.
  - Switch off subsystems
  - Reduce voltage, or frequency resp
- HW support via ACPI states, e.g. C- & P-States
- C-States (idle states)
  - C0: CPU active
  - C1, … ↗: power ↘ , latency [→ C0] ↗, residency ↗
- P-States (executing states)
  - Operating points (voltages & freqs) of a CPU core
  - P0, … ↗: freqs & power ↘, compute performance ↘
  - Intel DVFS technologies e.g.
    - Intel® EIST, Intel® Speed Shift Technology (HWP), Intel® Turbo Boost Technology





Intel Power Management



HWP control

- cpuid: HWP support
- pstate driver (intel_pstate) may expose attributes via sysfs
- Direct interaction via MSRs

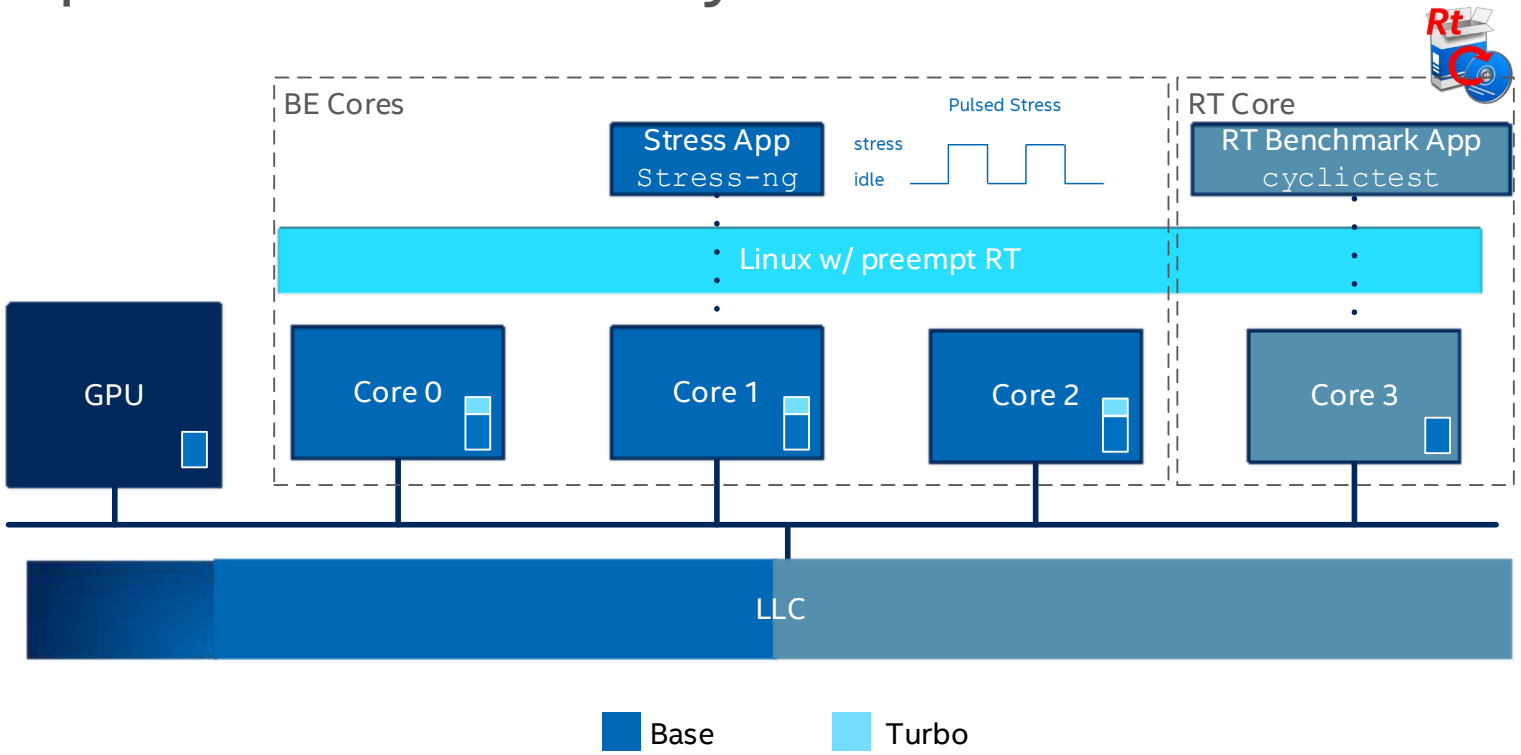
- IA32_HWP_CAPABILITIES:
  - HWP enabling
  - HWP performance range enumeration
  - OS performance hints
- IA32_HWP_REQUEST
  - Energy Performance Preference (EPP)
  - Min, max, and desired performance

intel_pstate: CPU Performance Scaling Driver (Linux)

- Control per logical CPU
- Sysfs exposure via
  - `/sys/devices/system/cpu/intel_pstate/`
  - `/sys/devices/system/cpu/cpu*/cpufreq/`
- Two modes: Active & Passive
- Active:
  - Default mode on HWP capable processors
  - Explicit enabling via cmdline "`-intel_pstate=active`"
  - HWP is automatically enabled
  - CPUs select P-States
  - 2 schemes for P-State selection provided:
    - "powersave": Selection logic favoring power optimization. CPU utilization ↗ ⇒ P state ↘
    - "performance": Selection logic favoring performance

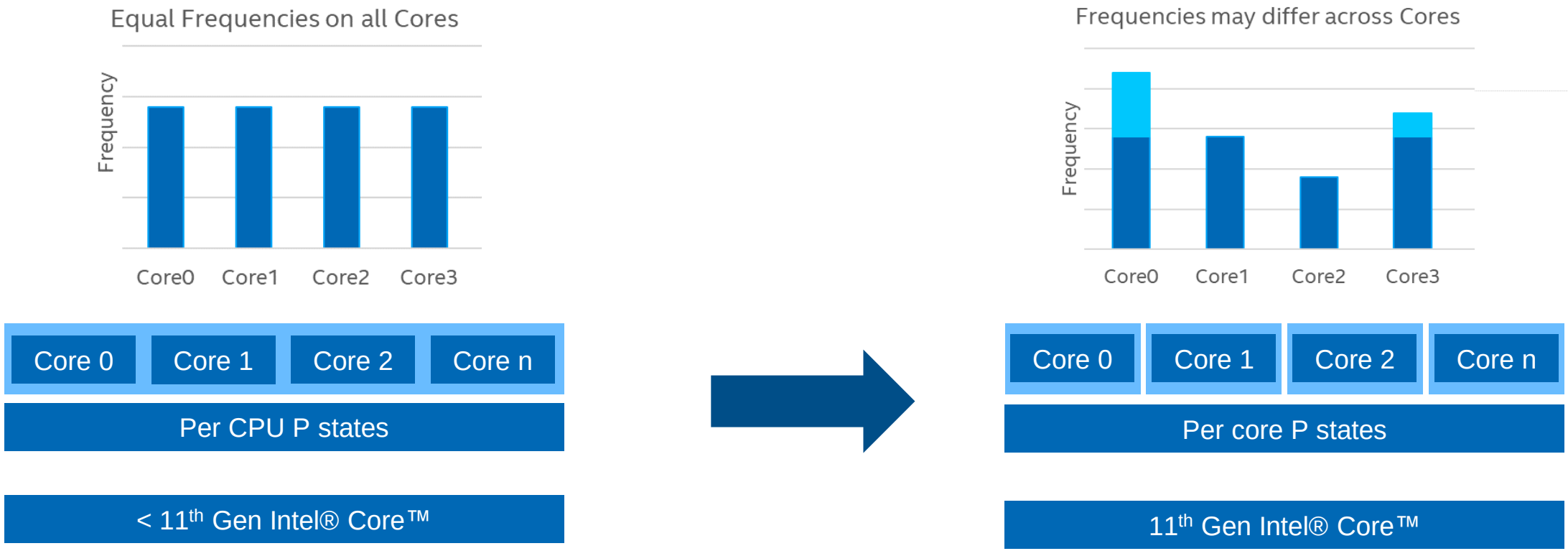
Analysis

Test Setup for the Analysis



- Pulsed CPU centric stress on best-effort core to enforce P-state transitions
  - `taskset -c 1 stress-ng -c 1 -l 100`
- cyclictest on isolated real-time core to measure the real-time performance
  - `cyclictest -t1 -a3 -D 12h -p99 -m -i250 -h400 -q -latency=1 >output`

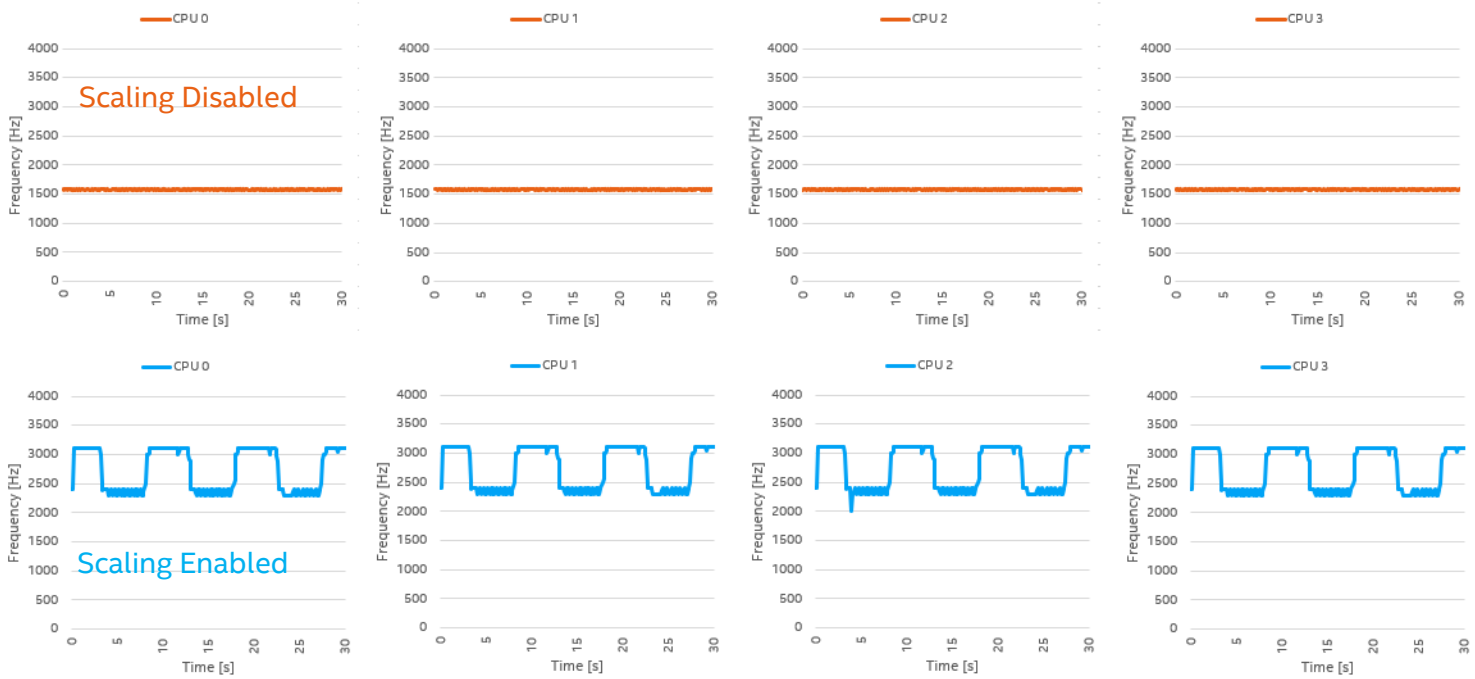
DVFS Feature Enhancements



11th Gen Intel® Core™ introduced

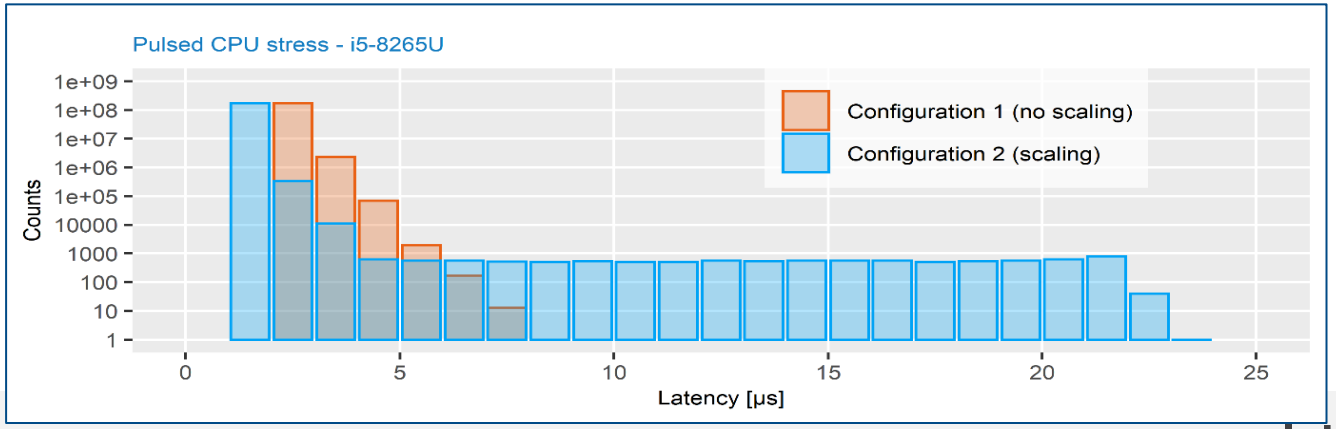
- Per core PLL ⇒ Per core P-States
- Digital PLLs with fast P-State transition latencies and local impact to core

DVFS impact on Intel® Core™ i5-8265U



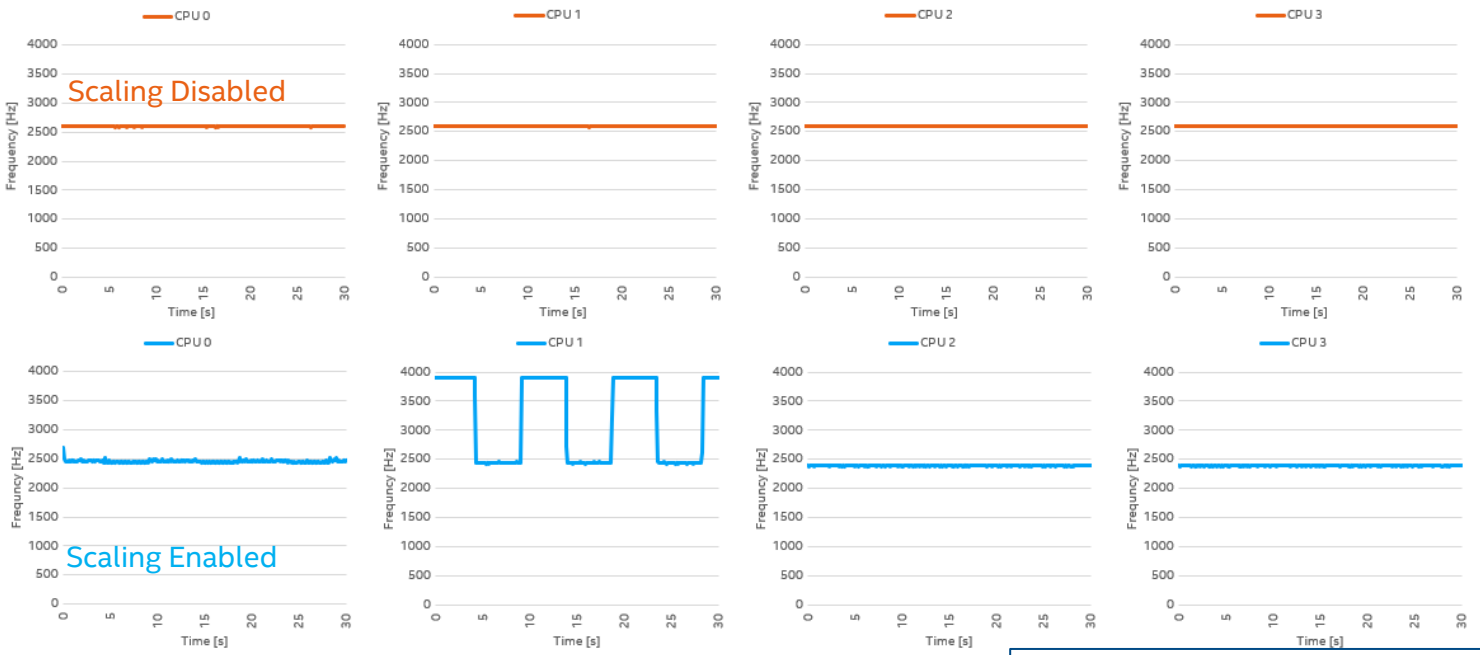
P-state transition on
< 11th Gen Intel® Core™ Processors:
Latency hit to all cores >10 us

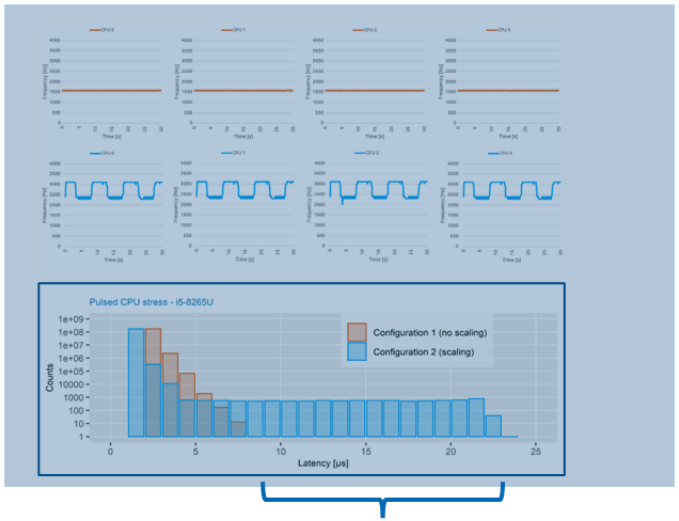
➔ Not acceptable for many RT apps



For workloads and configurations see backup

Results on Intel® Core™ i5-1145GRE



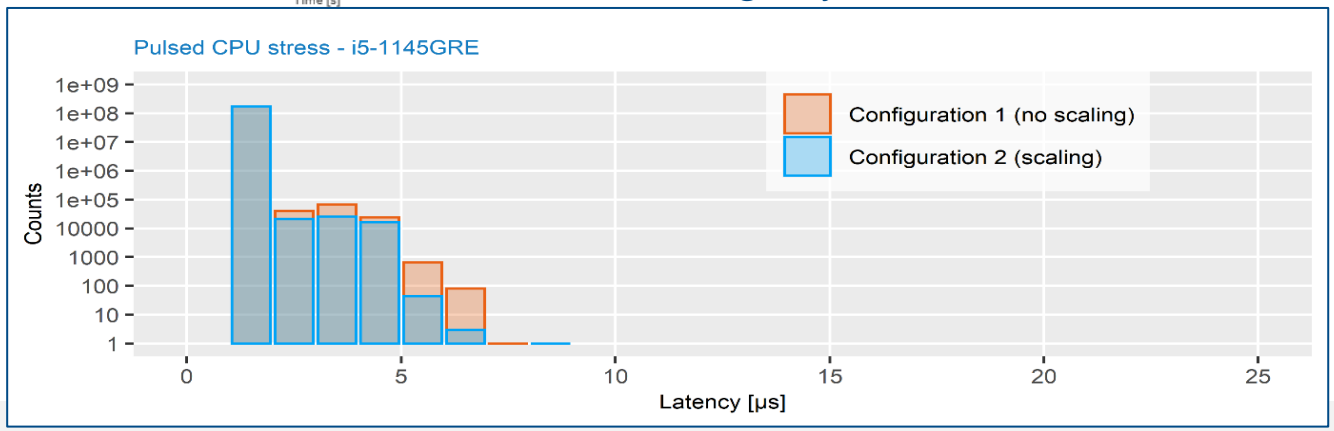


Results on i5-8265U

Higher jitter on 8th Gen Intel® Core

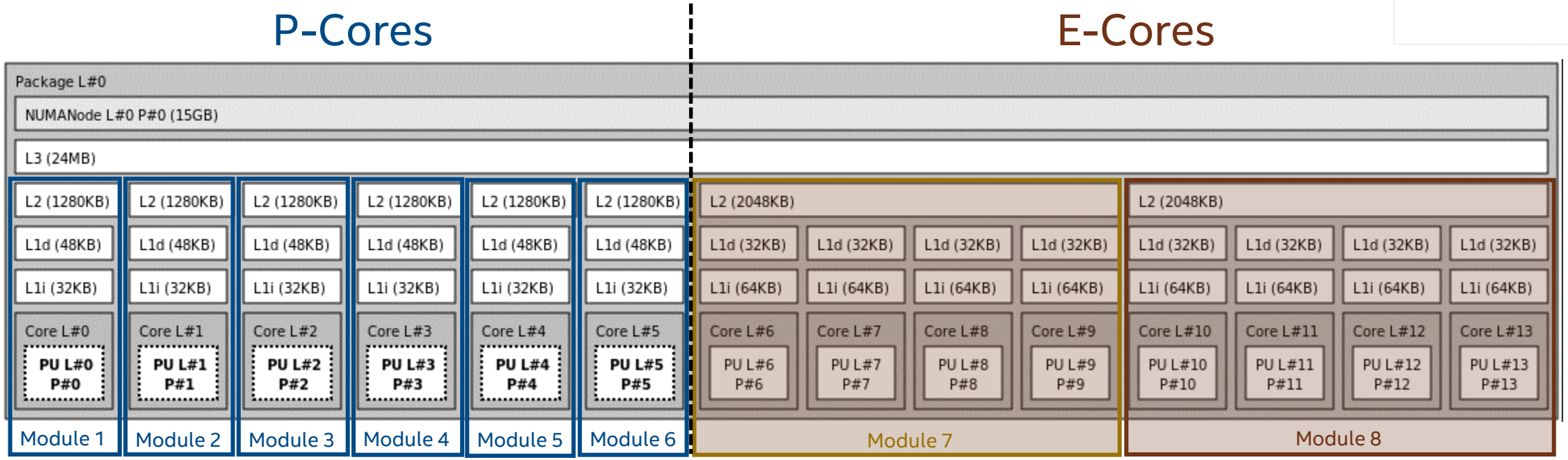
DVFS on BE cores w/o significant impact to RT performance of 11th Gen Intel® Core™ Processors

➔ More flexibility for Mixed Criticality RT system designs



For workloads and configurations see backup

Hybrid Architecture - Example

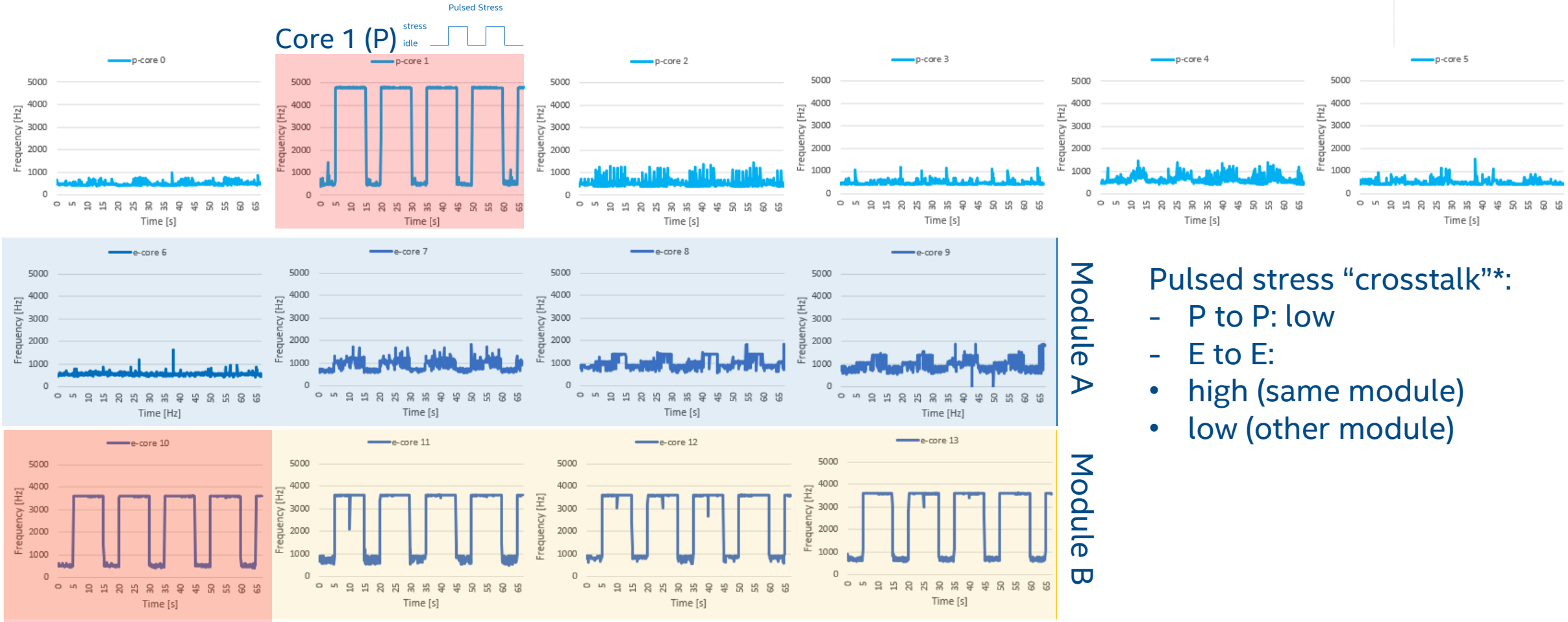


Hybrid architecture Platforms e.g. on 12th, or 13th Gen Intel® Core™

- Each Module: 1 Performance (P-) Core + 1 PLL => per Core P-State
- Each Module: 4 Efficiency (E-) Cores + 1 PLL => per Module P-State

* 13th Gen Intel(R) Core(TM) i7-1370PE

Impact of Frequency Changes to P- and E-Cores



Pulsed stress "crosstalk"*:

- P to P: low
- E to E:
  - high (same module)
  - low (other module)

*some effects may be OS related

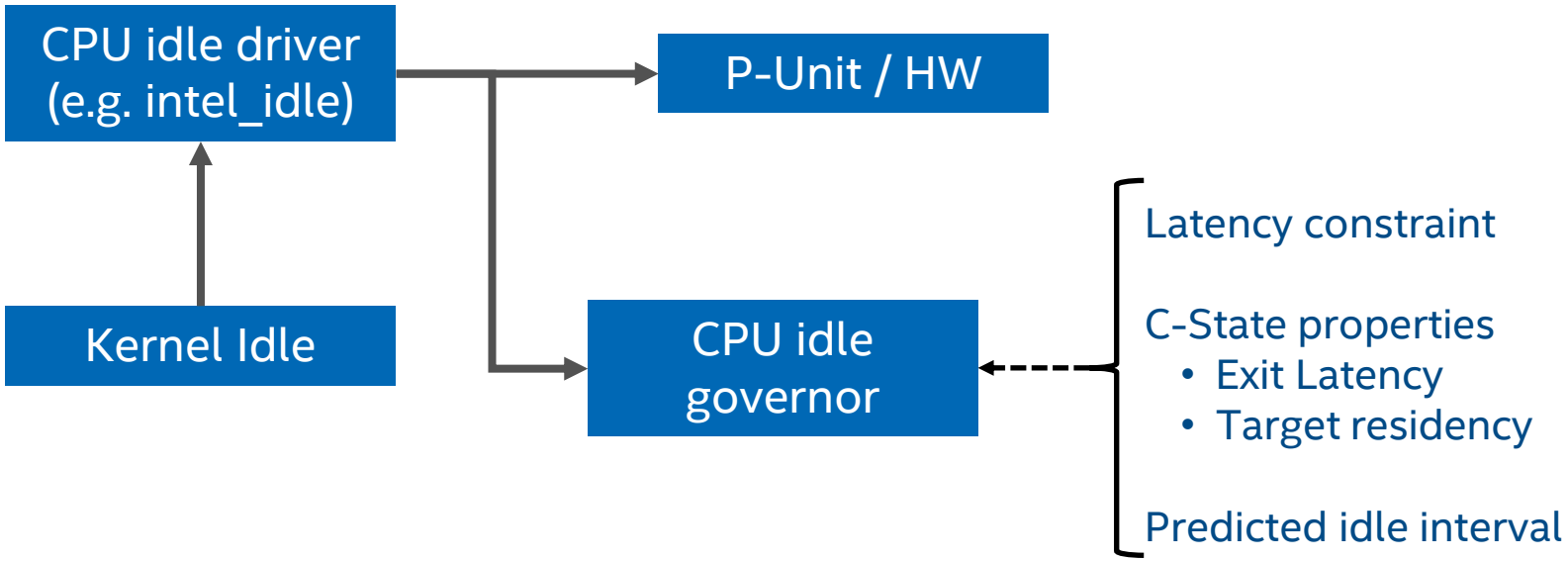
* 13th Gen Intel(R) Core(TM) i7-1370PE

C-States - recap

- C0:
  - Lowest C state: CPU active
- C1, C2, etc (CPU idle states):
  - C-States ↗ ⇒ “Deeper” C-States
  - C-States ↗ ⇒ Power savings ↗ ⇒
    - Transition latencies ↗ ⇒ ok: BE; nok: RT
    - Power budget for active cores ↗ ⇒ BE peak performance ↗
- Mixed criticality setup:
  - Benefits to BE
  - Low impact to RT

Per core C-States w/ low impact to RT performance on other cores

C-State selection Policy in Linux Kernel



C-State selection mostly driven by:

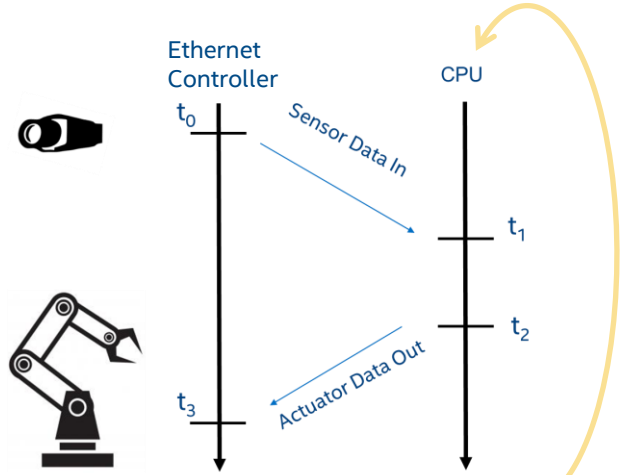
- C-State exit latency: Latency Cn → C0; i.e exiting idle state
- C-State target residency: ≤ estimated time core will stay in idle state

Example of Intel® 11th Gen Core™ Processors

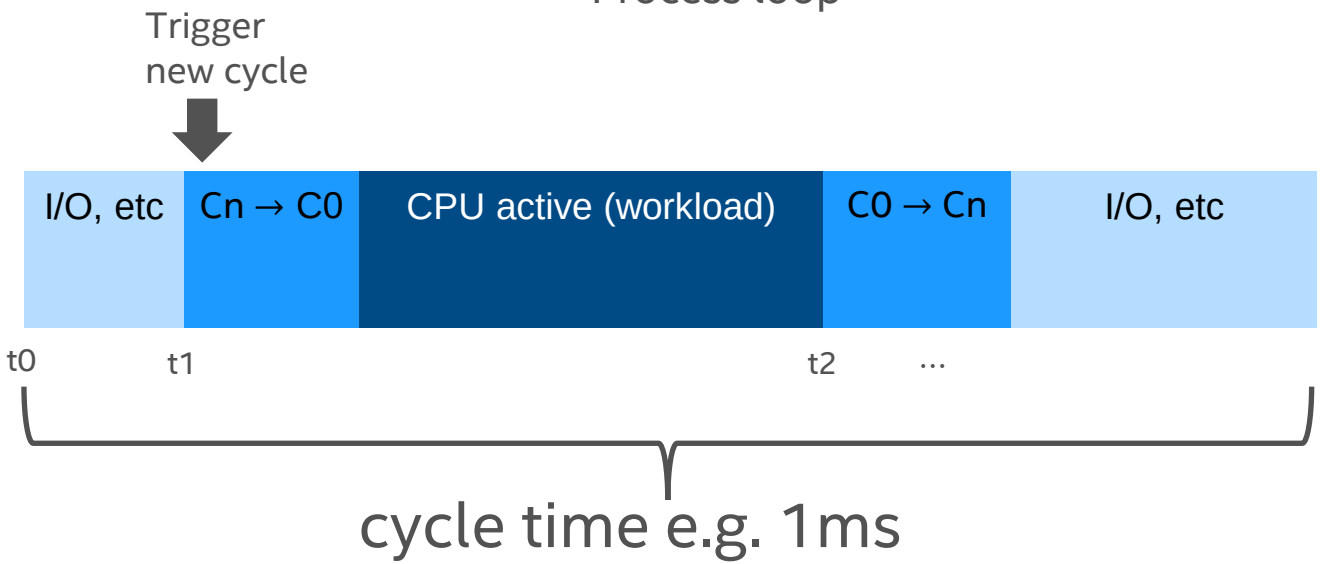
```
~# cd
/sys/devices/system/cpu/cpu<n>/cpuidle

~# ls
state0   state1   state2   state3

~# grep . */name          ~# grep . */disable
state0/name:POLL          state0/disable:0
state1/name:C1_ACPI       state1/disable:0
state2/name:C2_ACPI       state2/disable:0
state3/name:C3_ACPI       state3/disable:0

~# grep . */latency
state0/latency:0
state1/latency:1
state2/latency:253
state3/latency:1048

~# grep . */residency
state0/residency:0
state1/residency:1
state2/residency:759
state3/residency:3144
```

µs

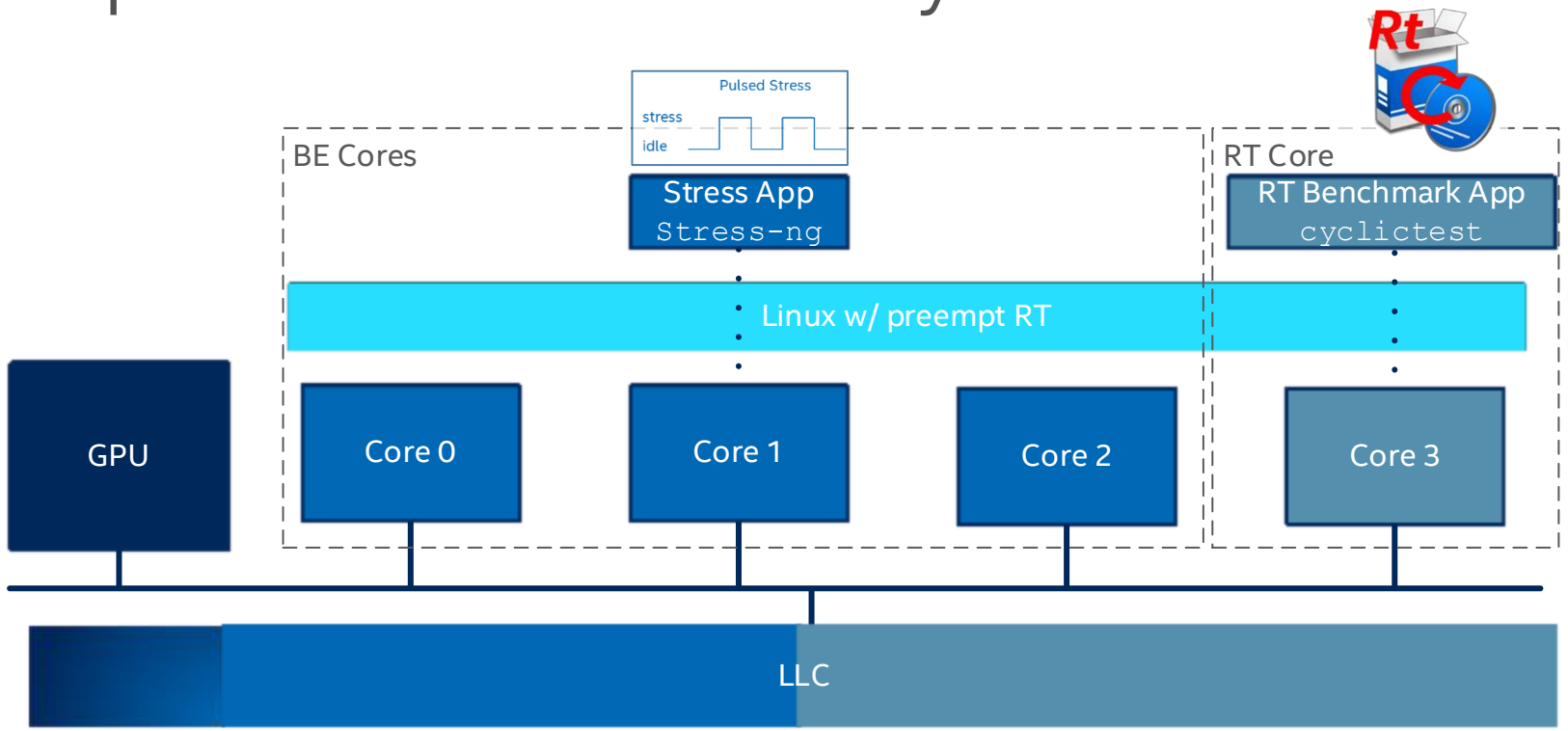
µs



Process loop

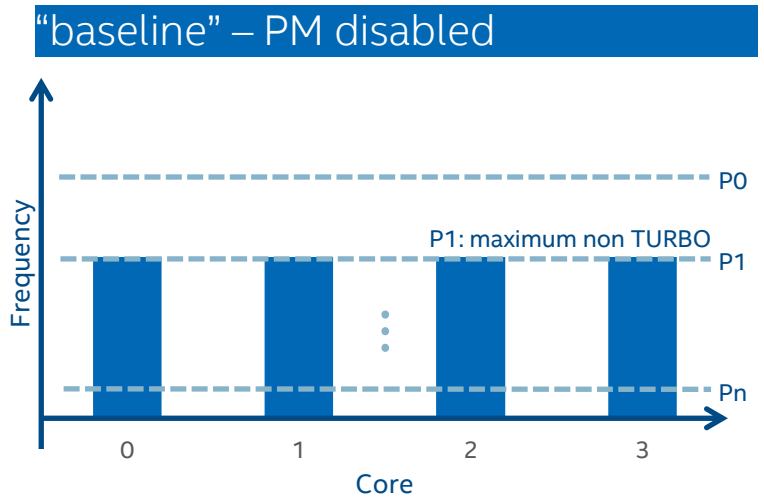


Test Setup for C-State Analysis

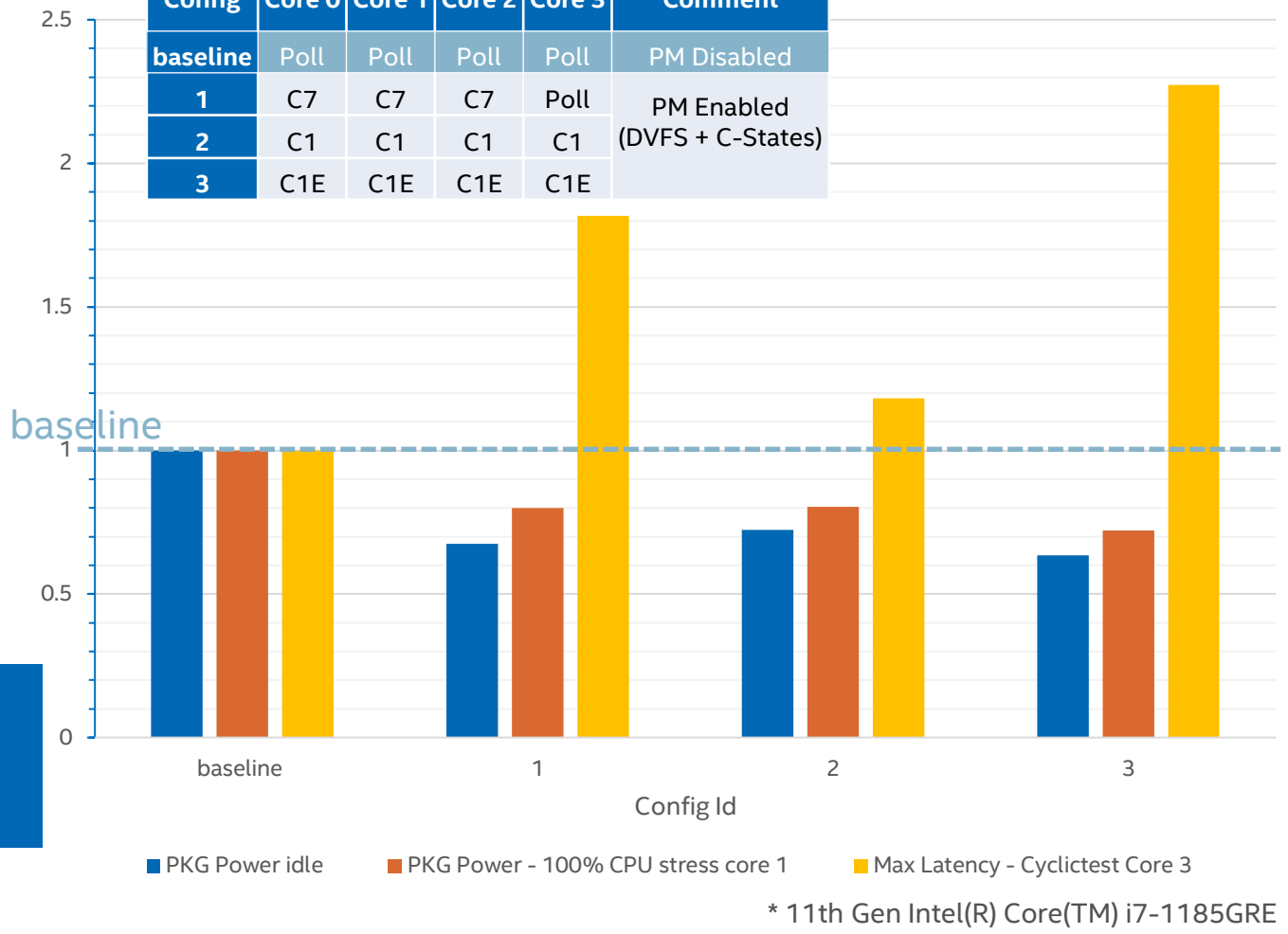


Config	Core 0	core 1	core 2	core 3
1	Poll	Poll	Poll	Poll
2	C7	C7	C7	Poll
3	C1	C1	C1	C1
4	C1E	C1E	C1E	C1E

C-States in mixed criticality scenarios: Power and RT Performance



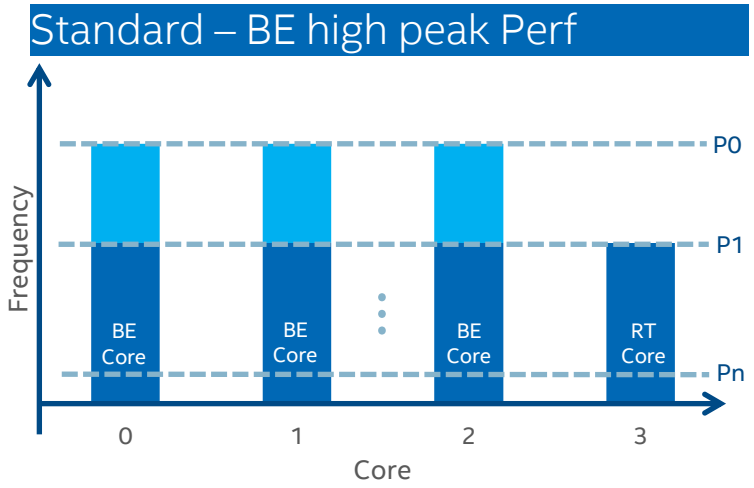
Config	Core 0	Core 1	Core 2	Core 3	Comment
baseline	Poll	Poll	Poll	Poll	PM Disabled
1	C7	C7	C7	Poll	PM Enabled (DVFS + C-States)
2	C1	C1	C1	C1	
3	C1E	C1E	C1E	C1E	



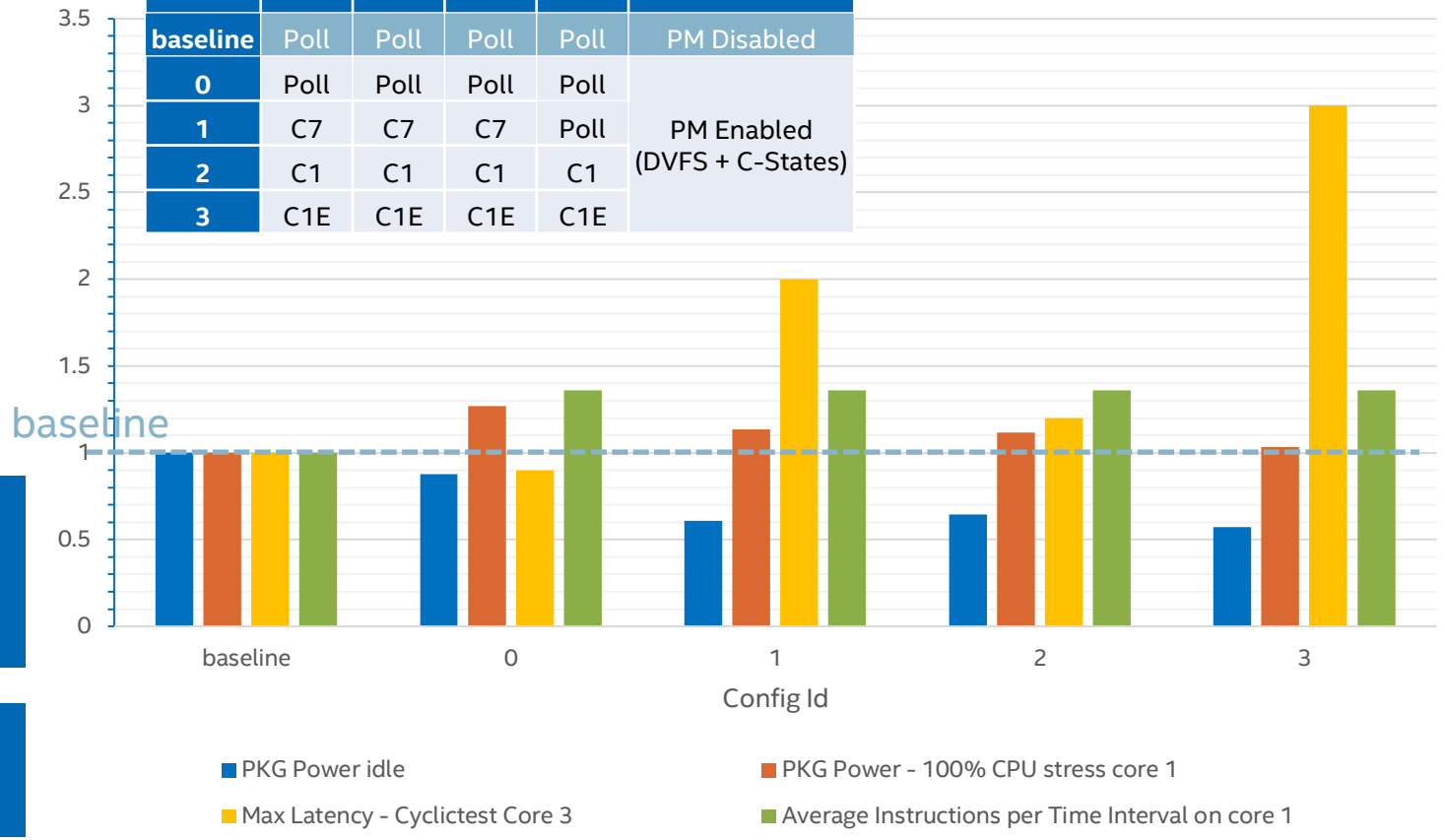
* 11th Gen Intel(R) Core(TM) i7-1185GRE

C-states in mixed criticality scenarios:
- Power consumption ↘
- low impact to RT performance

C- & P-States in mixed criticality scenarios: Power and RT Performance



Config	Core 0	Core 1	Core 2	Core 3	Comment
baseline	Poll	Poll	Poll	Poll	PM Disabled
0	Poll	Poll	Poll	Poll	PM Enabled (DVFS + C-States)
1	C7	C7	C7	Poll	
2	C1	C1	C1	C1	
3	C1E	C1E	C1E	C1E	



Lower Idle Power Consumption with low impact on RT performance

Higher Peak Performance for BE workloads

* 11th Gen Intel(R) Core(TM) i7-1185GRE

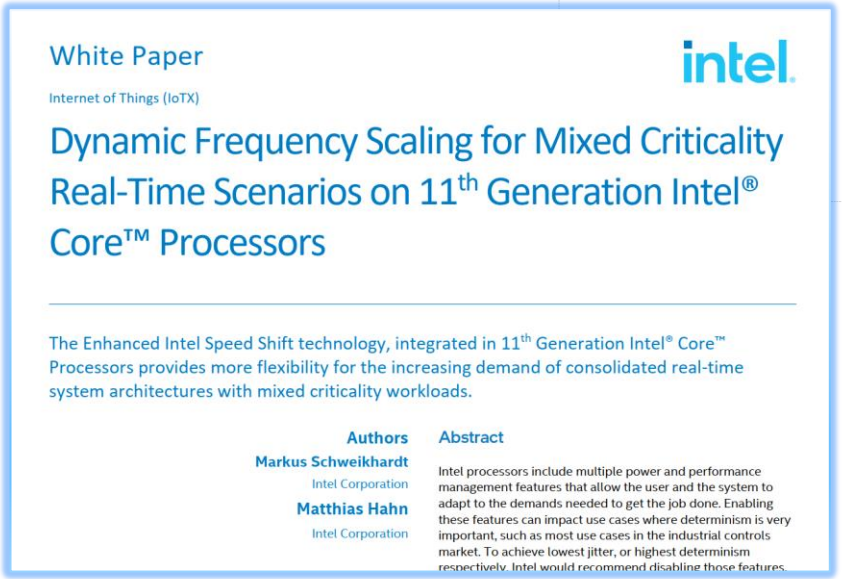
Summary and Conclusion

Conclusion

- For previous hardware generations: Real-Time & low jitter ⇒
  - No ACPI (P-States, C-States)
  - 24/7 fixed frequency
- DVFS
  - More flexibility
  - New use cases
- Get New Intel HW or check OSADL farm to see first results.

Further Reading

- *Dynamic Frequency Scaling for Mixed Criticality Real-Time Scenarios on 11th Generation Intel® Core™ Processors,* Markus Schweikhardt and Matthias Hahn, Intel Whitepaper 2022
- *Echtzeitverhalten bei dynamischen Taktänderungen,* Matthias Hahn und Markus Schweikhardt, computer&automation 2023



White Paper

intel

Internet of Things (IoTX)

Dynamic Frequency Scaling for Mixed Criticality Real-Time Scenarios on 11th Generation Intel® Core™ Processors

The Enhanced Intel Speed Shift technology, integrated in 11th Generation Intel® Core™ Processors provides more flexibility for the increasing demand of consolidated real-time system architectures with mixed criticality workloads.

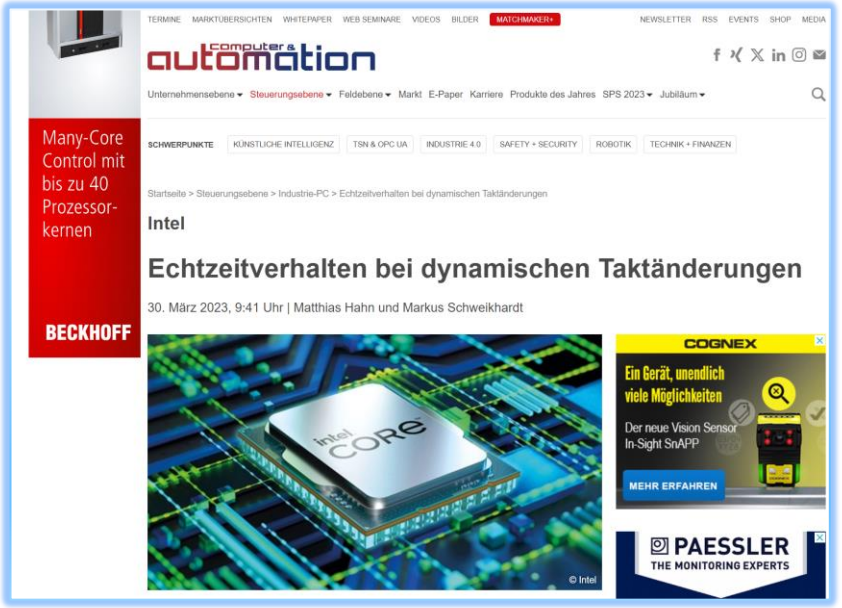
Authors

Markus Schweikhardt
Intel Corporation

Matthias Hahn
Intel Corporation

Abstract

Intel processors include multiple power and performance management features that allow the user and the system to adapt to the demands needed to get the job done. Enabling these features can impact use cases where determinism is very important, such as most use cases in the industrial controls market. To achieve lowest jitter, or highest determinism respectively, Intel would recommend disabling those features.



computer & automation

Many-Core Control mit bis zu 40 Prozessor-kernen

BECKHOFF

Intel

Echtzeitverhalten bei dynamischen Taktänderungen

30. März 2023, 9:41 Uhr | Matthias Hahn und Markus Schweikhardt

COGNEX

Ein Gerät, unendlich viele Möglichkeiten

Der neue Vision Sensor In-Sight SnAPP

MEHR ERFAHREN

PAESSLER

THE MONITORING EXPERTS


intel

Workloads and Configuration

Linux OS

Yocto Project -based Intel board support package (BSP) BKC MR3 release with Linux kernel 5.10.41-rt42 and preempt real-time patch.

"615079_11th_Gen_Intel_Core_Processors_BSP_Release_Notes_3.2_MR2," Intel, 07 2021. Available: https://cdrdv2.intel.com/v1/dl/getContent/617124?explicitVersion=true

kernel cmd-line parameters	DVFS disabled	DVFS enabled
processor.max_cstate	0	
intel_idle.max_cstate	0	
clocksource	tsc	
tsc	reliable	
nowatchdog		
intel_pstate	disable	active
idle	poll	
noht		
isolcpus	2,3	
rcu_nocbs	all	
rcupdate.rcu_cpu_stall_suppress	1	
rcu_nocb_poll		
irqaffinity	0	
i915.enable_rc6	0	
i915.enable_dc	0	
i915.disable_power_well	0	
mce	off	
hpet	disable	
numa_balancing	disable	
efi	runtime	
art	virtuallow	
iommu	pt	

Workloads and Configuration

Intel Reference BIOS configuration changes

DVFS disabled	DVFS enabled
Intel® TCC Mode Parameters manually set - see “Intel Time Coordinated Compute User Guide” section B2.4	
Intel SpeedStep® = Disabled	Intel SpeedStep® = Enabled
Intel® Speed Shift Technology = Disabled	Intel® Speed Shift Technology = Enabled
Intel® Turbo Mode= Disabled	Intel® Turbo Mode= Enabled

Intel TCC User Guide - https://cdrdv2.intel.com/v1/dl/getContent/786715?explicitVersion=true

cyclictest - version 1.6

https://git.kernel.org/pub/scm/utils/rt-tests/rt-tests.git

stress-ng – version 0.12.06

https://github.com/ColinIanKing/stress-ng.git

Embedded into a python script to generate pulsed stress.

```
Pseudocode
while true:
    run for 5 sec stress-ng
    sleep 5 sec
```

HW Platform used for the DVFS Tests

11th Gen Intel® Core™ i5-1145GRE	Intel® Core™ i5-8265U	13th Gen Intel® Core™ i7-1370PE
Intel® Customer Reference Board (CRB)	Maxtang® AX8265U	Intel® Customer Reference Board (CRB)
Intel Reference BIOS Version MR2 4225_01	AMI BIOS version WL10R109	Intel Reference BIOS Version 4115_04

HW Platform used for the C-State Tests

Intel® Core™ i7-1185GRE
Intel® Customer Reference Board (CRB)
Intel Reference BIOS Version MR2 4225_01