



Power and performance management features of Intel Xeon and Xeon Phi processors

Workshop on Energy-Aware High Performance Computing
ISC-High Performance

June 22nd 2017, Frankfurt, Germany

Andrey Semin

Principal Engineer, HPC EMEA

⁺ Some content of this presentation does not reflect the official opinion of the Intel Corporation. Responsibility for the information and views expressed in the therein lies entirely with the author⁺⁺.

⁺⁺ But the Law of conservation of energy is respected any case

Copyright © 2017, Intel Corporation

*Other brands and names may be claimed as the property of others.

Legal Disclaimer

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The cost reduction scenarios described in this document are intended to enable you to get a better understanding of how the purchase of a given Intel product, combined with a number of situation-specific variables, might affect your future cost and savings. Circumstances will vary and there may be unaccounted-for costs related to the use and deployment of a given product. Nothing in this document should be interpreted as either a promise of or contract for a given level of costs.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families: Go to:

[Learn About Intel® Processor Numbers](#)

All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

The products described in this document 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.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: <http://www.intel.com/design/literature.htm>

Intel, Intel Xeon, Intel Xeon Phi™ are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States or other countries.

Copyright © 2016, Intel Corporation, *Other brands and names may be claimed as the property of others.

Optimization Notice

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2®, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804



Outline

- Application performance dependencies
- Energy efficiency controls evolution
- Intel Turbo Boost and frequency
- Summary

Performance dependencies

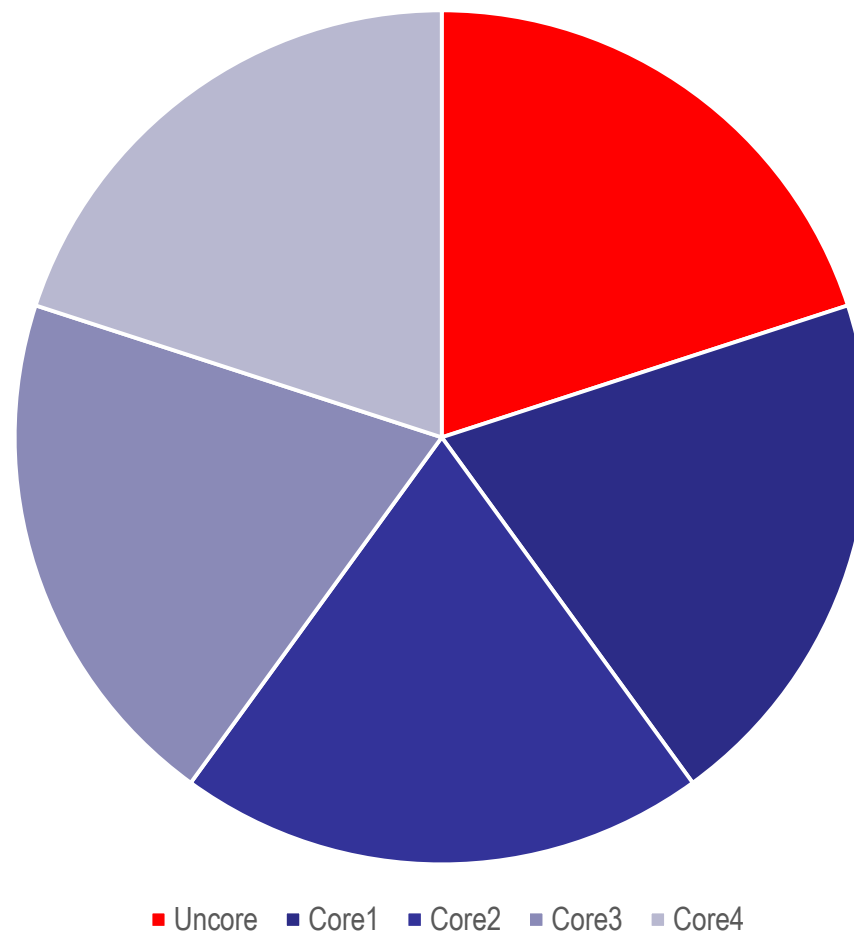
$$time = \#instr \cdot CPI \cdot \frac{1}{f}$$

- **#instr** - total number of instructions to be executed: application specific
- **CPI** (cycles per instruction) - depends on application and microarchitecture
- Cycle time/period ($\frac{1}{f}$)...

The frequency is constant, right? ... No ☹

Power Challenges for Highly Integrated Processors

Power distribution inside a 4 core processor



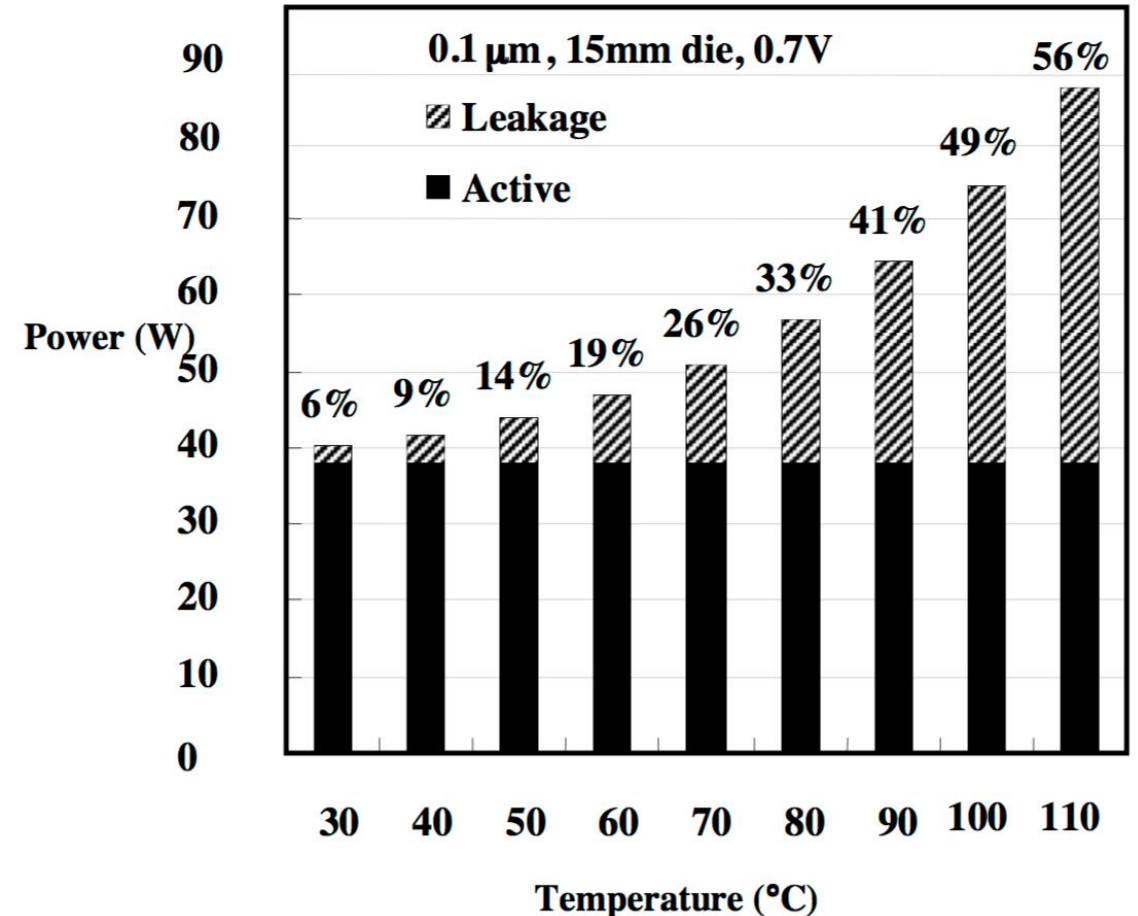
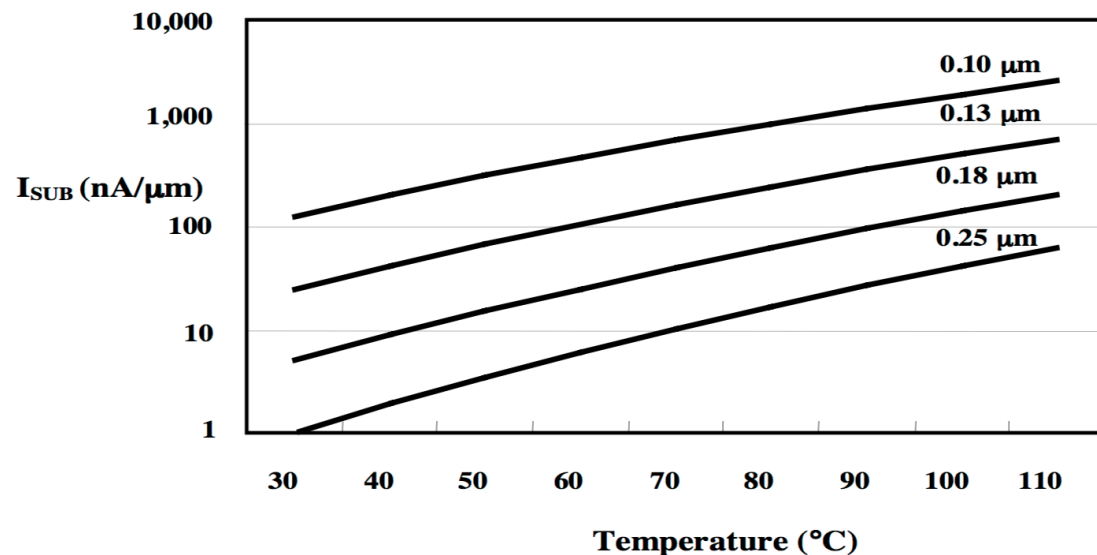
Power: the Good, the Bad, and the Ugly⁺

$$P_{total} = P_{dynamic} + P_{short\ circuit} + P_{leakage}$$

$$P_{dynamic} = ACV^2f$$

$$P_{short\ circuit} = tAVfI_{short}$$

$$P_{leakage} = VI_{leak}$$



Simplified for illustrative purposes

Figures source: Fallah, F, & Pedram, M 2005, 'Standby and active leakage current control and minimization in CMOS VLSI circuits', IEICE Transactions On Electronics, E88-C, 4, pp. 509-519

* The content of this publication does not reflect the official opinion of the Intel Corporation. Responsibility for the information and views expressed in the therein lies entirely with the author.

Energy efficiency controls evolution



Autotune: Turbo, RAPL policies, NodeManager, HWPM

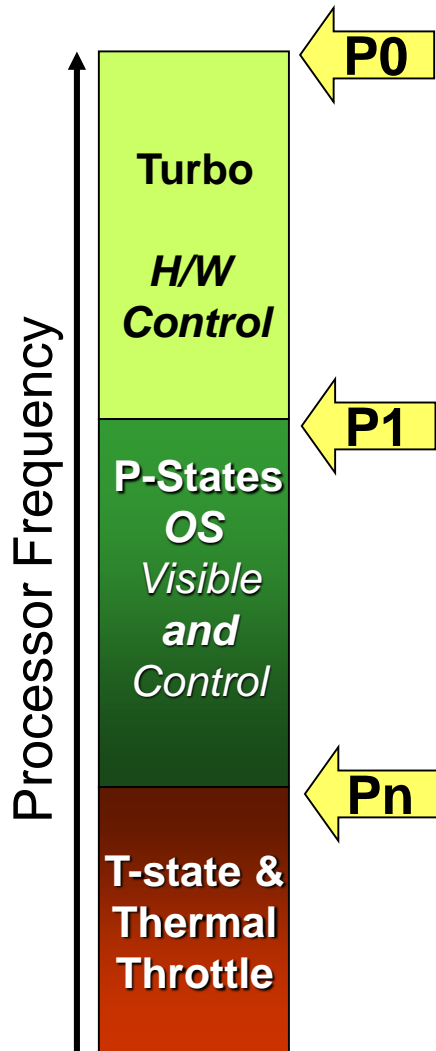
Coordination with OS/User: RAPL, NodeManager, P-states

Prevent Damage: TM, TM2, Hardware Duty Cycle (HDC)

Measure and monitor: DTS, RAPL energy counters



P-states Basics



P0 state is requested by the OS for maximum performance (Turbo boost request)

- Frequency opportunity to run above base frequency is based on number of active cores up to current, power, and temperature limits
- Some operating systems manage frequency between P1 and P0, others leave this entirely to hardware

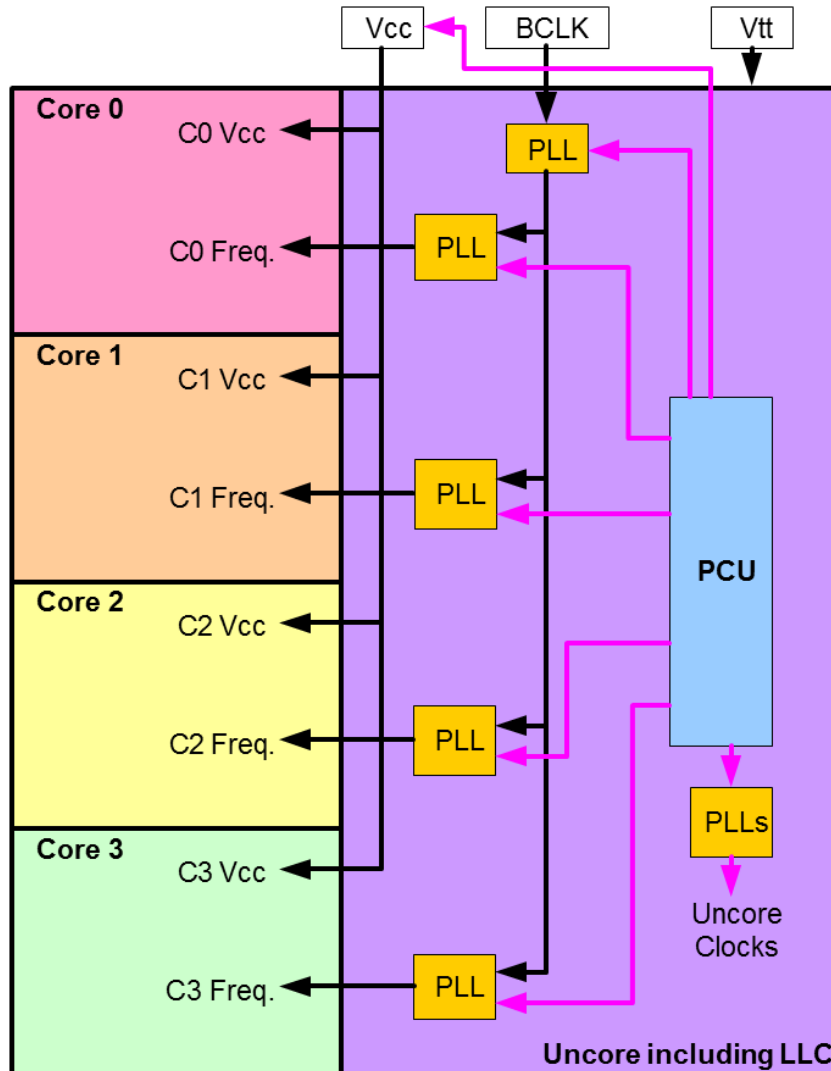
P1 is the processors rated frequency

- Operating system controls Pn through P1 state based on heuristic algorithm that monitors CPU utilization
- Processor continues to enforce specifications to stay within current, power and temperature limits

Pn is the lowest p-state

- Frequency below Pn is used only for thermal throttling and RAPL power throttling

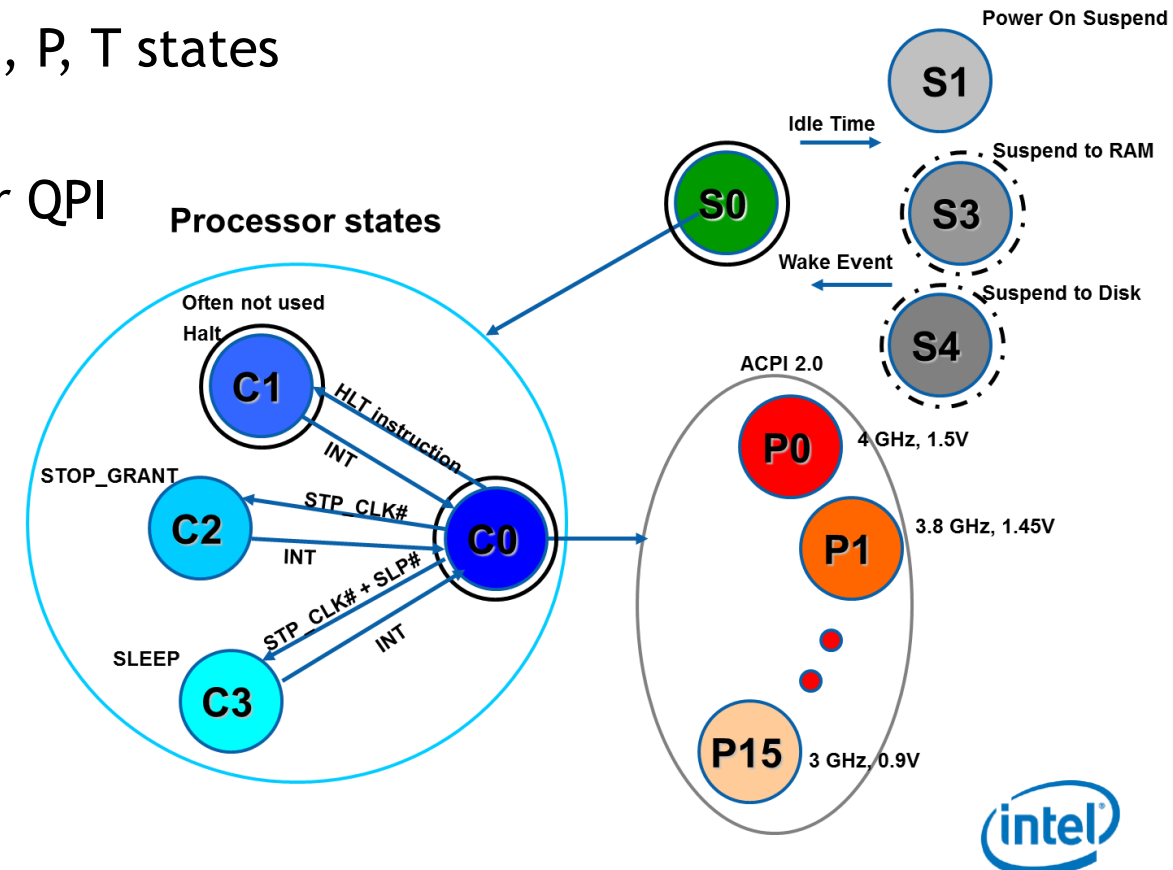
Power Control Unit (PCU)



Monitors voltage, temperature, power management requests

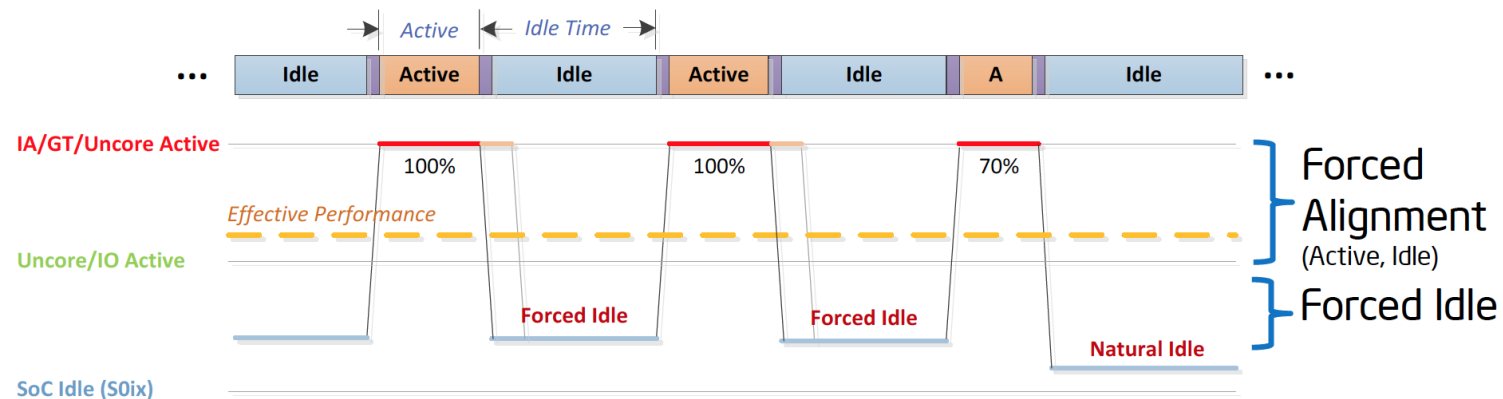
Controls

- Core/Package C, P, T states
- Memory states
- Power states for QPI

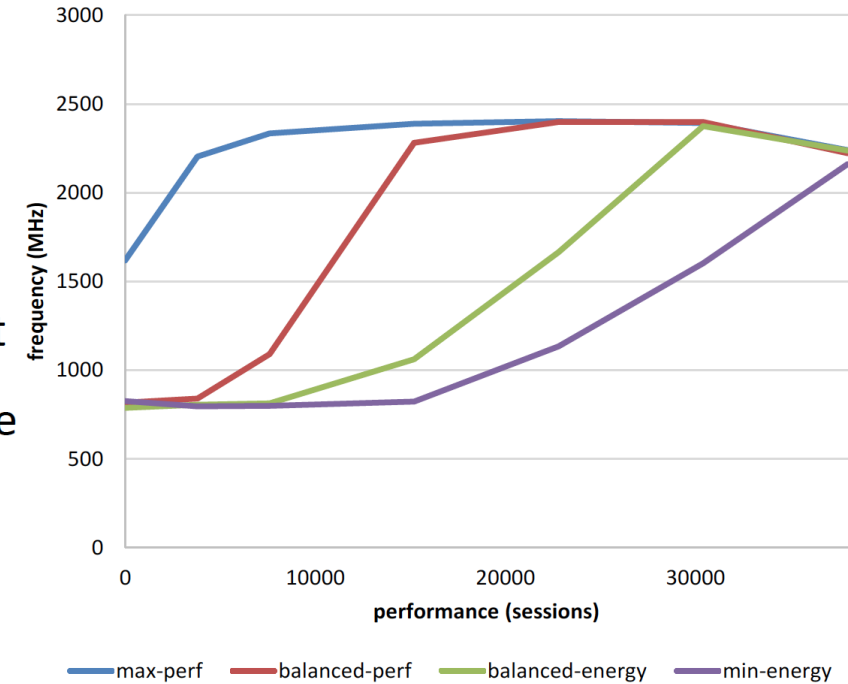


New in Broadwell Generation

- Hardware Duty Cycling (HDC)
 - Fine-grained, dynamic policy embedded in Intel Si
 - Minimizes power consumption at semi-active workload conditions
 - HW control knobs for HDC and SW policy algorithms that activate HDC control knob as necessary and defined by the platform



- Hardware Power Management (HWPM)
 - embeds frequency control capabilities in the CPU: no OS required
 - independent and cooperative modes
 - term HWP (hardware p-states) is used to describe the software interface: OS can optionally provide controls and hints



Intel® Turbo Boost Technology 2.0 Made Simple

- Bathtub analogy:

Flow from the faucet...

(CPU power)

can exceed drain flow...

(thermal solution capability)

for a short time...

(depending on capacity, level and flow)

if the bathtub...

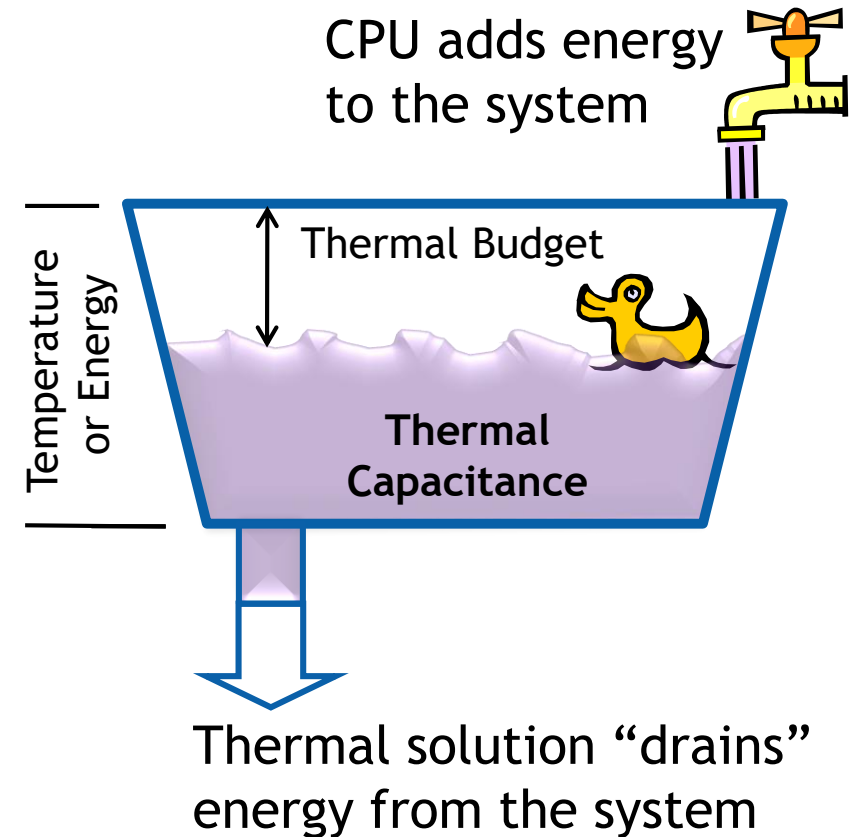
(energy or temperature)

is not full.

(at limits)

- Turbo 2.0 Implications:

- CPU can safely operate >TDP watts for short periods
- Potential for more time in Turbo mode, especially when system has been operating at low power

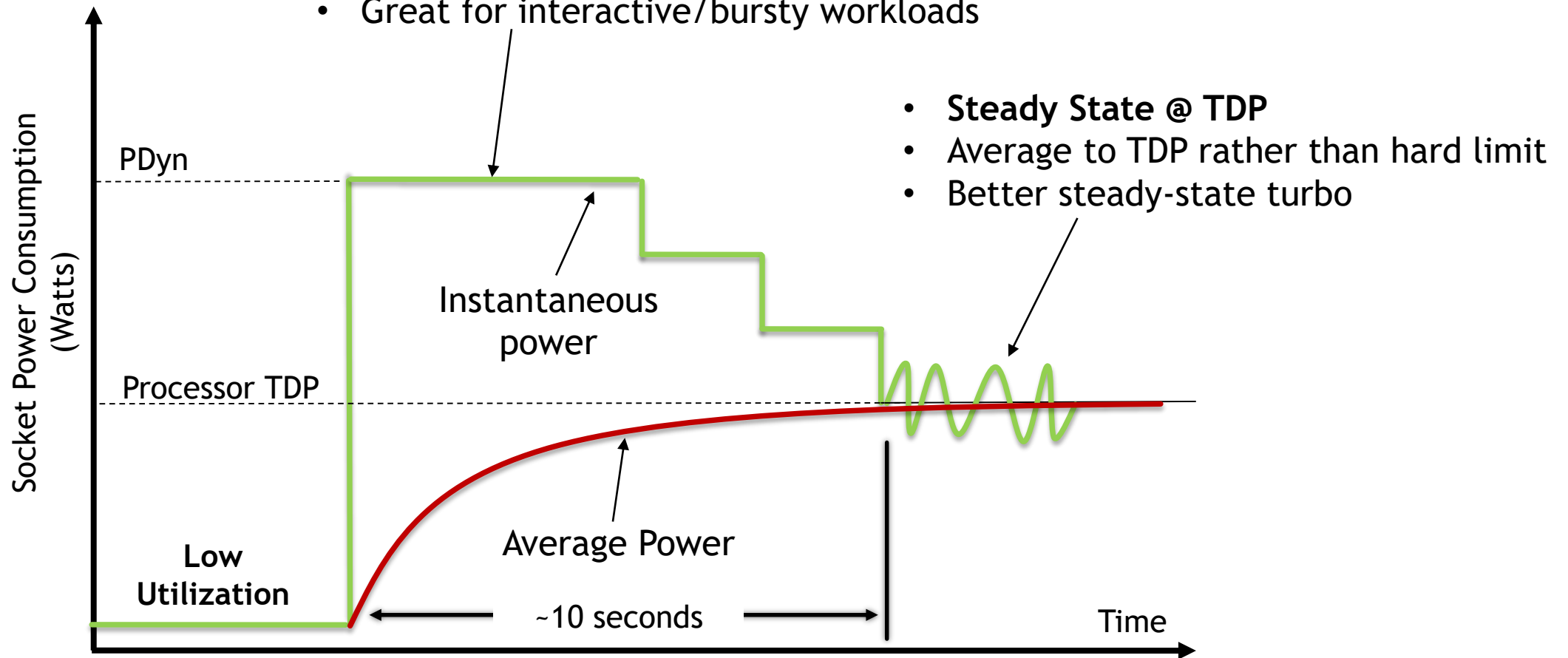


For more information, see www.intel.com/go/turbo

For illustrative purposes only

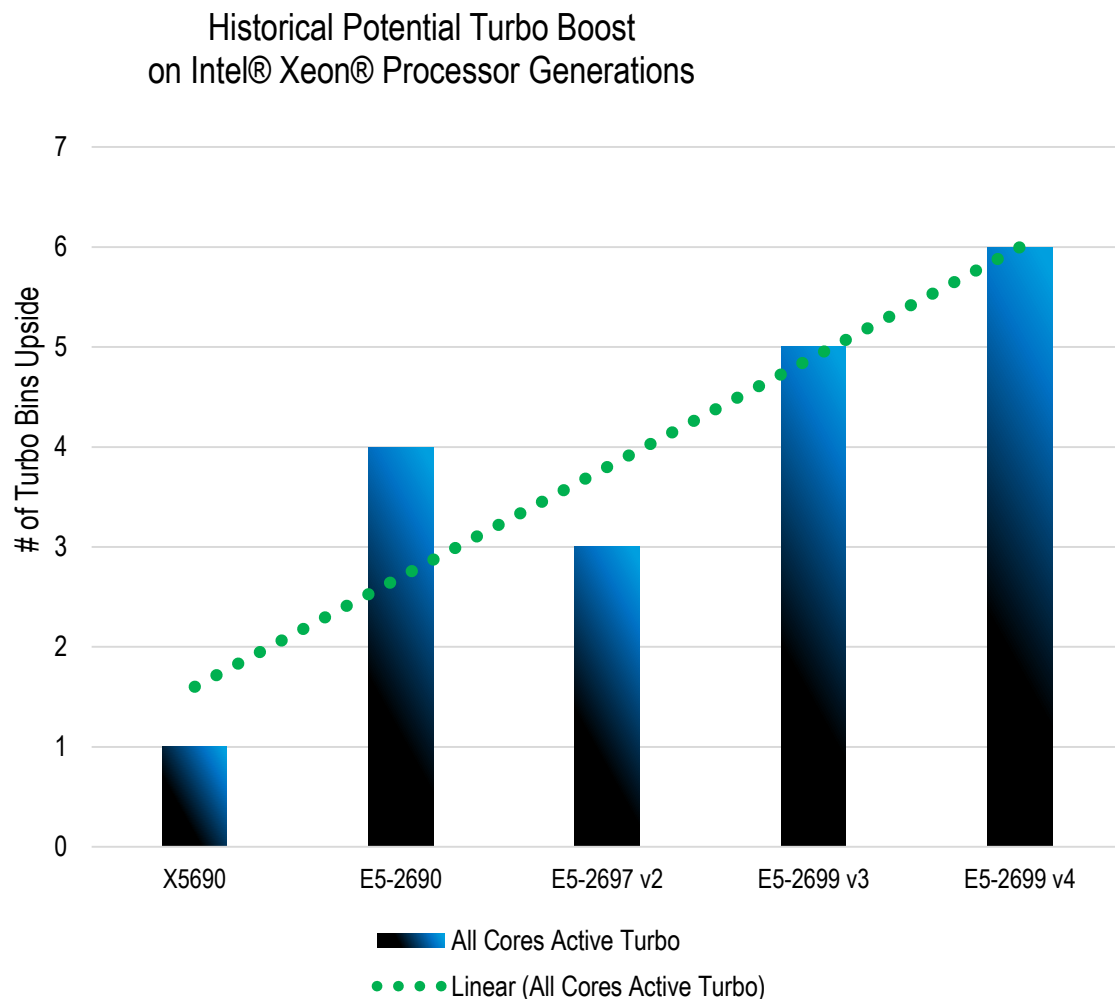
Intel® Turbo Boost Technology 2.0

- Power boost to P_{dyn}
- Possible after low utilization periods
- Great for interactive/bursty workloads



Intel Turbo Boost establishes dependency of frequency on power
(and temperature)

Frequency Upside Opportunity with Intel® Turbo Boost Technology



- Turbo delivers extra frequency on demand for higher application responsiveness and throughput for many types of workloads
- More turbo bins naturally leads to:
 - Increasing turbo bin upside potential
 - Creating opportunity for burst frequency
 - Processor frequency variability
- More turbo bins naturally leads to higher variability
 - All turbo influence factors within a given datacenter apply (power, thermals, etc.)



of Turbo Bins Upside = additional frequency based on number of 100 MHz increments versus a typical part (+/- 1 bin = +/- 100 or 133 MHz, etc.) depending on model

Broadwell-EP and Knights Landing Frequency Ranges

BDW-EP SKU	TDP (W)	Cores	Cache (MB)	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7
E5-2699 v4	145	22	55																										
E5-2698 v4	135	20	50																										
E5-2697A v4	145	16	40																										
E5-2697 v4	145	18	45																										
E5-2695 v4	120	18	45																										
E5-2683 v4	120	16	40																										
E5-2690 v4	135	14	35																										
E5-2680 v4	120	14	35																										
E5-2660 v4	105	14	35																										
E5-2650 v4	105	12	30																										
E5-2640 v4	90	10	25																										
E5-2630 v4	85	10	25																										
E5-2620 v4	85	8	20																										
E5-2609 v4	85	8	20																										
E5-2603 v4	85	6	15																										
E5-2650L v4	65	14	35																										
E5-2630L v4	55	10	25																										
E5-2687W v4	160	12	30																										
E5-2667 v4	135	8	25																										
E5-2643 v4	135	6	20																										
E5-2637 v4	135	4	15																										
E5-2623 v4	85	4	10																										

SKU:	TDP (W)	Cores	Mesh (GHz)	1.0	1.1	1.2	1.3	1.4	1.5	1.6
7250	215	68	1.7							
7230	215	64	1.7							
7210	215	64	1.6							

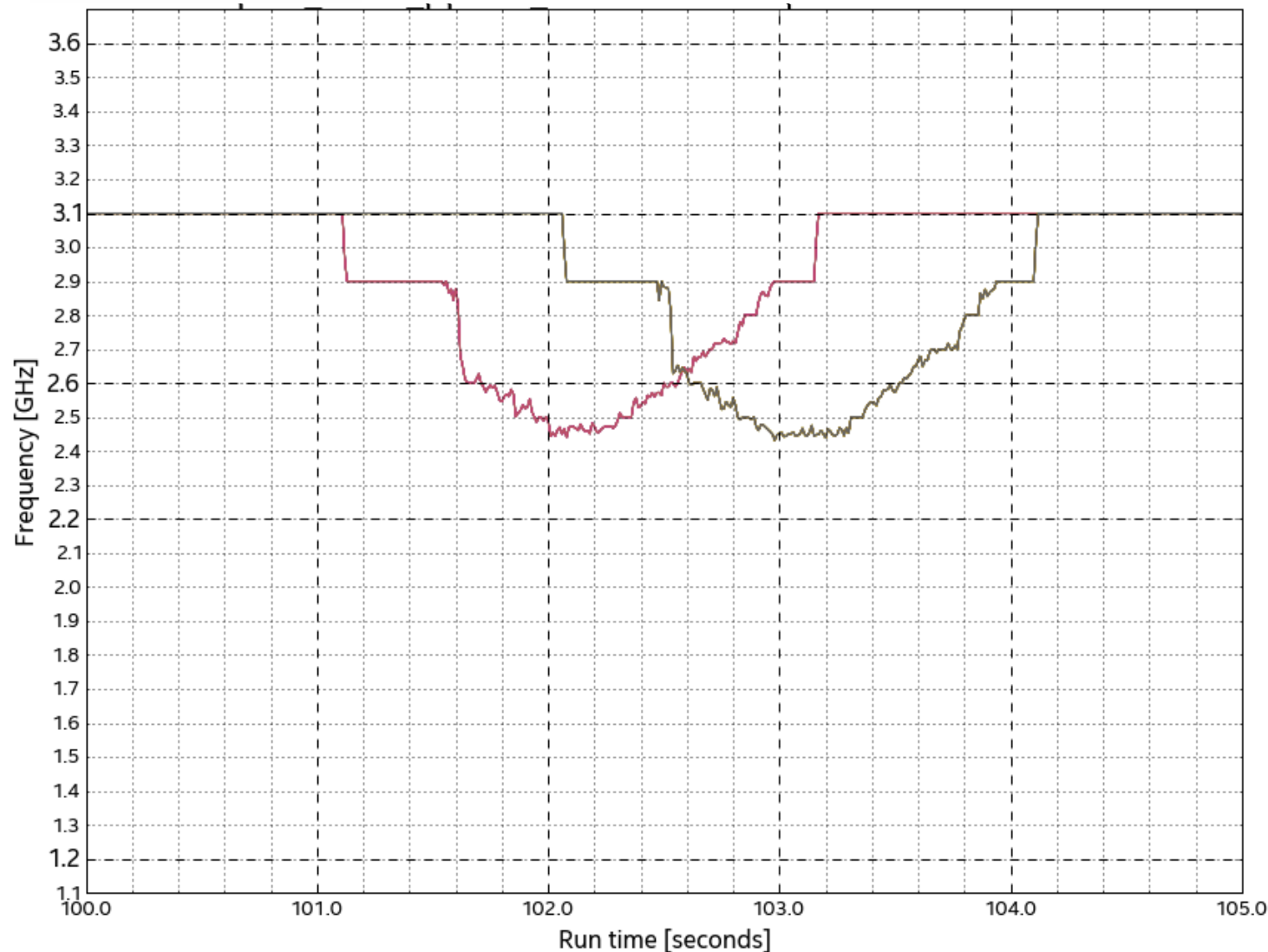
Legend:

-  non-AVX frequency range
-  AVX frequency range



For illustrative purposes only

Frequency observation⁺

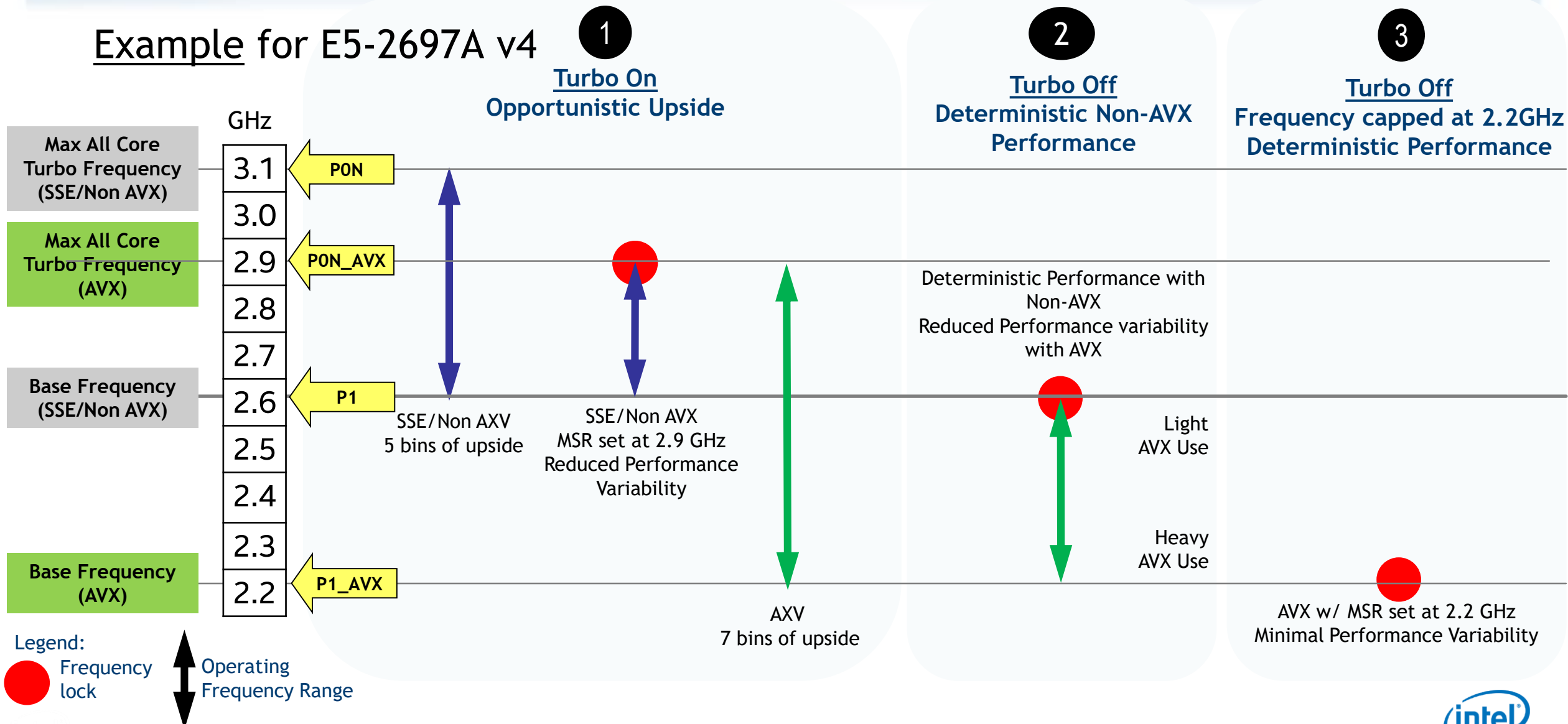


- Qbox: <https://asc.llnl.gov/CORAL-benchmarks/#qbox>
- Quantum molecular dynamics. Memory bandwidth, high floating-point intensity, collectives (alltoallv, allreduce, bcast)
- Job run on 16 nodes with two E5-2697A v4 and 64GiB DDR4-2400
- Frequency for one of the nodes is displayed: all cores
- Two MPI ranks per node (pinned to sockets) with 14 OpenMP threads each
- AVX2 optimizations applied
- Intel Emon-based frequency sampling with 10ms interval

⁺ The content of this publication does not reflect the official opinion of the Intel Corporation. Responsibility for the information and views expressed in the therein lies entirely with the author.

Frequency Boost on Intel Xeon E5-2600 v4 family processors

Example for E5-2697A v4



Summary and conclusions

- Temperature should be monitored to account for the total power in modern large scale systems
- Built-in technologies for power and energy-efficiency optimization has grown from simple temperature sensors to sophisticated control logic covering all aspects of the platform
- The amount of frequency upside with Intel Turbo Boost Technology has continued to grow from generation to generation
- Turbo implementations in Xeon and Xeon Phi are different with the latter providing more deterministic execution and lower frequency variability