Exploring RAPL as a Power Capping Leverage for Power-Constrained Infrastructures

Vladimir Ostapenco[1], Laurent Lefèvre[1], Anne-Cécile Orgerie[2], and Benjamin Fichel[3]

[1] EnsL, UCBL, CNRS, Inria, LIP, University of Lyon, Lyon, France
{laurent.lefevre,vladimir.ostapenco}@ens-lyon.fr
[2] Inria, CNRS, IRISA, University of Rennes, Rennes, France
anne-cecile.orgerie@irisa.fr
[3] OVHcloud, Brest, France

Abstract. Data centers are very energy-intensive facilities whose power provision is challenging and constrained by power bounds. In modern data centers, servers account for a significant portion of the total power consumption. In this context, the ability to limit the instant power consumption of an individual computing node is an important requirement. There are several energy and power capping techniques that can be used to limit compute node power consumption, such as Intel RAPL. Although it is nowadays mainly utilized for energy measurement, Intel RAPL (Running Average Power Limit) was originally designed for power limitation purposes. Some works use Intel RAPL for power limitation in a limited context without full knowledge of the inner workings of this technology and what is done behind the scenes to enforce the power constraint. Furthermore, Intel has not revealed any details about its internal implementation. It is unclear exactly how Intel RAPL technology operates and what effects it has on application performance and power consumption. In this work, we conduct a thorough analysis of Intel RAPL technology as a power capping leverage on a variety of heterogeneous nodes for a selection of CPU and memory intensive workloads. For this purpose, we first validate Intel RAPL power capping mechanism using a high-precision external power meter and investigate properties such as accuracy, power limit granularity, and settling time. Then, we attempt to determine which mechanisms are employed by RAPL to adjust power consumption.

Keywords: Power capping · Power limiting · Intel RAPL · Energy consumption reduction · Power control.

1 Introduction

The use of information and communication technologies (ICT) is constantly increasing. Data centers are key elements in this growth and are very energy-intensive facilities, estimated to represent around 1% of global electricity consumption [10]. Power provision of such facilities is challenging and subject to power bounds like the nominal delivered power. Yet, in some cases, these bounds may be lowered due to external conditions. Because computing nodes account for a significant portion of the total power consumption in modern data centers, the ability to limit their consumption is essential. Also, limiting

the power consumed by computing nodes indirectly reduces the need for cooling, and thus the power consumption required by the cooling units. In this article, we consider *power capping* or *power limiting* as the capability to set the maximum electrical power consumption for specific resource components (CPU, DRAM). Several techniques for limiting the power consumption of a computing node have been studied [13, 6, 15]. The most common technique for influencing a computing node's power consumption is using dynamic voltage and frequency scaling (DVFS). However, recent trends show that the DVFS software control is being deprecated, and the frequency and voltage control are increasingly being moved to hardware on recent processors [7]. Hopefully, the manufacturers of computing node components already offer additional interfaces such as RAPL (Running Average Power Limit), which allow software-controlled power limiting capabilities. Intel RAPL, introduced by Intel in 2011 [4], provides power limiting capabilities on certain parts of a computing node by enabling the specification of average power consumption over a time period. Even though RAPL was originally designed for power limitation purposes, it is now primarily used for energy measurement [9]. The adoption of Intel RAPL is additionally slowed by Intel withholding details about its internal implementation. Moreover, it is not clear how the RAPL power limiting mechanism influences application performance because there is no linear relationship between performance and power limit [7]. Furthermore, prior to using RAPL, it is also important to understand its accuracy, the time it takes to reduce power and stabilize, the power limits that can be used, and how these vary across various processor generations.

In this paper, in order to clarify previously discussed shortcomings, we conduct a thorough analysis of RAPL technology for power capping purposes on three distinct heterogeneous nodes for a variety of CPU and memory intensive workloads. For this purpose, we first validate the Intel RAPL power capping mechanism using a high-precision external power meter. After that, we investigate RAPL characteristics such as accuracy, the supported power capping values, the minimum granularity with which the power capping can be applied, and power limit settling time (the amount of time required until limit is reached). Finally, we attempt to determine which mechanisms are employed by RAPL to adjust the power consumption and meet the power limitation constraint.

The remainder of this paper is organized as follows. Section 2 covers related work regarding the RAPL power limiting mechanism. In section 3, we provide a detailed description of Intel RAPL technology and describe the methodology used to experimentally validate and understand the mechanism (section 4). In section 5, we analyze how Intel RAPL power limiting mechanism affects the infrastructure metrics. Later in the section 6, we examine various RAPL characteristics. Finally, in section 7, we discuss the lessons we learned about Intel RAPL power limiting mechanism before concluding in Section 8.

2 Related Works

Numerous previous works evaluate RAPL capabilities in terms of measurement, but only few works explore RAPL for power capping purposes. Zhang and Hoffmann [18] evaluate the Intel RAPL power limiting mechanism while executing a set of both

CPU intensive and memory intensive benchmarks. The work investigates the stability, accuracy, settling time, overshoot, and efficiency of Intel RAPL on Sandy Bridge architecture processors using consumption metrics provided by the Intel RAPL power sensing mechanism. Ostapenco *et al.* [14] evaluate Intel RAPL for power limiting purposes on a single node embedding a Cascade Lake-SP processor equipped with a high precision external power meter. This evaluation is then used for power capping at the scale of a data center. Cerf *et al.* [3] explore the application of control theory for dynamic power regulation using Intel RAPL power limiting mechanism. Rountree *et al.* [17] explore variations in power efficiency across processors as well as variations in performance under power limits on a large number of homogeneous nodes with Intel Sandy Bridge processors. None of these works evaluated, validated and characterized the Intel RAPL power limiting mechanism across multiple processor architectures for a variety of CPU and memory intensive workloads, relying not only on RAPL power sensing for power and energy consumption studies, but also on external measurements. Furthermore, these works focus on applying RAPL power capping or observing its performances, but not on understanding how it works.

3 RAPL (Running Average Power Limit)

RAPL interface (Running Average Power Limit) was introduced by Intel in 2011 in the Sandy Bridge architecture and provides power limiting capabilities by allowing the specification of average power consumption over a time period. In addition to power limitation, Intel RAPL has the ability to measure energy consumption.

Intel does not include any version information with RAPL technology. However, in this work, we recognize at least two versions of Intel RAPL. The first version is the one that was introduced with the Sandy Bridge architecture and uses a software power model in order to estimate energy usage [16]. The second version of RAPL, introduced with the Haswell architecture, is based on fully integrated voltage regulators and enables actual energy measurements, improving the accuracy of RAPL measurements and demonstrating nearly perfect correlation with an external power meter [5].

The power capping with Intel RAPL is achieved by enabling a dynamic specification of average power consumption over a specified time period. RAPL can be configured with two types of power constraints: long-term and short-term. The CPU and DRAM must comply with the long-term constraint for the majority of the time, but can consume more for a brief period while adhering to the short-term constraint. The duration of the constraint windows can be adjusted by user, but by default, the short-term constraint window counts in a few milliseconds and the long-term constraint window counts in a couple of seconds.

Intel RAPL supports multiple power domains that correspond to different CPU socket components. Each power domain reports its energy consumption and can be power limited over a specified time window. The availability of power domains differs between architectures and processor models. In general, RAPL supports the following power domains: *PKG (Package)* - Package domain refers to the entire CPU package; *PP0 (Core)* - Power plane 0 (Core power plane) domain represents all CPU package cores; *PP1* - Power plane 1 (Uncore graphic device plane) domain represents

the integrated graphics device if available; *DRAM* - DRAM domain represents the random access memory attached to the CPU memory controller; *PSys* - PSys domain, introduced in Skylake architecture, represents the entire platform, including not only CPU package but also other components such as PCH (Platform Controller Hub) and eDRAM [9]. Kashif Khan *et al.* [9] look into the availability of power domains and energy units on various processor architectures.

According to some works, Intel RAPL leverages the DVFS (Dynamic Voltage and Frequency Scaling) to control power consumption [19] as well as other techniques that force processor components into an idle state at low power levels [6]. The UFS (Uncore Frequency Scaling) leverage is available on recent processor architectures [5], and appears to be used by the Intel RAPL power control mechanism [20]. The RAPL memory power capping mechanism adjusts the back-to-back CAS timings and introduces delays to the issuance of commands by the memory controller to reduce the power consumed by DRAM [4].

4 Methodology

4.1 Experimental Environment

As previously stated, we recognize two major versions of Intel RAPL and anticipate differences in power capping mechanism operation across CPU architectures. Thus, we perform the experiments on three heterogeneous clusters of the Grid'5000 [2] large-scale test-bed for experimental research, which were chosen to include nodes with different CPU architectures and versions of Intel RAPL technology. Table 1 shows the specifications of the nodes in each selected cluster. For all nodes and experiments, we used a minimal version of Ubuntu 22.04 operating system available on the Grid'5000 with Hyper-Threading technology disabled. The power consumption of each node in each cluster is individually monitored by a high-precision external power meter Omegawatt [12] which is used with a sampling frequency of 50 Hz and has a precision of 0.1 W.

Table 1: Specifications of the nodes where the experiments were carried out.

	Taurus	**Gemini**	**Troll**
Model	Dell PowerEdge R720	Nvidia DGX-1	Dell PowerEdge R640
CPU	2 x Intel E5-2630	2 x Intel E5-2698 v4	2 x Intel Gold 5218
CPU Architecture	Sandy Bridge	Broadwell	Cascade Lake-SP
Memory	32 GiB	512 GiB	384 GiB 1.5 TiB PMEM
RAPL Version	1	2	2
Power domains	PKG, CORE, DRAM	PKG, DRAM	PKG, DRAM

During our experiments, we collect various metrics with Likwid tool (version 5.2.2)[4]. The majority of the experiments were carried out with a collection frequency

[4] https://github.com/RRZE-HPC/likwid/tree/master

of 1Hz, with a few performed at 10Hz to validate some observations. The gathered metrics include CPU core/uncore frequency and energy/power consumption retrieved from RAPL counters.

To enforce power capping, we use the Power Capping Framework[5] which exposes power capping devices to user space via `sysfs` in the form of files. In this work, we apply the power cap to various power domains by changing the power limit value in the long-term constraint file. The short-term constraint power limit is set to its maximum value.

We select two kernels of the NAS benchmark (version 3.4.2) [1] and the STREAM benchmark (version 5.10) [11] that represent a set of heterogeneous high-performance computing workloads. The first kernel of the NAS benchmark is the EP (Embarrassingly Parallel) kernel, which generates pairs of Gaussian random deviates, making intensive use of the CPU without memory operations. The second kernel is the MG (Multi-Grid) kernel, which executes a V-cycle multigrid algorithm and tests both short and long-distance data communication, making it memory intensive. STREAM benchmark is a synthetic benchmark that measures sustainable memory bandwidth and the computation rate for simple vector kernels exhibiting consistent memory-intensive behavior. For each benchmark, we adjust the size of the problem or number of iterations in order to have an execution time sufficient to draw conclusions.

4.2 Experimental Methodology

Before launching experiments, we enable all available power domains and set their power constraints to maximum values. Then, to assess Intel RAPL technology, we execute four sets of experiments.

In the first set, we investigate the impacts of applying the power limiting with Intel RAPL on CPU core and uncore frequency, and power consumption (retrieved from both RAPL counters and external power meter). The power limits are set for a one-minute duration in increments of 10 W from the maximum power limit value to the minimum value of 10 W. The goal of the second set of experiments is to determine the lowest granularity at which the power limit can be applied. For this purpose, we apply power limits with 0.065 W increments to the minimum value of 10 W on the Package power domain for 10 seconds while executing EP NAS benchmark. The third set of experiments was performed to examine the settling time, which is the amount of time required for Intel RAPL to apply a power limit and stabilize power consumption. For this purpose, for each benchmark and each power limit, we apply a power limit on the Package (PKG) power domain and measure how long it takes for the Intel RAPL to settle. To ensure the reliability of our results, the execution of all experiments as well as the processing of the results were automated, and each experiment is carried out five to ten times depending on the variability of the measurements. Each figure in this work shows the median values that are calculated from all the executions.

[5] https://www.kernel.org/doc/html/next/power/powercap/powercap.html

5 Impacts of RAPL on Infrastructure

5.1 Impacts on Power Consumption

Figure 1 illustrates how various power limits affect the power consumption of the power domains (PKG, Core, DRAM) as reported by RAPL power sensing mechanism while running the NAS EP benchmark. Even though it appears that Intel RAPL is able to apply some kind of hard power cap, we discover that the CPU can exceed the power limit because Intel RAPL, rather than implementing a hard power cap, attempts to maintain the specified average power consumption during a time window. Figure 2a, which illustrates the power profile when running the MG NAS benchmark, confirms that these overshoots are much more noticeable when running a memory-intensive benchmark having an unstable power profile.
Additionally, we can observe that every node has a particular power capping range where power is affected, and that range does not match the values found in the Intel RAPL registers. According to the RAPL registers, the Package power domain's minimum power limit for the Taurus node is 46 W and for Troll and Gemini nodes is 68 W. However, we are able to apply lower power limits, bringing the power down to near 20, 35W and 30W respectively.
Figure 2a demonstrates that the DRAM domain consumption is more important when executing the memory-intensive NAS MG benchmark. It can also be mentioned that the consumption of DRAM package is only affected by low power limit values. This leads to the conclusion that the DRAM power domain is only impacted when it is necessary to reduce power below what can be accomplished by affecting the non-DRAM part of the CPU. We observe that each other cluster exhibits a similar behavior.
Figure 2b shows the power consumption as reported by the external power meter and Intel RAPL power sensing mechanism for the Package domain of both CPU sockets during the NAS EP benchmark execution. We can mention that, in comparison to the power reported by Intel RAPL, the power decrease observed from the perspective of the external power meter is almost equivalent. Upon closer examination of the power decrease, we can note that the power reduction as perceived by the external power meter is slightly more significant. Given that the external power meter is positioned between the power supply unit (PSU) and the wall socket, this appears to be normal as it monitors the consumption of the entire node.
Even though we are only presenting here a dynamic power profile for NAS EP benchmark execution, the results are similar for all other studied benchmarks.

5.2 Management of CPU Frequency for Power Limiting

We observe that, for most power limits, CPU core frequency decreases in a manner similar to power consumption (illustrated in Figure 1). We can, however, point out a few exceptions. First, we notice a 10 W decrease in CPU Package power and a nearly unchanged CPU core frequency when the 115 W power limit is applied to the Troll node. We also observe that the CPU Package power is somewhat reduced without affecting CPU core frequency when the 15W power limit is applied to the Taurus node. We can therefore conclude that for most power limits, the CPU core frequency

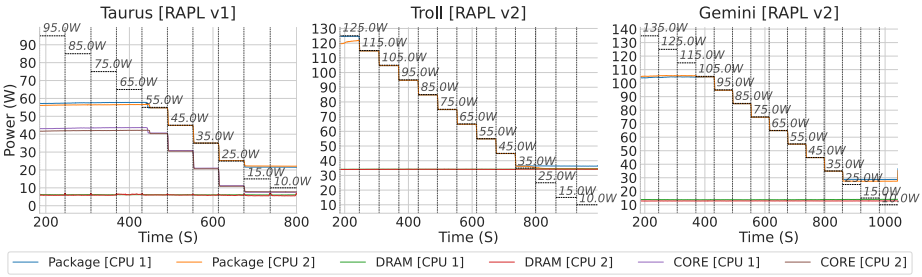
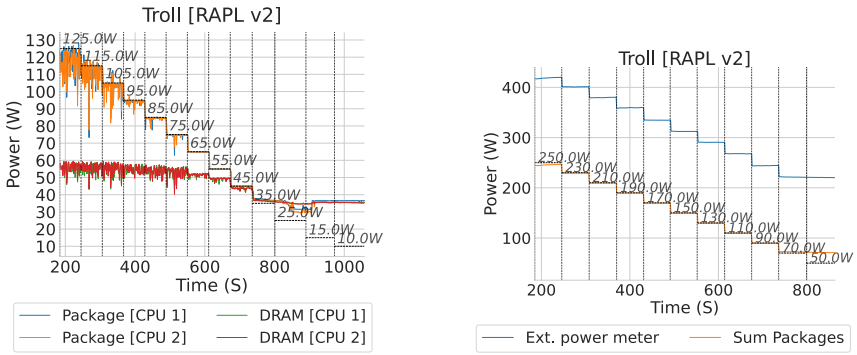


Fig. 1: Power profile when applying power limits on Package (PKG) domain for NAS EP benchmark.



(a) PKG and DRAM power domains - MG benchmark

(b) External power meter and sum of Package domains of both CPUs - EP benchmark

Fig. 2: Power profiles when applying power limits on Package (PKG) domain on Troll cluster node

modulation is used by Intel RAPL mechanism to respect power limits. Furthermore, the RAPL power limiting mechanism appears to rely on more than just CPU core frequency modulation, as in some cases it manages to reduce power consumption without changing the CPU frequency.

As Sandy Bridge processors share a common frequency for both core and uncore parts [5], the behavior of CPU core and uncore frequencies is identical. For Troll and Gemini nodes, the uncore frequency is managed independently of the core frequency [5], therefore Intel RAPL handles the uncore frequency differently while enforcing power limits. We observe that the uncore frequency is significantly reduced when applying the power limits of 115 W and 105 W to the Troll node executing CPU intensive NAS EP Benchmark. However, the uncore frequency is much less reduced when applying the same power limits while executing a memory-intensive STREAM Benchmark. This highlights that the Intel RAPL power limiting mechanism manages uncore frequency differently depending on the nature of the executed workload. The Gemini node exhibits similar behavior.

6 RAPL Characteristics

To assess the accuracy of Intel RAPL power limiting technology, we compute the mean absolute percentage error (MAPE) using the same methodology as [18]. We exclude from the accuracy study power limits that have no effect on power consumption, because their MAPE values are not representative. Even though MAPE values are almost always less than 2%, we found that MAPE is much less important for NAS EP and STREAM benchmarks for any cluster. Since the behavior of these two benchmarks is stable, we can conclude that Intel RAPL is more accurate for applications with a stable power profile. When we compare the MAPE values of clusters implementing different versions of Intel RAPL, we find that the second version is significantly more accurate for any given power limit value. For the NAS EP and STREAM benchmarks, the second version of RAPL shows a MAPE close to zero.

The power limit is specified in the RAPL model-specific registers (MSR) in "Power Units" and the default value of the "Power Unit" is 1/8 Watts [8], we validate that the application of a power limitation with a granularity lower than 0.125 Watts has no effect on power consumption. If we attempt to set the power limit using the Power Capping Framework with a lower granularity, the power limit value is rounded to the closest multiple of 0.125 Watts. Our experiments validate that power consumption can be influenced by 0.125 Watt power steps for each studied cluster.

We observe the "Settling time" which refers to the amount of time needed for Intel RAPL to apply a power limit and stabilize power consumption. The power limit is considered enforced and power consumption stable, if the power recorded by Intel RAPL is within 5% of the power limit for at least 5 measurements. We find that the maximum settling times observed for the two RAPL versions differ significantly. The nodes with newer CPU architectures (Gemini and Troll) that are implementing the second version of Intel RAPL have a much lower settling times (lower is better): between 0.5 to 1 second, than the node implementing the first version of Intel RAPL (Taurus) which takes more than 5 seconds to reach a stable state. This discrepancy may be caused by the fact that the default power constraint time windows of the Gemini and Troll nodes are substantially shorter (near 0.9 seconds) than those of the Taurus node (near 9 seconds), as well as by differences in the way the RAPL power limiting mechanism is implemented. Therefore, we can conclude that the second version of Intel RAPL with default time windows configuration is much more efficient in terms of settling time, and that the settling time is further improved on the newer CPU architectures.

7 Lessons Learned

First, we confirm that power limiting with RAPL can be used to comply with an average power consumption over a time window and should not be viewed as a hard power cap leverage, as the specified power limit can be exceeded, especially for memory-intensive benchmarks with an unstable power profile. We observe that for every node, we can enforce lower power limits than the minimum value found in the Intel RAPL registers. Furthermore, we confirm using an external power meter that the Intel RAPL power limiting mechanism effectively reduces a compute node's

power consumption while applying power limits.
Then, we discuss how CPU frequencies are changed when applying different power limits and confirm that the Intel RAPL power limiting mechanism primarily relies on CPU core and uncore frequency modulation to meet power limiting constraints. We find that the modulation of CPU core and uncore frequencies performed by Intel RAPL is dependent on the nature of the benchmark executed, as it differs for CPU- and memory-intensive benchmarks, and can be based on some utilization metrics.
We demonstrate by calculating the MAPE (mean absolute percentage error) between the power limit and power consumption that the power limitation with Intel RAPL is quite accurate (MAPE less than 2%). The accuracy is far more important for stable benchmarks, and steadily improves with subsequent CPU generations. We validate that the power limit can be applied in 0.125 Watt increments and that power limits with such granularity effectively impact power.
Finally, we find that the Intel RAPL mechanism's settling times with default constraint time window configuration vary depending on CPU architectures, ranging from less than 0.5 seconds for more recent CPUs to over 5 seconds for older CPUs.

8 Conclusion and Future Works

In this paper, we conduct a thorough study of the Intel RAPL power limiting mechanism for power capping purposes. We discuss existing power capping technologies and explain why Intel RAPL is one of the most prominent for power capping purposes. We describe Intel RAPL technology in detail, highlighting the lack of official information on its internal implementation. We conduct a series of experiments to determine how Intel RAPL works and validate the RAPL power limiting mechanism using a high precision external power meter at three nodes with various CPU architectures.

For future work, we intend to use and orchestrate the Intel RAPL power limiting leverage in the context of large-scale heterogeneous data center environments in order to respect global power constraints and reduce the overall infrastructure energy consumption. We also plan to validate the Intel RAPL power limiting leverage for a different set of workloads, including cloud workloads.

Acknowledgments

This work is supported by the "FrugalCloud" Inria and OVHcloud partnership. This work received funding from the France 2030 program, managed by the French National Research Agency under grant agreement No. ANR-23-PECL-0003. Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

References

1. Bailey, D.H., Barszcz, E., Barton, J.T., Browning, D.S., Carter, R.L., Fatoohi, R.A., et al.: The NAS parallel benchmarks. Tech. rep., RNR-94-007, NASA Research Center (1994)

2. Balouek, D., Carpen Amarie, A., Charrier, G., Desprez, F., Jeannot, E., Jeanvoine, E., et al.: Adding virtualization capabilities to the Grid'5000 testbed. In: Cloud Computing and Services Science. vol. 367, pp. 3–20. Springer (2013)
3. Cerf, S., Bleuse, R., Reis, V., Perarnau, S., Rutten, E.: Sustaining performance while reducing energy consumption: a control theory approach. In: Euro-Par: International Conference on Parallel and Distributed Computing. pp. 334–349 (2021)
4. David, H., Gorbatov, E., Hanebutte, U.R., Khanna, R., Le, C.: RAPL: memory power estimation and capping. In: ACM/IEEE Int. Symp. on Low-Power Electronics and Design (ISLPED). pp. 189–194 (2010)
5. Hackenberg, D., Schöne, R., Ilsche, T., Molka, D., Schuchart, J., Geyer, R.: An energy efficiency feature survey of the Intel Haswell processor. In: IEEE Int. Parallel and Distributed Processing Symp. workshops. pp. 896–904 (2015)
6. Haidar, A., Jagode, H., Vaccaro, P., YarKhan, A., Tomov, S., Dongarra, J.: Investigating power capping toward energy-efficient scientific applications. Concurrency and Computation: Practice and Experience **31**(6), e4485 (2019)
7. Imes, C., Zhang, H., Zhao, K., Hoffmann, H.: CoPPer: Soft Real-Time Application Performance Using Hardware Power Capping. In: IEEE International Conference on Autonomic Computing. pp. 31–41. IEEE (2019)
8. Intel: Intel® 64 and IA-32 architectures developer's manual: vol. 3B
9. Khan, K., Hirki, M., Niemi, T., Nurminen, J., Ou, Z.: RAPL in action: experiences in using RAPL for power measurements. ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS) **3** (Jan 2018)
10. Masanet, E., Shehabi, A., Lei, N., Smith, S., Koomey, J.: Recalibrating global data center energy-use estimates. Science **367**(6481), 984–986 (2020)
11. McCalpin, J.D.: Memory bandwidth and machine balance in current high performance computers. IEEE Computer Society Technical Committee on Computer Architecture (TCCA) Newsletter pp. 19–25 (1995)
12. OmegaWatt: Omegawatt (2018), http://www.omegawatt.fr, Accessed: 2023-01-04
13. Orgerie, A.C., D. Assunção, M., Lefèvre, L.: A survey on techniques for improving the energy efficiency of large-scale distributed systems. ACM Comp. Surveys **46**(4) (2014)
14. Ostapenco, V., Lefèvre, L., Orgerie, A.C., Fichel, B.: Modeling, evaluating, and orchestrating heterogeneous environmental leverages for large-scale data center management. The International Journal of High Performance Computing Applications **37**(3-4) (2023)
15. Petoumenos, P., Mukhanov, L., Wang, Z., Leather, H., Nikolopoulos, D.: Power Capping: What Works, What Does Not. In: IEEE International Conference on Parallel and Distributed Systems (ICPADS). pp. 525–534 (2015)
16. Rotem, E., Naveh, A., Ananthakrishnan, A., Weissmann, E., Rajwan, D.: Power-Management Architecture of the Intel Microarchitecture Code-Named Sandy Bridge. IEEE Micro **32**(2), 20–27 (2012)
17. Rountree, B., Ahn, D.H., De Supinski, B.R., Lowenthal, D.K., Schulz, M.: Beyond DVFS: A first look at performance under a hardware-enforced power bound. In: IEEE Int. Parallel and Distributed Processing Symp. Workshops (2012)
18. Zhang, H., Hoffman, H.: A quantitative evaluation of the RAPL power control system. Feedback Computing **6** (2015)
19. Zhang, H., Hoffmann, H.: Maximizing performance under a power cap: A comparison of hardware, software, and hybrid techniques. ACM SIGPLAN Notices **51**(4) (2016)
20. Říha, L., Vysocky, O., Bartolini, A.: Evaluation of DVFS and Uncore Frequency Tuning Under Power Capping on Intel Broadwell Architecture. IOS Press (2020)