Christian von Elm, Thomas Ilsche, Robert Schöne, Mario Bielert, Markus Schmidl

Investigating the Cause and Effect of an AMD Zen Energy Management Anomaly

Erstveröffentlichung in / First published in:

ICPE '21: ACM/SPEC International Conference on Performance Engineering, Virtual Event, 19. – 23.04.2021. ACM Digital Library, S. 103–106. ISBN 978-1-4503-8331-8.

DOI: https://doi.org/10.1145/3447545.3451193

Final edited form was published in "ICPE '21: ACM/SPEC International Conference on Performance Engineering", Virtual Event, 19. – 23.04.2021. S. 103–106. ISBN 978-1-4503-8331-8.
https://doi.org/10.1145/3447545.3451193

Investigating the Cause and Effect of an AMD Zen Energy Management Anomaly

Christian von Elm
Technische Universität Dresden
Center for Information Services and High Performance Computing (ZIH)
Dresden, Germany
christian.von_elm@tu-dresden.de

Thomas Ilsche
Technische Universität Dresden
Center for Information Services and High Performance Computing (ZIH)
Dresden, Germany
thomas.ilsche@tu-dresden.de

Robert Schöne
Technische Universität Dresden
Center for Information Services and High Performance Computing (ZIH)
Dresden, Germany
robert.schoene@tu-dresden.de

Mario Bielert
Technische Universität Dresden
Center for Information Services and High Performance Computing (ZIH)
Dresden, Germany
mario.bielert@tu-dresden.de

Markus Schmidl
Technische Universität Dresden
Center for Information Services and High Performance Computing (ZIH)
Dresden, Germany
markus.schmidl@mailbox.tu-dresden.de

ABSTRACT

This paper discusses an architectural anomaly observed on server processors of the AMD Zen microarchitecture: At a specific operating point, increasing the number of active cores reduces system power consumption while increasing performance more than proportionally to the additional cores. The occurrence of the anomaly is rooted in the hardware control loop for energy management and software-independent. Experiments show a connection to the AMD turbo frequency feature *Max Core Boost Frequency* (MCBF). In less efficient configurations, this feature could be employed from a processor's perspective, even though it is not necessarily used on any core. Voltage measurements indicate that the availability of MCBF leads to a higher voltage from mainboard voltage regulators, subsequently raising power consumption unnecessarily.

We describe the impact of this anomaly on the performance and energy-efficiency of several micro-benchmarks. The *reduced* power consumption when additional cores are enabled can lead to *higher* core frequencies and increased per-core-performance. The presented findings can be used to avoid inefficient core configurations and reduce the overall energy-to-solution.

ACM Reference Format:
Christian von Elm, Thomas Ilsche, Robert Schöne, Mario Bielert, and Markus Schmidl. 2021. Investigating the Cause and Effect of an AMD Zen Energy Management Anomaly. In *Companion of the 2021 ACM/SPEC International Conference on Performance Engineering (ICPE '21 Companion), April 19–23, 2021, Virtual Event, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3447545.3451193

ICPE '21 Companion, April 19–23, 2021, Virtual Event, France

ACM ISBN 978-1-4503-8331-8/21/04…$15.00
https://doi.org/10.1145/3447545.3451193

1 INTRODUCTION AND BACKGROUND

Modern processors come with complex energy management facilities to ensure efficiency and performance for a variety of utilizations. However, those facilities' energy management decisions are not always perfect as they can only utilize heuristics based on limited information. The number of execution cores in x86 processors is steadily rising, bringing in more and more components whose energy demands have to be balanced against each other, increasing the necessary complexity. Together, rising complexity and imperfect control mechanisms can lead to anomalies, where the power consumption and performance of processors behave unexpectedly. Such an anomaly is present in the AMD Zen architecture, where a system running an independent workload on 28 cores exhibit less power consumption and more performance than when utilizing 27 cores. This work aims to delimit the enabling factors of this anomaly. We examine the impact of the anomaly on parts of the microarchitecture. Finally, we show how careful maneuvering of the anomaly lead to increased energy efficiency.

Sec. 2 of this paper provides background on the AMD Zen microarchitecture and energy management facilities.In Sec. 3, we introduce our methodology and the used test system. Sec. 4 presents the experimental results and discusses the mechanics behind the anomaly, while Sec. 5 examines the impact of the anomaly on the performance and energy efficiency. Lastly, Sec. 6 concludes this paper and presents future research.

2 THE AMD ZEN MICROARCHITECTURE

An overview on the AMD Zen microarchitecture is given in technical manuals [1]. Details have been discussed previously by Singh et al. [7] and Burd et al. [4].

2.1 Structural Composition

Figure 1 shows the layout of the AMD Zen microarchitecture. The lowest interconnect level shown there is the Core Complex (CCX), which consists of up to four processor cores sharing one L3 cache. Up to two of the CCX constitute a NUMA-node, which connects to

Final edited form was published in "ICPE '21: ACM/SPEC International Conference on Performance Engineering", Virtual Event, 19. – 23.04.2021.
S. 103–106. ISBN 978-1-4503-8331-8.
https://doi.org/10.1145/3447545.3451193

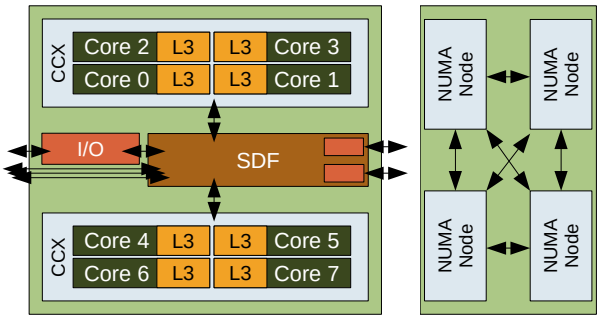
main memory and I/O via the Scalable Data Fabric (SDF) (cf. Figure 1a). For high core counts, up to four NUMA-nodes, each on one die called Zeppelin, are connected on one package (cf. Figure 1b).

2.2 Processor Voltages

The AMD Zen microarchitecture utilizes a complex set of voltage domains. According to Burd et al. [4], the voltage regulators on the mainboard supply the VDDCPU and VDDIO domains to the processor. The VDDIO domain covers the DRAM and I/O components of the package. The L3 caches are powered by the VDDCPU domain, which is also used as input of the core voltage regulators. Each core has a low-dropout regulator, controlled by the system management unit (SMU) of the respective die / NUMA-node. The SMUs control the voltages of their cores independently. They each receive the global maximum allowed frequency from the master SMU to enforce infrastructure limits "including package power, temperature, current, and voltage" [4, Section II-E]. Before the master SMU signals an increase in maximum frequency to the other units, it requests a raise of VDDCPU from the mainboard voltage regulators such that the higher voltage constraints of higher frequencies are met. If the maximum frequency is decreased, the master SMU first informs the other SMUs and then the mainboard voltage regulators to lower VDDCPU. The VDDCPU voltage is always higher than the voltage of any core, allowing the low-dropout regulators smooth out voltage dips produced by high current peaks on the VDDCPU rail. Finally, the remainder of the package components uses the VDDSOC and VDDSOC_S5 domains.

2.3 C-States

AMD Zen processors implement multiple idle states, which allows the operating system to apply power-saving measures like clock and power-gating. While these idle states are standardized by the Advanced Configuration and Power Interface (ACPI) [8] as power states or C-states, their implementation depends on the microarchitecture. AMD Zeppelin processors support three different C-states: the active state, where a core executes instructions (C0), and the idle states C1 and C2 [2]. C1 is the first sleep state in which the ACPI standard demands a negligible time-to-enter and time-to-exit. The deepest sleep-state supported in the AMD Zen architecture is C2, which corresponds to the ACPI C3 state. The ACPI C3 state is supposed to use power gating and have a non-negligible latency to exit or enter.



(a) NUMA-Node (b) 4-Die Package

Figure 1: Layout of the AMD Zen processors, based on [4]

2.4 Dynamic Voltage and Frequency Scaling

A critical parameter in processor design is the Thermal Design Power (TDP), initially defined as the maximum amount of heat that a cooling system designed for that processor has to dissipate. Modern microarchitectures such as AMD Zen have dynamic parameters such as voltages and frequencies that influence power consumption and performance. This variation aims to stay within the TDP and other temperature, power, and current limits while retaining a high energy efficiency, particularly during variable and partial load.

The most relevant feature of Dynamic Voltage and Frequency Scaling (DVFS) present in the AMD Zen microarchitecture regarding this paper is turbo frequencies. In general, turbo frequencies allow the processor or parts to increase their frequency beyond the nominal specification to achieve higher maximum performance when the thermal/power budget is not yet fully utilized.

Two primary frequency limits guide the turbo frequencies, i.e., the *All Core Boost Frequency* and the higher *Max Core Boost Frequency* (MCBF). With turbo enabled, a processor core can always utilize a frequency between the nominal frequency the All Core Boost Frequency as long as the thermal budget is not exhausted. A further increase in the core frequency, up to the MCBF, can only be achieved for cores in NUMA nodes where at least five cores reside in the C2 sleep state. "If any NUMAnode has no more than 3 cores in C0 or C1, then the core can boost further" [2, Section 6.3]. If there are three cores in C-states C1 or C0 in the first NUMA-node and four cores in C1 or C0 in the second NUMA-node, only the remaining cores of the first NUMA-node can utilize the MCBF.

3 METHODOLOGY

To isolate the anomaly and understand its impact, we systematically control the processor activity and measure various parameters. In Sec. 4, we investigate the influence of active core count on frequency, power consumption, and voltages. This is complemented by comparing configurations with the same number of tasks that are scheduled differently among NUMA-nodes. In Sec. 5, we then quantitatively demonstrate the impact of this anomaly on several benchmarks stressing different parts of the microarchitecture.

3.1 Configurations of Active Cores

As a straightforward way to control core activity, we vary the number of active cores, whereas cores are enabled linearly by their OS-given numbering. This way, first, the cores of the first CCX are activated, then of the second CCX of the same NUMA-node, followed by the other NUMA-nodes. To describe more complex configurations, we use a tuple that denotes the number of active cores of each NUMA-node. In the following setup, a core is active if it executes a workload thread pinned to one of its hardware threads. Alternatively, we keep a core active by disabling deep sleep states through the operating system without running a task.

3.2 Workloads and Benchmarks

To stress the cores in the first experiment, we used a simple compute loop that performs vector multiplications for a fixed amount of time.[1] To demonstrate the performance and efficiency impact

[1]The workload, measuring code and our raw results are provided in this repository: https://github.com/tud-zih-energy/zen-anomaly

Final edited form was published in "ICPE '21: ACM/SPEC International Conference on Performance Engineering", Virtual Event, 19. – 23.04.2021.
S. 103–106. ISBN 978-1-4503-8331-8.
https://doi.org/10.1145/3447545.3451193

Processor	AMD EPYC 7551P
Cores / Threads	32 / 64
NUMA Nodes / CCX	4 / 8
Nominal Frequency	2.0 GHz
All Core Boost Frequency	2.55 GHz
Max Core Boost Frequency	3.0 GHz
Operating System	Ubuntu 18.04.1 @ Linux 5.0.0-15

Table 1: Test System Details

of the anomaly, we use four benchmarks which are bottle-necked by different aspects of the architecture: STREAM (main memory throughput) [5], pointer chasing (L3 cache latency), NPB Embarrassingly Parallel (EP, core performance) and NPB Multi-Grid (MG, cross-package communication) [3]. All benchmarks perform a fixed amount of work, thus we use the runtime as performance metric.

3.3 System under Test and Measurements

We perform the experimental evaluation on a system with one AMD EPYC 7551p processor (see Table 1). We use `perf` to measure the `msr/aperf` and `msr/mperf` counters to compute the per-core frequencies. Using IPMI out-of-band commands, we collect the VDDCPU voltage domain samples provided by the test system's BMC. Additionally, we record the VID of the requested voltage on each active core using the undocumented model-specific register `0xC0010293` with the bitmask `0x3FC000` using a custom knob in the x86_adapt library [6]. Note that frequency and voltage values for inactive cores are not meaningful. For generally idle cores, the frequency as computed by 2 GHz × `aperf/mperf` only corresponds to short activity during interrupts and measuring the core voltage causes a wake-up and the value therefore doesn't represent voltage during actual idle. A calibrated LMG450 power meter [9] provides the full node power measurements. We do not use RAPL power readouts as they are likely based on a model on this system.

4 IMPACT OF ACTIVE CORES ON FREQUENCY, POWER, AND VOLTAGE

Figure 2 shows frequencies, power consumption, and voltages for the different number of active cores executing the simple compute loop. For this measurement, the system has turbo mode enabled on all cores. The anomaly between 27 and 28 active cores is clearly visible in almost all measurements. Consistent with the initial observation, the power consumption for 28 active cores is significantly lower than for 27 cores, even though it increases almost linearly in other cases. [2] The per-core frequency measurements show that up to three cores of a NUMA-node use higher turbo frequencies up to the MCBF of 3.0 GHz. This observation is consistent with the documentation discussed in Section 2.4. Nevertheless, due to significantly lower frequencies on other NUMA-nodes, the mean frequency of active cores remains below the All Core Boost Frequency of 2.55 GHz between 4 and 9 or 12 and 27 active cores. However, for 28 active cores and higher, all active cores reach ~2.55 GHz — raising the mean core frequency.

[2]There are noticeable but less significant plateaus between the third and fourth core of each NUMA-node. They are likely caused by respective voltage and frequency changes local to the NUMA-node.

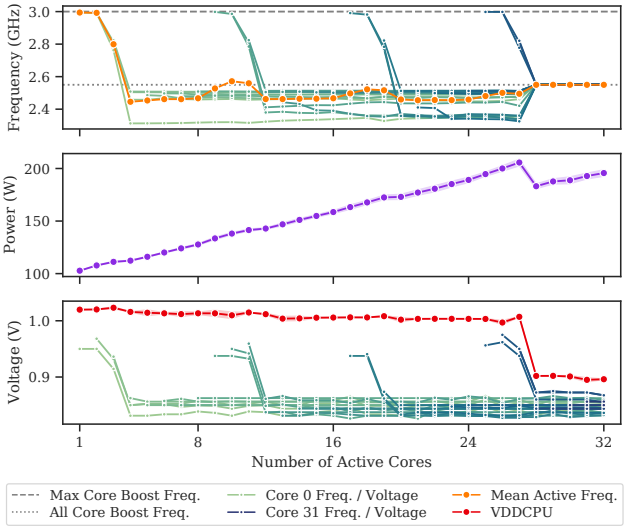


Figure 2: Frequencies (for each active core, mean of all active cores, specified Boost frequencies), power consumption (full system), and voltages (per-core and processor VDDCPU) for different numbers of cores running the simple independent compute task. All configurations show the mean of ten repetitions, the power and VDDCPU voltage charts further display the standard deviation as area. The frequency and voltage for inactive cores is not shown.

The core voltages follow a pattern similar to the core frequency with higher voltages for cores approaching the MCBF on underutilized NUMA-nodes. Contrary, the global processor voltage VDDCPU is high even for configurations where the MCBF is not used[3], and lower only for 28 and more active cores. The differences in global processor voltage result in a lower power loss in the low-dropout regulators and decrease the power consumption of the L3 cache. Therefore, this differences is most likely the reason for the reduced system power consumption at 28 and more active cores. Moreover, the reduced power consumption may have an indirect impact on frequency: Reduced power consumption from lower voltages could increase the available power and thermal budget of cores on a highly utilized NUMA-node[4], which allows all cores to reach the specified All Core Boost frequency. Note that at the higher VDDCPU, cores on highly utilized NUMA-nodes remain at lower frequencies. The significant differences across the frequency of cores can be explained by manufacturing variabilities resulting in different power consumption.

To further verify that this anomaly relates to the utilization of NUMA-nodes and the theoretical possibility of MCBF, we execute the NPB EP benchmark for different configuration pairs. Each configuration pair uses a fixed number of active threads/cores but varies the distribution of threads across NUMA-nodes. Table 2 shows that the configurations where all NUMA-nodes are highly utilized — MCBF is not possible — use consistently less power and runtime. None of the patterns discussed appear when repeating the same experiments at nominal core frequency with turbo mode disabled.

[3]i.e., between 4 and 8, 12 and 16, or 20 and 24 active cores
[4]A NUMA-node with four or more active cores is considered highly utilized.

Final edited form was published in "ICPE '21: ACM/SPEC International Conference on Performance Engineering", Virtual Event, 19. – 23.04.2021. S. 103–106. ISBN 978-1-4503-8331-8.
https://doi.org/10.1145/3447545.3451193

n	MCBF	no-MCBF	ΔPower	ΔRuntime	ΔEnergy
16	(4, 4, 5, 3)	(4, 4, 4, 4)	−7.2 %	−1.0 %	−8.1 %
17	(8, 3, 3, 3)	(5, 4, 4, 4)	−7.1 %	−1.4 %	−8.4 %
24	(8, 8, 8, 0)	(6, 6, 6, 6)	−9.1 %	−1.4 %	−10.2 %
27	(8, 8, 8, 3)	(8, 8, 7, 4)	−9.2 %	−1.7 %	−12.8 %

Table 2: Relative power consumption, runtime, and energy of NPB EP configuration pairs with the same number of threads (n). For each configuration pair, one could use the MCBF and the other cannot. Negative values indicate lower metrics for the no-MCBF configuration.

5 IMPACT ON PERFORMANCE AND ENERGY EFFICIENCY

Figure 3 shows the differences in power consumption, runtime, and energy of benchmarks running in the configuration (8, 8, 8, 3) where MCBF is possible and then with an additional core forced to stay outside of the C2, so that MCBF is not possible anymore. An apparent reduction in power consumption, comparable to the drop observed in Section 4, can be seen in the core-adjacent EP and pointer chasing benchmarks. However, the STREAM and MG benchmarks show a far less pronounced reduction or no reduction. This observation further indicates that the reduction in power consumption is due to less power loss at the low-dropout regulators, as the EP and pointer chasing benchmarks apply heavier load on those. The cross-package communication bound MG and main-memory bandwidth bound STREAM benchmark also involve the Infinity Fabric supplied from the VDDIO voltage domain. Furthermore, the runtime of the EP and pointer chasing benchmarks decreases slightly, increasing energy efficiency. This improvement is due to the slightly increased core frequencies that, as indicated in Section 4, could be due to increased power and thermal budget. As not the core frequency, but the communication with off-core components mainly limits the performance of the MG and STREAM benchmarks, the runtime of these benchmarks does not change noticeably, likewise the energy efficiency. As none of the four benchmarked components, i.e., cores, L3 cache, processor interconnect, and main memory, showed any significant drops in their performance, careful maneuvering of the anomaly seems feasible.

6 CONCLUSION AND FUTURE WORK

This paper presents an energy management anomaly in the AMD Zen microarchitecture and reveals its probable cause. The voltage measurements show a suboptimal configuration prompted by the system management units, which leads to a lower mean core frequency and higher power consumption of the package. This misconfiguration can only occur when at least one inactive core could use the MCBF feature upon its wakeup. Once no core can use this feature, the processor reaches a more energy-efficient operation point by lowering the VDDCPU voltage.

With a focused examination using four different benchmarks, we gather a preliminary impression on the impact of the anomaly on applications. While the energy impact is most pronounced in the benchmarks that heavily utilize the L3 cache and processor cores, the effect on communication-heavy and memory-bound benchmarks is negligible. Thus, mitigating the anomaly seems advantageous in general. It seems plausible that a firmware update

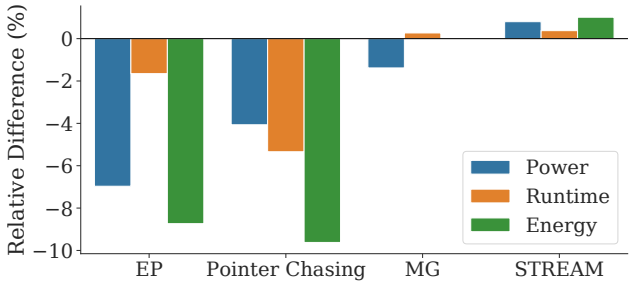


Figure 3: Relative difference in power consumption, runtime and energy of the EP, pointer chasing, MG, and STREAM benchmarks when the MCBF is not used.

of the system management units could fix this particular anomaly, i.e., by changing the requested voltage for VDDCPU when no core uses the MCBF. However, this anomaly could stem from a deliberate choice to reduce latency by keeping the voltage up, as long as it would be possible for an awakening core to use the MCBF immediately. In this case, but not limited to, further research into an anomaly-aware scheduler could lead to noticeably increased energy-efficiency in real-life workloads.

Further research should evaluate the impact on real-life applications, possibly with detailed instrumentation of the voltage regulation circuitry. We could not reproduce the anomaly on processors of the subsequent AMD Rome architecture.

ACKNOWLEDGMENTS

This work is supported in part by the German Research Foundation (DFG) within the CRC 912 - HAEC.

REFERENCES

[1] Advanced Micro Devices Inc. [n.d.]. *Processor Programming Reference (PPR) for AMD Family 17h Models 01h,08h, Revision B2 Processors.* https://www.amd.com/system/files/TechDocs/54945_3.03_ppr_ZP_B2_pub.zip
[2] Advanced Micro Devices Inc. 2018. *HPC Tuning Guide for AMD EPYC Processors.* http://developer.amd.com/wp-content/resources/56420.pdf
[3] David Bailey, E. Barszcz, Barton J.T, Browning D.S, Carter R.L, Dagum D, Fatoohi R.A, Paul Frederickson, Lasinski T.A, Robert Schreiber, Horst Simon, Venkat Venkatakrishnan, and Weeratunga K. 1991. The Nas Parallel Benchmarks. 5 (09 1991), 63–73. https://doi.org/10.1177/109434209100500306
[4] Thomas Burd, Noah Beck, Sean White, Milam Paraschou, Nathan Kalyanasundharam, Donley Gregg, Alan Smith, Larry Hewitt, and Samuel Naffziger. 2019. Zeppelin: An SoC for Multichip Architectures. 54, 1 (1 2019), 133–143. https://doi.org/10.1109/JSSC.2018.2873584
[5] John D. McCalpin. 1995. Memory Bandwidth and Machine Balance in Current High Performance Computers. (12 1995), 19–25.
[6] Robert Schöne and Daniel Molka. 2014. Integrating Performance Analysis and Energy Efficiency Optimizations in a Unified Environment. *Comput. Sci.* 29, 3–4 (Aug. 2014), 231–239. https://doi.org/10.1007/s00450-013-0243-7
[7] T. Singh, S. Rangarajan, D. John, C. Henrion, S. Southard, H. McIntyre, A. Novak, S. Kosonocky, R. Jotwani, A. Schaefer, E. Chang, J. Bell, and M. Co. 2017. 3.2 Zen: A next-generation high-performance x86 core. In *2017 IEEE International Solid-State Circuits Conference (ISSCC).* 52–53. https://doi.org/10.1109/ISSCC.2017.7870256
[8] Unified EFI Forum Inc. 2017. *Advanced Configuration and Power Interface Specification Version 6.2.* https://uefi.org/sites/default/files/resources/ACPI_6_2.pdf
[9] ZES Zimmer Electronic Systems GmbH. [n.d.]. *4-Kanal Leistungsmessgerät LMG 450.* https://www.zes.com/en/content/download/286/2473/file/lmg450_prospekt_1002_e.pdf