AMD EPYC™ 9004 AND 8004 SERIES CPU POWER MANAGEMENT

AMD
together we advance_data center computing

June 2024

AMD EPYC™ 9004 AND 8004 SERIES CPU POWER MANAGEMENT

CONTENTS

INTRODUCTION

When you choose AMD EPYC™ processors, you start with leadership performance and efficiency for a broad set of data center workloads. AMD EPYC processors have set more than 300 world performance records,[EPYC-022E] and they also power the most energy-efficient servers in the world.[EPYC-028D] This extraordinary combination of performance and efficiency is due, in part, to the effectiveness of our 'Zen 4' core design and to the 4th Gen EPYC processor's 5nm process technology.

The most important factor in how an EPYC processor performs is the workload it executes. Different workloads engage some aspects of the processor more than others. For example, some database workloads perform best with a small number of high-frequency cores; high-performance and technical computing applications often perform better when given a large amount of cache; and some applications thrive on high core density. The hybrid, multi-die architecture of AMD EPYC processors enables us to assemble different combinations of cores and I/O dies to create products whose strengths match the demands of the particular applications they were built to serve.

It's easy to see how the performance of any workload running on an EPYC processor is subject to the functioning of multiple components, including the memory controllers, the I/O subsystem, the AMD Infinity Fabric™ interconnecting multiple CPUs, CPU dies, the cores that run on them, and the Level-3 cache. Performance is a multivariate function of the contribution by each of these components on the overall workload performance.

Optimizing a processor's performance and efficiency is similarly a continuous multivariate integration of each component's operational parameters so that excellent performance and/or performance per watt can be achieved in real time. In AMD EPYC 9004 and 8004 Series processors, AMD Infinity Power Management achieves this multivariate integration through telemetry on each of the processor components. Our EPYC 9004 and 8004 Series processors are self-regulating so they can deliver the most performance while honoring the boundaries placed by the processor itself, the server infrastructure surrounding it, and the parameters set by the user to indicate power management preferences.

Without making any custom settings, AMD Infinity Power Management is designed to deliver excellent performance for a wide range of workloads with its default settings. However, you can adjust the many power management parameters to deliver the workload behavior that you desire. This white paper provides an overview of AMD Infinity Power Management in AMD EPYC 9004 and 8004 Series processors, and it provides a brief description of the various power management modes that you can set.

A NEW ERA IN POWER MANAGEMENT

The industry is on the threshold of significant change in data center operations. The need for performance and rollout of massive amounts of infrastructure is tempered by the desire to reduce power consumption at the same time. Data centers are transitioning to higher ambient temperatures, so efficient cooling both in the data center and within servers themselves become important factors in determining processor performance.

The closer you can match infrastructure to workloads, the higher the power efficiency you can achieve. Out of the box, AMD EPYC Infinity Power Management provides excellent, automated balancing of power consumption to deliver optimal performance. Some of these variables can be set through BIOS tokens, and some can be set in real time by operating systems through through collaborative processor performance control (CPPC), and they can exert fine-grained control through our unique host system management port (HSMP). This document describes power management in AMD EPYC 9004 and 8004 Series processors.

AMD EPYC SYSTEM MANAGEMENT UNIT

The CPU's system management unit (SMU) resides on the processor's I/O die, where it manages power consumption for the entire system on chip. It has agents residing on each CPU die that can autonomously implement power policies set by the SMU across the CPU die or on a per-CPU-core basis. The SMU constantly integrates the telemetry it obtains from these agents and allocates power based on its real-time observations. New to the AMD EPYC 9004 and 8004 Series is the addition of directly observed thermals to the list of variables it manages, including performance, power, voltage, amperage, and core idle states. More power generally supports more performance, however the CPU also monitors current and voltage so it does not exceed the boundaries of the chip or the supporting infrastructure.

The operating system controls processor activity from its base frequency (P0) to lower states P1 and P2 and ultimately into one of several sleep states beginning with C0 (Figure 1). When the operating system sleeps a core, the SMU powers it off but leaves it ready for a fast wakeup. Once the operating system sets state P0, and the frequency boost feature is enabled, the SMU adjusts

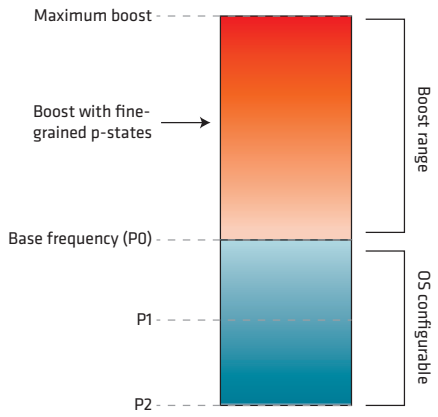


Figure 1: P-states higher than P0 are assigned by AMD Infinity Power Management when frequency boost is enabled

processor frequencies as quickly and as high as it can, ramping up power through a range of fine-grained p-states. When a state lower than P0 is set, the SMU likewise reduces frequencies as quickly as possible. The SMU also recognizes when it needs to boost power to the AMD Infinity Fabric interconnecting multiple processors, and to the memory controllers to make memory-intensive workloads perform well. By closely matching power allocation to actual needs, power is only used when it is needed and this helps to raise overall efficiency.

All of the instantaneous decisions the SMU makes are designed to maximize power efficiency while delivering maximum performance for any given power budget. It always honors boundaries set by the CPU's specified TDP or its software-configured cTDP. Cores may be accelerated to their maximum boost frequency so long as the CPU is kept within its specified TDP or configured cTDP.[EPYC-018] It also honors fixed limits such as the overall case temperature in addition to the fine-grained thermal telemetry the SMU has throughout the system on chip.

SELF-CONTAINED DECISION MAKING

One of the principles on which the SMU is based is that its power management is entirely self-contained, and decisions are made

using internal parameters while honoring external settings from the BIOS or operating system. AMD EPYC 9004 and 8004 processors, for example, cannot be overclocked, as all processor frequency decisions are made internally. This has implications that can be observed from the outside world. Because the CPU manages its own thermal conditions, using unit-less temperature control outputs from the processor, all processors are normalized to the same output to simplify thermal control across OPNs. Some processor models are designed to run at high temperatures and are rated at a 400W TDP even though they will never draw this amount of power. The higher TDP rating is to ensure a higher level of platform current, voltage, or cooling. Because of this TDP specification, it's possible to see CPUs running at high temperatures, or even at their all-core boost frequency without drawing maximum power.[EPYC-021] It's important to note that neither of these situations are cause for alarm: Because of the self-contained design philosophy of the SMU, it will not allow the CPU to run at temperatures that could damage the processor or shorten its life. As documented in our product specifications, we design our server CPUs to durably operate at the high temperatures necessary to support modern datacenters. The SMU is responsible for driving the CPU and it will prevent it from running off the road.

POWER MANAGEMENT FEATURES

There are a variety of mechanisms available for managing power consumption with AMD EPYC processors. BIOS settings provide a static approach for setting variables. AMD libraries for baseboard management controllers (BMCs) facilitate the setting and retrieval of tokens. For example, the advanced platform management link (APML) supports the BMC's out-of-band monitoring and control over power parameters. Operating systems can directly balance efficiency vs. performance preferences through collaborative processor performance control (CPPC), advanced configuration and power interface (ACPI), and the AMD Host System Management Port (HSMP). A partial list of the variables that influence how the AMD Infinity Power Management operates includes:

- **CONFIGURABLE THERMAL DESIGN POWER (cTDP)** can be used to set custom power-consumption boundaries. Some of our EPYC 8004 Series processors can be reduced to as low as 70W TDP.
- **FREQUENCY BOOST** enables core frequencies to be elevated above the P0 (base) frequencies. It can be set on or off.
- **PROCESSOR P-STATES** can be set for the base frequency (P0) and lower (P1 and P2).Setting P0 enables the SMU to boost the core frequency if frequency boost is turned on. States P1 and P2 can be used for lower-priority tasks that don't require boost.
- **ADAPTIVE IDLE CONTROL** through c-states manage multiple levels of sleep and are coordinated with cache management parameters to help ensure data consistency on return from sleep.
- I**NFINITY FABRIC PERFORMANCE SETTINGS** include setting, or allowing the SMU to set dynamically, AMD Infinity Fabric connection widths and frequencies.
- **I/O PERFORMANCE SETTINGS** allow automatic adjustments, including capping PCIe® frequencies to either Gen 4 or Gen 5. These features are a direct benefit of AMD EPYC processors being systems on chip (SoC) that have built-in PCIe connectivity. Networking, and disk drivers that would otherwise be on external chip sets and not part of the overall processor power management strategy.

PERFORMANCE AND POWER DETERMINISM

The determinism settings dictate which policies the SMU should use to manage the frequency of each core based on real-time input from the telemetry agents placed across the system on chip.

The SMU optimizes the variables of performance, power, voltage, current, thermals, and number of active cores to keep each variable within bounds. Power, for example, is not independent of its two components, current and voltage. So the SMU may demand more current, but it won't do so at the expense of violating platform boundaries imposed by the server power supplies.

Note that the SMU considers performance to be a variable. When performance determinism is set, performance is a constant, and the platform strives to maintain consistent performance based on the parameters of the chip specification so that each CPU with the same model number performs as consistently as possible. Consider, for example, a rack full of servers cooled from the floor up. The servers at the top of the rack may have higher air intake temperatures compared to those closer to the floor. Performance determinism aims to negate these relatively small differences in environmental parameters to deliver consistent performance. In contrast, the power determinism setting allows the performance variable to be unbounded, so the CPU's performance floor is the same as in performance determinism. If it is able to deliver higher performance given the bounds of its variable space, the SMU will allow this to happen. In the rack cooling example, you may observe higher performance from the servers lowest on the rack because they are cooled more effectively and the thermal part of the equation is relaxed and performance may be increased.

Our CPUs are designed with the capability to operate at high temperatures without damaging the processor. This involves rating some of them with higher TDP values than they will reach so that the server infrastructure—power supply and cooling—will help the chip achieve optimum performance.

- **PERFORMANCE DETERMINISM:** The goal of this setting is to provide a constant level of performance delivered by each CPU of the same model number as long as the platform is running within specification, including cooling, power, and current delivery. This

setting is often used by cloud service providers who wish to deliver consistent performance regardless of how different client workloads are distributed across servers in a cluster. Performance determinism is also used in high-performance computing where it's important to have the same performance regardless of the utilization level.

- **POWER DETERMINISM:** When consistent performance between CPUs is not important to the workload, this setting allows the SMU to adjust boost frequencies up to the boost limit, as long as the processor does not exceed its cTDP setting, its built-in maximum TDP, or infrastructure bounds on thermals, power, voltage, or current. This allows the CPU to deliver the best possible performance depending on variable workload and environmental conditions. This may allow the processor to deliver higher performance than in performance determinism mode. Because performance is dictated in part by thermal conditions, adequate cooling is important for achieving the highest possible performance from a part. We don't regard delivering the full performance possible from our CPUs as an opportunity to uplevel the CPU into a higher bin, rather it is our philosophy to give you the tools you need to achieve high performance given the boundaries of the SMU, and the most important variables: workload and environmental factors.

PERFORMANCE MODES

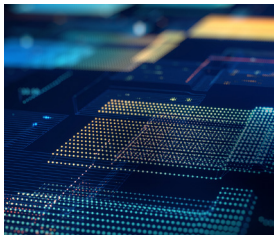
In addition to the capability to set individual power-management parameters, EPYC 9004 and 8004 Series processors offer performance modes that engage a set of behaviors in the CPU in order to support specific organizational goals. They set priority between core, I/O, and memory performance. For example, a workload that is memory-bandwidth bound may get the best results using the balanced-memory mode instead of the high-performance mode. The modes are detailed in Table 1:

- **HIGH PERFORMANCE:** Processor parameters are optimized for high performance, with memory speeds and Infinity Fabric width and speed change dynamically to match workload needs. All remaining power is allocated to cores, and if that need is met without reaching TDP, the memory and fabric speeds are set to their maximums.
- **HIGH EFFICIENCY:** This reflects the settings used to optimize performance for SPECpower® benchmarks. Core boost frequencies are constrained, and Infinity Fabric speeds are increased conservatively.
- **I/O PERFORMANCE:** This mode sets the Infinity Fabric speeds to the top two frequencies while allocating remaining TDP budget to increasing core frequencies.
- **BALANCED MEMORY PERFORMANCE:** This mode seeks to balance memory and CPU performance, allowing power to shift from accelerating memory and Infinity Fabric connections to CPU acceleration when bandwidth and latency needs are predicted to be low.

Table 1: Performance modes in EPYC 9004 and 8004 CPUs

MODE	BEHAVIOR
High Performance	• Memory speeds set to match workload needs • Infinity Fabric width and speeds adjusted to match bandwidth and latency needs • The TDP budget not allocated to the memory and fabric is allocated to run active cores at the highest possible frequency • If TDP is not reached by the cores and memory subsystem, memory speeds and Infinity Fabric width and speed are set to maximum
High Efficiency (Settings optimized using SPECpower® benchmarks)	• Reduce Infinity Fabric speeds based on the number of cores running at base frequency • Reduce Infinity Fabric width range to quarter-to-half instead of the default quarter-to-full • Set core boost frequency to base frequency plus a small delta • If some cores are 100% active, allow remaining cores to consume remaining TDP not used by the memory subsystem • If all cores are less than 85% active, lower core clock frequency
High I/O Performance	• Set Infinity Fabric width and speed to the maximum allowed by other parameters • Maintain the fabric clock at the top two frequencies • Manage core performance same as "High Performance"
Balanced Memory Performance	• Set memory speed and Infinity Fabric speed and width to match workload bandwidth and latency needs, biased toward moving to slower and narrower states if a minor performance impact is projected • Manage core performance same as "High Performance"

CONCLUSION



The most important factor in performance is the workload that the CPU is asked to execute. We help you match processors to workloads through a comprehensive set of CPU products that vary our hybrid, multi-chip architecture to provide the resources that meet different workload needs.

Once you match a CPU to your workload, the next step is to vary the power allocated to individual subsystems—cores, memory controllers, and Infinity Fabric—to enhance workload performance. AMD Infinity Power Management constantly integrates across a multivariate space to make real-time adjustments to deliver the performance that is needed to propel your workload.

Out of the box, AMD Infinity Power Management does an excellent job at managing performance through power allocation. If you wish to tune the power parameters so that the CPU responds exactly as you wish, we provide a wide range of knobs that can influence how the SMU adjusts power across the multiple CPU components. Performance Modes can be used to dictate operational philosophies such as high performance or high efficiency.

We believe in providing all of the performance our CPUs are capable of delivering, even if it means taking advantage of the inherent variabilities between parts. In Power Determinism mode, the CPU is free to deliver all that it can, as long as it operates within its power and thermal parameters. When you choose AMD EPYC 9004 and 8004 processors, you gain all of the performance that our 'Zen 4' cores, memory, I/O subsystem and Infinity Fabric has to offer, all while enabling high levels of efficiency.

END NOTES

For details on the footnotes used in this document, visit www.amd.com/en/legal/claims/epyc.html.

EPYC-018: Max boost for AMD EPYC processors is the maximum frequency achievable by any single core on the processor under normal operating conditions for server systems.

EPYC-021: All-core boost for AMD EPYC processors is the average frequency of all processor cores running in performance mode while utilizing a low activity workload. Actual achievable all-core boost will vary based on hardware, software, workloads and other conditions

EPYC-022E: For a complete list of world records see amd.com/worldrecords.

EPYC-028D: SPECpower_ssj® 2008, SPECrate®2017_int_energy_base, and SPECrate®2017_fp_energy_base based on results published on SPEC's website as of 2/21/24. VMmark® server power-performance / server and storage power-performance (PPKW) based results published at https://www.vmware.com/products/vmmark/results3x.1.html?sort=score. The first 105 ranked SPECpower_ssj®2008 publications with the highest overall efficiency overall ssj_ops/W results were all powered by AMD EPYC processors. For SPECrate®2017 Integer (Energy Base), AMD EPYC CPUs power the first 8 top SPECrate®2017_int_energy_base performance/system W scores. For SPECrate®2017 Floating Point (Energy Base), AMD EPYC CPUs power the first 12 SPECrate®2017_fp_energy_base performance/system W scores. For VMmark® server power-performance (PPKW), have the top 5 results for 2- and 4-socket matched pair results outperforming all other socket results and for VMmark® server and storage power-performance (PPKW), have the top overall score. See https://www.amd.com/en/claims/epyc4#faq-EPYC-028D for the full list. For additional information on AMD sustainability goals see: https://www.amd.com/en/corporate/corporate-responsibility/data-center-sustainability.html. More information about SPEC® is available at http://www.spec.org. SPEC, SPECrate, and SPECpower are registered trademarks of the Standard Performance Evaluation Corporation. VMmark is a registered trademark of VMware in the US or other countries.

© 2024 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, EPYC, Infinity Fabric, and combinations thereof are trademarks of Advanced Micro Devices, Inc. PCIe® is a registered trademark of PCI-SIG Corporation. SPEC® and SPECpower® are trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information. Other product names used in this publication are for identification purposes only and may be trademarks of their respective owners. Certain AMD technologies may require third-party enablement or activation. Supported features may vary by operating system. Please confirm with the system manufacturer for specific features. No technology or product can be completely secure. LE-91402-00 06/24

READY TO CONNECT? Visit explore.amd.com/server-newsletter/sign-up