AMD EPYC Brings New RAS Capability

Increasing reliability, availability, & serviceability in the latest AMD design

Executive Summary

With each new enterprise platform transition, the reliance on technology increases. Today, businesses are looking to strategies like digital transformation to reshape their operations, but these new approaches to business will put more stress and reliance on servers. By digitizing, instrumenting, and capturing untold amounts of data, businesses can get a better handle on their operations by turning data into insight. But as this reliance on data increases, protecting data assets—both in flight and at rest—will need to increase dramatically.

AMD's latest generation of enterprise platform architecture is bringing new functionality, new performance, and most importantly new RAS capabilities with its EPYC CPU formerly codenamed "Naples". EPYC CPUs will feature a host of new reliability, availability, and serviceability features that are designed to enable the CPU to better handle the critical workloads driving the next generation of enterprise applications.

Today's Customer Needs

For many years, a server meant one operating system and one application. Period. Because of the relatively inexpensive cost of x86 servers in relation to larger proprietary systems, IT was happy to deploy a server for every application, partitioning applications and creating self-contained failure zones. While simple and cost effective, this containment strategy led to server sprawl. Virtualization solved that sprawl issue but also changed this failure dynamic, creating an environment where a CPU failure could impact multiple applications or environments. This change drove the need for more reliability, availability, and serviceability (RAS) in processors to help maintain better application uptime and help lower total cost of ownership (TCO) for businesses. Each new generation of CPU brought new levels of RAS as business criticality increased.

For today's always-on business world that demands real-time analysis and insight, boosting server uptime pays dividends back to organizations by enabling them to be more agile and flexible. However, this environment also pushes the limits, as systems need constant availability. Application design has moved from smaller, self-contained applications to a more services-oriented environment that drives a greater CPU RAS need; outages in a horizontal service can now have a more dramatic impact on multiple applications as the failure zones have increased, as their impact is multiplied.

Looking towards future trends, as more compute moves toward the network edge and away from the datacenter, servers need every advantage to help ensure continuous operation, as they are not under the direct eye of IT, nor are they always in a tightly controlled environment. Because of these new dynamics, businesses are looking for more reliability, availability, and serviceability in their CPU and platform designs.

EPYC Brings a New Generation of CPU Technology

EPYC is a new generation of AMD CPU technology that is designed around AMD's "Zen" x86 core. As a standard practice, AMD creates core designs that can be leveraged across a wide range of platforms, from the most intensive enterprise datacenter deployments all the way down to lightweight client devices. In this design process, it is important to consider that the most demanding environments, those enterprise deployments, will require the most stringent designs and most advanced features and functionality. As AMD must design for this broad range of products, it creates the most robust design possible, targeted at the highest end needs, and then creates its sub-designs for less critical client environments by removing or not using features and capabilities as they move down the stack. This "design for the enterprise" methodology is the opposite of starting small and then bolting capabilities on while moving up the stack, which can lead to a suboptimal design for enterprises.

Each EPYC CPU will feature up to 32 cores and up to 8 channels of DDR-4 memory, giving them tremendous scalability along with the ability to address massively scalable applications and services. One of the new introductions to this architecture is the addition of simultaneous multithreading (SMT) which will enable each EPYC CPU to handle 64 concurrent threads. Because of this massive scalability, AMD will be focusing EPYC toward applications that are heavily threaded or must handle highly concurrent operations. In environments like high performance computing, technical computing, cloud computing, database, and transaction processing, these products should do well in keeping pace with the scaling needs of modern applications. In newer workloads like big data analytics (especially real-time), cloud computing, or IoT applications, the flexibility of the architecture to scale up and down easily and quickly is designed to bring an unprecedented level of control to help businesses anticipate and move with agility.

But with all the intended capabilities that this CPU will bring to the market, the exposure that has been opened truly demands more RAS features to help protect data in the compute cycles. With more simultaneous threads, a larger memory footprint, and more I/O bandwidth, there is more data available for potential exposure. The system needs to be able to safeguard that critical data as it moves through each processing cycle. The

24/7 nature of today's applications and the constant stream of telemetry from devices, systems, applications, and networks means that AMD needs to work harder to protect all this data in flight, because EPYC is handling more complexity than ever before. To ensure EPYC can handle the pressing needs of businesses, AMD focused much time on the reliability, availability, and serviceability of this new CPU design.

AMD DESIGNS EPYC FOR INCREASED RELIABILITY

Reliability must be engineered into the product; it simply cannot be "bolted on" later. Server silicon designs are created for the most stringent environments, because these CPUs will need to work with much higher duty cycles than conventional client CPUs. EPYC CPUs includes both features and processes that are designed to increase reliability of the CPU. Additionally, reliability is tied to data integrity—with regards to environmental factors—ensuring that the data being stored, acted on, or dispatched is correct and has not changed at all except when being modified properly by authorized applications.

To begin with, there are features that help avoid problems in the design and manufacturing of the CPU. FinFET Transistors, for instance, are used because they are less susceptible than bulk transistors to soft errors from external factors like radiation. Throughout the core of the CPU, soft-error tolerant devices are used in critical locations, because they can retain their state, even when being hit by high-energy particles.

AMD recognizes that some errors will occur in any environment, so it is important to be able to both detect and then recover from these situations. EYPC was designed with features that enable parity checking on the major components like pipelines, buffers, queues, and data buses. Through parity, errors can be detected, contained, and in some instances corrected, enabling systems to operate in a more reliable manner.

With each new generation of CPU, the size of the cache generally increases, primarily because cache is tied to cores, and the core count is escalating. Because caches hold the most recent and most relevant data to the current processing, it is critical that this data be accurate. To enable this, AMD has designed EPYC with multiple tiers of cache protection. The level 1 data cache includes SEC-DED ECC, which can detect two-bit errors and correct single-bit errors. Through parity and retry, L1 data cache tag errors and L1 instruction cache errors are automatically corrected. The L2 and L3 caches are extended even further with the ability to correct double errors and detect triple errors.

With 8 channels of memory and support for larger memory footprints than ever before, EPYC uses its memory protection features to help ensure data is not lost or corrupted

within the DRAM memory pools. Memory manufacturers control the memory RAS features; AMD works closely with ecosystem partners to ensure that EPYC supports the latest RAS features in the most current memory specifications. DRAM protections (Error Checking and Correcting (ECC) with Chipkill, DRAM Address / Command Parity with Replay, DRAM Write Data CRC with Replay) are all designed to check and correct errors with the DRAM that might impact applications. The addition of Row Hammer Protection enables the system to help protect against a single memory row being accessed repeatedly with high frequency, which could create issues in nearby memory locations in some cases. NVDIMM support means non-volatile DIMMs can be used in the server. These DIMMs feature non-volatile flash memory along with standard volatile memory, enabling protection for data in DRAM in the event of a power loss.

With the high core counts and large memory footprints, a lot of data will need to be moved in and out of the CPUs, all through the I/O buses. EPYC will push the limits of I/O, moving data between other CPUs, network, and storage devices, so it includes new features designed to help drive better reliability for this data in motion. Parity is now included on all internal data buses and Link Packet CRC with Retry is included on the global memory interconnect, the external global memory interconnect, and the PCIe bus. If an error is detected but cannot be corrected, Global Uncorrectable Error Containment will capture that error and prevent incorrect data from being propagated to non-volatile storage. Additionally, the PCIe bus is protected by Advanced Error Recovery (AER) as well as Enhanced Downstream Port Containment (eDPC) which both help isolate errors on the PCIe links and enable the system to recover from them.

These reliability features help prevent issues from occurring, or minimize the opportunity for errors to be introduced into the system. But as inevitable errors will occur, availability features help minimize their impact.

AVAILABILITY FEATURES

Continuous operation is critical in today's datacenters and enterprise environments. The availability features in EPYC are designed to ensure that servers are continuously running and that there is no interruption to processing. With IT measuring platform availability in terms of uptime, several features will help address this.

Machine Check Recovery delivers process-level containment of uncorrected errors, enabling software recovery for hardware-uncorrectable errors that would normally take a system offline. By localizing some errors, they can be contained, limiting their impact on the overall system. This feature enables the system to deal with an errant process or virtual machine, shutting down just that portion before it can impact the overall server.

Errors are a part of any system that is constantly running and often under stress. Through the Error Thresholding that is part of the Predictive Failure Analysis, error data can be used to identify patterns or suggest that a failure is impending. This enables online repair opportunities, system reconfiguration to avoid faulty resources, or administrators to schedule service at a more convenient time when users are less impacted (like a weekend) instead of having a failure event occur at random.

Using Watchdog Timers, the system can track the progress of execution through the CPU, detecting conditions where the processor can no longer make forward progress on executing. When discovered, the Watchdog Timer will reset the system instead of just letting the process hang.

If a single core has an issue, Core Disable can enable a server to remove that core from use in the system, allowing the server to be brought back online without fear that another core error could bring the system down again. This feature requires a system restart, but once that restart has happened, the system can remain online longer before the CPU needs to be replaced, allowing service to be scheduled at convenience.

Using Non-Transparent Bridging (NTB) on the PCIe bus, a CPU can see the memory in a remote node's PCIe memory space. For mission-critical applications, a heartbeat between two servers can be established, enabling a secondary node to monitor the state and condition of the primary node. The primary node can update state information on the secondary node on a timely basis, enabling the secondary node to take over should the primary node fail.

The goal of these availability features is ensuring that applications stay online and accessible to users and that systems require less servicing. However, servicing is a part of any system lifecycle, so it should be minimized as well, as much as possible.

SERVICEABILITY FEATURES

When addressing serviceability in the server world, most assume that the primary focus is around the physical form factor (like plugability and hot swapping). But there are aspects of the CPU that can also help accelerate servicing time and get the system back online as well. The primary goal of serviceability is threefold: first, enable more of the servicing to happen at preplanned times; second, accelerate troubleshooting; and third, minimize the time down, bringing the system back on line as quickly as possible. AMD has included a host of serviceability features in EPYC.

Error handling has been broken down into DOER and SEER roles, depending on whether the error is supposed to prompt the system into acting upon the error or use that error information for deeper analysis into the system resources. The new Scalable MCA initiative is designed for future growth of the system architecture in relationship to troubleshooting and TCO. The Machine Check Architecture Extensions (MCAX) are added extensions to the MCA architecture that improve scalability by expanding the number and content of MCA banks and provide ownership of the banks all the way down to a single core. With each bank being "owned" by a single core, the MCA registers can only be accessed from the MCA master thread, helping guarantee that the interrupts from a machine check bank are routed to the appropriate thread. Through architectural support for up to 256 MCA banks, each with 16 memory registers, the expanded MCA architecture can hold more information regarding system health that can be queried to accelerate troubleshooting. A key goal is First Error Diagnosability—capturing enough data on the error so that the error does not need to be recreated to be understood. With dedicated Deferred Error Registers, any time data poisoning happens, the error is stored in dedicated registers. Should a higher-level error also occur which might normally mask the lower level error, that lower level error will still be accessible. Additionally, EPYC can generate Deferred Error Interrupts that allow the operating system for more accurately and timely assessments of the overall server health.

To assist in the serviceability, many system repairs can happen online, helping reduce the amount of time required to recover from hard errors. L2 and L3 caches, for instance, contain spare bits which can be configured to replace failed bits by the CPU automatically through self-test. Once the system reboots from the error, it is automatically corrected through that reboot process. To help the system ride out DRAM memory failures, DDR4 Post Package Repairs can be done. When DRAMs or rows experience either soft or hard errors, the system firmware can automatically detect and reconfigure the DRAM after a quick reboot, mapping out the problem and removing the need for an immediate service event.

EPYC was designed with an Advanced Platform Management Link (APML) that is used to enable sideband communication for systems management tasks. This management construct includes the capability for both remote notification and remote management of the CPU, capturing data about CPU registers, temperature, and a variety of other areas that are important to systems management tasks.

Because the firmware is both intelligent and at the lowest level, it can identify errors that can be handled at its level, so that higher levels of the system do not need to spend extra cycles addressing them. Platform First Error Handling (PFEH) enables the system

to identify and handle issues at the firmware level, before they move up to the hypervisor or operating system level. Platform firmware, which is supplied by the system vendor, can more accurately assess hardware service needs, helping reduce service costs.

Memory errors are a large part of the total errors that servers will face, often resulting in issues that drive systems offline for servicing. To detect and correct latent memory errors, the DRAM Scrubber will perform periodic read-modify-write operations on memory locations. This feature helps increase reliability by removing correctable errors before a second error occurs that would force the system to mark the data as uncorrectable. To help with software testing and debug, servers with EPYC CPUs can use Address-based DRAM Error Injection. This feature enables a user to inject both correctable and uncorrectable errors into any identified memory space, helping speed the testing process, so memory errors can be addressed quickly and efficiently.

These features are designed to help improve the serviceability of the system. The enhancements help OEMs create a more robust system with fewer service events and help businesses focus on maintaining the online application status.

TABLE 1: AMD EPYC RAS FEATURES

Reliability	Availability	Serviceability

• FinFET Transistors
• Parity Checking on Pipelines, Buffers, Queues, & Data Buses
• SEC-DED ECC Cache Protection
• Support for DRAM RAS
  ○ ECC with Chipkill
  ○ Row Hammer Protection
  ○ Support for NVDIMM
• Link Packet CRC with Retry
• Global Uncorrectable Error Containment
• Advanced Error Recovery
• Enhanced Downstream Port Containment
	• Machine Check Recovery
	• Predictive Failure Analysis with Error Thresholding
	• Watchdog Timers
	• Core Disable
	• Non-Transparent Bridging
		• Scalable MCA Initiative
		• Machine Check Architecture Extensions
		• First Error Diagnosability
		• Deferred Error Registers
		• Deferred Error Interrupts
		• DDR4 Post Package Repairs
		• Advanced Platform Management Link
		• Platform First Error Handling
		• DRAM Scrubber
		• Address-Based DRAM Error Injection

AMD designs are used throughout high performance computing, avionics, medical, and financial markets, as well as others that require high quality and reliability. The company uses rigorous and stringent production standards to deliver products that are designed to meet the needs of those demanding customers.

CALL TO ACTION

Today's applications are more critical than ever, leading businesses to demand higher levels of RAS to protect not only their data but also their bottom line. Applications no longer run in a single domain, instead favoring interconnected environments and complex multi-tier applications that share data from multiple sources. This makes higher RAS an integral component of any platform. Interruptions can impact multiple applications as failure domains increase with application complexity; a single server failure can have multiple downstream impacts in this interconnected environment.

AMD has engineered EPYC to increase reliability, availability, and serviceability through an exhaustive set of new features and functions. These features build on the already available RAS capabilities from previous generations that have proved themselves out in some of the largest server environments like massive cloud deployments, split-second real-time financial applications, or critical government research clusters. Based on the new RAS capabilities in EPYC, Moor Insights & Strategy recommends that businesses consider AMD EPYC CPUs for their server projects, including those that are highly critical.

IMPORTANT INFORMATION ABOUT THIS PAPER

AUTHOR
John Fruehe, Senior Analyst at Moor Insights & Strategy

PUBLISHER
Patrick Moorhead, Founder, President, & Principal Analyst at Moor Insights & Strategy

EDITOR / DESIGN
Scott McCutcheon, Director of Research at Moor Insights & Strategy

INQUIRIES
Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS
This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

LICENSING

DISCLOSURES
This paper was commissioned by Advanced Micro Devices. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.