Intel® Xeon® Processor E7 Family: Reliability, Availability, and Serviceability

Advanced data integrity and resiliency support for mission-critical deployments

EXECUTIVE SUMMARY

Today's businesses increasingly depend on Intel® Xeon®-based servers to run data-intensive and mission-critical applications. Server reliability, availability, and serviceability (RAS) are crucial issues for modern enterprise IT shops that deliver mission-critical applications and services, as application delivery failures can be extremely costly per hour of system downtime. Furthermore, the likelihood of such failures increases statistically with the size of the servers, data, and memory required for these deployments. . The Intel® Xeon® processor E7 family offers an extensive and robust set of RAS features in silicon to provide error detection, correction, containment, and recovery in all processors, memory, and I/O data paths. This feature set is a powerful foundation upon which hardware and software vendors can build higher-level RAS layers to provide overall server reliability across the entire hardware-software stack from silicon to application delivery and services. The Intel Xeon processor E7 family delivers all these features at a highly competitive price point and power consumption level compared to traditional RISC-based solutions in the market.

A failure of a single core business application can easily cost hundreds of thousands to millions of dollars per hour.

Data Center Group
Intel Corporation

An Introduction to Reliability, Availability and Serviceability

Mission-critical applications such as database, enterprise resource planning (ERP), customer resource management (CRM), and business intelligence (BI) applications require newer technologies at better price points to deliver to their promise. Many of these applications also need to be available 24/7 on a wide area or global basis. While clustering and virtualization can help meet availability requirements, they are not adequate solutions for very large databases, BI, and high-end transactional systems. A failure affecting a single core business application can easily cost hundreds of thousands to millions of dollars per hour. All this evidences a need for scalable and highly resilient servers that are well suited for critical business applications and large-scale consolidation.

Traditional approaches to server RAS limited their scope to only those errors that could be dealt with at the hardware level. In the traditional approach, an unrecoverable hardware error brings down an entire server and the applications running on it, causing major service disruption and costing users and businesses alike. This traditional approach is no longer sufficient. Today's crucial business challenges require the handling of unrecoverable hardware errors, while delivering uninterrupted application and transaction services to end users.

Modern approaches strive to handle unrecoverable errors throughout the complete application stack, from the underlying hardware to the application software itself (Figure 1). Such solutions involve three components: (1) reliability, how the solution preserves data integrity; (2) availability, how it guarantees uninterrupted operation with minimal

Table of Contents

degradation; and (3) serviceability, how it simplifies proactively and reactively dealing with failed or potentially failed components.

Reliability. Data integrity concerns the protection of data through detection and correction of errors, or if they cannot be corrected, the containment of these errors. Error detection ensures that errors are identified at the data and instruction level. Error correction addresses a detected error by restoring the erroneous data bit or bits to their correct value. Error containment ensures that compromised data is flagged as such across all major components and data pathways so that subsystems other than the failing one can take appropriate action should they encounter such errors.

Availability. Modern approaches improve system availability by providing mechanisms that enable uninterrupted operation even in the presence of uncorrectable errors. These mechanisms include multiple levels of redundancy (spare processors, memory DIMMs, and I/O resources), automated failover at the silicon and hardware levels, and software-assisted error recovery in the various layers of the software stack, from the OS and virtual memory manager (VMM) at the lower end to the database, transactional, or application layers at the higher end. All of these provide server resiliency in the presence of uncorrectable hardware failures.

Serviceability. Today's approaches enhance serviceability using predictive failure analysis to identify problematic components before they cause uncorrectable errors or actual downtime, thereby simplifying replacement of components in the case of hard failures. System partitioning is used to further isolate workloads affected by uncorrectable errors from other active workloads running on the same server infrastructure, and facilitates maintenance.

Hardware Errors and Self-healing

Hardware errors can affect computed results, data stored in memory, and data in transit between components. Such errors affect the accuracy, reliability, and integrity of computations. Hardware errors fall into two categories: *soft* or *transient*

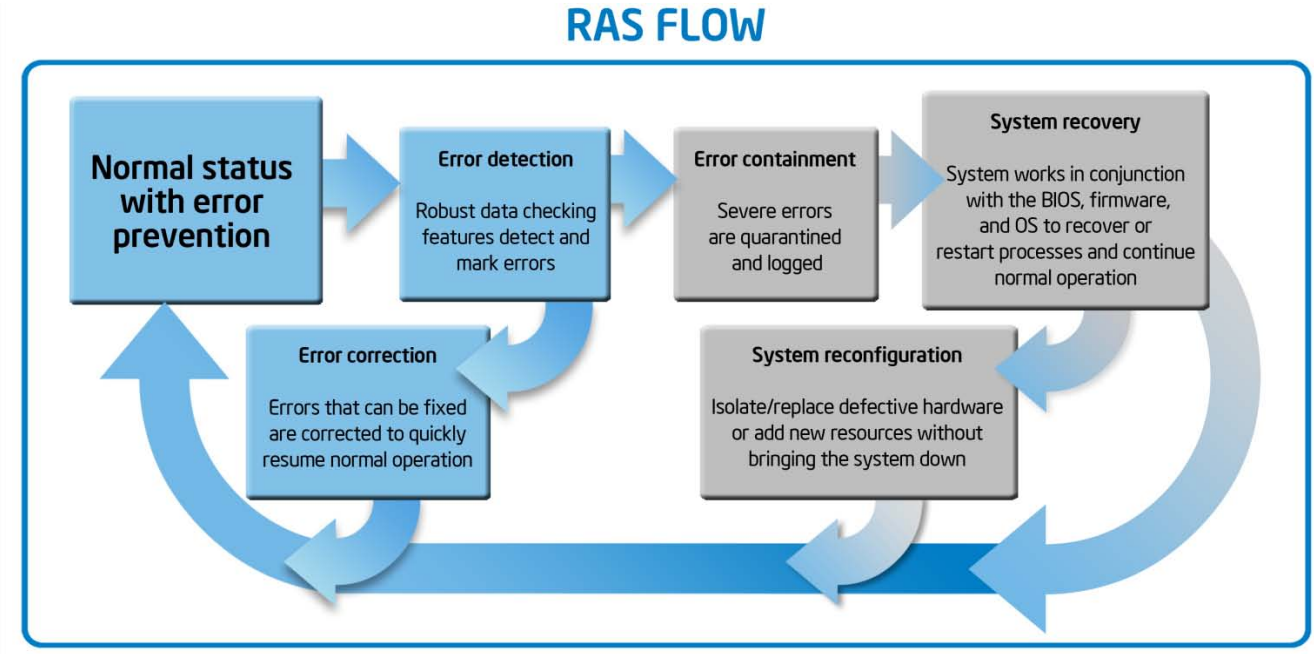

Figure 1. Advanced RAS flow in the hardware/software stack.

An important goal in system architecture has been to design and implement self-healing systems, which can diagnose themselves and recover automatically from hardware component-level failures.

errors and *hard* or *permanent* errors. Soft errors mostly occur because of random events affecting electronic circuits at the molecular level, such as alpha particles or cosmic rays dislodging electrons and therefore moving charges from one part of a circuit to another. Such events usually change the logic behavior of one or more gates. Soft errors seriously affect dynamic random access memory (DRAM). Hard errors are permanent physical failures at the hardware level, e.g., a stuck bit in a data bus, a bad bit in a dual inline memory module (DIMM), or a faulty internal circuit in a processor. Soft errors, if detected, can be corrected by circuitry that changes the logic state of a failing bit. Hard errors can be corrected only by physical replacement, which involves removing the failing component and installing a new one, or by logical replacement, which involves disabling the failing component and activating an online spare.

An important goal in system architecture has been to design and implement self-healing systems, which can diagnose themselves and recover automatically from hardware component-level failures. Self-healing requires robust failure detection, diagnosis, failure isolation, and failure-handling capabilities.

Error Detection and Correction

Handling errors requires first detecting that an error has occurred and then correcting that error. The current process for ensuring reliable hardware performance is to *detect and correct* errors where possible, *recover* from uncorrectable errors through either physical or logical replacement of a failing component or data path, and *prevent* future errors by replacing in a timely fashion components most likely to fail.

Error correcting codes (ECCs) were devised decades ago to enable the detection of single bit hardware errors and in certain cases the correction of multiple bit errors. One ECC in common use is SECDED (single error correct double error detect), which, as its name indicates, allows the correction of one bit in an error or detection of a double-bit error in a memory block. Hardware errors can be classified as either (1) detected and corrected errors (DCE) or (2) detected but uncorrected errors (DUE). Handling DCEs is done in silicon using ECCs and can be made transparent to system components. Handling DUEs, however, can require collaboration from higher layers in the hardware-software stack, from silicon to virtual memory manager (VMM), to the operating system (OS), and sometimes even the application layer.

For many years, ECC was the primary technology for silicon-level error detection and correction. Unfortunately the effectiveness of ECC protection decreases as memory capacity rises.

Memory Errors

Memory errors are among the most common hardware causes of machine crashes in production sites with large-scale systems. The typical response to memory failures is to replace any affected memory modules, which makes memory modules among the most commonly replaced server components. These system failures and their correction are very costly. ECC and advanced ECC memory debuted in industry-standard servers in the early 1990s for this use. Memory systems in servers employ ECC technologies that detect and correct errors in one or more adjacent bits. Using these codes reduces fatal memory errors that cause crashes, maintaining the illusion of operating with error-free memory as long as the error rate is low enough.

More advanced ECC technologies can increase reliability. These include IBM® Chipkill™ technology, which can correct and detect more error bits in a single memory word than traditional ECCs. Chipkill can correct up to eight adjacent error bits, and thus can correct a broken four or eight-bit-wide DRAM chip. Chipkill is frequently combined with sparing, so that if a memory chip fails (or exceeds a threshold of correctable bit errors), a spare memory chip is used to replace it. Chipkill technology lets servers withstand errors that could crash a system with less advanced error correction. Although IBM introduced the technology, today other vendors have equivalent technologies. Sun Microsystems, now a subsidiary of Oracle Corporation, calls its approach Extended ECC, and HP calls its approach Chipspare. Intel introduced Single Device Data Correction (SDDC), in the 1990's which uses ECC to protect against single x4 or x8 DRAM device failure. Intel also provides a more advanced memory correction technology called Enhanced Double Device Data Correction (DDDC), which allows recovery from two sequential DRAM failures on the memory DIMMs, as well as recovery from a subsequent single-bit soft error on the DIMM.

ADVANCED RAS AND THE INTEL XEON PROCESSOR E7 FAMILY

The Intel Xeon Processor E7 Family

The Intel Xeon processor E7 family delivers top-of-the-line performance for the most data-demanding workloads and mission-critical applications. Built on the 32 nm Intel process technology, the Intel Xeon processor E7 family offers an increased core count of up to 10 cores and 20 hardware threads in either four- or eight-socket configurations. For massive

A STUDY OF MEMORY ERRORS

Google® Inc. researchers conducted a two-year study of memory errors in Google's server fleet (see Google Inc., "DRAM Errors in the Wild: A Large-Scale Field Study," in the References section). All of the servers in the study were protected by ECC, in the form of either SECDED or Chipkill. Researchers found that ECC fixed most of the memory errors. In that study, Chipkill technology was more effective in correcting errors by a factor of 4 to 10 over the less-powerful SECDED codes.

Researchers observed more than 8 percent of DIMMS and about one-third of the machines in the study were affected by correctable (DCE) errors per year.

Researchers observed more than 8 percent of DIMMS and about one-third of the machines in the study were affected by correctable (DCE) errors per year, and DIMMs averaged nearly 4,000 correctable errors per year. The study found FIT rates (failures in time per billion device hours) of 25,000 to 70,000 per Mbit. The annual percentage of detected, uncorrected errors (DUE) was 1.3 percent per machine and 0.22 percent per DIMM.

DCEs can be handled by ECC and therefore are largely transparent to the running application. As the Google paper concludes, the DUE rate makes a crash-tolerant layer indispensable for large-scale server farms.

The Google research studied whether errors could be used to detect future errors. It found that DCEs in DIMMs provide a strong indication of future DCEs and DUEs. For the analysis, the researchers sorted the servers into six platform groups, defined by motherboard and memory generation. For each tested platform, they found that in 85 percent of cases, a DCE was followed by at least one more DCE in the same month, making the various platforms from 13 to 90 times more likely to have DCEs later in the month compared to an average month. DCEs in the previous month also increased the probability of DCEs. DUEs were 27 to 400 times more likely in a month with DCEs, and from 9 to 47 times more likely if there had been DCEs the previous month.

The research also found similar, though much slighter, correlations among failures in DIMMs in the same system, with systems that experienced failures on any DIMMs more likely to soon have failures on other DIMMs.

The Google researchers analyzed external factors related to errors. They found that memory utilization and age of DIMMs strongly correlated with memory errors. Correctable error rates increased almost logarithmically as a function of utilization levels. The incidence of

For each tested platform, they found that in 85 percent of cases, a DCE was followed by at least one more DCE in the same month,

correctable errors increased with DIMM age, while the incidence of uncorrectable errors decreased with age (due to replacements of DIMMs with those errors). DIMM capacity and operating temperature, two factors previously thought to affect errors, did not correlate with error. The study found no evidence of worse error behavior among newer generation, higher-capacity DIMMs. Within the range of temperatures of the tested production systems, temperature also did not correlate with errors.

scalability, OEMs can use node controllers to further scale up to 256 sockets per server. At least fifteen different server platforms shipping today have eight or more sockets. The Intel Xeon processor E7 family offers up to 30 MB of last-level cache to hold more data in fast access memory, and can hold very large amounts of data in memory, supporting 16GB and 32GB DIMMs and providing up to 2 TB of memory per four-socket server, or 4 TB of memory per eight-socket server. While the four-socket E7 configuration is ideal for virtualization, the powerful eight-socket E7 configuration is ideal for the deployment and consolidation of very large monolithic databases, ERP and business intelligence applications. Compared to the previous-generation Intel Xeon processor 7500 series, the Intel Xeon processor E7 family provides 25 percent more cores, 25 percent more last-level cache, and twice the memory capacity—all within the same power envelope.

RAS Features of the Intel Xeon Processor E7 Family

The Intel Xeon processor E7 family implements a powerful collection of RAS features built around the philosophy of continuous self-monitoring and self-healing (Figure 2). Self-monitoring enables a system to actively and passively monitor all its key interconnects, data stores, data paths and subsystems for errors. Self-healing features enable the server to proactively and reactively repair known errors and minimize future ones by acting automatically based on configurable DCE thresholds. By collaborating with hardware, the OS, virtual machine monitor (VMM), and application vendors, Intel enables tight integration and broad support for new, silicon-based RAS features across the hardware and software stack. (See Table 1.)

The Intel Xeon processor E7 family's RAS feature set accomplishes three main goals: it *protects data, increases system availability, and minimizes planned downtime.*

PROTECT DATA

Preserving data integrity is the foundation for RAS processing, from both computational and data management perspectives. The Intel Xeon processor E7 family provides advanced support for error detection, correction, and containment across all major components and communication pathways. It does this by reducing circuit-level errors, detecting and correcting data errors, and containing

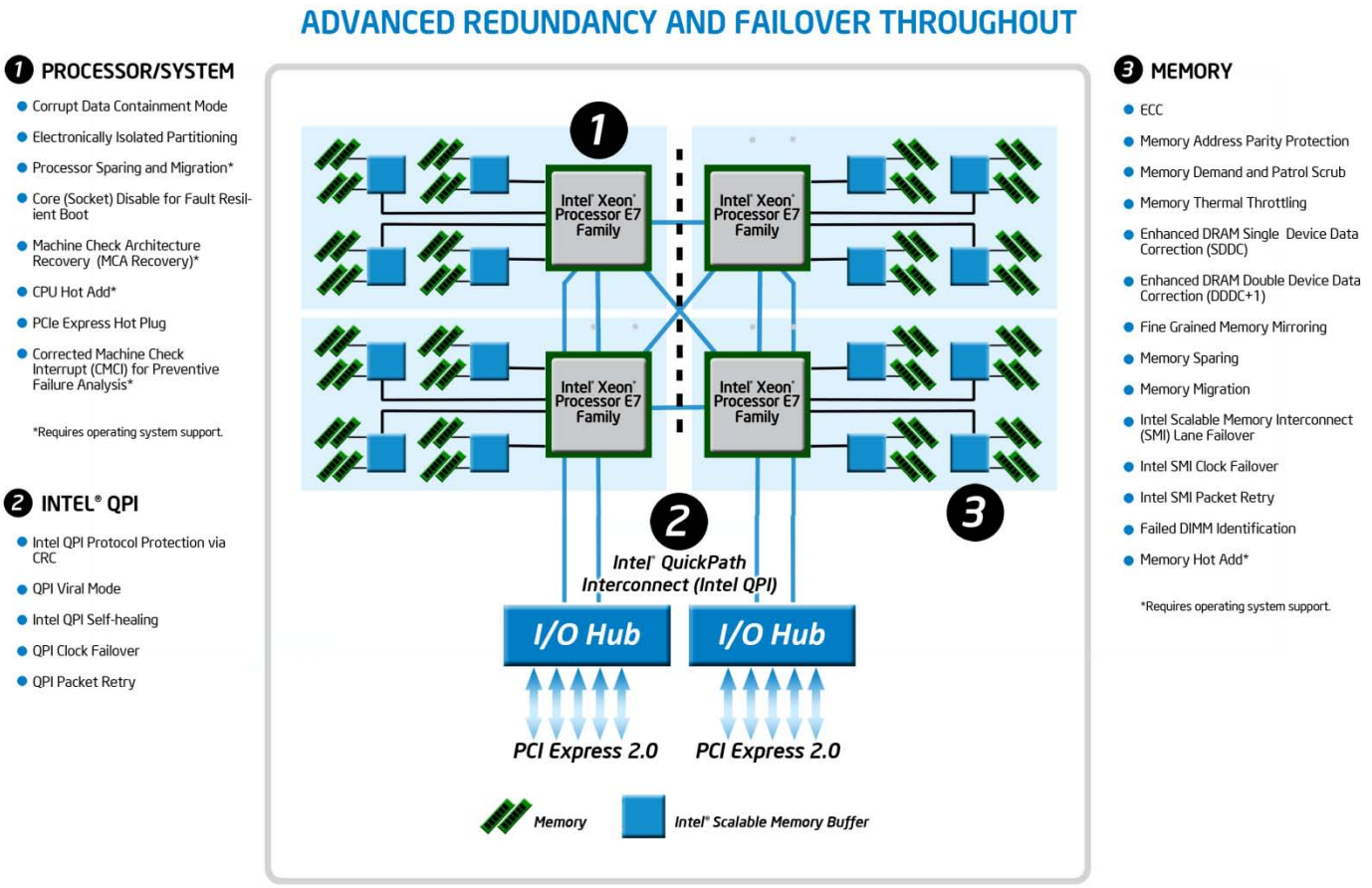

Figure 2. Intel Xeon processor E7 family advanced RAS features.

uncorrectable errors, thus limiting their impact across the system. Ensuring data integrity helps avoid computation and database corruption. For a full description of these features, see page 12.

INCREASE AVAILABILITY

For modern enterprise systems, the need for uninterrupted 24/7 operation is critical. As companies strive to become more responsive to their customers, they move toward real-time business processes, which make high availability more important than ever. For many businesses, the failure of a core business application has significant financial implications, ranging from hundreds of thousands to even millions of dollars per hour. Increased availability is especially important for high-volume transactional systems, customer support, and real-time business analysis operations. The Intel Xeon processor E7 family supports high levels of system availability through the use of multiple levels of redundancy and OS-assisted recovery from certain uncorrectable errors that would have brought down previous-generation servers.

Memory

To further increase fault-tolerance, the Intel Xeon processor E7 family includes mechanisms for memory entity sparing, mirroring, and failover. Sparing and mirroring are two RAS features that allow on-the-fly failover from a failing component to another component. Sparing allows failover to a physical spare in the same memory controller, and mirroring preserves data in the case of DRAM component failure. Failures can also occur in internal data paths, such as channel transaction and clock errors, and transient errors in the DDR channel address lines In these cases soft errors in the internal communication fabric can be handled by retries and hard errors are handled by having extra "failover" lanes in the data paths, which are activated upon the detection of such errors. Refer to page 13 for a complete list of memory RAS features.

Processor/ Socket

At the CPU and socket level, the Intel Xeon processor E7 family provides internal on-die error protection to protect processor registers from transient faults, and enables dynamic processor sparing and migration in the case of a failing processor. This set of features also handles errors at the Intel QuickPath Interconnect (QPI) and the PCIe channels. This includes QPI packet retries (similar to SMI packet retries) as well as transient error detection and recovery on QPI links. Intel QPI connects each processor to any other processors in the system and to the I/O Hub. QPI Self-healing reduces the width of a specific QPI link in response to persistent errors in order to keep the system running until repairs can be made, as shown in Figure 3. Refer to page 14 for a complete list of processor/socket RAS features.

Server

At the highest level, the Intel Xeon processor E7 family supports interactions with the operating system, VMM, and application software running on the server to enable recovery from errors that cannot be corrected by hardware. This capability is discussed in more detail later in this paper. Refer to page 14 for a complete list of server-level RAS features.

MINIMIZE PLANNED DOWNTIME

As server density increases in the data center, it becomes increasingly important to reduce operating costs, particularly in the areas of routine service and maintenance. The Intel Xeon processor E7 family provides enhanced serviceability through support of predictive failure analysis and

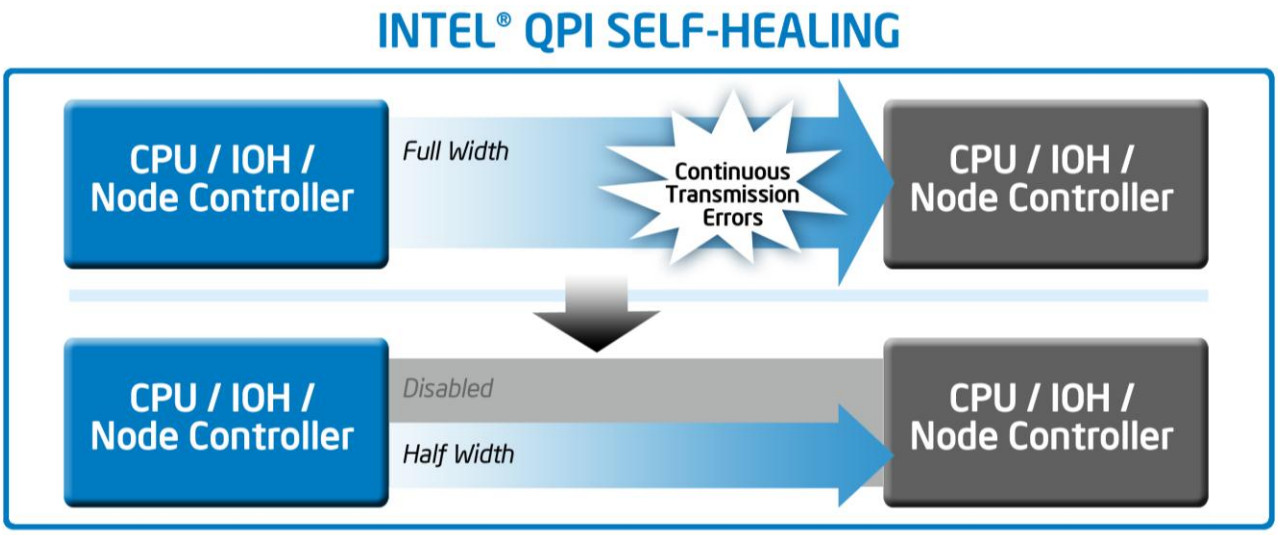

Figure 3. Adaptive QPI bus-width degradation in the presence of errors.

electronically isolated hardware partitioning, as shown in Figure 4. By reporting hardware-corrected errors up the software stack to the OS and management layers, the Intel Xeon processor E7 family allows analysis of error patterns that can be used to predict that a component will fail before the actual failure occurs and thus enables preventative maintenance during planned downtime. Likewise, the Intel Xeon processor E7 family allows IT to maintain isolated partitions instead of systems, which can reduce administration costs and enhance the ease of delivery to customer SLAs. For a full description of these features, see pages 14-15.

SOFTWARE-ENHANCED ERROR RECOVERY AND CONTAINMENT

Limitations of Other x86 Processor-based Platforms

Errors in data or processing in any computing system can impact data reliability and system availability. As they seek to maintain data

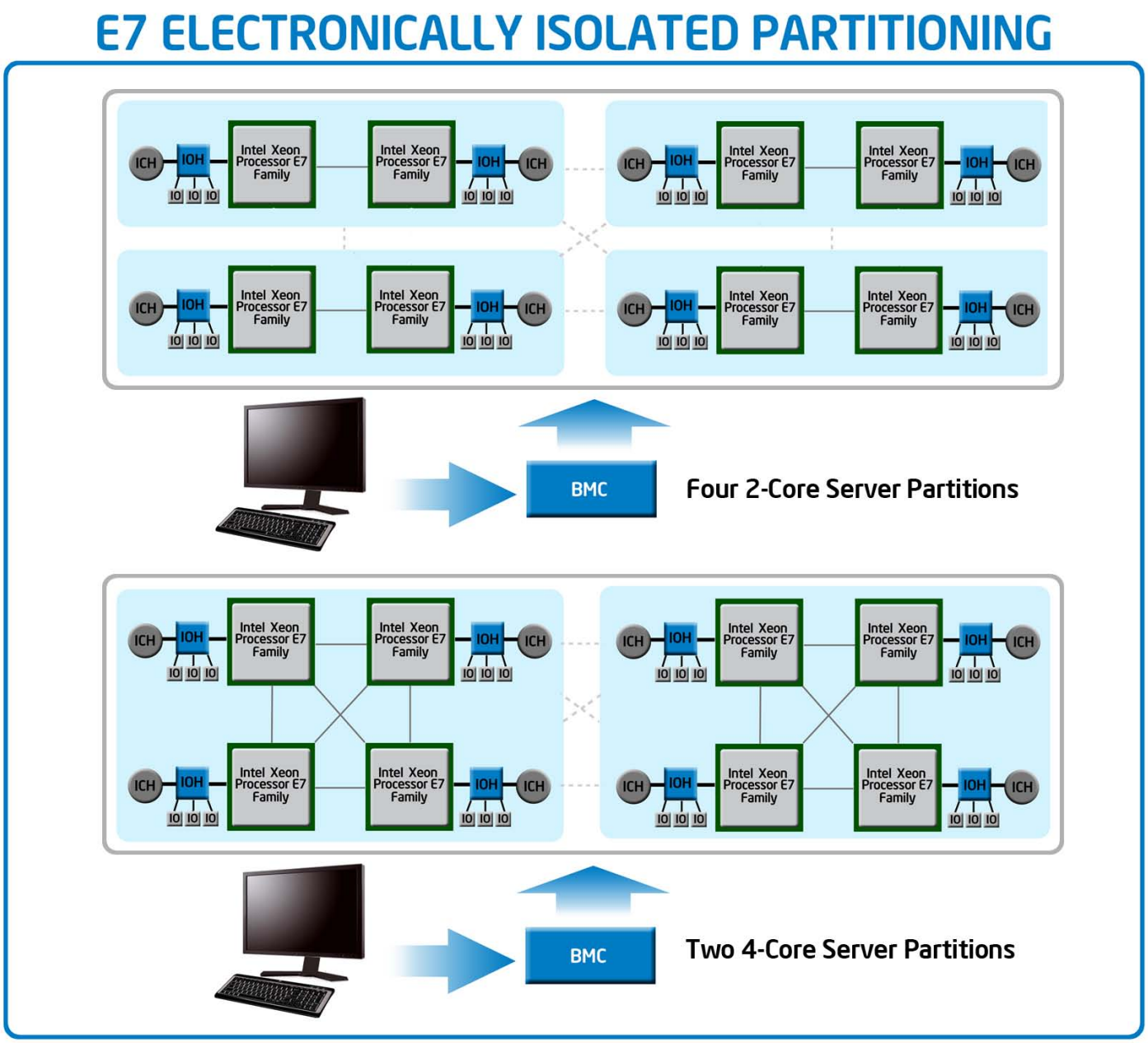

Figure 4. Electronically Isolated Partitioning provides increased reliability and service levels. Partitioning support in servers is vendor-dependent. Please check with your server vendor to determine support.

integrity and keep mission-critical services online, companies must consider both the soft and hard errors defined earlier.

Where older servers based on previous-generation Intel architecture and other x86 processors rely on server redundancy combined with software-level solutions for reliability, the Intel Xeon processor E7 family takes RAS support to the next level, bringing high-end capabilities that are unprecedented in open-system server architectures. Silicon-based RAS features, such as those of the Intel Xeon processor E7 family, overcome many of the limitations of server redundancy and software-based approaches.

In older processor technologies and other x86 processors, uncorrectable memory and hardware errors are almost always fatal, bringing system operation to a halt and causing major problems for those with mission-critical servers. In order to mitigate this limitation, the RAS features of the Intel Xeon processor E7 family are designed to cooperate and communicate with the software stack, from the BIOS and VMM/OS layers to the application layers, to handle uncorrectable errors in ways that were not previously possible in x86 servers. RAS features such as Machine Check Architecture (MCA) recovery, failover capabilities, electrically isolated partitioning, and hot sparing greatly improve reliability by enabling software-assisted recovery from uncorrectable errors that would have brought down previous-generation x86 servers. Data-intensive, mission-critical computing environments need such features to increase availability and minimize planned downtime.

Operating System Support for RAS Features

In order to obtain the full benefits of the Intel Xeon processor E7 family RAS feature set, the processor silicon layer needs to collaborate with the OS and software stack. Major OS vendors support Machine Check Architecture (MCA) recovery, as well as Corrected Machine Check Interrupt (CMCI) for logging corrected errors and doing predictive failure analysis. Other features where OS support is important are processor/memory on-lining and socket/memory migration. These latter features allow self-healing in the presence of processor and memory faults. Table 1 below shows available support for RAS features from major operating system vendors such as Microsoft® and its implementation of Windows Server® 2008 R2, Red Hat and its Red Hat® Enterprise Linux® (RHEL) 6 implementation, Novell® SUSE®'s implementation of its Linux Enterprise Server (SLES) 11 SP1, and Oracle®'s Oracle Solaris® 10 implementation.

Software-Assisted Extensibility of Machine Check Architecture (MCA) Recovery

MCA recovery is an Intel Xeon processor E7 family feature that allows the silicon layer to enlist the support of OS, VMM, and even application software to enable recovery from errors that cannot be corrected in the hardware. As Figure 5 shows, when a correctable error is detected at the silicon level, the Intel Xeon processor E7 RAS error-correction mechanisms will automatically repair the error and notify the OS that a correctable error was encountered. The OS can then log the error for further preventive maintenance analysis, and system operation can return to normal. If on the other hand, an uncorrectable error is detected, where previously the system would have been shut down, now a machine check is signaled to the OS layer. At that point, if the OS determines that the memory page where the uncorrectable error occurred is not in use, then that page is unmapped from storage and marked for repair.

If the faulty memory page is in use, the OS layer signals the enabled application that is using that page that an unrecoverable machine check has been encountered and the location at which the error has occurred. The application has the opportunity to attempt to recover from the data error. If the affected data can be

Table 1. Intel Xeon processor E7 family RAS Features OS support summary. Additional features will be supported in upcoming OS releases. Please contact OS vendors for additional details.

	Microsoft® Windows Server®	Linux		Oracle® Solaris®
E7 RAS Features requiring OS Support	WS2008 R2®	Red Hat® RHEL6®	Novell® SLES11 SP1®	Solaris 10, Open Solaris®
Corrected Machine Check Interrupt (CMCI) for Predictive Failure Analysis	WS2008R2 WHEA Support	RHEL6	SLES11SP1	Open Solaris Solaris 10 FMA
Machine Check Architecture (MCA) Recovery	WS2008R2	RHEL6	SLES11SP1	Open Solaris Solaris 10
Processor Hot Add	WS2008R2	RHEL6		Open Solaris
Memory Hot Add	WS2008R2	RHEL6		Open Solaris
Processor Sparing w/OS Assisted Socket Migration	WS2008R2			

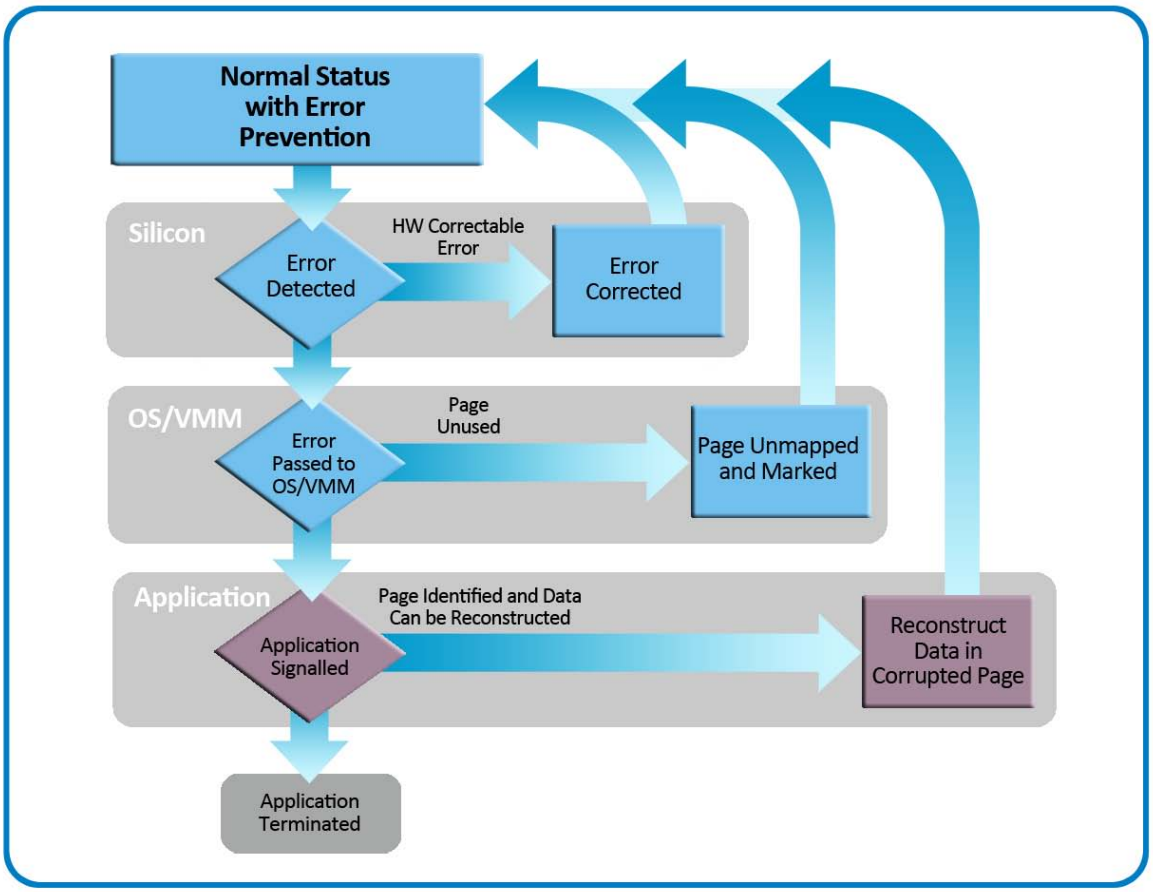

Figure 5. Software-assisted MCA recovery process.

reconstructed, the application proceeds to reconstruct the data in the corrupted page in a different memory area. Once this has been done, and the failing page has been unmapped and marked for repair, the application and the system can return to normal operation.

Finally, if the application cannot reconstruct the corrupted data, or if the application is not enabled for MCA recovery, the operating system will terminate the application and take appropriate action at that point.

We should note that the E7 family MCA extensibility enables a level of collaboration between hardware and software in the isolation and repair of uncorrectable errors that was impossible up to now. This escalation from silicon to OS to application where each system layer uses the specific knowledge it has about the impact of the error on the overall system and application/service operation, is a very powerful model for collaborative hardware-software uncorrectable error recovery.

The section on SAP® HANA discusses an actual case of how an application can collaborate closely with the OS to perform higher-level software recovery from uncorrectable data errors.

MCA recovery in a virtualized environment

In a virtualized environment, the Virtual Machine Manager (VMM or also "hypervisor") is also an actor in the MCA recovery process. In a VMM, multiple virtual machines share the silicon platform's resources, with each virtual machine (VM) running an OS and applications. In systems without MCA recovery, an uncorrectable data error would cause the entire system and all of its virtual machines to crash, disrupting multiple applications.

However, with the Intel Xeon processor E7 family, when an uncorrectable data error is detected, the system can isolate the error to only the affected VM. Here the hardware notifies the VMM, which then attempts to retire the failing memory page(s) and notify affected VMs and components. If the failed page is in free memory then the page is

retired and marked for replacement, and operation can return to normal. Otherwise, for each affected VM, if the VM can recover from the error it will continue operation; otherwise the VMM restarts the VM. In all cases, once VM processing is done, the page is retired and marked, and operation returns to normal. It is possible for the VM to notify its guest OS and have the OS take appropriate recovery actions, and even notify applications higher up in the software stack so that they take application-level recovery actions.

MCA recovery at the application layer - SAP HANA in-memory database

According to Philip Winslow of Credit Suisse "The industry stands at the cusp of the most significant revolution in database," with the shift towards in-memory databases (see Credit Suisse, "The Need for Speed: How In-Memory and Flash Could Transform IT Architectures and Drive the Next 'Killer Apps'," in the References section). Most notably, SAP, the market leader for enterprise software, is porting its complete product portfolio to their in-memory database, SAP HANA. This is possible because memory has become large enough to contain all data. Keeping data in memory allows orders of magnitude faster access to data. For recovery purposes, a backup of the database is kept on flash disks; however the complete reconstruction of the database in memory is still a significant operation. It is therefore crucial for uninterrupted business that even in the case of uncorrectable memory errors applications are not terminated by the operating system but are able to continue operation.

In order to satisfy this requirement, in collaboration with Intel, SAP has successfully applied MCA recovery to SAP HANA to ensure maximum resiliency of in-memory data stores. SAP HANA is an excellent example of an application that benefits from the Intel Xeon processor E7 family's MCA recovery features. It provides in-memory computing through an in-memory row and columnar data store that provides transactional capabilities, together with an engine that implements persistent views of data, a unified data modeling environment, and a data management service. The very large main memory in Intel Xeon E7 processors enables SAP HANA to store and process multi-terabyte transactional data stores completely in memory.

If an unrecoverable memory error is encountered, the processor issues an MCA recovery signal to the OS. The OS in turn determines which software applications are currently using the offending memory. Traditionally, the OS would at best have to halt any application that uses that memory, or at worst, stop all processing and halt the system. With Intel's MCA recovery, it is possible to use the intelligence of each layer in the hardware-software stack to handle the error and quickly return to normal processing. The OS can notify the application, which, armed with knowledge of its internal data structures, can decide what course of action to take to repair the effects of the memory error.

If an unrecoverable memory failure occurs during the SAP HANA appliance operation, and the corrupt memory space is occupied by one of the SAP in-memory tables, the OS signals the SAP infrastructure software, which in turn responds by reloading associated tables. SAP HANA analyzes the failure and determines whether it affects other stored or committed data, in which case it uses snapshots (built during SAP HANA operation and kept in flash memory) to recover and reconstruct the committed data in a new working memory location.

In all cases, SAP HANA takes appropriate action at the level of its own data structures to ensure a smooth return to normal operation without disruption in continuity of service or loss of information. Extending the functionality of MCA recovery makes the collaboration between silicon, OS, and the application layer in the Intel Xeon processor E7family possible.

CONCLUSION

The Intel Xeon processor E7 family is accelerating mission-critical transformation from traditional RISC environments. Processors in the family are extremely well suited to the mission-critical needs in business processing, decision support, and large-scale consolidation workloads. They provide up to 10 cores, with 2 threads per core and 30 MB on-die cache, and support up to 2 TB of DDR3 memory with a maximum of 64 memory slots and 6 to 11 I/O slots in a four-socket configuration. Their consolidation and virtualization characteristics make them even more attractive in such contexts.

We have shown how the Intel Xeon processor E7 family offers a unique set of silicon-based advanced RAS features that enable a level of data integrity, system reliability, and improved serviceability throughout the whole hardware-software stack impossible to achieve until now in x86 based servers. The advanced RAS features of the Intel Xeon processor E7 family include a combination of CPU, memory, QPI, and system RAS features that together deliver *new levels of reliability and high availability to x86 mission critical deployments.*

These Intel Xeon processor E7 family RAS features provide a powerful and unique silicon foundation upon which manufacturers can build enhanced self-healing solutions at all levels of the hardware/software stack, from silicon through the operating system and virtual memory manager layers to the database and application layers at the top.

APPENDIX A – INTEL XEON PROCESSOR E7 FAMILY RAS FEATURES BENEFITS TO IT

Benefits for IT	RAS Silicon Features of the Intel Xeon Processor E7 Family
Protects Data	
• Reduces circuit-level errors • Detects data errors across the system • Limits the impact of errors	Error Correction Code (ECC)
	Memory Address Parity Protection
	Intel® QuickPath Interconnect (Intel® QPI) protocol protection via Cyclic Redundancy
	Memory Demand and Patrol Scrub
	QPI Viral Mode
	Corrupt Data Containment Mode
	Electronically Isolated Partitioning
Increases Availability	
• Heals failing connections • Supports redundancy and failover for key system components • Recovers from uncorrected data errors	Memory Thermal Throttling
	Single Device Data Correction and Enhanced DRAM Double Device Data Correction (DDDC)
	Fine Grained Memory Mirroring
	Memory Sparing
	Memory Migration
	Intel Scalable Memory Interconnect (SMI) Lane Failover
	Intel SMI Clock Failover
	Intel SMI Packet Retry
	Processor Sparing and Migration
	Socket Disable for Fault Resilient Boot
	Intel QPI Self-healing
	Intel QPI Clock Failover
	Intel QPI Packet Retry
	Machine Check Architecture (MCA) recovery
Minimizes Planned Downtime	
Helps IT • Predict failures before they happen • Maintain partitions instead of systems • Proactively replace failing components	Failed DIMM Identification
	CPU Hot Add
	Memory Hot Add
	PCIe Express Hot Plug
	Electronically Isolated Partitioning
	Corrected Machine Check Interrupt (CMCI) for Preventive Failure Analysis

APPENDIX B – DETAILED DESCRIPTIONS OF THE INTEL XEON PROCESSOR E7 FAMILY RAS FEATURES

Protect Data

Detect and Correct Errors

ECC	ECC is used to protect processor registers, processor caches, and system memory from transient faults that can corrupt program data without damaging the hardware. The increasing density of modern processors increases the likelihood of such faults.
Memory Address Parity Protection	Enables the correction of many transient errors on the address lines of the Double Data Rate (DDR) channel. Traditional parity is limited to detecting and recovering single-bit errors. Memory Lockstep provides protection against both single-bit and multi-bit errors. Memory Lockstep lets two memory channels work as a single channel, moving a data word two channels wide and providing eight bits of memory correction.
Intel QPI Protocol Protection via CRC	Detects transient data errors. It can use checksum of either 8-bits or 16-bits rolling, and is capable of detecting 1, 2, 3, and odd numbers of bit errors and errors of burst length up to 8. When it detects errors, it retries the packet. The 16-bit rolling CRC transmits 8 bits of CRC on each flow control digit (flit), and approaches the level of protection provided by a true 16-bit CRC scheme.
Memory Demand and Patrol Scrub	These features provide the ability to find and correct memory errors, either reactively (demand) or proactively (patrol) addressing memory problems. In all cases, whenever the system detects an ECC error, it will attempt to correct the data and write it back, if possible. When correcting the data is not possible, as is the case with a permanent memory error, the corresponding memory is tagged as failed or "poisoned." Demand scrubbing is the attempt to correct a corrupted read transaction. Patrol scrubbing involves proactively sweeping and searching system memory and attempting to repair any errors found. Patrol scrubbing errors may activate the Machine Check Architecture Recovery (MCA recovery) mechanism described later.

Contain Uncorrected Errors

QPI Viral Mode	Viral mode notifies the system of an uncorrectable error, with all packets having the viral bit set to indicate the presence of such errors. Viral mode causes the CPU and QPI to go into viral state, blocking QPI to PCIe messages. Software can detect this condition and respond to it appropriately. The system configuration agent will stay in that state until software changes the state or is reset.
Corrupt Data Containment Mode	Corrupt Data Containment mode, or data poisoning, works in tandem with viral mode to prevent corrupt data from spreading through the system. Data poisoning, i.e., tagging data that comes from a corrupt memory location, limits the corrupt data to the process currently running, thus preventing the data from affecting the rest of the system. The receiver of the data can check the poison tag and detect whether the data is corrupted.
Electronically Isolated Partitioning	Provides the ability to divide a single physical server into a set of independent units or partitions. It is supported with complete electronic isolation, thus providing strong workload isolation and more efficient maintenance cycles. The partitions can boot independently of each other, and each partition runs its OS and applications in isolation from the others. You can reconfigure partitions as necessary, splitting a single partition into smaller partitions or combining existing partitions into fewer, larger partitions. In addition to flexibility, electronically isolated partitioning also provides protection from hardware or software failures in other partitions, as well as a high degree of security between partitions. Partitioning changes require a server re-boot to take effect.

Increase Availability

Memory

Memory Thermal Throttling	Can prevent DIMMs from overheating while balancing power and performance. The processor monitors memory temperature and can temporarily slow down the memory access rates to reduce temperatures if needed.
Enhanced DRAM Single Device Data Correction (SDDC) **Enhanced DRAM Double Device Data Correction (DDDC+1)**	Protect the system from memory chip failure. SDDC can correct any single memory chip failure as well as multi-bit errors from any portion of a single memory chip. It can reconstruct memory contents on the fly, even in the event of the complete failure of one chip. DDDC enables a memory DIMM to continue operation even in the event of two sequential DRAM device hard-errors. Enhanced DDDC (DDDC+1) adds the capability to detect and correct an additional single bit error on top of DDDC. DDDC+1 is a new feature unavailable in previous-generation processors. The ability to recover from two DRAM failures improves uptime and extends the time between service calls, lowering overall service costs.
Fine Grained Memory Mirroring	A method of keeping a duplicate (secondary or mirrored) copy of the contents of select memory that serves as a backup if the primary memory fails. The Intel Xeon processor E7 family supports more flexible memory mirroring configurations than previous generations allowing the mirroring of just a portion of memory, leaving the rest of memory un-mirrored. The benefit to IT is more cost-effective mirroring for just the critical portion of memory versus mirroring the entire memory space. Failover to the mirrored memory does not require a reboot, and is transparent to the OS and applications.
Memory Sparing	Allows a failing DIMM or rank to dynamically failover to a spare DIMM or rank behind the same memory controller. When the firmware detects that a DIMM or rank has crossed a failure threshold, it initiates copying the failing memory to the spare. There is no OS involvement in this process. If the memory is in lockstep, the operation occurs at the channel pair level. DIMM and rank sparing is not compatible with mirroring or migration.
Memory Migration	Moves the memory contents of a failing DIMM to a spare DIMM, and reconfigures the caches to use the updated location so that the system can coherently use the copied content. This is necessary when a memory node fails or the memory node ceases to be accessible. The act of migrating the memory does not affect the OS or the applications using the memory. Typically, this operation is transparent to the OS. Because in some cases OS assistance improves performance, OS-assisted memory migration is also available.
Intel Scalable Memory Interconnect (SMI) Lane Failover	Can detect channel transaction errors, affect automatic hardware recovery, and retry the transactions after. Upon detection of an error, SMI will map the failed data lane to a spare lane. Even though a wire has failed, this remapping allows the system to keep running, without compromising fault detection, until it is eligible for repair. Intel SMI provides an additional lane for both southbound and northbound links. Persistent CRC errors on the channel trigger this mapping, and using the spare lane enables failover without compromising CRC protections.
Intel SMI Clock Failover	Directs forwarded clocks to the clock failover lane in the case of a forwarded clock failure. As with lane failover, this allows for uninterrupted operation until the system can be repaired (refer to the explanation of QPI clock failover below).

Intel SMI Packet Retry	Restarts a cycle when SMI detects a transient failure on the link. When the CPU detects a CRC error, SMI retries the request at the link level. Intel SMI supports retry on both northbound and southbound transient errors. Note that repeated CRC errors can activate viral mode.

Processor/Socket

Processor Sparing and Migration*	Enables the dynamic and proactive reassignment of a CPU workload to a spare CPU in the system in response to failing memory or CPU components. The migration, which requires OS assistance, configures the state of a spare CPU socket to match the processor and memory state of the failing CPU. Once the migration is complete, the system can force the failing CPU offline for replacement in the next maintenance cycle.
Core (Socket) Disable for Fault Resilient Boot	This feature disables a failing core (or socket) at boot time, allowing the system to power on despite the core (socket) failure.
Intel Quick Path Interconnect (QPI) Self-healing	Intel QPI connects each processor to any other processors in the system and to the I/O Hub. QPI Self-healing reduces the width of a QPI link in response to persistent errors. This dynamic width reduction allows the system to continue operation in a degraded mode until repairs can be made. When detecting persistent errors, a full-width port reduces to a half-width port. If necessary, a half-width port can reduce further to a quarter-width port. IT can set the error threshold at which QPI will enter self-healing mode.
QPI Clock Failover	Directs forwarded clocks to one of the two dual use lanes. Lanes 9 and 10 normally function as data lanes, but, in the event of a failure, one of these data lanes is used as the clock lane.
QPI Packet Retry	Automatically retransmits packets containing errors. This supports recovery from transient errors on QPI links. Persistent failures will enter half-width mode.

* Requires operating system support.

Server

Machine Check Architecture recovery (MCA recovery)*	Allows higher-level software, such as the hypervisor, OS, or MCA recovery-aware application to recover from some data errors that cannot be corrected at the hardware level. Memory Patrol Scrub or Last Level Cache Write Back detect these errors. MCA recovery reports the location of the error to the software stack, which then takes the appropriate action. For example, the OS might abort the task owner in response to an error, allowing the system to continue running.

* Requires operating system support.

Minimize Planned Downtime

Failed DIMM Identification	Identifies specific failing DIMM(s), enabling IT support to replace only those DIMMS.
CPU Hot Add*	Allows the addition of a physical CPU module to a running system at QPI interface. A new CPU can immediately replace a failing CPU via migration or be brought on-line later.
Memory Hot Add*	Physical memory can be added while the system is running. Added memory can immediately replace failing memory via migration or be brought on-line later.

PCIe Express Hot Plug	Allows addition or removal of a PCIe card while the system is running.
Electronically Isolated Partitioning	Electronically Isolated Partitioning, discussed earlier in the Protect Data features table, enables isolation of a processor partition from hardware and software failures in other partitions in the same server. This provides added security, reduces maintenance costs, and helps ensure SLAs.
Corrected Machine Check Interrupt (CMCI) for Preventive Failure Analysis*	Sends a report to the OS about hardware errors that the RAS features of the Intel Xeon processor E7 family corrected. Although these errors have no effect on the running system, reporting corrected errors to the BIOS or OS allows predictive failure analysis to anticipate and avoid future problems.

* Requires operating system support.

REFERENCES

Credit Suisse; "The Need for Speed: How In-Memory and Flash Could Transform IT Architectures and Drive the Next 'Killer Apps' ". Credit Suisse Equity Research. March 30, 2011.

Dao, T.T.; Fairchild Camera and Instrument Corporation. "SEC-DED Nonbinary Code for Fault-Tolerant Byte-Organized Memory Implemented with Quaternary Logic." IEEE Transactions on Computers v. C-30 Issue 9 (1981). Accessed May 12, 2011, http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1675864. doi: 10.1109/TC.1981.1675864.

Google Inc.; Schroeder, Bianca; Pinheiro, Eduardo; and Weber, Wolf-Dietrich. "DRAM Errors in the Wild: A Large-Scale Field Study." Paper presented at SIGMETRICS/Performance '09, Seattle, WA, June 15-19, 2009.

Hewlett-Packard Development Company, L.P. "Memory technology evolution: an overview of system memory technologies technology brief, 9th edition." December 2010.

Hewlett-Packard Development Company, L.P. "Advanced Memory Protection technologies technology brief, 5th edition." April 2008.

Hewlett-Packard Development Company, L.P. "RAS Features of the Mission-Critical Converged Infrastructure." June 2010.

IBM Corporation; Mitchell, Jim; Henderson, Daniel; Ahrens, George; and Villarreal, Julissa. "IBM Power Platform Reliability, Availability, and Serviceability (RAS)." June 5, 2009.

IBM Corporation; Neaga,Gregor; Buratti, Pierluigi; Kellermann, Helmut ; Klinkert, Peter; Labauve, Christian ; Raffel, Gordon. "Continuous Availability Systems Design Guide." December 1998. Accessed on May 12, 2011, http://www.redbooks.ibm.com/redbooks/pdfs/sg242085.pdf.

Intel Corporation; Kumar, Mohan; Demshki, Michael; and Shiveley, Robert. "Advanced Reliability for Intel Xeon Processor-based Servers." March 2010.

Red Hat, Inc.; Heublein, Alex, and Barooah, Vedanta. "Accelerating IT Migration Success with a Rock-Solid HP and Red Hat Enterprise Linux Platform." Paper presented at Red Hat Summit 2010-JBoss World 2010, Boston, Massachusetts, May 3-6, 2010.

Sun Microsystems, Inc.; Tang, Dong; Carruthers, Peter; Totari, Zuheir; and Shapiro, Michael W. "Assessment of the Effect of Memory Page Retirement on System RAS Against Hardware Faults." Paper presented at the International Conference on Dependable Systems and Networks, Philadelphia, Pennsylvania, June 25-28, 2006.