Case Study

Data Center

Intel® Memory Resilience Technology

Intel & SK hynix: Memory Failure Analysis and Prevention in Data Centers

Maximizing Memory Reliability in Data Centers Through Artificial Intelligence-Assisted Failure Analysis

Servers rely heavily on dynamic random-access memory (DRAM) as the primary memory source for their speed and cost efficiency. However, DRAM failures can lead to computational errors, which can have a direct impact on the reliability, availability, and serviceability (RAS) of servers, potentially disrupting data center continuity. These memory failures often go unnoticed until a server crashes.

To address this issue, Intel® Memory Resilience Technology was developed to provide system administators with an early detection tool for identifying and preventing potential memory failures before they occur.

The Challenges of Memory Reliability

Memory faults can lead to a variety of correctable errors (CEs) including single bit errors, single row errors, and multi-array errors, each with their frequency patterns (see Figure 1). These faults can also have their own victim patterns, with some having a higher risk of becomming uncorrectable errors (UEs). Some memory faults are intermittent and difficult to trace, while others can be replicated. Currently, there is no one-size-fits-all solution for addressing memory errors. For example, random single bit errors can be corrected using Error Correction Code (ECC), while other types of memory errors require different technologies such as System ECC, Single Data Device Correction (SDDC), Post-Package Repair (PPR), and Intel® Memory Resilience Technology.

The team at Intel and SK hynix have identified a small batch of DDR4 memory DIMMs with memory faults that can be replicated. This allowed Intel to conduct a deep dive analysis to better understand their failures. Additionally, Intel was able to gather large-scale data from its own data centers to complete a comprehensive memory failure analysis.

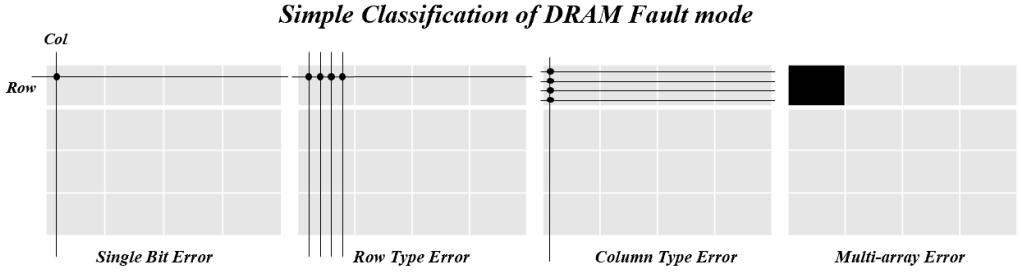

Figure 1. A Simple classification of memory fault modes

Analyzing a Small Sample of Faulty DDR4 DIMMs

To trace memory faults, the areas of the defective memory DIMMs were translated into hardware addresses. Then, using an Intel® Xeon® Scalable platform, the error characteristics were recorded both with and without Intel® Memory Resilience Technology enabled. The goal was to study memory failure patterns to determine if Intel® Memory Resilience Technology can mitigate errors caused by unreliable memory DIMMs.

The root causes of memory errors are defects in the manufacturing for memory DIMMs, while the errors themselves are symptoms of these defects. These defects, such as row, column, or bank faults, can affect multiple memory pages in the operating system that share the daulty physical DRAM address. Additionally, simply counting the number of CEs per page does not fully capture the complexity of cross-page faults.

The challenge is further exacerbated by the fact that traditional OS page offlining solutions lack knowledge of platform specific ECC implementations and DRAM-specific memory failure characteristics. ECC is the error correction capability provided by the CPU, and DRAM-specific failure characteristics depend on the microarchitecture of the DRAM. Furthermore, not all faults or pages within a certain rate of CEs are equally likely experience future UEs. The rate CEs in the past is not a reliable indicator of future UEs.

Intel® Memory Resilience Technology

Intel® Memory Resilience Technology enables data center operators to proactively predict potential memory failure risks, ensuring data center operation and workload continuity. This includes prediction-based memory page offline, replacement of unreliable memory DIMMs, and critical workload migration.

By virtue of its multidimensional model and algorithms, Intel® Memory Resilience Technology examins DIMM errors at the micro-level to assign health scores and detect potential future failures in real-time. It utilizes AI to create a predictive by analyzing thousands of memory error logs from the field. By comparing this model to scans from the operator's data center, Intel® Memory Resilience Technology can pinpoint potential issues.

Results and Analysis

The following results and analysis were obtained using a memtester tool, a user space tester that stress-tests the memory subsystem. This tool is particularly effective at identifying intermittent and non-deterministic faults.

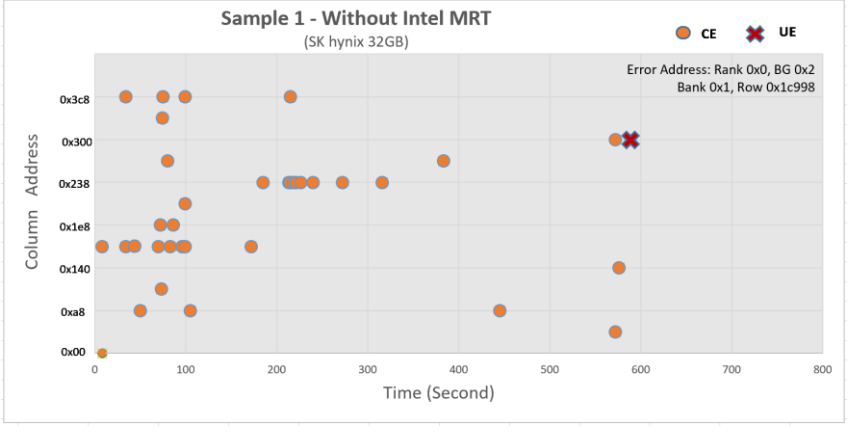

Figure 2. Unreliable DIMM Sample 1 without Intel® Memory Resilience Technology

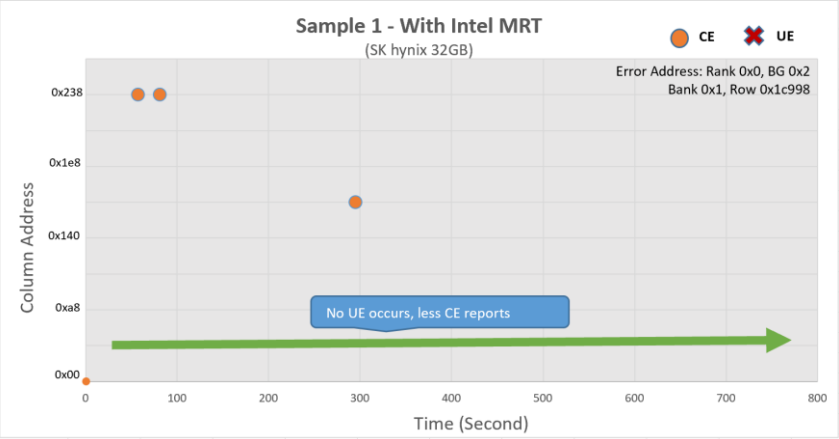

Figure 3. Unreliable DIMM Sample 1 with Intel® Memory Resilience Technology

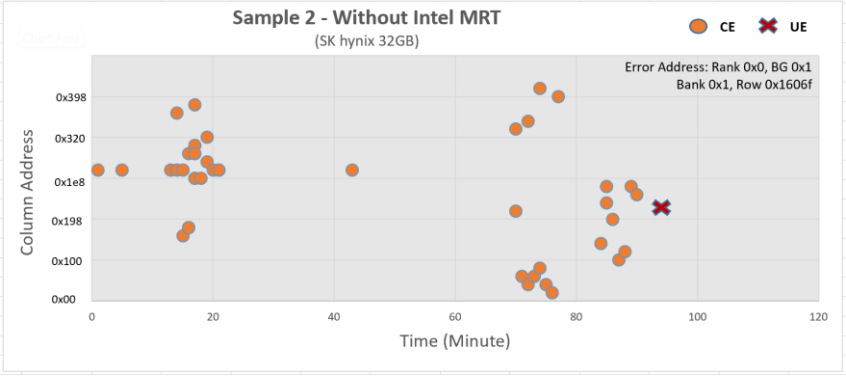

Figure 4. Unreliable DIMM Sample 2 without Intel® Memory Resilience Technology

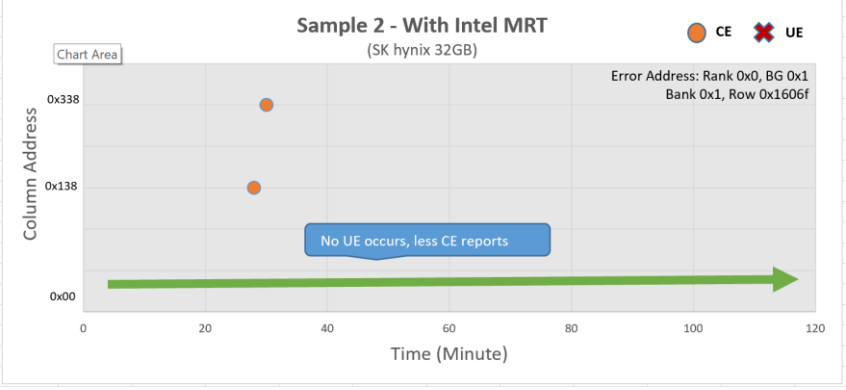

Figure 5. Unreliable DIMM Sample 2 with Intel® Memory Resilience Technology

The DIMM Sample 1 was installed in a Intel® Sky Lake server with Intel® Memory Resilience Technology. First, Intel® Memory Resilience Technology was turned off, (see Figure 2). The OS legacy Predictive Failure Analysis (PFA) page offline feature was enabled with 10/24 policy, (meaning that if OS detected 10 CEs in 24 hours, the page would be offlined by the OS). Many CEs were reported from row address 0x1c998, and column addresses 0x00 to 0x3c8. After 34 CEs occurred in 9 minutes, a fatal UE caused the system crash. The system log confirmed that one page was offlined by the OS PFA.

Next, Intel® Memory Resilience Technology was turned on, and the test was repeated (see Figure 3). Intel® Memory Resilience Technology detected the memory error patterns, triggered associated pages to be offlined, and no more CEs were collected afterwards. In total, 8 pages were offlined, and the memtester finished successfully after ~3 hours.

From the analysis above, It is clear that Intel® Memory Resilience Technology can detect memory error patterns, automatically offline associated faulty pages in advance, and keep the system running intact. Even if the failed DIMM is in a different environment (both hardware and software) and the memory characteristics change, Intel® Memory Resilience Technology can detect high-risk errors using memory error pattern.

Similar to Sample 1, for memory DIMM Sample 2, Intel® Memory Resilience Technology detected the high-risk memory errors, triggered the page offline, and ensured that the system continued to run smoothly (see Figure 4 and Figure 5).

Both Sample 3 and Sample 4 were severely damaged and were unable to pass the BIOS memory test. Intel® Memory Resilience Technology is not suitable for such cases. In general, severely damaged memory DIMMs should identified and removed from deployment.

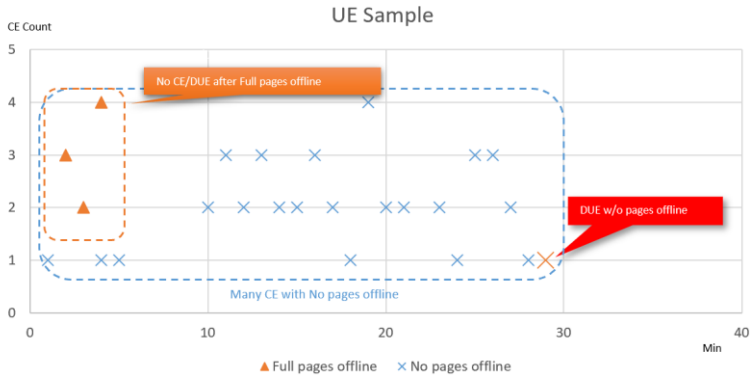

Figure 6. Test results provided by SK hynix

In Figure 6, the X-axis represents time and the Y-axis represents the number of memory errors. The blue-dotted rectangle represents a test where no pages were offlined, and after many CEs, a UE occurred within 30 minutes. In contrast, the orange-dotted rectangle represents a second test where all associated pages were offlined, and no CE or UE occurred.

These test results demonstrate that a UE can occur within minutes to hours after a CE, which highlights the need for real-time failure analysis and swift preventative action.

Large-Scale Testing

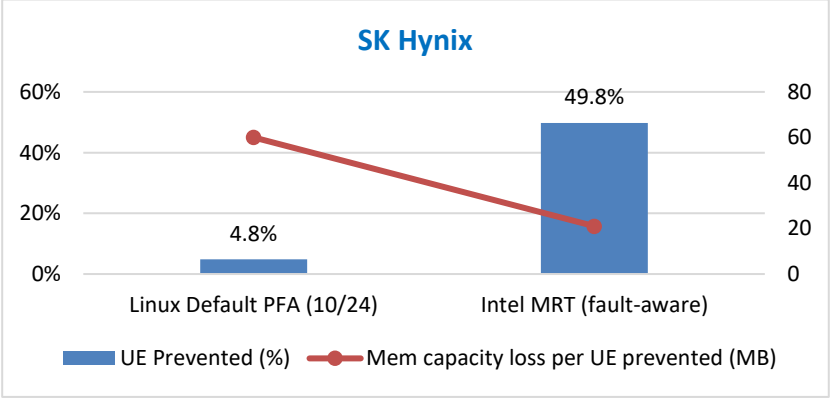

Figure 7. Large Scale Data Test

PAGE OFFLINING POLICY	UE PREVENTION (%)	MEM CAPACITY LOSS PER UE PREVENTION (MB)
Linux Default PFA (10/24)	4.8%	60
Intel® Memory Resilience Technology (fault-aware)	49.8%	21

Table-1

Intel collaborated with SK hynix to compare the performance of Intel® Memory Resilience Technology with Linux's default PFA on thousands of servers (Intel SkyLake and Cascade Lake systems) from June 2022 to November 2022. The collected data was filtered and summarized in Figure 7 and Table 1 indicating that 49.8% of memory failures were prevented by Intel® Memory Resilience Technology, at the cost of less than 21 MB memory space per UE prevented. In contrast, Linux's default PFA could only prevent 4.8% of memory failures at the cost of 60 MB memory space. The UE prevention rate varied depending on the DRAM part. The UE prevention for different SK hynix parts ranged from 27.6 % to 84.3% with an average of 49.8%. Therefore, Intel® Memory Resilience Technology shows varying effectiveness depending on the DRAM fault modes and parts, and can have a high fault detection capability and UE prevention performance for some DRAM fault modes and parts.

Conclusion

SK hynix demonstrated the effectiveness of UE prevention by taking offline all pages, including all defective row addresses on the platform. It is crucial to retire all pages associated with memory errors in advance.

Intel® Memory Resilience Technology can detect the memory error patterns of the samples provided by SK hynix, trigger pages to be offlined, and keep the system running smoothly.

Intel® Memory Resilience Technology cannot manage errors that are too severe to pass the BIOS memory test. DIMMs with such errors need to be filtered by a memory vendor test or server manufacturing test in advance.

Intel® and SK hynix are collaborating on developing new methods to predict memory errors and retirement, not only for DDR4, but also DDR5.

SK hynix Inc., headquartered in Korea, is the world's top-tier semiconductor supplier offering Dynamic Random Access Memory chips ("DRAM"), Flash memory chips ("NAND Flash"), and CMOS Image Sensors ("CIS") for a wide range of distinguished customers globally. The Company's shares are traded on the Korea Exchange, and the Global Depository shares are listed on the Luxemburg Stock Exchange. Further information about SK hynix is available at www.skhynix.com and https://news.skhynix.com/

Where to Get More Information

For more information on Intel® Memory Resilience Technology, visit www.intel.com/mrt.