Case Study

Data Center
Intel® Memory Resilience Technology

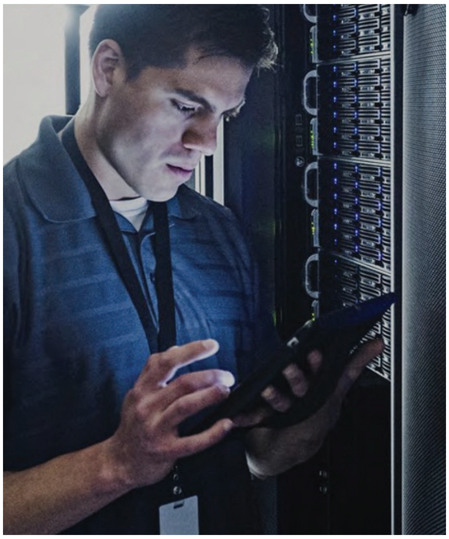
Intel & Samsung: Improving Memory Reliability at Data Centers

Reduce memory errors at data centers with predictive failure analysis and intelligent prevention using fine grain telemetry and artificial intelligence.

SAMSUNG

Business:

Samsung Electronics Co. Ltd inspires the world and shapes the future with transformative ideas and technologies. The company is redefining the worlds of TVs, smartphones, wearable devices, tablets, digital appliances, network systems, and semiconductor and LED solutions. For the latest news, please visit the Samsung Newsroom at news.samsung.com.



Background

Servers frequently have memory errors that go unnoticed by users or operators. Very often, these errors are considered inconsequential, but in large groups can impact performance, reducing server reliability and potentially disrupting data center continuity. As one of the top hardware failures that occur in data centers, memory failures have a direct impact on server reliability, availability and serviceability (RAS). Sometimes memory failures are uncorrectable and can cause an unexpected server crash. In fact, when data center operators don't receive early enough warning to take preemptive action and prevent a future outage, memory failures can result in severe damage. For example, at the data center of one global cloud service provider (CSP), it was determined that ~50% of hardware-triggered server downtime was caused by memory failures.

Legacy Solutions to Memory Failure: OS Page Offlining

Page offlining is an error-prevention mechanism implemented in modern operating systems. Traditional offlining policies are based on the correctable error (CE) rate of a page in a past period. However, CEs are just the observations while the underlying causes are memory circuit faults. A certain fault, such as a row fault, can impact quite a few pages. Meanwhile, not all faults are equally prone to uncorrectable errors (UEs).

Challenges of Legacy Solutions

As mentioned, memory errors are just the observations while the underlying causes of the errors are memory circuit faults. Faults such as a row fault, a column fault, or a bank fault can impact many pages that share the same underlying circuit. Counting CEs per page does not comprehend the nature of the cross-page faults.

The traditional OS page offlining solution has no knowledge of platform specific Error Correction Code (ECC) implementation, which is the error correction capability provided by CPU, and DRAM specific memory failure characteristics.

Furthermore, not all the faults (or the pages with the CE rate satisfying a certain condition) are equally prone to future UEs. The CE rate in the past period is not a good predictive indicator of future UEs.

Intel & Samsung: Improving Memory Reliability at Data Centers

Intel & Samsung Joint Research on Memory Failures

As determined by joint research on memory failures by Intel and Samsung, to effectively prevent UEs and reduce CEs using page offlining requires a holistic solution. This optimal solution incorporates fine-grained failure analysis on underlying DRAM faults and platform specific ECC knowledge; multiple page offlining to prevent continuous CEs and UEs which share the same defective circuit; and checking the memory mapping between the OS physical memory address to DRAM memory location.

A short-term solution is to lower the OS page offlining threshold option from 10 CEs per 24 hours to 2 CEs per 24 hours. To break through the limitations of the traditional approach to memory failure, the long-term solution is to incorporate artificial intelligence (AI).

Intel® Memory Resilience Technology

With Intel® Memory Resilience Technology and its multidimensional model and algorithms, DIMM errors are mined at the micro-level to assign health scores and identify future failures in real time. By comparing thousands and thousands of memory error logs on DDR4 from the field, Intel® Memory Resilience Technology uses AI to create a model of predictive patterns. Intel® Memory Resilience Technology then compares this model with scans from an operator's data center to determine where problems may exist. Customers are able to reasonably predict potential memory failure risks and ensure data center operation and workload continuity. This helps customers managing critical workload migration, unreliable memory DIMM replacement, and prediction-based memory page offlining.

Why Offline Multiple Pages?

When a row failure occurs on the DRAM, to guarantee faulty row isolation, multiple pages having the same DRAM row address should be offlined at the same time. The total memory size impact due to one row failure is limited from tens to hundreds of KB, depending on DRAM model. Figure 1 illustrates estimated memory size impact. Figure 2 illustrates the estimated results of percentage of UE prevented and memory size impact per UE prevented for traditional OS page offlining policy, enhanced OS page offlining policy with reduced CE rate threshold, and predictive page offlining based on Intel® Memory Resilience Technology. A UE might be reported by CPU from patrol scrub on the location where the page has been offlined, while it has no harmful effects to the system.

32GB DIMM Loading (CLX / NUMA / 4KB Page)

# of CPU	Loading	Open Page		Closed Page	
		Number of Offline Pages	Offline Memory Size	Number of Offline Pages	Offline Memory Size
1	1 CH 1DPC	4 pages	16KB	128 pages	512KB
	2 CH 1DPC	8 pages	32KB	128 pages	512KB
	3 CH 1DPC	12 pages	48KB	128 pages	512KB
	6 CH 1DPC	24 pages	96KB	128 pages	512KB
	6 CH 2DPC (12 DIMMs)	24 pages	96KB (0.00002% of Total DRAM Size)	128 pages	512KB (0.0001% of Total DRAM Size)
2	6 CH 2PDC (24 DIMMs)	24 pages	96KB (0.00001% of Total DRAM Size)	128 pages	512KB (0.00006% of Total DRAM Size)

X Page Size: 4KB

X Page Size: 4KB

Figure 1. Estimated Memory Size Impact for One Row Failure

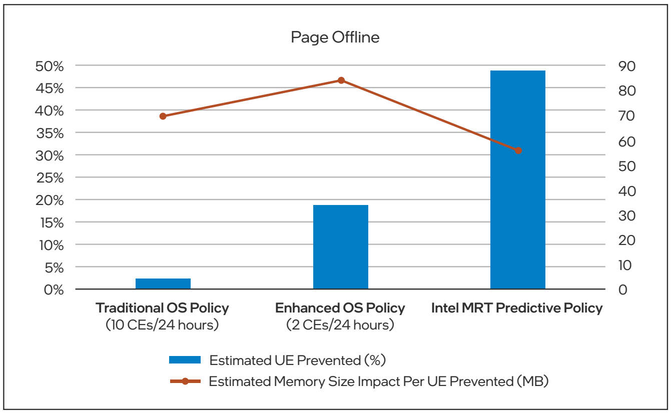


Figure 2. Estimated Results of Page Offlining

Page Offline	Estimated UE Prevented (%)	Estimated Memory Size Impact Per UE Prevented (MB)
Traditional OS Policy (10 CEs/24 hours)	2.44%	70
Enhanced OS Policy (2 CEs/24 hours)	18.70%	84
Intel MRT Predictive Policy	48.78%	56

Conclusion

Memory errors are just the observations while the underlying causes of the errors are memory circuit faults. Faults such as a row fault, a column fault, or a bank fault can impact many pages that share the same underlying circuit. Counting CEs per page does not comprehend the nature of the cross-page faults. Joint research by Intel and Samsung shows that to effectively prevent UEs and reduce CEs using page offlining requires a holistic solution that incorporates fine-grained failure analysis on underlying DRAM faults and platform specific ECC knowledge; multiple page offlining to prevent continuous CEs and UEs for the row faults; and checking the memory mapping between the OS physical memory address to DRAM memory location. A short-term workaround is to lower the OS page offlining threshold option from 10 CEs per 24 hours to 2 CEs per 24 hours. The long-term solution is to incorporate AI. Intel® Memory Resilience Technology uses AI to create a model of predictive patterns by comparing thousands and thousands of memory error logs from the field, then compares this model with scans from an operator's data center to determine where problems may exist to support data center operation and workload continuity. In fact, Intel® Memory Resilience Technology is an optimal solution for the Whitley platform and beyond.

Where to Get More Information

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

Memory resilience technology results provided through the use of Intel® Memory Resilience Technology are estimated and may vary based on differences in system hardware, software, or configuration. Results are derived using multi-dimensional models and algorithms to predict potential memory failures and do not constitute a representation or guarantee regarding memory failure.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

©2022 Intel Corporation Printed in USA ♲ Please Recycle