WHITE PAPER

The Row Hammer Effect:
Enhancing Memory RAS

with SGI® MEMlog™ Memory Error Manager and Logger

TABLE OF CONTENTS

1.0 Introduction 1

2.0 The Need for a New Approach for Memory Errors 1

3.0 Introduction to Memory Error Handling 2

- 3.1 Error Detection and Correction 2
- 3.2 Memory Resiliency 4

4.0 MEMlog Memory Error Management 5

- 4.1 Detailed Memory Error Logging 5
- 4.2 Predictive Failure Analysis with Proactive Actions 6
- 4.3 Alerts to Schedule Service 7
- 4.4 Advantages of SGI MEMlog Software 8
- 4.5 Software Availability 8

5.0 Conclusion 9

6.0 References 9

1.0 Introduction

Individual memory circuits have become increasingly vulnerable to transient errors as memory has become faster and more dense. This problem is exacerbated by the **Row Hammer Effect** in which repeated rapid activation of the same row within a device causes memory cells to leak charge. SGI developed MEMlog software to provide more comprehensive monitoring and logging of all corrected memory errors; the software goes far beyond the standard capabilities of memory controllers with detailed memory error logging down to the cell level, predictive failure analysis, proactive page retirement, and automated alerting. A proprietary transient error filter distinguishes between transient errors and hard failures that require device replacement. As a result, MEMlog software reduces costs, improves performance, and increases uptime versus systems using standard methods alone.

2.0 The Need for A New Approach to Memory Errors

Steady advancements have been made to increase memory speed and density over the years while decreasing voltage. As a result, individual circuits on a DRAM memory die have become more vulnerable to cosmic radiation, leakage current, and noise that can result in transient errors.

Distinguishing transient errors from hard failures that require device replacement can be difficult. Recently, this problem has been exacerbated by the row hammer effect in which repeated rapid activation of the same row within a device causes memory cells to leak their charge and interact with neighboring cells, producing numerous transient errors. For example, at an SGI customer's computing center, one dual inline memory module (DIMM) saw a burst of more than 2,400 corrected errors in a 95-minute period. After the row-hammer-inducing job completed, no further errors were seen despite a continued heavy workload from other jobs and repeated memory tests. In the field, a memory module experiencing this many errors would typically be replaced as a matter of course.

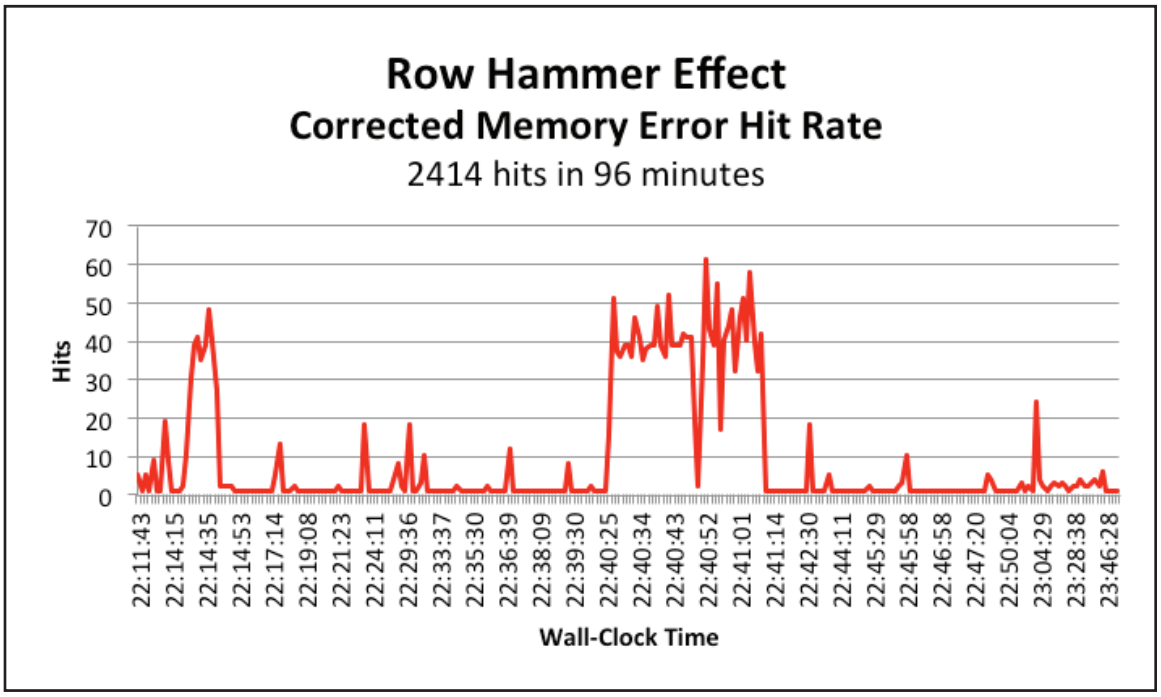

Figure 1. High error rate caused by the row hammer effect. This DIMM experienced no more errors after the job that caused the row hammer effect ended

With the total number of deployed DIMMs in customer environments steadily rising, the problem isn't just a matter of academic interest. As a major supplier of high-performance computing (HPC), high-performance data analytics (HPDA), and other computing systems, SGI ships more than 500,000 DIMMs each year. SGI recognized the need to better characterize memory errors almost 10 years ago. The SGI MEMlog utility was first released in 2008 to help system administrators improve monitoring of memory health in their SGI systems and more accurately plan memory replacement activities during regularly scheduled downtime.

In 2015, SGI added a transient error filter which reliably distinguishes between transient and hard failures and further reduces the number of memory components requiring replacement, thereby providing additional cost savings (time and replacement memory) for the datacenter.

This paper explores memory reliability, availability, and serviceability (RAS) technologies, and describes how SGI's MEMlog memory error manager and logger improves customer satisfaction using predictive failure analysis and proactive action. This solution is the only memory resiliency tool available that enables full bandwidth memory performance, while reducing the need for planned downtime and cutting the cost of sparing or replacing memory.

MEMlog Software Detects Row Hammer Effect on Large Cluster

An earth and space sciences customer has an 1,100-node SGI Rackable cluster with almost 19,000 DDR4 DIMMs. Initially, the team saw a significant number of memory errors requiring repeated rounds of scheduled maintenance for DIMM replacement. After replacement, memory errors continued from the same DIMM locations, and the returned parts passed all factory tests.

SGI MEMlog software identified the problem as a row hammer condition; MEMlog transient error filtering has now drastically reduced the number of DIMM replacements.

3.0 Introduction to Memory Error Handling

3.1 Error Detection and Correction

A typical error correcting code (ECC) capable DIMM is shown in Figure 2. Each DIMM has one or more "Ranks" of 18 dynamic random access memory chips (DRAMs), 16 chips store data, two store ECC information. Each DRAM is further subdivided into "Banks" with memory cells arranged in a grid within each Bank. (There are 8 banks in DDR3 memory and 16 banks in DDR4.) Individual cells are uniquely identified by their DIMM slot, Rank, Bank, Row, Column, and bit(s).

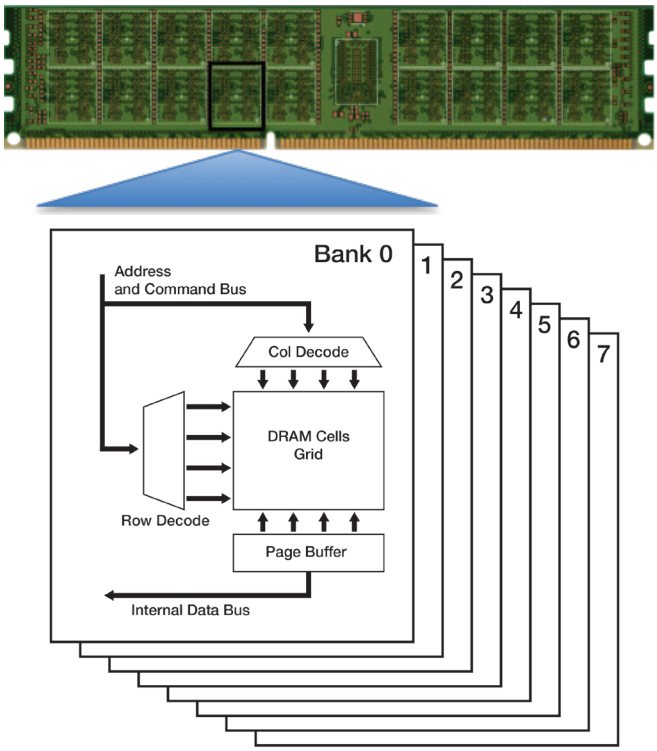

Figure 2. Logical organization of a typical ECC DIMM

Memory errors are typically tracked by ECC-capable memory controllers only to the level of the memory Rank on the DIMM. When error correction detects a memory error, the memory controller increments a counter associated with the DIMM Rank to log the error. If a DRAM in a particular Rank experiences frequent failures, the counter goes up. In some cases, the memory controller may track overall error counts for individual DRAMs within a Rank.

The memory resiliency system is triggered when a memory controller error counter hits a previously set threshold level. The threshold for these counters is commonly between 10 and 40; on some SGI systems this is adjustable via the BIOS. Because some memory errors are transient, the memory controller uses a simple aging system for error counters. Over a period of days, error counters are slowly decremented based on the assumption that some of the previous errors were transient ones.

This hardware-based method has several deficiencies. First, the memory controller cannot distinguish if an error is transient or persistent. By “leaking” errors while continuing to count up to the trigger threshold, the controller is in effect placing a bet on what type of error is occurring.

The second deficiency is the low specificity of the error counters. A transient error burst from one or more rows within a DRAM may be lumped in with a persistent error in an unrelated area of the same DRAM, or a transient error from one DRAM might be lumped in with a persistent error from another DRAM in the same Rank. Generally, the entire error counting and threshold mechanism assumes a small number of memory errors. As we’ve already seen, the emergence of the row hammer effect can generate a substantial number of transient memory errors in a short time period and is therefore likely to trigger a memory RAS event.

3.2 Memory Resiliency

The standard approach to correcting memory errors with ECC uses full-device data correction. ECC is used to protect against a single DRAM chip failure on a DIMM, and, in some implementations, one additional bit error. Memory controllers support several additional options that help protect memory resiliency. The following options are available on Intel processors. (The discussion in this white paper is primarily applicable to Intel processors and memory controllers. Similar options may or may not be available on memory controllers for other processor types.)

- **DIMM or Rank Sparing.** When a Rank exceeds its error threshold, the memory controller copies all data to a spare device that has been held in reserve and ceases to use the "failing" Rank.
- **Memory Mirroring.** All DIMMs in the system are mirrored. Failures are corrected by reading the alternate DIMM.
- **DDDC+1.** (Double Device Data Correction +1) Two DIMMs operate in lock step mode, allowing recovery from two sequential DRAM chip failures plus an additional bit error.

These resiliency methods result in different behaviors when threshold error counts are reached, and they can affect the amount of usable memory available to the system as well as memory performance. For instance, memory mirroring cuts usable memory capacity in half, while the lock step mode used by DDDC+1 reduces memory bandwidth. These effects are summarized in Table 1.

When the row hammer effect occurs, the outcome resulting from methods using spare DIMMs is far from ideal. Each node in a server or cluster has a single spare DIMM. In a Rank or DIMM sparing scenario, a row hammer that runs up the error count in a short period will trigger a spare switchover. For example if the threshold is set to 40, and a burst of errors in the hundreds takes place, when the threshold count is reached the memory controller switches to the spare Rank. Because of the change in memory configuration, it is impossible to predict whether the row hammer will continue. However, with the spare activated, there is no fall back.

The spare was needlessly consumed for a transient burst, and–without a way to distinguish transient from persistent failures–the DIMM that was spared will be scheduled for replacement. To reverse the sparing process, a power cycle would be required.

DDDC +1 uses a spare DRAM and has similar behavior. If a row hammer event occurs, the spare DRAM will be consumed with no way to fall back without power cycling.

	DIMM or Rank Sparing	Memory Mirroring	DDDC+1	Full Device Data Correction[1]
Memory Capacity	Some memory is held in reserve until a failure occurs	Usable capacity is half the total	Full capacity	Full capacity
Bandwidth Loss	Up to 20% after sparing	Up to 50% vs non-mirrored system	Some loss due to lock step	None
Response to Transient Errors	Sparing when threshold reached	Corrects transient errors	Sparing when threshold reached	Reports errors and heals
Response to Hard Failures	Sparing when threshold reached	Reads from mirror	Sparing when threshold reached	Reports errors only

Table 1. Memory resiliency options

[1] Also known as single device data correction (abbreviated SDDC or SDDC+1)

4.0 MEMlog Memory Error Management

SGI MEMlog software is supported on all SGI systems. It provides comprehensive monitoring and logging of all memory errors corrected by ECC. It goes far beyond the standard capabilities built into memory controllers. The functionality of MEMlog software can be divided into three categories:

- Detailed memory error logging
- Predictive failure analysis with proactive actions
- Alerts to schedule service

4.1 Detailed Memory Error Logging

MEMlog software provides a detailed description of a system's DIMM configuration. It tracks DIMMs by serial number, and maintains a separate history of corrected errors for failure analysis. Each log entry includes comprehensive information for each corrected memory error including Rank, Bank, Row, and Column. Sample MEMlog output is shown in Figure 3.

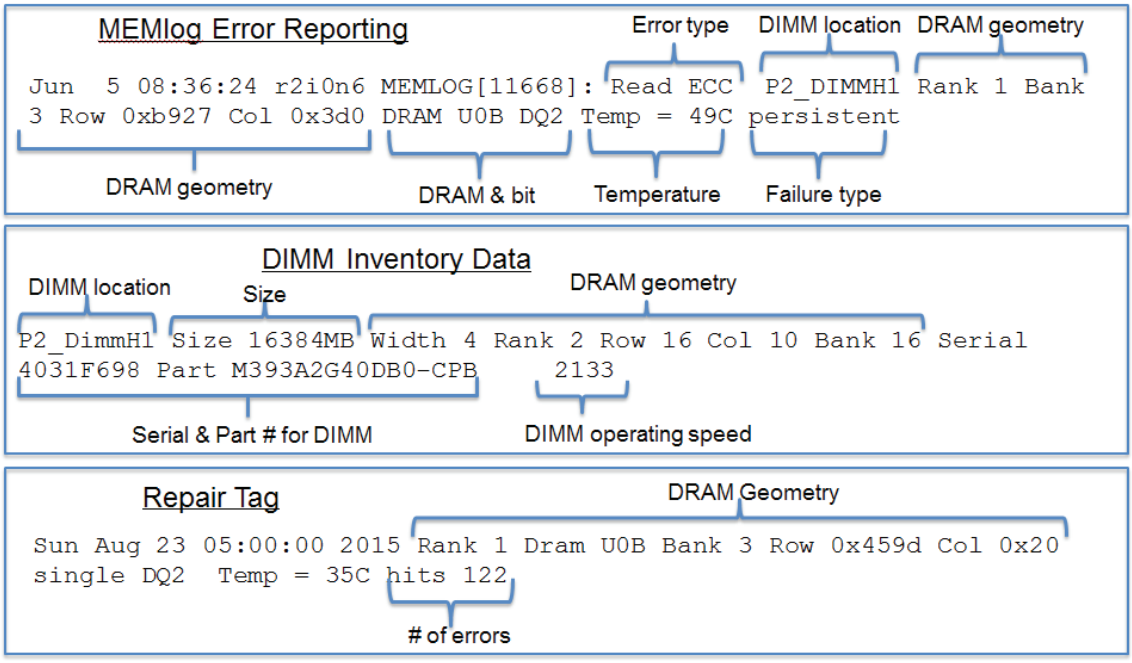

Figure 3. Sample MEMlog software output

A record of the failure history is maintained for each DIMM and is used to determine if a DIMM failure is the result of a single storage cell or a more widespread problem. Should a pattern of failures emerge, MEMlog issues an alert that the DIMM has reached a critical threshold and should be replaced.

4.2 Predictive Failure Analysis with Proactive Actions

Memory errors can occur with little warning and failures can cause unplanned downtime and lost productivity for production systems. SGI MEMlog software not only logs memory errors, it analyzes the error data and determines error resolution without system administrator intervention.

One of the key elements of MEMlog software is a ***proprietary transient error filter*** that identifies and resolves memory problems without triggering downtime. By analyzing logged memory data, MEMlog software is able to distinguish correctable transient errors from correctable hard failures and take action accordingly.

When the software detects a persistent DIMM failure, it passes memory error addresses to the kernel's page retirement interface to remove the affected page from use. Data is moved from the failing location to a good memory location without interrupting any running jobs, or requiring a system shutdown. A recent research paper, *Cosmic Rays Don't Strike Twice: Understanding the Nature of DRAM Errors and the Implications for System Design*[2], supports the idea that page retirement is an effective means of masking DRAM errors in production systems.

Because 4K memory pages are re-mapped at a logical level, the remapping is transparent and non-disruptive to running jobs. No bandwidth-robbing memory imbalances are created by this process, and memory hardware is never unnecessarily spared. The detailed failure analysis also makes possible predictable and accurate physical memory maintenance during scheduled maintenance windows.

Memory Error Management on Big Memory Systems

A large laboratory undertook a careful study of memory errors, noting that because of the amount of memory in its large systems, the total area of the memory circuitry exposed to cosmic radiation is much greater than the total area of circuitry in processors.

The customer uses MEMlog software on some SGI systems, and has noticed that the software reduces the DIMM replacement rate, keeping them from replacing DIMMs too early and at the same time identifying failing DIMMs more quickly.

[2] *Andy A. Hwang, Ioan Stefanovici, and Bianca Schroeder. Cosmic Rays Don't Strike Twice: Understanding the Nature of DRAM Errors and the Implications for SystemDesign. ASPLOS 2012. http://www.cs.toronto.edu/~bianca/papers/ASPLOS2012.pdf*

4.3 Alerts to Schedule Service

The SGI Remote Services tool provides 24x7 remote monitoring and data gathering for customer systems. This enables SGI to be proactive, resolving problems quickly and efficiently for improved uptime and system availability. On systems with the SGI Remote Services tool enabled, MEMlog software sends alerts to SGI Service first so that any necessary actions can be scheduled. If the SGI Remote Services tool is disabled, notifications are sent to the system administrator.

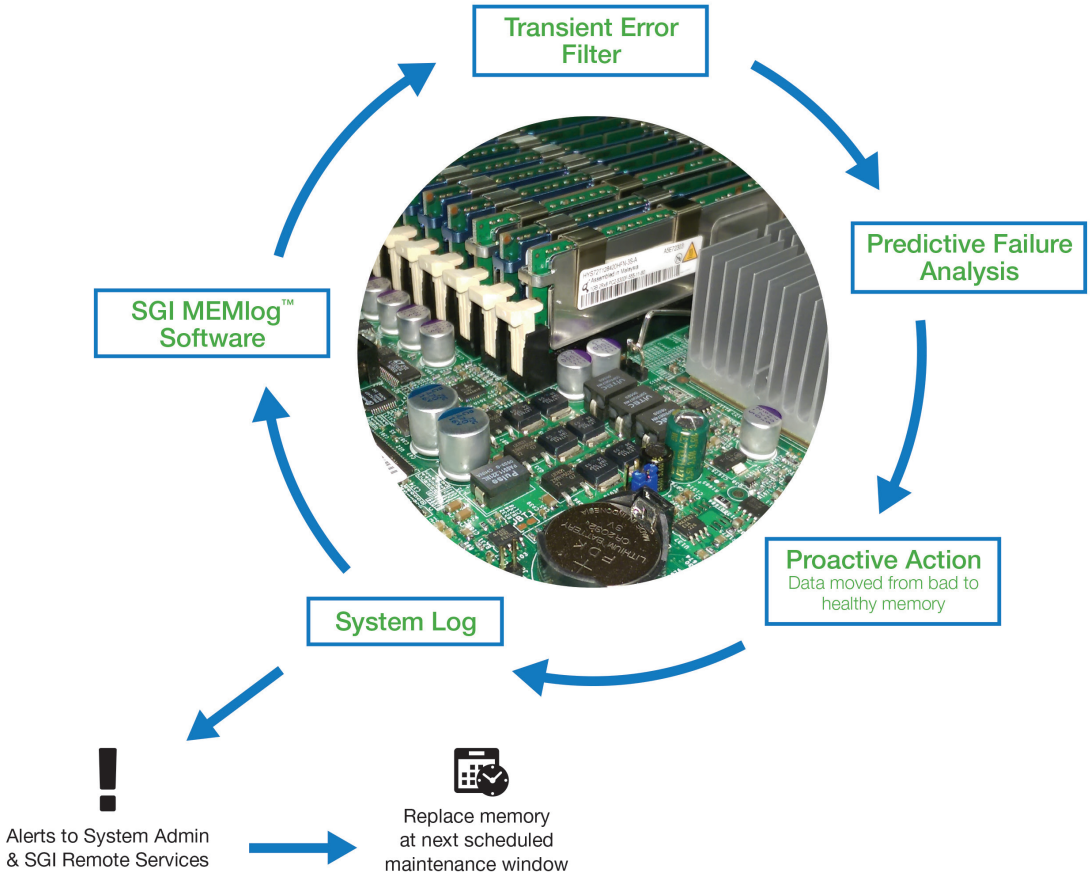

Figure 4. Functions of SGI MEMlog software

Many of today's datacenters have small teams who rely on proactive tools and remote services to get the job done. MEMlog software works in conjunction with SGI Management Center, which provides provisioning, health management, and other services for SGI platforms. Because SGI Management Center uses the open-source Nagios architecture to collect and analyze data, the environment is easily extensible to address unique requirements.

4.4 Advantages of SGI MEMlog Software

System vendors largely depend on the memory resiliency features provided by modern processors. SGI MEMlog software significantly enhances memory RAS above this baseline.

SGI MEMlog software is a software-centric solution that preserves maximum memory bandwidth with proactive resolution of memory errors while reducing unscheduled downtime.

In conjunction with SDDC+1:

- Transient errors are healed with no limits and no spares required.
- Hard failures are mitigated via page migration; data is moved to a healthy DIMM, and the page is retired until future replacement, avoiding the memory imbalances that occur with physical sparing.

The result is better memory performance with greatly reduced costs for sparing and replacing memory. By distinguishing transient errors from hard failures, MEMlog software reduces the need for unscheduled downtime for memory replacement, allowing administrators to adhere to downtime schedules.

4.5 Software Availability

SGI MEMlog runs on all SGI platforms, including SGI® Rackable™, SGI® ICE™, and SGI® UV™. Supported operating system are:

- SUSE® Linux Enterprise Server 11, 12
- Red Hat® Enterprise Linux 6, 7
- CentOS 6

MEMlog Influences Alabama Supercomputer Center Decision

"MEMlog played a key role in our decision to purchase an SGI UV system, as MEMlog addresses a stability concern for long running simulations."

David Young Ph.D., HPC Group Leader, Alabama Supercomputer Center

When the Alabama Supercomputer Center needed to reduce unplanned system interruptions due to memory failures, it turned to SGI. The team cites SGI MEMlog software as a key factor in its decision to choose SGI UV. Using the MEMlog error manager and logger, the center is able to accumulate a list of bad DIMMS for replacement and maintain its six-month maintenance windows.

5.0 Conclusion

Before the addition of the transient error filter to MEMlog memory error management, distinguishing between transient and hard memory failures required significant guesswork. Given the continued evolution of memory devices, the rapid increase in the number of deployed DIMMs in systems of all sizes, the increased rate of transient errors, and the potential impact of the row hammer effect, MEMlog software has quickly become a must have for managing memory errors. SGI customers have learned that MEMlog's transient error filtering is the only way to break the endless loop of memory replacement in big memory systems. Transient error filtering combined with the use of page retirement enables jobs to continue running with sustained memory bandwidth and without interruption, which improves user productivity.

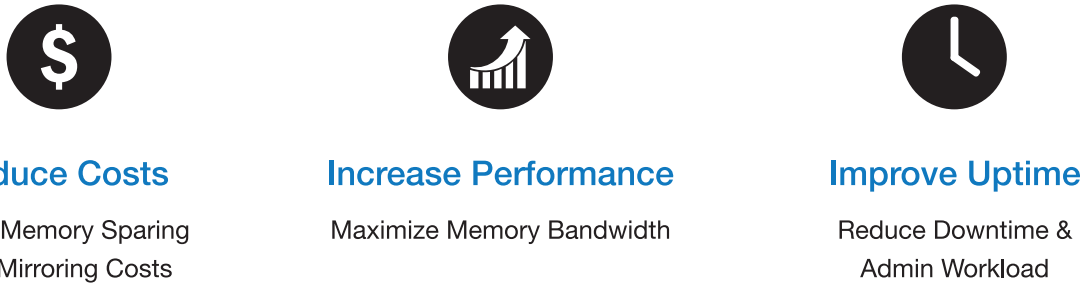

Figure 5. Advantages of deploying SGI MEMlog software

6.0 References

The Known Failure Mechanism in DDR3 Memory Referred to as Row Hammer, Barbara P. Aichinger, September 2014.

- http://www.ddrdetective.com/files/6414/1036/5710/The_Known_Failure_Mechanism_in_DDR3_memory_referred_to_as_Row_Hammer.pdf

  Rowhammer attack exploits shrinking process size in DRAM, Ryu Connor, March 2015.

- http://techreport.com/blog/27936/rowhammer-attack-exploits-shrinking-process-size-in-dram

  Cosmic Rays Don't Strike Twice: Understanding the Nature of DRAM Errors and the Implications for SystemDesign. Andy A. Hwang, Ioan Stefanovici, and Bianca Schroeder. ASPLOS 2012.

- http://www.cs.toronto.edu/~bianca/papers/ASPLOS2012.pdf

  Additional Row Hammer resources.

- http://www.ddrdetective.com/row-hammer/

Global Sales and Support: sgi.com