Soft Memory Errors and Their Effect on Sun Fire™ Systems

Sun Microsystems, Inc.
4150 Network Circle
Santa Clara, CA 95054 U.S.A.
650-960-1300

Part No. 816-5053-10
April 2002, Revision A

Send comments about this document to: docfeedback@sun.com

Please
Recycle

Soft Memory Errors and Their Effect On Sun Fire Systems

This document is intended to explain the behavior of Sun Fire™ systems in the presence of soft memory errors. It discusses what soft memory errors are and why they happen, the behavior of the Solaris™ Operating Environment, and when memory should be serviced in the presence of soft memory errors.

Semiconductor Memory Errors and ECC Concepts

Soft Errors

Any non-persistent storage device, whether it is a dynamic random access memory (DRAM) used for main memory or a static random access memory (SRAM) used for caches, is subject to occasional incidences of data loss due to the impact of energetic alpha particles or cosmic rays. This data loss manifests itself in the changing of the value stored in the memory location affected by the collision. Typically only a single bit (or memory cell) is affected, but there is a small probability (<10%) that multiple cells can be upset. When a bit flips due to this phenomenon, it is referred to as a soft error because the correct data can be written back to the cell and a subsequent read operation will produce the correct results. This distinguishes it from a hard error resulting from faulty hardware where further attempts to write and read the correct data will produce the same error. These soft errors happen at a rate, called the *soft error rate*, that can be predicted as a function of the memory density, the memory technology, and the geographic location of the memory system.

Alpha particle contamination has been virtually eliminated in DRAM devices. With the rare exception of package or process contamination, soft errors rates from alpha particles will not be observed. However, high energy and thermal neutrons are a

result of cosmic rays colliding with the earth's atmosphere and are difficult or impossible to shield. The rate of neutrons impacting DRAM memory is dependent on many environmental factors—solar cycle, the earth's magnetic field, altitude and building materials.

Generally speaking, cosmic ray soft errors occur in DRAM memory at a rate of ~10 to 100 FIT/MB (1 FIT = 1 device fail in 1 billion hours). So a system with 10 GB of memory should show an ECC event every 1,000 to 10,000 hours, and a system with 100 GB would show an event every 100 to 1,000 hours. However, this is a rough estimation that will change as a function of the effects outlined above.

ECC Detection and Correction

ECC was invented many years ago to facilitate the survival of these naturally occurring losses of data. The concept is that every addressable word of data stored in memory also has check information stored along with it. This combination of both data and check information is often called a check word.

The check information serves two purposes. First, when a check word is read out of memory, the check information can be used to detect if any of the bits of the data have changed. Additionally, the check information can be used to determine if just a single bit has changed or more than one bit has changed. This is known as ECC detection. Second, in the event that only a single bit has changed, the check information can be used to determine which bit in the data changed and therefore facilitate the correction of the data by flipping this bit back to its complimentary value. This is known as ECC correction.

When an ECC detection mechanism has detected that one or more bits in a word of data has changed, this is broadly categorized as an ECC error. These errors can be further categorized as a function of the number of bits in error. Because ECC can correct single bit flips, single bit errors are referred to as Correctable Errors (CEs). Multi-bit errors are referred to as Uncorrectable Errors (UEs). ECC detection and ECC correction mechanisms can be implemented by either hardware or software. Sun Fire systems ECC *detection* mechanisms are all implemented in hardware. Sun Fire systems ECC *correction* mechanisms have instances of both hardware and software correction.

Sun's memory designs use a scheme called interleaving—the bits in each word are composed of cell addresses that are physically separated or interleaved on the DRAM chip. Therefore, no two nearest neighbor memory cells that are upset will end up in the same check word. So *multi-cell* cosmic ray events can only create *single bit* errors that will be ECC correctable.

It is important to note that any bit flips that occur in memory are not detected until the affected word of data is read out of memory and presented to the ECC detection mechanism. These undetected errors are referred to as latent errors. It is also

important to note that, strictly speaking, ECC correction applies only to the copy of the affected word of data. The data as it resides in memory still contains the flipped bit. If this flipped bit in memory is not corrected as well, there is an exposure to another bit flipping in the same word of data. This will result in a UE, which would then result in loss of services. It is important, therefore, that the system provide functionality that goes back and corrects the flipped bit in memory. On Sun Fire systems, this functionality is provided by the Solaris software.

Hard Faults

Physical failures within a semiconductor, or its package, are called hard faults. Examples of hard faults are broken solder joints, defects in the transistor oxide, breaks in the metal interconnect lines, and so forth. If a hard fault affects just one bit of a checkword, it will only cause correctable ECC errors. These errors will be repeatable, though, thereby allowing for the diagnosis of a hard fault.

Weak Bits

There is a class of DRAM failures that belong in the *hard fault* category but display characteristics that are very similar to *soft errors*. They occur when there is a minor defect in the gate oxide or transistor junction of the memory cell. This defect leads to a very small leakage current that slowly “drains” the cell of its stored memory. Most of the writes and reads to this defective cell will proceed without error. However, under very special conditions (such as temperature and voltage variation), the read cycle will produce an error—hence the term *weak bit*. Most of the weak bits are screened at the DRAM vendors using an accelerated aging process called *burn-in*. This eliminates infant mortality from the DRAM population. But latent defects that are not caught at burn-in can grow in the DRAM over time and operation to create new weak bits. This failure rate is very low, but can be comparable to other hard fails.

System Noise and Test Escapes

Noise in the system due to poor / deteriorating connections or marginal system design can create errors in the write or read cycles to memory. This condition is extremely difficult to diagnose in the field. In addition, the DRAM manufacturers test the memory under various voltages, temperatures, and frequencies to ensure error free operation. However, there can be extremely complex conditions in customer applications that are not replicated in the DRAM vendor's tests or Sun's factory tests. This class of error is called *test escapes*. Once we can reproduce the particular customer conditions that creates this error, Sun works closely with the

DRAM vendors to implement new test screens that eliminate this source of failures. The DIMMs captured as test escapes and returned to the vendors for failure analysis play a key role in this ongoing quality improvement process.

Distinguishing Weak Bits from Soft Errors

Since soft errors from cosmic rays and weak bits from latent defects have very similar ECC signatures, it is important to develop discrimination techniques to distinguish the two. Due to the nature of cosmic rays, the observed soft errors should be randomly distributed in time and DIMM location. The one caveat is that soft errors can occur linearly in time but remain latent in memory until they are detected. So a system that is idle (or utilizing very little memory) will be accumulating soft errors. Then when a high memory utilization application is invoked (such as a scrub), latent soft errors will be detected. The appearance will be that the soft errors are correlated to specific periods of time. The reality is that the soft errors are *occurring* randomly but being *detected* at specific times.

Weak bits, system noise and test escapes on the other hand, are **not** random. They are device or location specific. If a certain DIMM is failing more often than its counterparts, this cannot be explained by cosmic rays. If a particular DIMM is failing more often than the others, it should be replaced.

Solaris Operating System Behavior

Handling of Correctable Errors

As discussed earlier, when a processor detects a CE as a result of a read to main memory, it will correct the incoming data and continue its operation. The error will be logged in the processor's asynchronous fault status register (AFSR) and the faulting physical address will be logged in the asynchronous fault address register (AFAR). The processor will then take a trap so that the error information can be logged.

As part of handling the error, the Solaris software will proceed to log a fair amount of diagnostic information. One such event log, taken from a Sun Fire 6800 system running Solaris 8 software, appears below:

```
Oct 25 09:06:25 wpc26 SUNW,UltraSPARC-III: [ID 796192 kern.notice]
NOTICE: [AFT0]
Corrected system bus (CE) Event on CPU18 at TL=0, errID
0x0000c9b9.19d92690
Oct 25 09:06:25 wpc26     AFSR 0x00000002<CE>.00000097 AFAR
0x00000001.04bdf7d0
Oct 25 09:06:25 wpc26     Fault_PC 0x10024a74 Esynd 0x0097 /N0/SB5/P3/
B0/D2 J16500
Oct 25 09:06:25 wpc26 SUNW,UltraSPARC-III: [ID 154767 kern.notice]
[AFT0] errID 0x0000c9b9.19d92690 Corrected Memory Error on /N0/SB5/P3/
B0/D2 J16500 is Persistent
Oct 25 09:06:25 wpc26 SUNW,UltraSPARC-III: [ID 682217 kern.notice]
[AFT0] errID 0x0000c9b9.19d92690 Data Bit 3 was in error and corrected
Oct 25 09:06:25 wpc26 SUNW,UltraSPARC-III: [ID 422650 kern.info]
[AFT2] errID 0x0000c9b9.19d92690 E$tag PA=0x00000000.00bdf7c0 does not
match AFAR=0x00000001.04bdf7c0
Oct 25 09:06:25 wpc26 SUNW,UltraSPARC-III: [ID 904800 kern.info]
[AFT2] errID 0x0000c9b9.19d92690 PA=0x00000000.00bdf7c0
Oct 25 09:06:25 wpc26     E$tag 0x00000000.01000001 E$state_7 Invalid
Oct 25 09:06:25 wpc26 SUNW,UltraSPARC-III: [ID 895151 kern.info]
[AFT2] E$Data (0x00) 0x5a8d0016.00000a20 0x20202020.37333231 ECC 0x128
Oct 25 09:06:25 wpc26 SUNW,UltraSPARC-III: [ID 895151 kern.info]
[AFT2] E$Data (0x10) 0x39062c00.5a8d0010 0x00000a20.20202020 ECC 0x03d
Oct 25 09:06:25 wpc26 SUNW,UltraSPARC-III: [ID 895151 kern.info]
[AFT2] E$Data (0x20) 0x37333330.32062c00 0x5a8f000c.00000a20 ECC 0x1f6
Oct 25 09:06:25 wpc26 SUNW,UltraSPARC-III: [ID 895151 kern.info]
[AFT2] E$Data (0x30) 0x20202020.37333330 0x34062c00.5a8f000d ECC 0x1fc
Oct 25 09:06:25 wpc26 SUNW,UltraSPARC-III: [ID 929717 kern.info]
[AFT2] D$ data not available
Oct 25 09:06:25 wpc26 SUNW,UltraSPARC-III: [ID 335345 kern.info]
[AFT2] I$ data not available
```

It is important to recognize that all of the above output is the result of one single CE event. Each of the messages is tagged with an asynchronous fault tag (AFT) to identify the data being logged. Continuation messages begin with four spaces. The different AFT tag values are:

- AFT0 is used for correctable errors.
- AFT1 is used for uncorrectable errors as well as for errors that result in a panic.
- AFT2 and AFT3 are used for logging diagnostic data and other error related messaging.

The extracts below were taken from the previous example:

```
[AFT0]Corrected system bus (CE) Event on CPU18 at TL=0, errID
0x0000c9b9.19d92690

AFSR 0x00000002<CE>.00000097 AFAR 0x00000001.04bdf7d0

Fault_PC 0x10024a74 Esynd 0x0097 /N0/SB5/P3/B0/D2 J16500

[AFT0] errID0x0000c9b9.19d92690 Corrected Memory Error on /N0/SB5/P3/
B0/D2 J16500 is Persistent

[AFT0] errID0x0000c9b9.19d92690 Data Bit 3 was in error and corrected
```

- `errID` is a timestamp of the event. This is very useful for correlating multiple errors that occurred at the same time.
- `AFSR` and `AFAR` are the asynchronous fault status and address registers.
- `Fault_PC` is the value of the PC at the time of the fault and is dependent upon the fault type as to whether the value is valid.
- `Esynd` is the ECC syndrome captured.
- `/N0/SB5/P3/B0/D2` is the identifier of the memory module which corresponds to the faulting address.
- `J16500` is the J number for that memory module.
- Solaris software describes this event as *Persistent*. The Solaris software error handling code provides a disposition code as a result of the scrub operation. This disposition is one of Intermittent, Persistent, or Sticky. The definition of each of these codes is:
  - Intermittent means the error was not detected on a reread of the affected memory location.
  - Persistent means the error was detected again on a reread of the affected memory location but the scrub operation corrected it.

- Sticky means that the error still exists in memory even after the scrub operation. These events should be investigated further to determine if some hardware replacement is necessary since this is indicative of a hard failure.

The error described can be categorized as a correctable memory error on memory module /N0/SB5/P3/B0/D2. The scrub operation successfully cleared the error in memory.

It is worth noting that as of the Solaris 8 KU-9 release, all Sun Fire and Ultra Enterprise™ system platforms log all correctable memory errors to both the console and error log. Prior to Solaris 8 KU-9 release, Ultra Enterprise mid-range systems would not log correctable errors unless a single DIMM experienced more than five errors. But, the Sun Enterprise 10000 has always logged all correctable errors.

Scrubbing

The Solaris software includes a memory scrubber. The purpose of the scrubber is to traverse all of physical memory, as seen by the domain, to reduce the likelihood that multiple transient errors will lead to an uncorrectable memory error.

The scrubber is implemented as a kernel thread, which periodically wakes up and traverses a portion of physical memory. The scrubber is setup so that it will traverse all of physical memory within 12 hours. To be as unobtrusive as possible, the scrubber will read 8 MBs of pages. The read operation is accomplished with the block load hardware to maximize read bandwidth and thus reduce the time needed to read a span.

Any correctable errors encountered during the read of the span are handled by the Solaris software as described above. The read operation is also done under kernel protection. This is done to avoid the potential of panicking on an uncorrectable error. Normally, an uncorrectable error encountered by the kernel will result in a panic.

Improvements in Messaging

With the Sun Fire product line, the Solaris software takes a conservative approach and logs all errors. This was done to fully characterize the error event when it occurs as well as to provide a history of events. For more sophisticated users, this can be too much information to parse.

The Sun Fire Solaris team is investigating alternatives that will provide more discrimination in messaging, the ability to apply advanced algorithms to error histories to determine acceptable versus unacceptable situations, and provide better correlation of errors to the field-replaceable unit that is in error.

Servicing Memory in the Presence of Soft Errors

Errors Categorized as Persistent

As indicated in the example in the previous section, the Solaris software uses the term Persistent to label correctable ECC events that are consistent with naturally occurring cosmic ray effects. Determining that a series of correctable ECC events are due to a defective DIMM and not simply a cosmic ray effect is an important question, because unnecessary changing of DIMMs due to cosmic ray soft errors will **decrease** the reliability of the system due to excessive handling and human intervention. While the long term plan is to enhance the Solaris environment to make service recommendations based on the internal tracking of event rates, it is currently necessary for Sun's service organization to follow guidelines with respect to servicing memory DIMMs. As of the writing of this document, the current published guidelines indicate that a single DIMM must experience three correctable ECC events, labeled Persistent, in a 24-hour period before service should be performed. If a DIMM experiences correctable ECC events with any less frequency, it is almost certainly due to cosmic ray events and therefore does not warrant replacement.

Errors Categorized as Sticky

If, during the handling of a CE event, the Solaris software cannot correct the flipped data bit in memory, it will label the event Sticky instead of Persistent. This disposition is fundamentally different in that it is not consistent with naturally occurring events. Rather, it is an indication of a hard fault. **Any components that have been diagnosed with a hard fault should be replaced as soon as possible.**