UCS Enhanced Memory Error Management

Tech Note

Updated: February 2015

Contents

Introduction

Ongoing trends within the computer industry have resulted in increasing rates of memory errors in servers. As a result, strong error correcting codes (ECC) are employed to prevent uncorrected errors that can crash a system, and sophisticated algorithms are needed for dealing with correctable errors to minimize their impact on server uptime.

This paper provides an overview of memory errors, why trends in server memory systems lead to increases in memory errors, and how Cisco UCS servers are well equipped to address memory errors.

Background on Memory Errors

Memory errors are encountered when an attempt is made to read a memory location. The value read from the memory does not match the value that is supposed to be there.

Classification of Memory Errors

Detected vs. Undetected Errors

In a system without ECC memory, there is no hardware error detection. Hence, memory errors will lead to silent data corruption, incorrect execution of operating system or application, and eventually system crashes. Cisco's UCS Servers use ECC memory. Therefore, powerful error correcting codes such as those provided by the Intel Xeon processors in UCS servers can detect memory errors so that silent data corruption does not occur.

Hard vs. Soft Errors

Errors that are caused by a persistent physical defect are traditionally referred to as "hard" errors. A hard error may be caused by an assembly defect like a solder bridge or cracked solder joint, or may be due to a defect in the memory chip itself. Rewriting the memory location and retrying the read access will not eliminate a hard error. This error will continue to repeat.

Errors caused by a brief electrical disturbance, either inside the DRAM chip, or on an external interface, are referred to as "soft" errors. Soft errors are transient and do not continue to repeat. If the soft error was due to a disturbance during the read operation, then simply retrying the read may yield correct data. If the soft error was due to a disturbance that upset the contents of the memory array, then rewriting the memory location will correct the error.

Hard errors are typically detected by memory tests run by the UCS BIOS at boot time, and any DIMMs containing hard errors are mapped out so that they cannot cause errors during runtime. UCS servers employ memory patrol scrubbing to automatically detect and correct soft errors during runtime.

Correctable vs. Uncorrectable Errors

Whether a particular error is correctable or uncorrectable depends on the strength of the ECC code employed within the memory system. Dedicated hardware is able to fix correctable errors when they occur with no impact on program execution. Uncorrectable errors generally cannot be fixed, and may make it impossible for the application or operating system to continue execution.

Error Correcting Codes

Traditional "SECDED" Error Correcting Codes

ECC codes on memory systems are traditionally applied across 64 bit (8-byte) data words protected by 8 check bits, to form a 72-bit code word. Such Single Error Correct, Double Error Detect (SECDED) ECC codes could correct any single bit error, and detect any double bit error. Through the use of Intel's Xeon processors, Cisco's UCS systems enhance the traditional SECDED features with more advanced error correction mechanisms such as those listed below.

UCS Error Detection and Correction

UCS servers built from Intel Xeon "EP" class processors employ ECC codes that not only correct any single bit error, but can also correct any number of errors that are confined to a single x4 DRAM chip, and detect errors in up to 2 devices. This capability is known as Single Device Data Correction (SDDC). Additionally, when operating in lockstep mode, which spreads the ECC code words across a pair of memory channels, SDDC is extended to correct errors in any x8 bit DRAM chip (or adjacent pair of x4 DRAM chips). To provide even higher levels of reliability and availability, UCS servers built from Xeon "EX" class processors can correct errors in any (not necessarily adjacent) pair of x4 devices, and detect errors in up to 3 devices. This capability is known as Double Device Data Correction (DDDC).

Trends in Server Memory Systems

Memory systems are a key area of ongoing innovation in servers. One trend is larger memory system capacity as this improves application performance by reducing the time spent waiting for slower disk accesses. In addition, increasing bandwidth is another trend, because high memory system bandwidth improves application performance through faster access to instructions and data needed by high core count processors. Also, as memory systems grow, operating voltages have been reduced to support denser, faster designs while improving power efficiency.

Increasing Capacity

A primary driver of increased error rates is the fact that memory systems are rapidly getting larger. As more and more bits of memory are added to the system, the likelihood of any one of them encountering an error increases. Such increases in system memory capacities are due to shrinking DRAM geometries (i.e. the ability to pack more bits on a single die). Since 2008, DRAM capacities have increased 16x from 512Mbit to 8Gbit. As chip capacity has increased, individual bit cells have been getting smaller. As the bit cell gets smaller, the number of stored charges per bit decreases, making it more difficult to distinguish between a stored "1" and "0". The basic storage element, or bit cell, in a DRAM chip is a tiny capacitor. DRAM bit cells are inherently leaky; thus, smaller bit cells storing fewer chargers are less tolerant of this leakage. Additionally, smaller bit cells are more easily upset by external sources like alpha particles or cosmic rays.

Today's advanced DRAM technologies deliver up to 8 Gbits of memory on a single die, and up to 64 GBytes of memory on a single memory module (DIMM). In addition, today's CPUs incorporate multiple memory channels on each processor socket, and multiple DIMMs on each channel.

Increasing Bandwidth

Memory system bandwidth has also been increasing steadily. In addition to the multiple memory channels on each processor socket, the speed of those channels has increased. Just a few years ago the top speed for DDR2 memory interfaces was 800 Mtps. Using advanced DDR4 memory, Cisco's B200-M4 supports memory channels operating at 2133 Mtps. Ever increasing operating frequencies, while providing higher bandwidth, also result in smaller bit times. As individual bit times decrease, timing margin also decreases, making it more difficult for receiving circuitry to separate each bit from those that precede and follow it.

Lower Operating Voltages

Another underlying technology trend is an ongoing reduction in operating voltages to lower power and cooling requirements, and to accommodate the smaller transistors associated with advances in process technology. DRAM voltages have decreased over the years, going from 2.5V to 1.8V to 1.5V to 1.35V to 1.2V as the industry has shifted from DDR to DDR2 to DDR3 and DDR4. As the operating voltages decrease, the available noise margin also decreases, making it more difficult for receivers and sense amps to distinguish between a "1" and a "0".

Why Error Rates are Increasing

As explained previously, increased error rates are attributed to multiple factors:

- Larger memory systems contain more bits
- Higher capacity DRAM chips require smaller bit cells which result in fewer stored charges per bit
- Lower operating voltages can lead to reduced noise margin
- Higher operating speeds can lead to reduced timing margin

UCS Memory Error Management

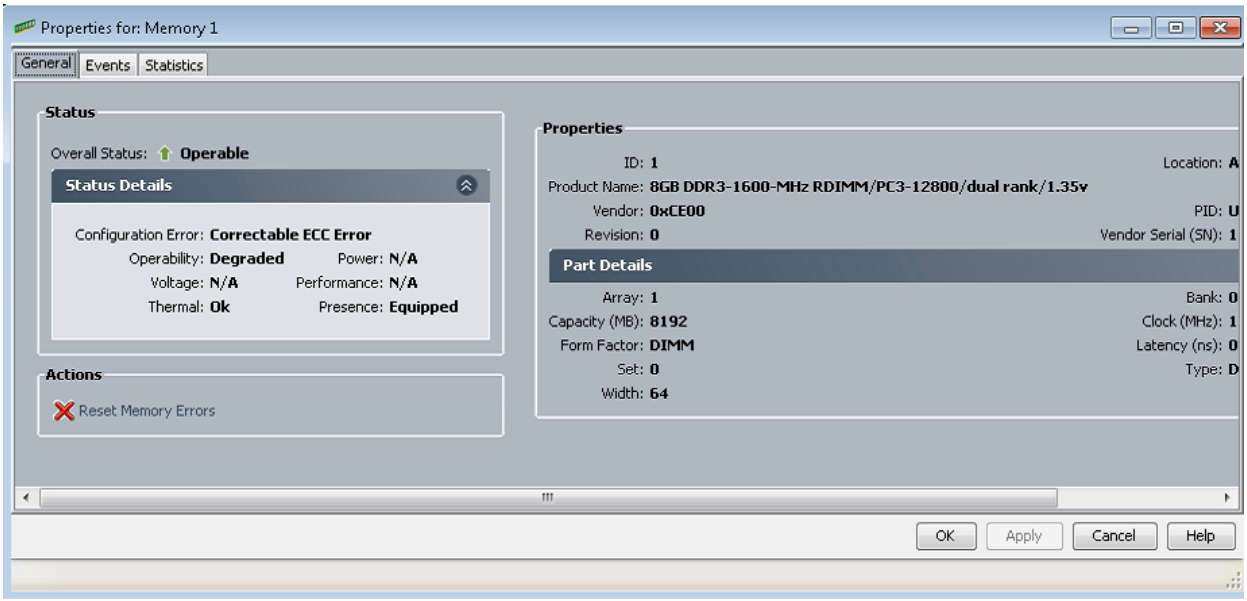
Error Reporting Mechanisms

When correctable memory errors occur, they are counted in registers which can be accessed by the Cisco Integrated Management Controller (CIMC). The CIMC tracks these errors and uses an IPMI sensor for each DIMM to enforce a failure threshold. When new correctable errors are detected, the CIMC will generate an entry in the System Event Log (SEL) indicating which DIMM encountered the error.

If a DIMM exceeds the pre-defined correctable ECC threshold, a subsequent entry will be made in the SEL indicating the threshold was exceeded. Once a DIMM crosses the threshold, it is marked as "Degraded" and will remain in this state until memory errors are manually reset, or the DIMM is replaced. Even though the DIMM is marked as "Degraded", the overall DIMM status is "Operable" and the DIMM will continue to operate as normal with no impact on performance (see Figure 1). The "Degraded" state is to serve as a notification that further investigation should be done to see why the DIMM crossed the threshold.

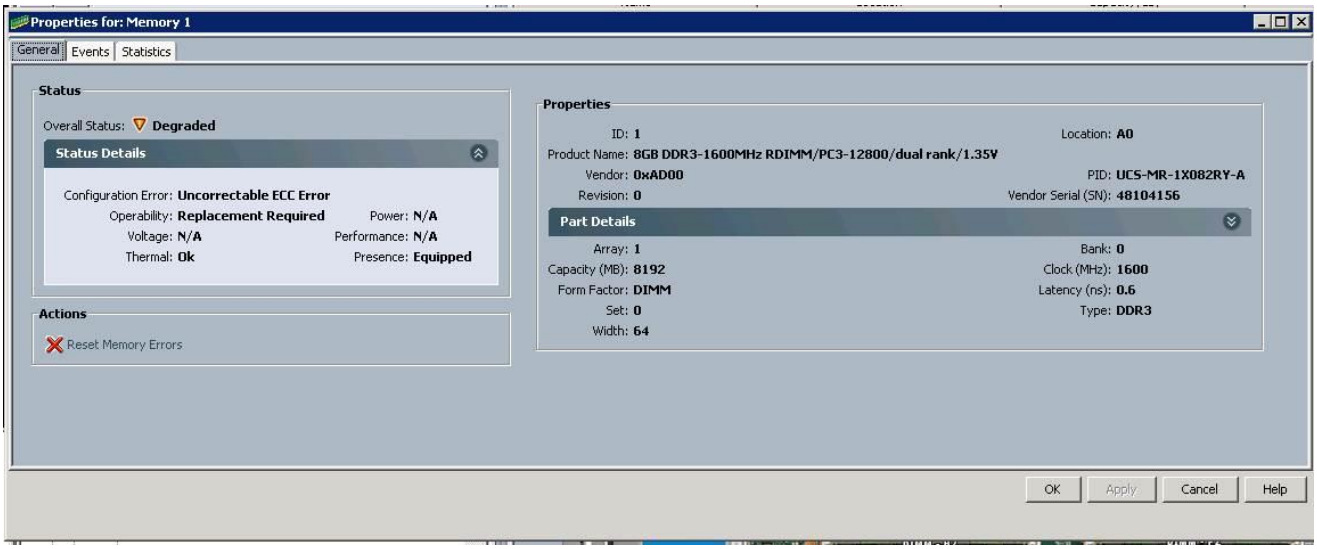
As a result of these SEL events, UCSM will trigger a series of faults pertaining to the individual DIMM state, the Server Health LED Status, and the Overall Server Operability. For more details on different ways to view memory error statistics, please refer to Appendix A.

Figure 1: For correctable ECC errors, DIMM overall status is "Operable" and continues to function with no impact on performance



Uncorrectable memory errors are reported by the BIOS which creates an entry in the CIMC SEL. A single uncorrectable error will result in a fault for the DIMM and the server, both indicating an overall status of "Degraded". The status details for an individual memory module reveal that a DIMM requires replacement when it encounters an uncorrectable ECC error (see Figure 3).

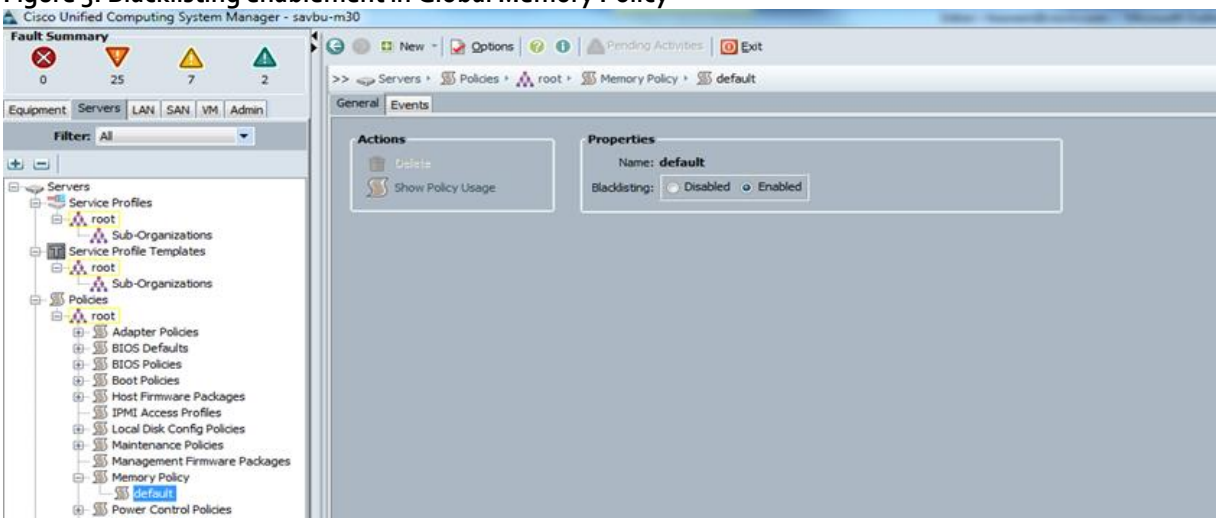
Figure 2: Uncorrectable ECC errors require replacement of the module; thus, DIMM overall status is reported as "Degraded"



DIMM Blacklisting

UCSM version 2.2 introduced the opt-in feature DIMM Blacklisting to help prevent repeat uncorrectable memory errors. When using this feature UCSM will "blacklist" DIMMs which have encountered uncorrectable errors during OS runtime. This mechanism prevents repeated crashes due to additional uncorrectable errors on the same DIMM before troubleshooting or corrective maintenance can occur. DIMM Blacklisting can be enabled in the default Global Memory Policy (see Figure 3).

Figure 3: Blacklisting enablement in Global Memory Policy



Enhanced Memory Error Management

As memory error rates have increased, false positives of DIMM replacements due to correctable errors have also increased. Replacing these DIMMs can be expensive, time consuming, and results in unwanted system downtime. In many cases, the DIMMs that were replaced had encountered soft errors, and when analyzed, resulted in a diagnosis of "No Trouble Found". Those DIMMs could have continued to operate in the system without increasing the likelihood of an uncorrectable error.

In response to those false positives, Cisco has developed an Enhanced Memory Error Management algorithm that detects those DIMMs that put the system at increased risk of encountering an uncorrectable error, and recommends replacement of only those DIMMs. The algorithm filters out DIMMs which encounter some correctable errors, but which present a negligible risk of causing an uncorrectable error. The algorithm takes into account the difference between hard and soft errors. Additionally, the sophisticated algorithm factors in the robust ECC code that can correct 4-bit symbol errors, not just single-bit errors, and the automatic patrol scrubbing that takes place in the background. It has been validated by extensive data collection and analysis of the correlation between correctable errors and uncorrectable errors in Cisco's own data centers and elsewhere.

The result of employing Cisco's Enhanced Memory Error Management is reduced operating costs and improved system availability due to fewer unnecessary DIMM replacements.

Software Supported

Support for Enhanced Memory Error Management is available on the below releases:

UCS Manager Patch Release 2.2(1b) and newer

UCS Manager Patch Release 2.1(3c) and newer

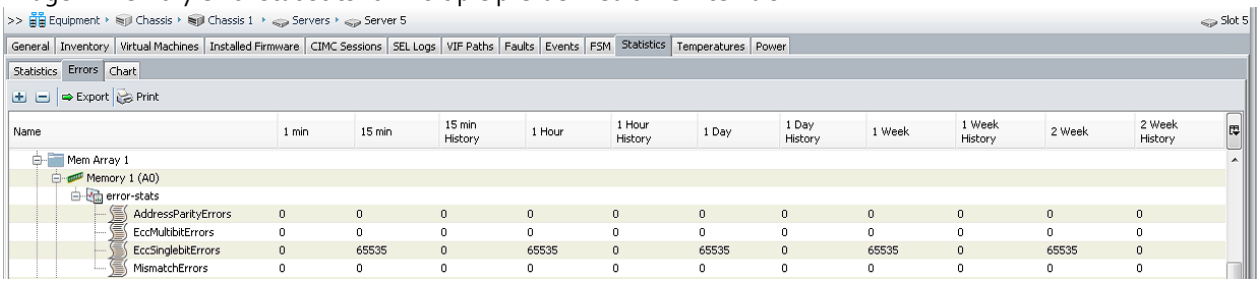
Conclusion

In brief, industry demands of increase capacity, increase bandwidth, and lower operating voltages leads to increase memory error rates. Extensive data analysis of this trend has resulted in Cisco's Enhanced Memory Error Management to refine error detection and protection for UCS Servers. With Enhanced Memory Error Management customers can experience a reduction in false positive error reports, reduction in unnecessary DIMM replacements, increase server uptime, and a reduction in overall server maintenance.

Appendix A: Additional Memory Error Reporting

Memory error statistics can be found for each individual DIMM within UCSM. UCSM presents these errors under each physical server for various time intervals (see Image 1).

Image 1: Memory error statistics for multiple pre-defined time intervals.



In addition to the UCSM graphical interface, memory errors statistics and details can also be checked via Command Line Interface (CLI).

```
Pod-1-UCS-6-B # scope server 1/6
Pod-1-UCS-6-B /chassis/server # scope memory-array 1
Pod-1-UCS-6-B /chassis/server/memory-array # show stats

Memory Error Stats:
    Time Collected: 2014-04-16T16:36:58.968
    Monitored Object: sys/chassis-1/blade-6/board/memarray-1/mem-1/error-stats
    Suspect: No
    Address Parity Errors: 0
    Mismatch Errors: 0
    Ecc Multibit Errors: 0
    Ecc Singlebit Errors: 65535
    Thresholded: 0

    Time Collected: 2014-04-16T16:36:58.968
    Monitored Object: sys/chassis-1/blade-6/board/memarray-1/mem-2/error-stats
    Suspect: No
    Address Parity Errors: 0
    Mismatch Errors: 0
    Ecc Multibit Errors: 0
    Ecc Singlebit Errors: 65535
    Thresholded: 0
```

As a result of SEL events, UCSM will trigger faults pertaining to the individual DIMM state, the Server Health LED Status, and the Overall Server Operability. These faults may be different depending on which UCSM release is being used. For UCSM patch release 2.2(3a) and newer, refer to Image 2 and 3 for correctable error faults and Image 4 and 5 for uncorrectable error faults.

Image 2: The Server Health LED is "Normal" and Overall Status is "Ok" with a correctable error.

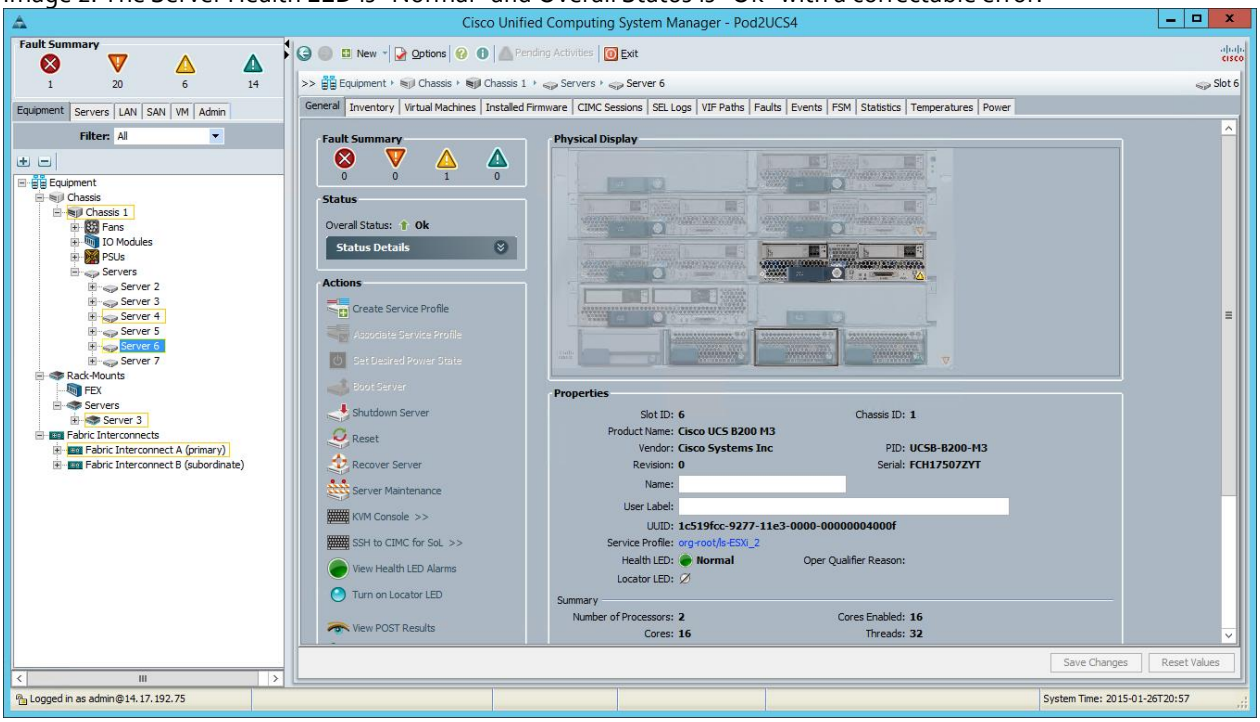


Image 3: The "Faults" tab under each server has a detailed view of the faults currently impacting the server. The "Minor" fault was generated when the correctable ECC error threshold was crossed.

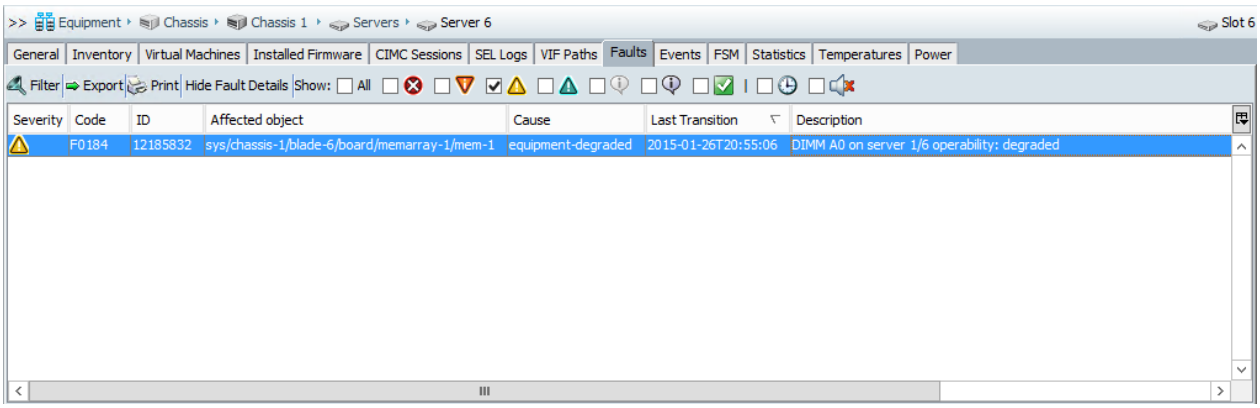


Image 4: Server Overall Status is "Degraded" as a result of an uncorrectable error. Server Health LED is set to "Minor" as a result.

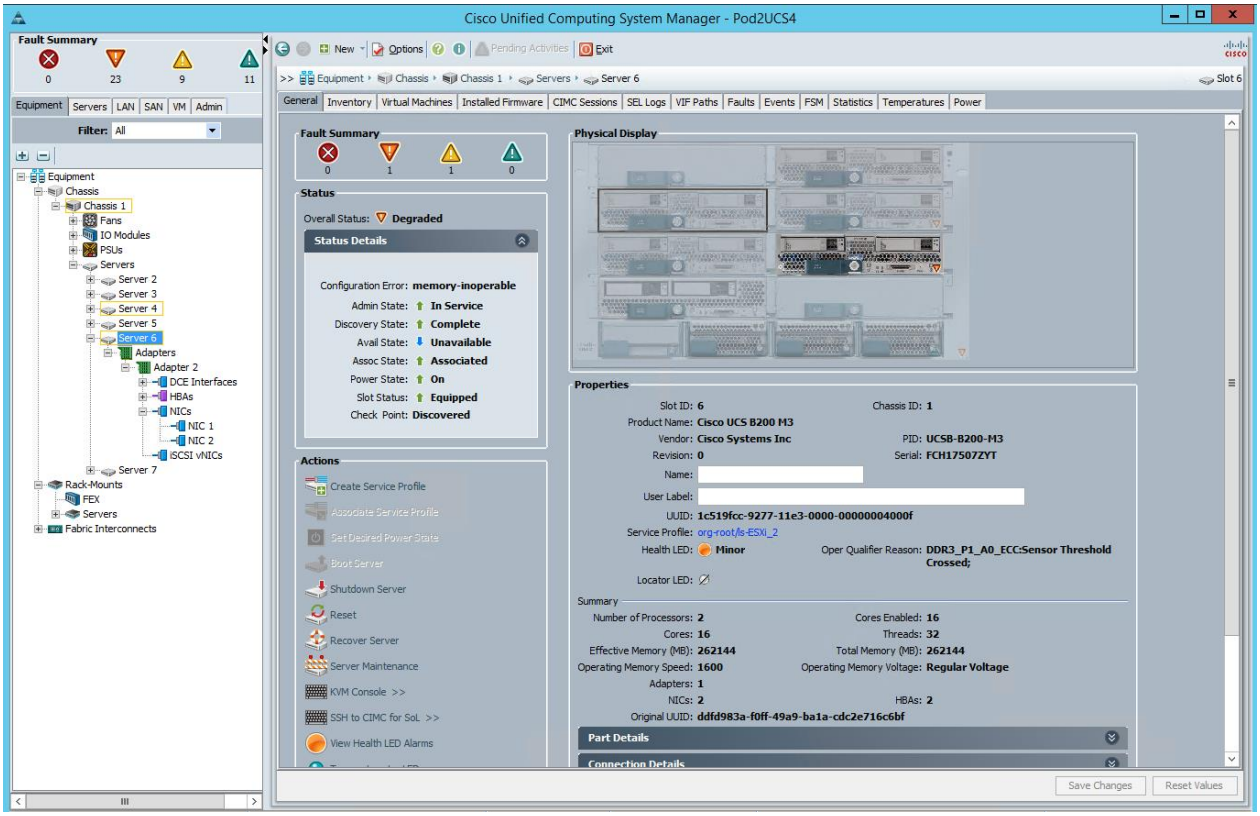
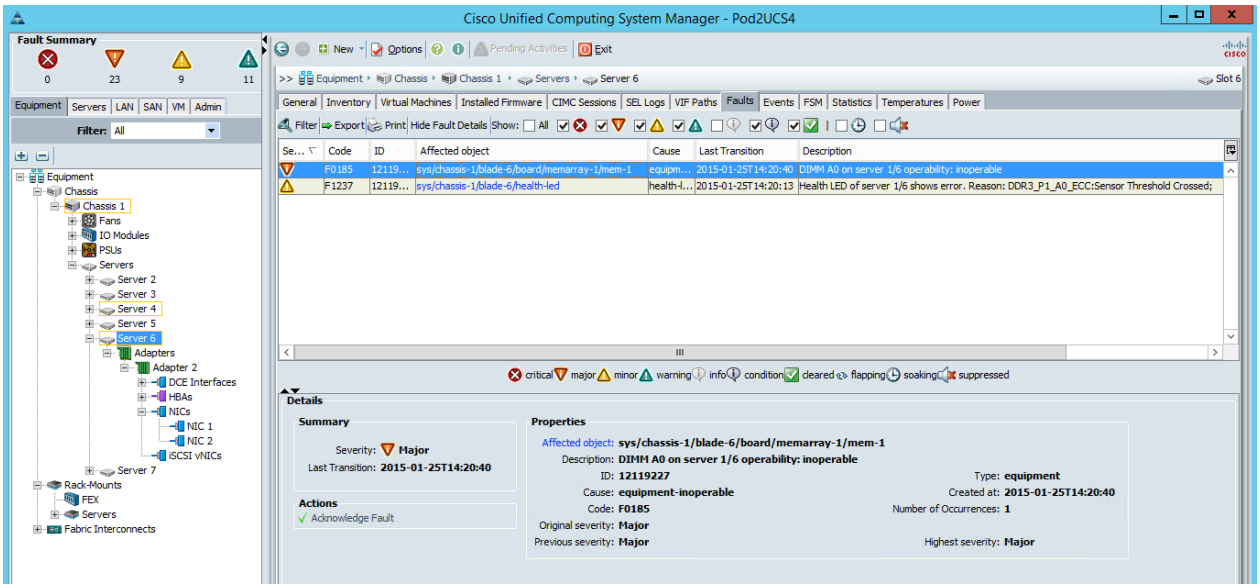


Image 5: The "Faults" tab under each server has a detailed view of the faults currently impacting the server. "Major" and "Minor" faults were generated after an uncorrectable error.



For UCSM patch releases older than 2.2(3a), refer to Image 6 and 7 for correctable error faults and Image 8 and 9 for uncorrectable error faults.

Image 6: Server Overall Status is "Ok" with a correctable error.

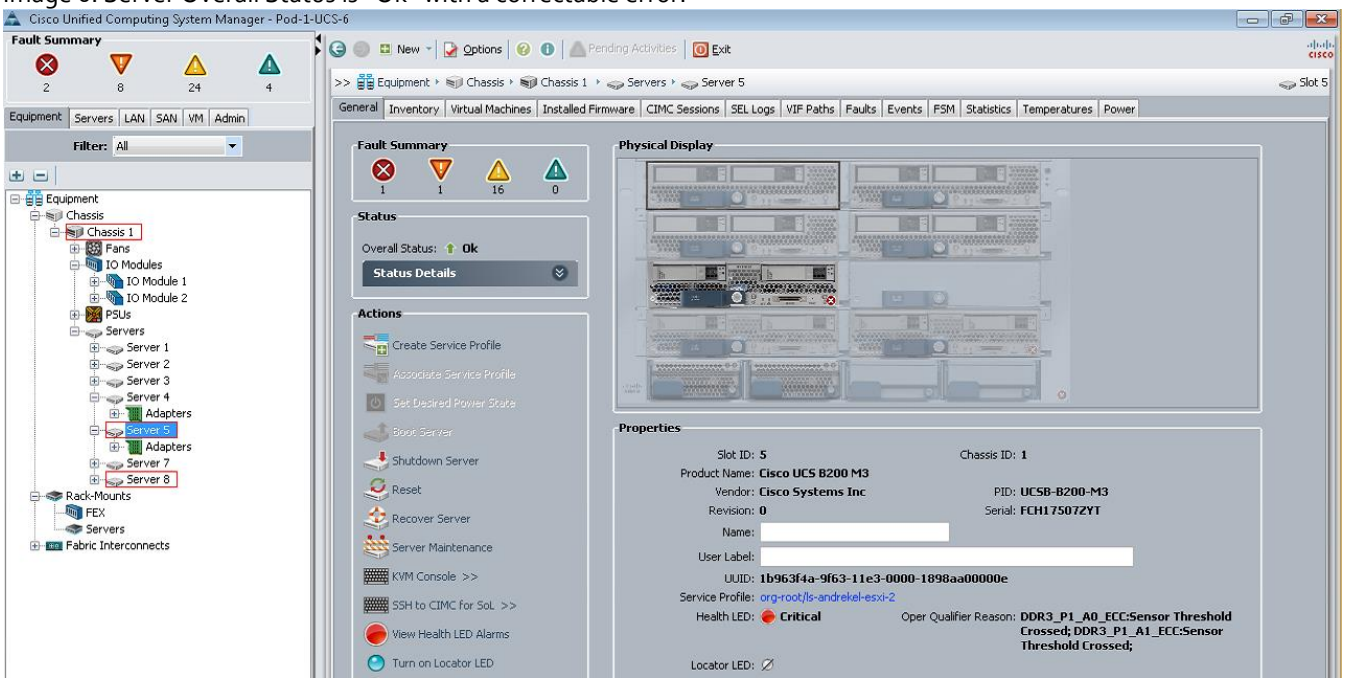


Image 7: The "Faults" tab under each physical server has a detailed view of the faults currently impacting the server. Shown here is a fault for a correctable error.

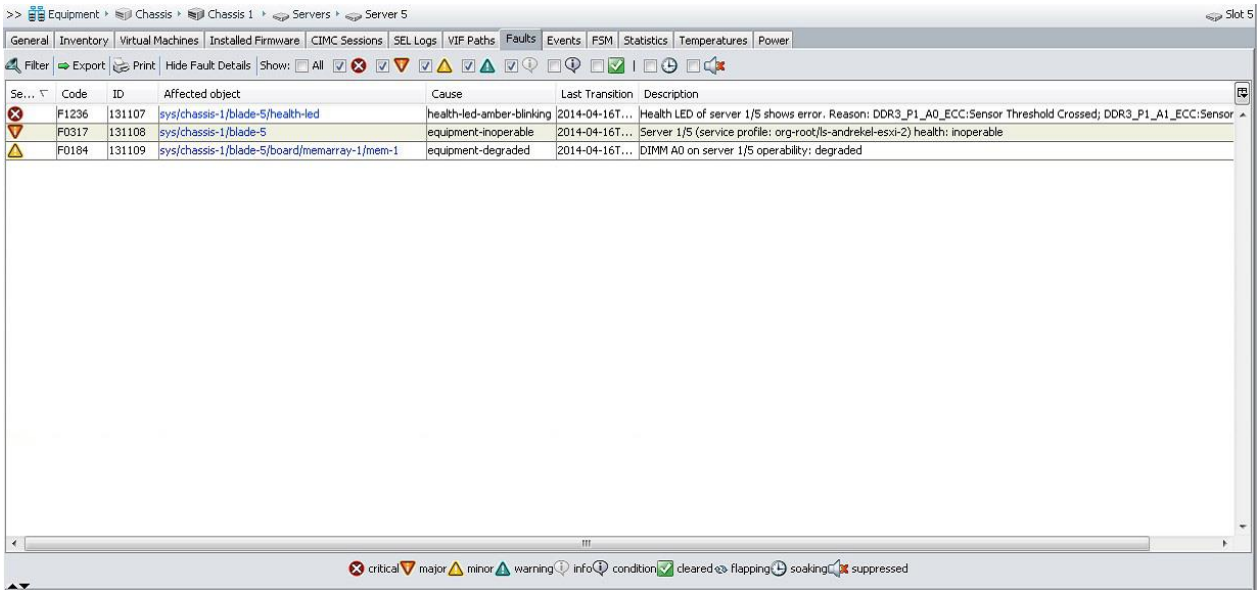


Image 8: Server Overall Status is "Degraded" as a result of an uncorrectable error.

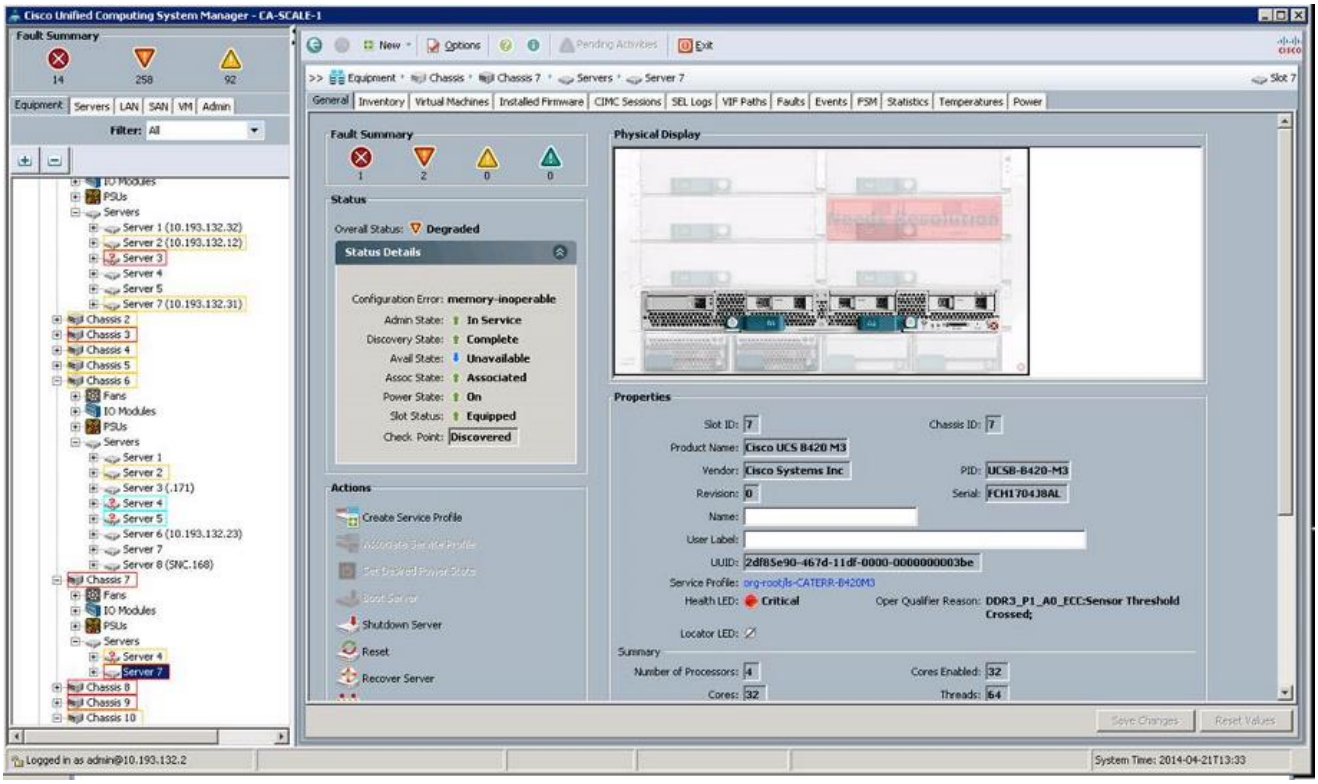
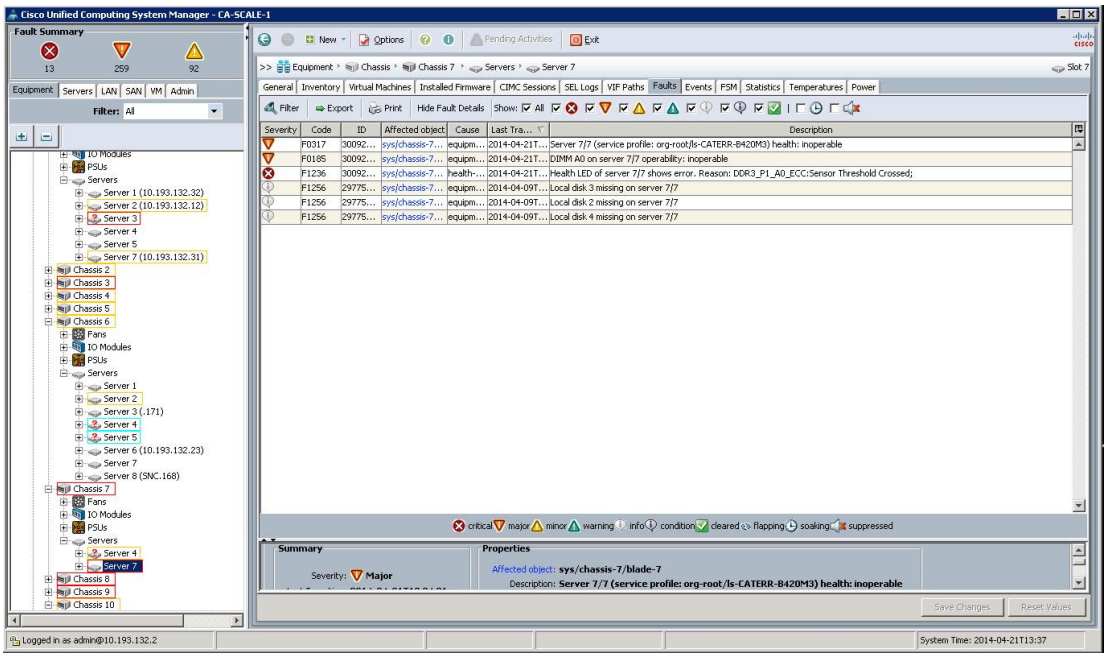


Image 9: The "Faults" tab under each physical server has a detailed view of the faults currently impacting the server. Shown here is a fault for an uncorrectable error.



Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**