CS6400 Reliability, Availability, and Serviceability (RAS) Features

S. Sridhar, CRI Technical Support Group, San Diego, CA

ABSTRACT: *The Cray CS6400's Reliability, Availability and Serviceability (RAS) features are unique among UNIX enterprise servers. The RAS features of CS6400 are described in this paper. Presented first in the paper is an overview of RAS, including definitions, importance and levels of RAS. Next, the goals of CS6400's RAS design and the strategy for achieving the goals are given. This is followed by details of CS6400's implementation of RAS features. Two of the key RAS features, dual pathing for I/O and system board replacement in a running system are then described. Next, typical system interrupt scenarios are compared. The paper concludes with a summary section.*

1 Introduction

The CRAY CS6400 system is a scalable symmetric multiprocessing (SMP) system designed and tuned for commercial and technical production environments. The CS6400 runs in a UNIX (SVR4) environment and is based on SPARC/Solaris technology - it runs all software developed for Solaris without recompilation.

Primary application areas for the CS6400 system include decision support, data warehousing, online transaction processing (OLTP), and text retrieval. In industries such as financial services, telecommunications, retail, and transportation, leading organizations are selecting the CS6400 system to support their enterprise-level business operations.

A key issue within most industries is system availability - or uptime - a critical requirement for many environments. The CS6400 system has been designed for superior uptime with advanced product features providing industry-leading reliability, availability, and serviceability (RAS). Together these features make the CS6400 system ideally suited for commercial and technical environments requiring high levels of availability in equal measure with unsurpassed performance.

This document describes some of the key issues organizations must consider when balancing the need for continuous system operation against a degree of acceptable risk. It discusses the implementation of the RAS features in the CS6400 system to maintain the highest productivity levels on critical shared resources.

2 Overview of RAS

2.1 Definitions

Reliability is simply the ability of system components to to function without failure over long periods of time. The overall system reliability is derived from a number of factors: care taken in the design process, attention paid to faithfully executing the hardware and software design, selection of high-quality components, and maintaining high quality standards during the manufacturing process; for example, electrostatic discharge protection, quality standards, and clean rooms.

Availability, also referred to as uptime, is the amount of time a system, and its users, can work productively. It is a critical concern of customers because of the fact that productivity in a computer-based business is very directly impacted by availability (also called uptime). The overall system reliability is a function of many factors: use of reliable and redundant hardware, use of proven application software, and the ability to recover quickly from any unscheduled interruptions.

Serviceability addresses issues such as the ability to detect and recover from error conditions, isolation and replacement of failed hardware, and upgrading system hardware and software. The goal of serviceability is to minimize the amount of time the system is unavailable because of scheduled and unscheduled interruptions.

Reliability, availability, and serviceability are interrelated features. The following formula expresses the relationship between the three RAS factors:

AVAILABILITY
= [[reliability - repair time] * 100] / reliability

The reliability of a computer system is often called the mean time to interrupt (MTTI), and is influenced by three types of interruptions:

- Solid failures that eventually require hands-on service of the machine. The time between these failures is the hardware mean time between failures (MTBF).
- Random transient failures, which are hardware-related and can be difficult to trace.
- Software failures are most often seen in development systems, where the environment is continually changing. They are less common in robust production environments.

The repair time for a computer system is expressed as the mean time to repair (MTTR), which is a function of the following factors:

- The system hardware and software design which impact the amount of time it takes to revive a down system.
- The board and parts sparing strategy and response time.
- Remote diagnosis and support procedures.

Rewritten for MTTI and MTTR, the availability equation can be expressed as:

AVAILABILITY = [[MTTI - MTTR] * 100] / MTTI

2.2 Levels of RAS

The SMP UNIX-based servers can be broadly classified into 4 groups, based on the percentage of time that the system is available for use, i.e., the "level" of RAS, as follows:

Fault Tolerant.....greater than 99.999%
High Availability with Failover .....99.990 to 99.995%
High Availability.....99.90 to 99.95%
Normal Commerccial .....99.0 to 99.5%

Fault Tolerant systems are characterized by essentially 100% availability, providing continuous processing through a combination of duplicated hardware and complex software. Such a system can absorb any single failure and some simultaneous multiple failures. However, designing for fault tolerance comes at a high cost. Fully fault-tolerant systems are extremely expensive, and are cost-justified only when an application cannot afford any downtime.

The Failover feature can be thought of as fault tolerance at the system level; not at the component or subassembly level. The configuration consists of a primary system and an alternate system. The two systems are identically configured from the point of view of hardware, operating system, and application software. Failover software resides in both systems. Should a problem develop with the primary system, a rapid and seamless switchover occurs to the alternate system.

A high availability system is characterized by availability in the range of 99.90 to 99.95%, i.e., an order of magnitude improvement over normal commercial systems. Though single failures can interrupt the system, the down time is minimized by a number of features: built-in hardware resiliency, automatic interrupt recovery, ability to replace hardware in a running system, and superior serviceability features. In contrast with fully fault tolerance, high availability is achievable at a reasonable cost.

A majority of the systems fall into the normal commercial availability category. Here when components fail, the system is down and repair is required. An expected availability of up to 99.5 percent seems impressive, but it adds up to 44 hours of downtime per year.

2.3 Importance of RAS

Organizations depending on computer-based systems are well aware that system failures cause serious consequences. System downtime can result in such consequences as:

- Customer and user dissatisfaction
- Lost productivity
- Lost revenue
- Increased overtime costs
- Potential for incurring penalties or fines, or worse

High levels of availability are essential for organizations whose goals are to make information instantly available to users across the enterprise. A high level of reliability, availability and serviceability is a crucial requirement for an enterprise servers.

3 CS6400 RAS Goals and Strategy

The CS6400 system is a "high availability" system. Further, it can be set up in a failover configuration, to achieve an even higher level of RAS. The primary goals of the CS6400 design were:

- Maximize system availability
- Maintain and protect the integrity of customer data

To achieve this, the strategy adopted was to focus on the following key areas:

- Redundancy of critical hardware
- Redundancy of paths to critical data storage devices (disks)
- Prompt failure detection and isolation
- Rapid, intelligent and automatic recovery
- Hardware replacement in a running system

4 CS6400 RAS Implementation

The CS6400 system, a large-scale data processing system, is composed of multiple, discrete components, including processors, memory banks, power supplies, system interconnect buses, I/O buses, disk drives, and cooling fans. There are external connections to networks such as Ethernet, ATM, and FDDI, and to the system service processor (SSP). The networks provide

user connections to the CS6400 system as well as information exchange between the CS6400 system and other systems.

Presented here are the key system features which together make the CS6400 system a high availability system.

4.1 Redundancy

A CS6400 system contains up to 16 system modules, each of which includes four processors, random access memory (RAM), and four I/O controller ports. Each system module contains a system bus to interconnect these components. The system bus extends over a centerplane to connect all the modules in the CS6400 system.

The system is a symmetric multiprocessing system with the memory spread over the system modules. Each processor has access to all of the memory in the system. The key point is that all processors and all sections of memory are equal. The CS6400 system can continue processing with less than the total number of processors or memory configured. The system bus is actually four separate buses, so there continues to be connection between the processors and memory should one or more buses fail.

Protection against failure of an I/O controller is achieved by using a duplicate controller to a provide a dual or an alternate path to the I/O device (disk) or to the network. Typically, the duplicate controller is configured on a different system module from the primary controller to protect against the system module being taken out of service or failing. The CS6400 system's operating system software has facilities to assist the system administrator to cleanly switch from one I/O controller to the other. Dual pathing is described elsewhere in the paper.

Although the CS6400 system was not designed as a fully fault-tolerant system, some areas lend themselves to fault tolerance. The CS6400 system is air-cooled using four blowers. Should one fail, there is sufficient reserve cooling to allow the machine to continue to operate, thus allowing replacement to be scheduled for a more convenient time, without adversely affecting availability. The system has comprehensive and fail-safe temperature monitoring to ensure that there is no temperature stressing of components in the event of a cooling failure.

Each system module has a corresponding plug-in power supply providing necessary voltage. These power supplies all have parallel outputs and, should one fail, the surviving ones can carry the load. Swap-out of the failed supply and insertion of the replacement are accomplished without service interruption.

4.2 Data Integrity

One of the techniques used in the CS6400 to maintain integrity of data is parity checking and correction. All data and control buses are protected by parity checks right out to the data on the disks. These checks ensure that errors are contained. All single-bit memory errors are automatically corrected, and double-bit (or worse) errors are detected. All parity errors are logged.

At the lowest level, memory is composed of multiple data storage chips called DRAMs (dynamic random access memories). The mapping of words to DRAMs is designed such that, should a complete DRAM fail, four successive memory words will have single-bit (correctable) errors; this is clearly superior to a four-bit failure in one word which could result in a system crash.

The other feature, called "memory scrubbing", is used to protect against idle segments of memory "growing" errors. Under control of the operating system, each memory word is accessed and rewritten once every 24 hours. Should there be any single-bit soft errors lurking, they are automatically corrected. In this way, there is less possibility of later having to deal with double-bit errors. All corrections during scrubbing are logged.

In the CS6400, the primary means of providing protection against catastrophic loss of data is via the technique of disk mirroring. This technique can be used in conjunction with dual pathing, i.e., use of duplicate I/O controllers. Disk mirroring is described elsewhere in the paper.

4.3 System Service Processor (SSP)

The System Service Processor (SSP), a dedicated UNIX workstation, is the key vehicle providing the CS6400 with superior serviceability features. It is the system administrator's interface to the CS6400 system. The SSP performs the following primary functions:

- Continuously monitors the CS6400
- Detects any failure, hardware and software, in the CS6400
- Sends out e-mail notification of a system interrupt
- After a system interrupt, logically disconnects or de-configures a failed hardware component
- Initiates system recovery, i.e., automatically reboots the system so that the operations can resume, possibly in a degraded configuration

The SSP has a graphical user interface (GUI) allowing easy administration of the CS6400 system. The GUI, called Hostview, makes the CS6400 system easy to manage. All CS6400 system error messages are logged to the SSP, which is useful for running diagnostics, should a failure occur, and predicting and preventing potential future failures. For example, an indication of a future failure would be excessive automatic correction of single-bit memory errors indicating that remedial action is required to avoid hard errors later.

The automatic reboot time mentioned above is, of course, dependent upon the hardware configuration on the CS6400 system. For a typical CS6400 system of 16 processors and 1 Gbyte of memory, the reboot time is less than 10 minutes, including integrity checking of all data on disk. This checking, and subsequent repair, if necessary, is virtually instantaneous, as all previous transactions have been journaled by the logging file system.

Another key functionality provided by the SSP is the remote support capability. A Cray customer service representative can

remotely, via Internet or modem, log into the SSP and diagnose and troubleshoot CS6400 problems. The remote support capability is a valuable serviceability feature.

4.4 System Board swap in a running System

The most significant serviceability feature of the CS6400 system is the ability to replace system modules online. When a hardware failure occurs, the failed components are de-configured from the system during the automatic reboot and recovery process. The ability to replace the failed component at a later time without shutting down the system is a significant contribution to achieving higher availability. The "Dynamic Reconfiguration" and "Hot-swap" features provide the ability to replace a system board in a running system, with negligible disruption to the operations. These features are described in a separate section of the paper.

A byproduct of the Dynamic Reconfiguration feature concerns upgrades to existing hardware. Additional processors, memory, or I/O controllers can be inserted into the system without shutting down the system

4.5 Domains

A feature that contributes to the RAS goals for the CS6400 system is that of domains - `systems within a system.' The CS6400 system may be logically divided into multiple domains. For example, one domain could be composed of four processors. The remaining processors would form another domain. Each section runs its own copy of the operating system and has its own peripherals and network connections.

Domains are useful for testing new applications or operating system updates on the smaller sections, while production work continues on the remaining, usually larger, domain. There is no adverse interaction between the domains, and customers can gain confidence in the correct operation of their applications without disturbing production work. When testing work is complete, the CS6400 system can be rejoined logically. There are no physical changes when using domains.

4.6 Failover Feature

As described earlier, a primary and a duplicate CS6400 system can be set up in a failover configuration, thus providing fault tolerance at the system level. Should a problem develop with the primary system, a rapid and seamless switchover occurs to the alternate system. The failover feature significantly enhances the high availability of the CS6400.

5 Dual Pathing for I/O

It is crucial that enterprise servers such as the CS6400 have built-in protection against catastrophic loss of data stored in disks. The CS6400 system offers discrete disks packaged 6 to a tray and disk arrays with up to 30 disks arranged in a matrix. Following are brief descriptions of techniques which are available in the CS6400 for protection against catastrophic loss of data and which guarantee a working path to the disk(s).

5.1 Disk Mirroring

For discrete disks, which are typically single-ported, mirroring (or RAID 1) protects against the failure of a disk by duplicating the information on an alternate disk. To protect against failure of the actual disk tray, the mirrored disks are installed in separate trays. Though not a requirement, each mirrored disk can be controlled by a separate disk controller on a separate system board. In fact, for single-ported disks, the disk mirroring is the only vehicle for achieving I/O path redundancy. Further resiliency can be incorporated by allocating "hot spare" disks, so that if one of the mirrored disks fails, the hot spare is automatically brought into service to take the place of the failed disk. Figure 1 shows an example of mirrored disk configuration having I/O path redundancy.

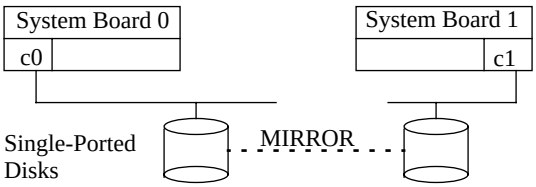

Figure 1 - Disk Mirroring with dual pathing

Disk arrays also support RAID 5 disk configurations. While mirroring requires twice as many disks to protect against failure, RAID 5 typically only requires 20 percent additional disk space. For instance, for every five disks used to store operational data, there is an extra parity disk that allows the data to be reconstructed, should one of the data disks fail. In practice, the data and the parity are distributed among all six disks, thus achieving better performance than having dedicated parity disks. If any one disk in a group should fail, the missing information is reconstructed from the surviving disks.

5.2 Alternate Pathing (AP)

In the case of dual-ported disk arrays, the Cray-developed Alternate Pathing (AP) feature is available for providing redundant paths to the disk array. Figure 2 shows an example of an Alternate Pathing configuration. Disk mirroring can also be used in conjunction with Alternate Pathing. It should be noted that, when mirroring is not a requirement, AP provides for redundant I/O pathing without having to incur the expense of duplicated disk drives.

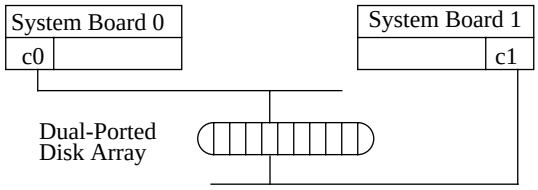

Figure 2 - Alternate Pathing

Disk mirroring and/or Alternate Pathing permit the matching of the customer's availability requirements with the degree of

protection against loss of data and/or loss of an I/O path to the disk drive(s).

6 System board swap in a running system

The capability to replace a system board in a running system, i.e., without shutting down the system, is a major contributor to the high availability of the CS6400 system. This capability is the result of two features: Dynamic Reconfiguration (DR), which deals with the software aspects of the process, and Hot-swap which deals with the hardware/electrical aspects of the process. The replacement of a system module is accomplished by the system administrator, or service provider, working through Hostview GUI on the SSP.

The first step in the process is the logical detachment of the system module, via the DR-detach feature. The Solaris operating system scheduler is informed that, for the board in question, no new processes should start. Any running processes and I/O operations are allowed to complete while memory contents are rewritten into other CS6400 system memory banks. A switchover to alternate I/O paths then takes place so that when the system module is removed, its I/O controllers are not missed.

The next step in the process is to connect the board to an external hot-swap power supply in preparation for its removal from the system. The operating system is then quiesced, i.e., all activities on the system bus are suspended for a few seconds. Once the operating system is quiesced, the system board can be removed from the system. System operations can then be resumed.

The removal sequences are all controlled by the SSP, and the system administrator simply follows the instructions given by Hostview. Timeouts ensure that the CS6400 system is not left in a suspended state.

The second half of process is the re-insertion of the system module into the CS6400 system. The process is simply the reverse of the removal sequence: attachment of the replacement board to the external hot-swap power supply, quiescing of the operating system, insertion of the board into the system, and finally, logical attachment of the board to the system, via the DR-attach feature.

Through a combination of dynamic reconfiguration and hot swap, the CS6400 system can be repaired or upgraded with minimal user inconvenience. There are two brief pauses during the quiescing of the operating system when a system module is actually being removed and replaced. The pauses seen by the user are measured minimal; measured in minutes.

7 System Interrupt Scenario

To highlight the high availability of the CS6400, Figures 3, 4, and 5 show three system interrupt scenarios as follows:

Figure 3 - no automatic reboot and no dynamic reconfiguration Figure 4 - automatic reboot and no dynamic reconfiguration Figure 5 - automatic reboot and dynamic reconfiguration

Comparisons of the figures 3, 4, and 5 clearly shows that the total time lost due to a system interrupt is minimal in the system such as the CS6400, which has both the automatic reboot and the dynamic reconfiguration features.

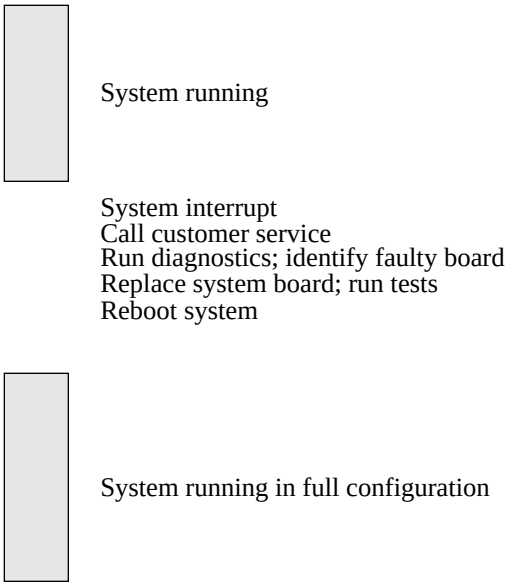

Figure 3 - No Automatic Reboot
No Dynamic Reconfiguration

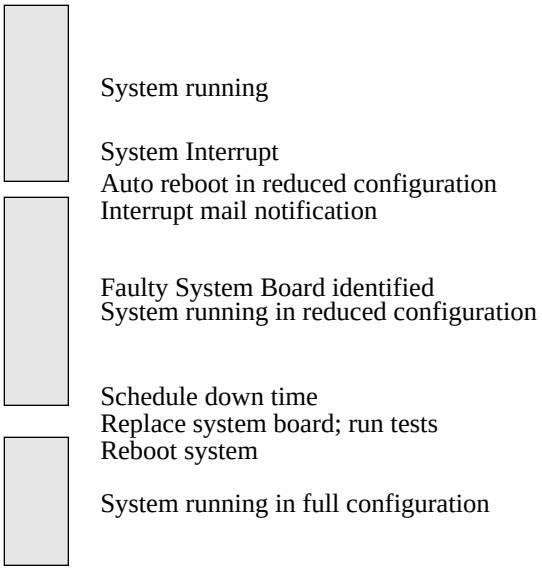

Figure 4- With Automatic Reboot
No Dynamic Reconfiguration

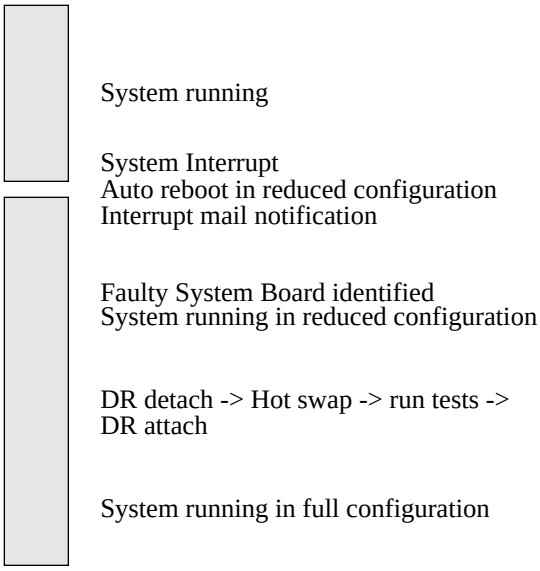

Figure 5- With Automatic Reboot
and Dynamic Reconfiguration

8 Summary

1. High levels of Reliability, Availability and Serviceability (RAS) are crucial for enterprise servers.
2. CS6400 is a high availability system fully capable of supporting business-critical applications.
3. CS6400 RAS features are unique among UNIX servers; in particular, the automatic reboot and the dynamic reconfiguration features.
4. The Failover feature of the CS6400 can take the system to a very high level of availability approaching full fault tolerance.

9 References

1.“RAS: Reliability, Availability, and Serviceability,” Technical White Paper, Cray Research Business Systems, 1995.