TECHNOLOGY BRIEF

August 1999

Compaq Computer Corporation

Prepared by
ISSD Technology Communications

CONTENTS

Compaq RAID on a Chip Technology

EXECUTIVE SUMMARY

Compaq pioneered RAID technology in 1989 with the introduction of the Compaq SMART Array Controller, and has led the industry ever since in expanding and improving hardware RAID functionality. Compaq's latest RAID innovation is a highly reliable, single-chip PCI RAID solution that provides excellent performance for a small numbers of drives. It is ideal for data center servers in which internal storage is optimized for operating systems and swap space, while more powerful array controllers and external storage subsystems are used for the data store.

This technology brief describes the technology, architecture, and functionality of the Compaq RAID on a Chip (ROC) subsystem and provides references for obtaining additional information. This brief will benefit those desiring an overview of ROC technology and those considering implementing RAID on a Chip in their servers.

This brief is written with the assumption that the reader already has a basic understanding of redundant arrays of independent disks (RAID).

Please direct comments regarding this communication to the ISSD Technology Communications Group at this Internet address: TechCom@compaq.com

COMPAQ

NOTICE

Compaq RAID on a Chip Technology
First Edition (August 1999)
Document Number 0191-0899-A

INTRODUCTION

Redundant Array of Independent Disks (RAID) is the industry-standard technology for ensuring data availability and reliability in high-volume servers. The value of RAID lies in the ability of the array controller to perform automatic data recovery in the event of a disk drive failure. For example, in a RAID 5 fault-tolerance configuration, if a drive fails and a read request is received for the missing data, the controller will automatically rebuild missing data from parity information.

Compaq pioneered RAID technology in 1989 with the introduction of the *Compaq SMART Array Controller* and has led the industry ever since in expanding and improving hardware RAID functionality. Continuing that leadership tradition, Compaq has produced the industry's first RAID on a Chip (ROC) solution, through collaborative development efforts with LSI Logic Corporation.

Compaq ROC is an embedded hardware-based RAID solution that enhances system reliability and improves processor utilization, input/output (I/O) efficiency, data integrity, and data recovery. It is ideal for data center servers in which internal storage is optimized for operating systems and swap space, while more powerful array controllers and external storage subsystems are used for the data store.

This technology brief describes the technology, architecture, and functionality of the Compaq ROC subsystem. In this brief, the term *host processor* refers to the server's central processing unit (CPU) to differentiate it from the embedded processor in the ROC subsystem.

ROC SUBSYSTEM TECHNOLOGY

By combining several discrete components and connections into one integrated circuit, Compaq has developed a fully integrated hardware RAID solution. The ROC subsystem consists of a processor, memory controller and hardware RAID engine, SCSI chip, PCI bridge, and an internal bus structure, all embedded on a single substrate (Figure 1).

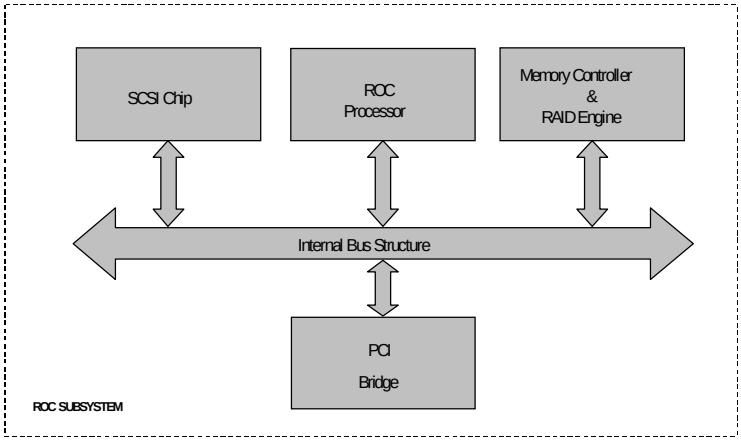

Figure 1. Simplified block diagram of Compaq ROC subsystem

Processor

The ROC processor is an ARM7 32-bit RISC core that controls the RAID implementation of the ROC subsystem. The product-specific firmware is read into the ROC memory from an external flash read-only memory (ROM) and executed within ROC, independently from the host processor. The host processor can perform other tasks while the ROC subsystem performs the RAID functions.

RISC: reduced instruction set computer.

SCSI Chip

The SCSI chip consists of two internal Ultra-2 SCSI channels. One channel is intended to support the server's internal disk drives. The second channel is intended for use with a SCSI tape drive. The SCSI chip is designed to provide optimal performance when supporting up to six internal drives and one SCSI tape drive. Although the second channel can be used to support an external storage unit instead of a tape drive, the added load will reduce overall I/O performance. For high I/O performance of configurations including external storage systems, Compaq recommends the use of a high-performance *Compaq Smart Array Controller* to support external storage systems.

PCI Bridge

The PCI bridge provides the interface between the internal bus structure of the ROC subsystem and the server's PCI bus. The host processor communicates with the ROC processor through the PCI bridge.

Internal Bus Structure

The ROC internal bus structure consists of several embedded buses that provide the communication paths between the internal components of the ROC subsystem. This structure enables highly reliable message and data control between the processor, memory, and storage in a dense environment. The internal bus structure is connected to the host processor through the PCI bridge.

Memory Controller and RAID Engine

The ROC memory controller has built-in interfaces providing access to dynamic random access memory (DRAM), flash ROM, and nonvolatile static random access memory (SRAM) (Figure 2) located on the motherboard. The RAID engine controls parity generation for RAID level 5 implementation.

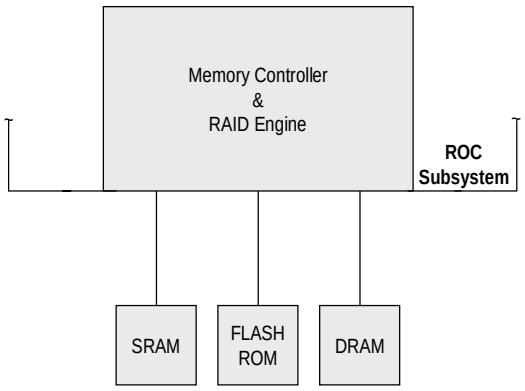

Figure 2. Block diagram of ROC memory controller showing interfaces to memory resources on the motherboard.

SRAM

The use of the SRAM is determined by the ROC program code. The SRAM is a nonvolatile memory that records the status of the subsystem during data rebuild. If power is lost during a rebuild, the SRAM enables the subsystem to continue the rebuild process from where it left off at the time of the power failure.

Flash ROM

Flash ROM provides initialization code to the host processor and program code to the ROC processor.

DRAM

DRAM is used to store ROC program code and operating data and to serve as a read-ahead cache.

The cache uses an intelligent, read-ahead algorithm that can anticipate data needs and reduce wait time. It can detect sequential read activity on single or multiple I/O threads and predict that sequential read requests will follow. It can read ahead, or pre-read data, from the disk drives before the data is actually requested. When the read request does occur, it can then be serviced out of high-speed cache memory at nanosecond speeds rather than from the disk drive at millisecond speeds.

This adaptive read-ahead scheme provides excellent performance for sequential small block read requests. At the same time it creates no penalty for random read patterns, because read-ahead is disabled when nonsequential read activity is detected.

ROC SUBSYSTEM FUNCTIONALITY

RAID implementation enhances data integrity and recovery through data striping and parity generation for the data rebuild process. The ROC subsystem improves host processor utilization by performing all RAID functions, thus freeing the host processor for other tasks. In most implementations, the ROC subsystem is fully integrated and embedded on the motherboard of the server. This implementation improves I/O efficiency and frees up a PCI slot for another controller. The ROC subsystem can queue tasks and data, thereby increasing performance.

When the ROC subsystem is equipped with memory (DRAM, Flash ROM, and SRAM) and LVD SCSI disk drives, it functions as a typical hardware RAID controller (Figure 3). The following paragraphs describe the functionality of the RAID controller.

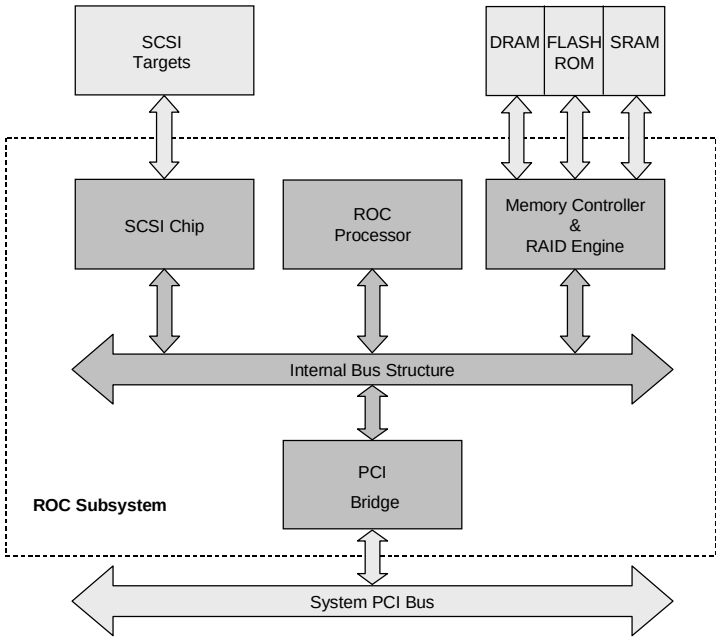

Figure 3. Block diagram of a typical embedded RAID controller

Modes of Operation

The ROC subsystem can be configured to initialize as a nonintelligent dual-channel SCSI controller, an intelligent dual-channel SCSI controller, or an intelligent dual-channel RAID controller. The mode of operation is determined at power-on.

If the ROC processor senses that no ROC memory hardware is present, the ROC processor selects the nonintelligent dual-channel SCSI controller mode. If the ROC processor senses that the ROC memory system is present, the ROC processor selects either the intelligent dual-channel SCSI controller mode or the intelligent dual-channel RAID controller mode, depending on the state of the SRAM address lines. The address lines are controlled by the ROC firmware, which resides in the ROC flash ROM.

Nonintelligent Dual-Channel SCSI Controller

In this mode of operation, each SCSI channel is controlled by the host processor through the PCI bridge. Messages are generated and controlled by the host processor for each SCSI channel.

Intelligent Dual-channel SCSI Controller

In this mode of operation, the ROC processor receives request messages from the host processor, processes them, and sends reply messages back to the host processor. The ROC processor and firmware together manage the transaction from start to finish without host processor intervention. No RAID functions are implemented in this mode.

Intelligent Dual-channel RAID Controller

In this mode of operation, the ROC processor receives request messages from the host processor, processes them, and sends reply messages back to the host processor. The ROC processor and firmware together manage the transaction from start to finish without host processor intervention. RAID functions are implemented in this mode.

Several levels of RAID are possible; each provides different fault-tolerance options to protect data. Each RAID level offers a different mix of performance, reliability, and cost effectiveness to the storage environment. Depending on the number of internal hard drives in the server, the RAID controller may be configured for RAID 0, 1, 0+1, or 5. RAID 1 and 5 support data recovery in case of a drive failure. RAID 1 provides data redundancy by mirroring data. RAID 5 allows the controller to rebuild lost data from the parity information on the remaining drives in the server. This is described under *Fault Tolerance and Recovery* later in this document.

The RAID levels supported by ROC are illustrated in Figure 4 and explained in more detail below.

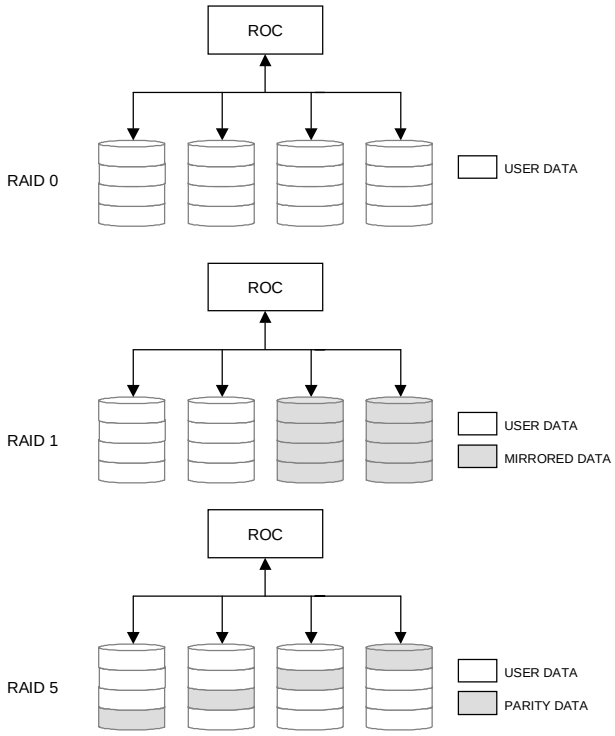

Figure 4. RAID levels supported by ROC

- RAID 0 (no fault tolerance): RAID 0 provides no fault protection against data loss. RAID 0 is known as "stripe sets" because data is simply striped across all drives in the array. This configuration provides high performance at low cost with excellent data availability. However, risk is incurred with possible data loss. RAID 0 may be assigned to drives that require large capacity (in some cases full capacity of the disks) and high speed, where loss of data due to disk failure can be tolerated. Typically, users of RAID 0 employ tape backup to safeguard critical data.
- RAID 1 (disk mirroring): This configuration of mirrored sets of data uses 50 percent of drive array storage capacity to provide greater data reliability by storing a duplicate of all user data on a separate disk drive. Therefore, half of the drives in the array are duplicated or "mirrored" by the other half. RAID 1 provides a very high level of fault tolerance, but drive cost doubles because this level requires twice as many disk drives to store the same amount of data. This may not be cost effective in some storage environments.
- RAID 0+1 (striping and mirroring): This configuration, sometimes called RAID 10, is a combination of RAID 0 and RAID 1. Compaq uses RAID 0+1 in any installation where RAID 1 is chosen by the user for 4 or more drives. The net results are the higher performance of RAID 0 and the higher protection of RAID 1.
- RAID 5 (distributed data guarding): This is the most popular RAID configuration. It is sometimes called "stripe sets with parity." The data is divided into blocks, and parity is generated for each block. Then, the data blocks are written in "stripes" to the disk drives, with one stripe on each disk saved for the parity data. This method is cost effective. The total amount of disk space used for redundancy is equivalent to the capacity of a single drive. The overall cost of this fault-tolerance method is lower than RAID 1. In RAID 5, if a drive fails,

the controller uses the parity data on the remaining drives to reconstruct data from the failed drive. This allows the system to continue operating with slightly reduced performance until the failed drive is replaced.

For more information on RAID technology, see "RAID Technology Overview" in the Compaq white paper *Configuring Compaq RAID Technology for Database Servers*, document number ECG011/0598.

Data Transactions

Once configured, the ROC processor receives request messages from the host processor, processes them, and sends reply messages back to the host processor. The ROC processor and firmware together manage the transaction from start to finish without host processor intervention. This relieves the host processor of performing many of the RAID functions.

Messages are decoded by the ROC processor into local actions, usually involving the transfer of data. Data may be moved between the main system memory and the SCSI targets (disk drives or tape), between main system memory and ROC memory system, or between the ROC memory system and the SCSI targets, depending on RAID level chosen (Figure 5).

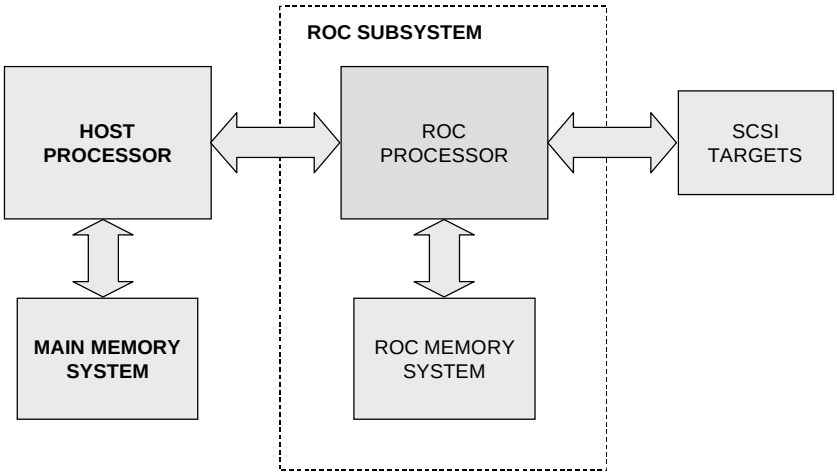

Figure 5. Simplified block diagram of ROC data flow

Fault Tolerance and Recovery

The ROC subsystem architecture provides hardware RAID protection that ensures data integrity and availability. When the ROC subsystem is configured for RAID 5 and prior to any possible drive failure, the subsystem proactively generates parity data so that it can keep all data available and the server running during replacement of any failed drive.

Hot-Plug Drive Support with Automatic Rebuild

The ROC subsystem supports SCSI hot-plug drives. If the storage subsystem contains hot-plug drives, users can insert or remove drives from fault-tolerant configurations while the system is up and running.

The ROC subsystem detects when a failed drive is removed and replaced. Then from parity information retrieved from remaining drives in the volume, it automatically rebuilds the data that was stored on the failed drive onto the replacement drive. When the rebuild operation is complete, data can again be read directly from the drive and no longer needs to be regenerated.

Online Spares

The ROC subsystem supports online spare disk drives. The number of spare drives is determined by the specific product that incorporates the subsystem. During server operation, the spare drives remain up and running but not active; that is, no I/O operations are performed to them during normal array operation. Spare drives are held in reserve in case one or more of the active drives should fail.

An online spare (Figure 6) differs from a parity drive. Parity drives are active or involved with all I/O operations. Online spares power up, but they remain in standby mode until needed. If an active drive fails during system operation, the ROC subsystem automatically and immediately begins a data rebuild operation onto the spare drive. Once the rebuild operation is complete, the system is again fully fault tolerant. The failed drive can be replaced at a convenient time. Once a replacement drive has been installed, the ROC subsystem will restore data to that replacement drive. The original online spare will return to standby mode and again be available as an online spare.

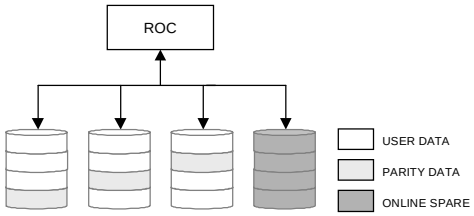

Figure 6. RAID 5 with online spare

Power Failure during Rebuild

The nonvolatile SRAM accessed by the embedded RAID controller stores the current state of the drive rebuild process. If a power failure occurs during the rebuild, the SRAM remembers the rebuild status. When power is restored to the controller, the SRAM restores the ROC subsystem to the state at power loss and continues the rebuild from that point.

CONCLUSION

As customer requirements for data availability, integrity, and security continue to rise, Compaq remains the industry leader in providing effective hardware RAID solutions. The Compaq ROC subsystem is an ideal RAID solution for data center servers in which internal storage is optimized for operating systems and swap space, while more powerful array controllers and external storage subsystems are used for data storage. As the world's leading server manufacturer, Compaq will continue to listen to customers and meet their computing needs with innovative, industry-standard enterprise solutions.