MINIMIZING MEMORY COST IN SET-TOP BOXES

Jeff Mitchell
Rambus Inc.

Abstract

Memory is the largest contributor to the cost of a set top box, typically comprising up to half or more of the component cost. This paper examines methods of minimizing this cost using commodity memory parts consumed in large volumes by the PC industry and by taking advantage of new developments in high density, high bandwidth DRAMs.

Architectural methods of extracting bandwidth from conventional DRAM are examined, along with performance requirements of emerging set-top box architectures. Alternatives are presented for meeting product performance goals while preserving low cost.

THE EFFECT OF MEMORY ON SYSTEM PERFORMANCE

Digital set top boxes require tremendous amounts of bandwidth in order to deliver the interactive experience desired by consumers. MPEG–2 decoding, stereo audio, computational 3D graphics and interactive user interfaces demand performance levels previously unheard of in consumer products. Yet the cost constraints of the consumer market demand that inexpensive, relatively low performance Dynamic Random Access Memory (DRAM) be used for memory. In the past, computer designers have used many DRAMs in parallel in order to increase memory bandwidth. This approach works well in PC's or workstations where there are many megabytes (MB) of memory, but is not effective in a set top box where the total memory size may only be as little as two to four MB. Two or four MB requires only one or two DRAM devices, whereas four or more devices are typically needed in order to provide sufficient bandwidth from conventional DRAMs. This makes a difficult choice for the system designer - add more memory and increase cost or accept lower performance.

Recently there have been new developments in the DRAM industry that offer solutions to the memory size/bandwidth dichotomy. Two new types of DRAM devices, Rambus™ DRAM (RDRAM®) and Synchronous DRAM (SDRAM) offer improved bandwidth over traditional devices and yet still maintain an inexpensive DRAM cost structure. Both are designed to be used in personal computer main memory applications, the largest user of DRAM memory. Over half of the world's annual DRAM shipments go into PC's. Interactive set top boxes which use these new high bandwidth types of memory can meet their performance objectives and take advantage of the PC cost/volume curves. This allows the consumer to have the best of both worlds - high performance at low cost.

MEMORY ALTERNATIVES

DRAM has always been the memory type of choice in cost-sensitive applications. Although DRAM performance is poor relative to other types of memory devices such as Static RAM (SRAM), DRAM has the virtue of being both dense and cheap. System designers go to great lengths to get the performance they need from DRAM and avoid using other types of memory in order to keep product cost down.

The conventional method of increasing performance in DRAM systems is to put several devices in parallel. If a single device cannot provide the required bandwidth, two devices used in parallel will double the performance. This also doubles the total amount of memory and nearly doubles the number of interface pins on the memory

controller. The effect on the system of using DRAMs in parallel for increased performance is shown in TABLE 1. This example uses 16 Megabit (Mb) page mode DRAM devices in a 16 bit wide organization (1M x 16).

DRAMs	*Bandwidth (MB/s)*	*Controller Pins*	*Total Memory*
1	66	40	2 MB
2	133	60	4 MB
4	266	120	8 MB

TABLE 1. SYSTEM EFFECTS OF OPERATING DRAMS IN PARALLEL

Memory Granularity

In order to achieve a desired level of bandwidth, some number of DRAMs must be used in parallel. The minimum amount of memory required to be used in parallel is called the memory granularity. As devices are paralleled to increase performance the total system memory grows proportionately. Large memory granularity can be a problem in applications where cost requires that the total memory be kept to a minimum.

In past generations of DRAM technology memory granularity was usually not an issue because DRAMs were less dense, with fewer total memory bits in each device. Four 4 Mb devices in parallel yields the same total bandwidth as shown in TABLE 1, but the memory granularity would only be 2 MB. This is a factor of four less than with 16 Mb DRAMs.

So why not just use lower density devices to get the needed performance? Besides the increased board space, power, and the sheer number of pins needed on the memory controller, DRAM economics dictate that the cheapest price per bit will be found on the densest device.

When designing a new product, the most cost effective memory to use is the densest DRAM technology available. This will yield the lowest total memory cost. FIGURE 1 shows the relative cost over time of various densities of DRAM. For new products it is best to target the densest device that will be cost effective in the time frame that the product reaches mass production.

For example, a product starting design today would target the 16 Mb generation while also giving consideration of how to use 64 Mb devices should the anticipated lifetime of the product extend beyond 1998 or 1999.

Considering that a 64 Mb DRAM is 8 megabytes of memory in a single chip, it is

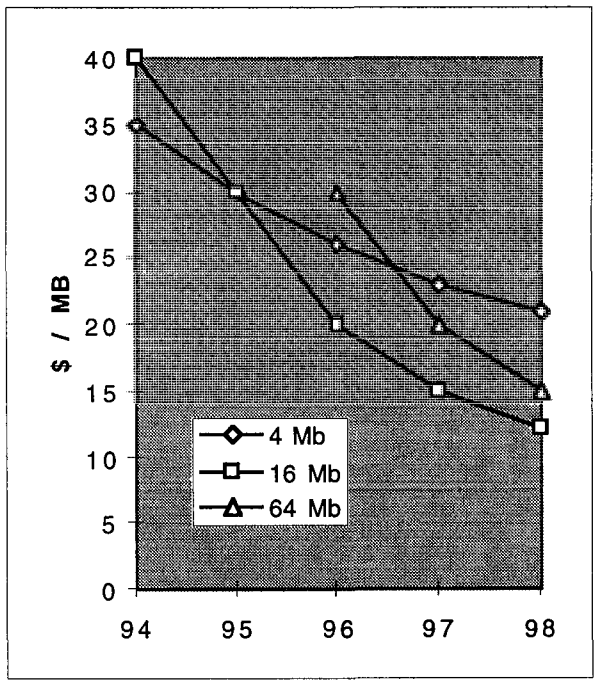

FIGURE 1. DRAM PRICING TRENDS

easy to see that the conventional practice of using DRAMs in parallel to obtain more bandwidth has become impractical to do in small memory systems. There are two alternatives to this practice - wider DRAMs and faster DRAMs.

Wide DRAMs

One alternative to using several DRAMs in parallel is for the DRAM

manufacturers to simply make wider devices with data bus widths of 16, 32 or 64 bits. This provides the same benefit as parallelism but without increasing the total system memory size. This approach works up to a point, but then becomes both financially unattractive and technically difficult.

As a DRAM is made wider the die size increases and the package gets larger and more costly. With more I/O pins, it also becomes more expensive to test. These factors tend to negate the cost advantages of DRAM.

A wide DRAM also cannot be operated as fast as a narrower device. The increased number of output pins causes more noise and ground bounce. The remedy to this problem is to run the device slower, which offsets the performance advantage of being wider. Wider parts must also have more pins providing power and ground connections, which again increases cost.

Wider DRAMs are a partial solution to achieving higher performance, but at some width around 32 bits this approach reaches diminishing returns.

Fast DRAMs

An alternative to making a wider DRAM is to make a faster DRAM. Here the objective is to keep the device width down to a manageable size, but increase the speed at which it operates.

There are two types of 'fast' DRAMs, Synchronous DRAM (SDRAM) and Rambus DRAM (RDRAM). These two DRAM derivatives are similar in that they both use a conventional memory core and run the external interface at a high speed. This provides the economic advantages of a conventional DRAM while providing much higher performance.

Synchronous DRAM (SDRAM)

An SDRAM is a conventional DRAM mated to a synchronous interface. The synchronous interface aligns data transfers into and out of the part with an external clock reference.

Synchronizing the data transfer to a clock allows for tighter timing parameters and therefore a higher operating speed. SDRAMs can run in systems at speeds up to 66 MHz, about double the speed of a conventional DRAM. Doubling the interface speed means that only half as many devices are needed for a given bandwidth. This reduces the memory granularity to half that of a conventional DRAM.

Rambus DRAM (RDRAM)

As with SDRAM, RDRAM is a conventional DRAM mated to a synchronous interface. An RDRAM has a 64 bit wide internal bus running at 75 MHz. The RDRAM connects to a memory controller, which also has a 64 bit wide internal bus running at 75 MHz.

These wide internal busses narrow to only 8 bits externally without any impact on performance. This gives an RDRAM the performance of a 64 bit wide DRAM while retaining all of the cost advantages of a narrow 8 bit external bus.

Type	*Bus Width*	*Package Pins*	*Bandwidth (MB/s)*
DRAM	16 bits	42	66
DRAM	32 bits	100	133
SDRAM	16 bits	50	133
RDRAM	8 bits	32	600

TABLE 2. PIN/BANDWIDTH COMPARISON

TABLE 2 summarizes the number of signals, package pins, and relative bandwidth of DRAM, wide DRAM, SDRAM and RDRAM.

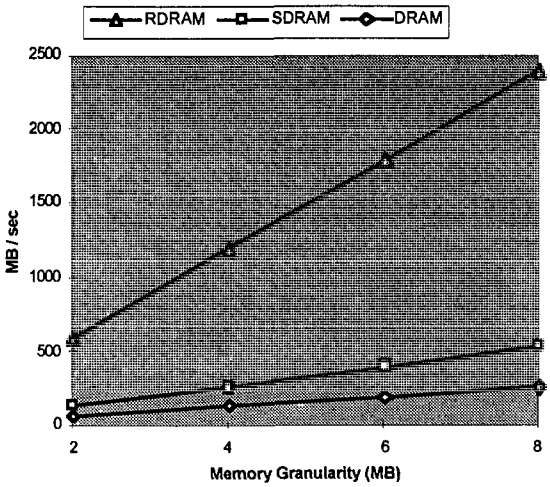

FIGURE 2. MEMORY GRANULARITY VS. BANDWIDTH

With each type of DRAM there is a straight-line relationship between bandwidth and memory granularity. This relationship makes it straightforward to approximate the memory granularity for a given level of system performance If the system performance requirement lies above the line shown in FIGURE 2 for a particular type of DRAM, either another type of DRAM will have to be used or the bus width will have to be increased, with a corresponding increase in memory granularity and system cost.

For example, an application which requires 250 MB/s of system bandwidth can be designed one of three ways, depending upon whether DRAM, SDRAM, or RDRAM is used. FIGURE 3 shows block diagrams of example systems comparing total bandwidth, memory granularity, and number of controller pins required for each of the three solutions.

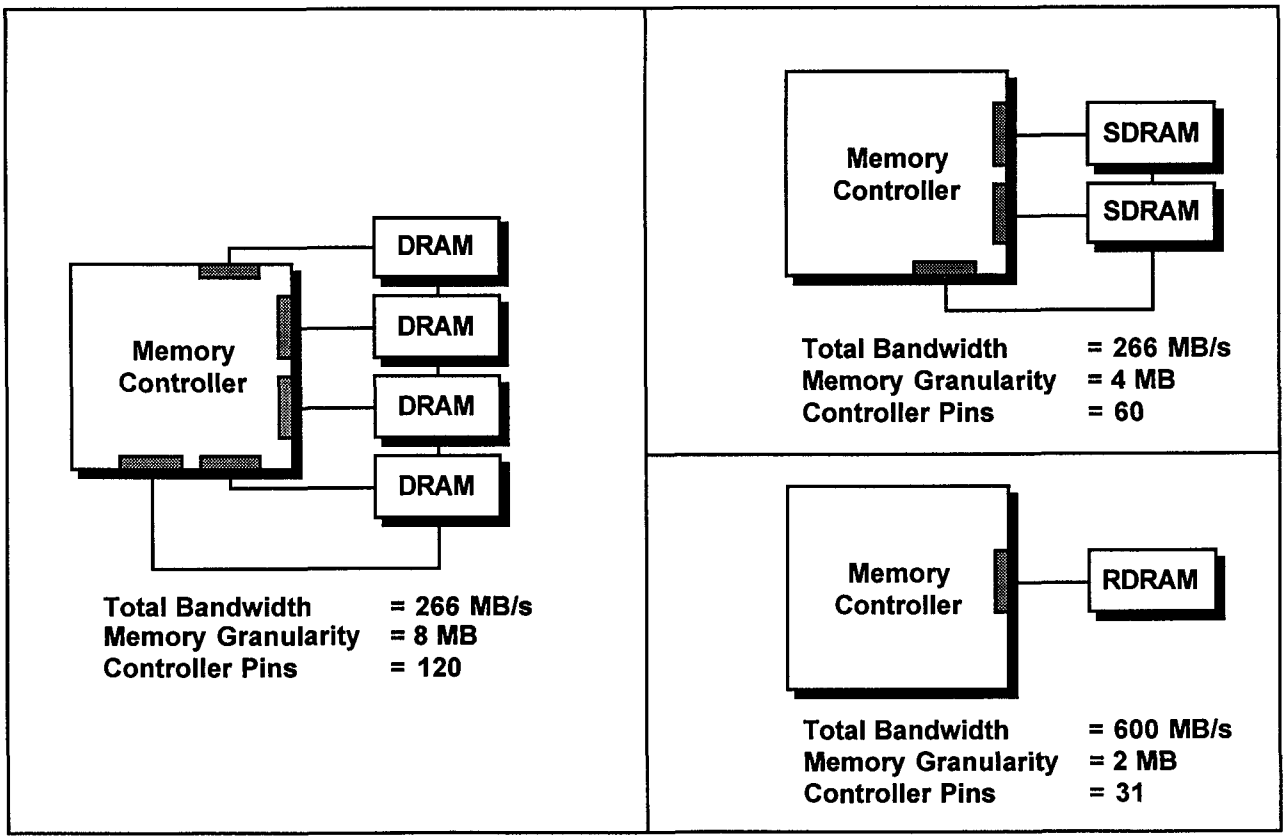

FIGURE 3. HIGH BANDWIDTH MEMORY SYSTEMS

SET-TOP BOX PERFORMANCE REQUIREMENTS

Depending upon the targeted application, a set-top box can have a broad range of memory bandwidth requirements. TABLE 3 lists the approximate bandwidth required for several common functions.

Function	*Min B/W*	*Max B/W*
Video	15	30
CPU	5	100
MPEG–2	100	200
2D Graphics	50	200
3D Graphics	100	300
Sound	10	50

TABLE 3. BANDWIDTH REQUIREMENTS (MB/s)

At one end of the spectrum a simple analog decoder has very modest bandwidth requirements. At the other end of the spectrum, a fully digital web-capable system with 2 channels of MPEG–2 for picture-in-picture and a fully interactive 3D user interface could easily require over 500 MB/s of memory bandwidth. This much bandwidth would require 16 MB of DRAM, 8 MB of SDRAM, or 2 MB of RDRAM. Clearly, it is difficult if not impossible to make a high performance, cost effective consumer product using conventional DRAM. Estimating system bandwidth requirements from TABLE 3 and comparing to FIGURE 3 gives an idea of what memory options are available for a cost optimized product.

Distributed vs. Unified Memory

There are two architectural methods to obtain the required system memory bandwidth. The conventional method has been to attach the required amount of memory to a chip performing a specific function. For example the microprocessor would have some memory connected directly to it, the video decoder would have some more memory separate from the CPU memory, and so on with each separate chip in the unit. This approach works well as long as the memory can be cleanly partitioned and there are no problems with memory granularity.

An alternative method is to unify the memory and have all functions operate directly out of the same block of memory. While this eases the memory granularity problem by combining many small pieces of memory into one large pool, it also increases the bandwidth that is needed from that one pool.

A move toward unified memory architectures is being motivated by the increasing integration of functions. Integrating previously separate functions into a single chip forces unification of the memory for all of those functions.

IMPLEMENTATION EXAMPLES

To demonstrate how system performance and memory granularity interact, two example systems are profiled below. These cover the entire range of set-top box functionality ranging from a simple dedicated decoder to a high end fully interactive system. The examples assume a unified memory architecture since that provides the potential for the lowest cost system by utilizing the densest memory devices.

Dedicated MPEG–2 Decoder

In this example a single NTSC MPEG–2 stream is being decoded. In addition there is a simple user interface generated by the CPU. From TABLE 3, the total system bandwidth can be estimated:

Video	15 MB/s
CPU	10 MB/s
MPEG–2	100 MB/s
TOTAL	125 MB/s

This bandwidth can be provided by either two DRAMs, one SDRAM, or one RDRAM. The memory granularity in this

example is 4 MB if DRAM is used, or 2 MB if SDRAM or RDRAM is used.

Fully Interactive Terminal

An advanced interactive terminal may consist of a complete Internet web browser in addition to a multiple stream MPEG–2 decoder (for picture-in-picture or faster response to channel surfing) with surround stereo along with a fully interactive 3D user interface. Such a terminal would need a significant amount of memory bandwidth. Again estimating total system bandwidth from TABLE 3:

Video	30 MB/s
CPU	100 MB/s
3D Graphics	300 MB/s
Sound	50 MB/s
MPEG–2	200 MB/s
TOTAL	680 MB/s

To provide this bandwidth from DRAM would require over 20 MB of memory! Digital systems are generally designed to support memory systems in binary increments, so a 128 bit data bus with 32 MB would be the memory granularity for a DRAM based system. An implementation using DRAM obviously would be too costly to be a consumer product.

A memory subsystem built from SDRAM would require a 64 bit data bus and 16 MB of memory in order to achieve the needed 680 MB/s of bandwidth. Design compromises could get the bandwidth down to 533 MB/s which would require only 8 MB of memory. Not cheap, but getting there.

Using RDRAM would require only 4 MB of memory which would provide well over the required 680 MB/s of bandwidth. Over 500 MB/s of spare bandwidth would be available for other functions or future performance improvements. Alternatively, if it were possible to put all of these functions into 2 MB of memory, then as with the SDRAM example the system could be re-engineered to get the required bandwidth down to 600 MB/s. This bandwidth can be satisfied by a single RDRAM.

A set-top box with this kind of high end functionality is not likely to become commercially viable for several years, at which point the most cost effective DRAM will be a 64 Mb device. The higher density DRAM will exacerbate the memory granularity problem. Using 64 Mb devices the minimum DRAM system would be 128 MB, SDRAM 32 MB, and RDRAM 8 MB (a single RDRAM device).

COST REDUCTION THROUGH INTEGRATION

Electronic products become less expensive every year. This is due to improving manufacturing yields of the electronic components and higher functional integration. The key to cost competitiveness in consumer products is taking advantage of increasing levels of IC integration.

Integration and unified memory architectures are complementary. Functions that require several chips, each with their own memory space, are becoming integrated into a single IC. When this is done the separate memories must also be integrated into a single space.

Integration of components has several benefits. Functions that completely reside in a single chip do not have to communicate with each other through I/O pins. An integrated device has a smaller total die area and fewer package pins than the same functions spread across several chips.

However, it may still have too many pins to be in a cost effective package if excess pins have to be used to get the needed bandwidth from the memory system. This again points to the benefit of high pin-bandwidth memory devices.

The cost benefits of integration can be very compelling. A consumer product that is similar in functionality and implementation to a set-top box is a video game console. In this marketplace there is an excellent example of the advantages of integration and unified memory architecture.

Sega Saturn

FIGURE 4 is a block diagram of a Sega Saturn. This is a 32 bit game console designed to support high performance 3D graphics.

The Saturn has a very distributed system architecture. There are several microprocessors, each with their own memory subsystem. The 3D graphics subsystem is spread across two chips, each of which is connected to its own private memory. Even the audio subsystem has a separate dedicated memory.

Because each of these memory systems is small, they are implemented using older technology 4 Mb DRAMs. This is much less cost effective than using current generation 16 Mb DRAM. In a distributed system such as this, it is impossible to use more cost effective higher density DRAM without increasing the total memory capacity tremendously.

Adding to the system cost is the large number of interconnects between the components. Several four layer printed circuit boards are required for connecting the devices together.

Nintendo 64™

FIGURE 5 is a block diagram of the Nintendo 64, a 64 bit game console designed, as was the Sega Saturn, to support high performance 3D graphics. The component integration level in the Nintendo 64 is substantially higher than in the Saturn design. Except for the game cartridge, the only memory in the system is 4 MB of Rambus DRAM - two devices.

The high bandwidth and low pin count interface of the RDRAM allow all of the 3D graphics, sound generation, and CPU control to be integrated into the Reality Coprocessor ASIC. The only other components in the system are the RISC CPU and some small glue chips.

The Reality Coprocessor in the Nintendo 64 provides the same functionality as the Saturn's two video display processors and audio processor. This high level of integration allows all of the memory that has been distributed in small pieces throughout the several subsystems in the Saturn to be collected into a single pool of Rambus DRAM. The Nintendo 64 takes advantage of 16 Mb DRAM technology for maximum cost effectiveness.

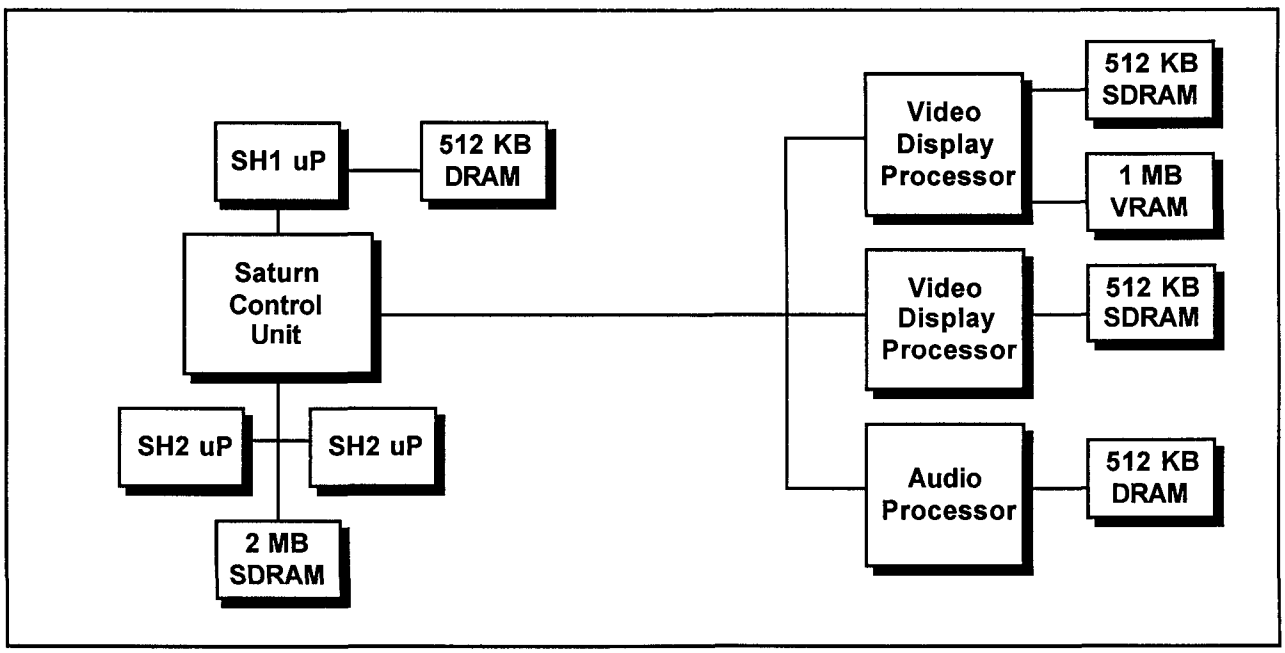

FIGURE 4. SEGA SATURN BLOCK DIAGRAM

The entire Nintendo system fits in a 6" x 6" form factor, which due to the simplicity of the design can be implemented on a single low cost two layer printed circuit board. The cost savings on the PC board alone is $5.00[1].

MINIMIZING SYSTEM COST

Consumer electronic products achieve cost reduction primarily through integration. This has two effects on system memory. The first is that there are physically fewer chips and pins to connect to the memory devices. The same amount of memory bandwidth therefore has to flow through fewer I/O connections. Second, the memory devices themselves become more integrated, packing more bits of memory into a single device. Again, the same amount of memory bandwidth has to flow through fewer I/O connections.

These two effects have a common result, that is fewer I/O pins connected to memory. This provides a cost savings, but can adversely affect system performance unless compensated for by using a higher bandwidth memory device.

The low individual component cost of standard 4 Mb DRAM can be deceiving. To minimize system cost in consumer products attempts should be made to take advantage of higher component integration levels and the lower cost per bit of 16 Mb DRAM. The high bandwidth Synchronous and Rambus DRAMs make such high levels of system integration technically feasible.

References:

[1] "The Ultra 64 Joypad", Interview with Genyo Takeda, "Next Generation" magazine, pp. 38-40, February 1996

Trademarks:

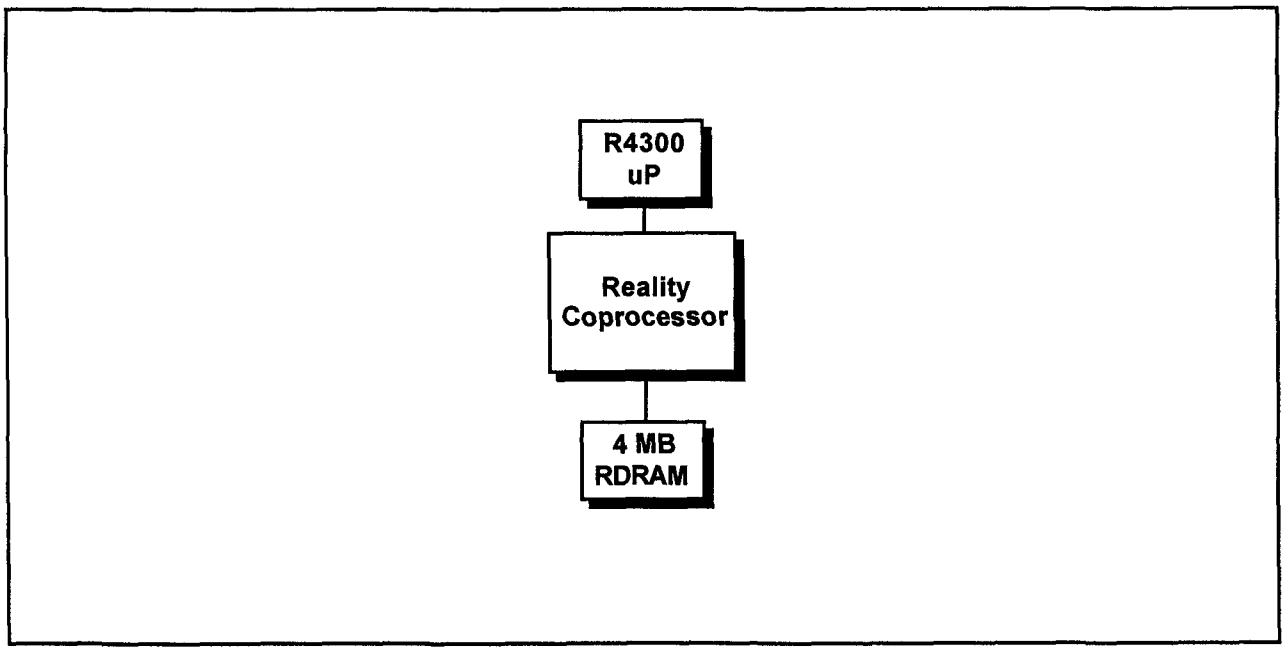

FIGURE 5. NINTENDO 64 BLOCK DIAGRAM