

# THE RAM GUIDE

By DEAN KENT

## Overview

Three years ago, there wasn't much to say about system RAM. Almost all PCs came with fast page mode (FPM) DRAM, which ran at speeds between 100ns and 80ns. However, escalating CPU and motherboard bus speeds outstripped the ability of FPM DRAM to deliver data in a timely manner. Nowadays there are a lot of different memory designs.

## Introduction

Due to cost considerations, all but the very high-end (and very expensive) computers have utilized DRAM for main memory. Originally, these were asynchronous, single-bank designs because the processors were relatively slow. Most recently, synchronous interfaces have been produced with many advanced features. Though these high-performance DRAMs have been available for only a few years, it is apparent that they will soon be replaced by at least one of the protocol-based designs, such as SynLink or the DRDRAM design from Rambus, Inc. and Intel.

## Basic DRAM Operation

A DRAM memory array can be thought of as a table of cells. These cells are comprised of capacitors, and contain one or more 'bits' of data, depending upon the chip configuration. This table is addressed via row and column decoders, which in turn receive their signals from the RAS and CAS clock generators. In order to minimize the package size, the row and column addresses are multiplexed into row and column address buffers. For example, if there are 11 address lines, there will be 11 row and 11 column address buffers. Access transistors called 'sense amps' are connected to the each column and provide the read and restore operations of the chip. Since the cells are capacitors that discharge for each read operation, the sense amp must restore the data before the end of the access cycle.

The capacitors used for data cells tend to bleed off their charge, and therefore require a periodic refresh cycle or data will be lost. A refresh controller determines the time between refresh cycles, and a refresh counter ensures that the entire array (all rows) are refreshed. Of course, this means that some cycles are used for refresh operations, and has some impact on performance.

A typical memory access would occur as follows. First, the row address bits are placed onto the address pins. After a period of time the RAS\ signal falls, which activates the sense amps and causes the row address to be latched into the row address buffer. When the RAS\ signal stabilizes, the selected row is transferred onto the sense amps. Next, the column address bits are set up, and then latched into the column address buffer when CAS\ falls, at which time the output buffer is also turned on. When CAS\ stabilizes, the selected sense amp feeds its data onto the output buffer.

## Asynchronous Operation

An asynchronous interface is one where a minimum period of time is determined to be necessary to ensure an operation is complete. Each of the internal operations of an asynchronous DRAM chip are assigned minimum time values, so that if a clock cycle occurs

# THE RAM GUIDE

By DEAN KENT

any time prior to that minimum time another cycle must occur before the next operation is allowed to begin.

It should be fairly obvious that all of these operations require a significant amount of time and creates a major performance concern. The primary focus of DRAM manufacturers has been to either increase the number of bits per access, pipeline the various operations to minimize the time required or eliminate some of the operations for certain types of accesses.

Wider I/O ports would seem to be the simplest and cheapest method of improving performance. Unfortunately, a wider I/O port means additional I/O pins, which in turn means a larger package size. Likewise, the additional segmentation of the array (more I/O lines = more segments) means a larger chip size. Both of these issues mean a greater cost, somewhat defeating the purpose of using DRAM in the first place. Another drawback is that the multiple outputs draw additional current, which creates ringing in the ground circuit. This actually results in a slower part, because the data cannot be read until the signal stabilizes. These problems limited the I/O width to 4 bits for quite some time, causing DRAM designers to look for other ways to optimize performance.

## Page Mode Access

By implementing special access modes, designers were able to eliminate some of the internal operations for certain types of access. The first significant implementation was called Page Mode access.

Using this method, the RAS\ signal is held active so that an entire 'page' of data is held on the sense amps. New column addresses can then be repeatedly clocked in only by cycling CAS\ . This provides much faster random access reads, since the row address setup and hold times are eliminated.

While some applications benefit greatly from this type of access, there are others that do not benefit at all. The original Page Mode was improved upon and replaced very quickly so you will likely never see any memory of this type. Even if you do, it wouldn't be worth even getting it for free, considering the advantages of later access modes.

## Fast Page Mode

Fast Page mode improved upon the original page mode by eliminating the column address setup time during the page cycle. This was accomplished by activating the column address buffers on the falling edge of RAS\ (rather than CAS\ ). Since RAS\ remains low for the entire page cycle, this acts as a transparent latch when CAS\ is high, and allows address setup to occur as soon as the column address is valid, rather than waiting for CAS\ to fall.

Fast Page mode became the most widely used access method for DRAMs, and is still used on many systems. The benefit of FPM memory is reduced power consumption, mainly because sense and restore current is not necessary during page mode access. Though FPM was a major innovation, there are still some drawbacks. The most significant is that the

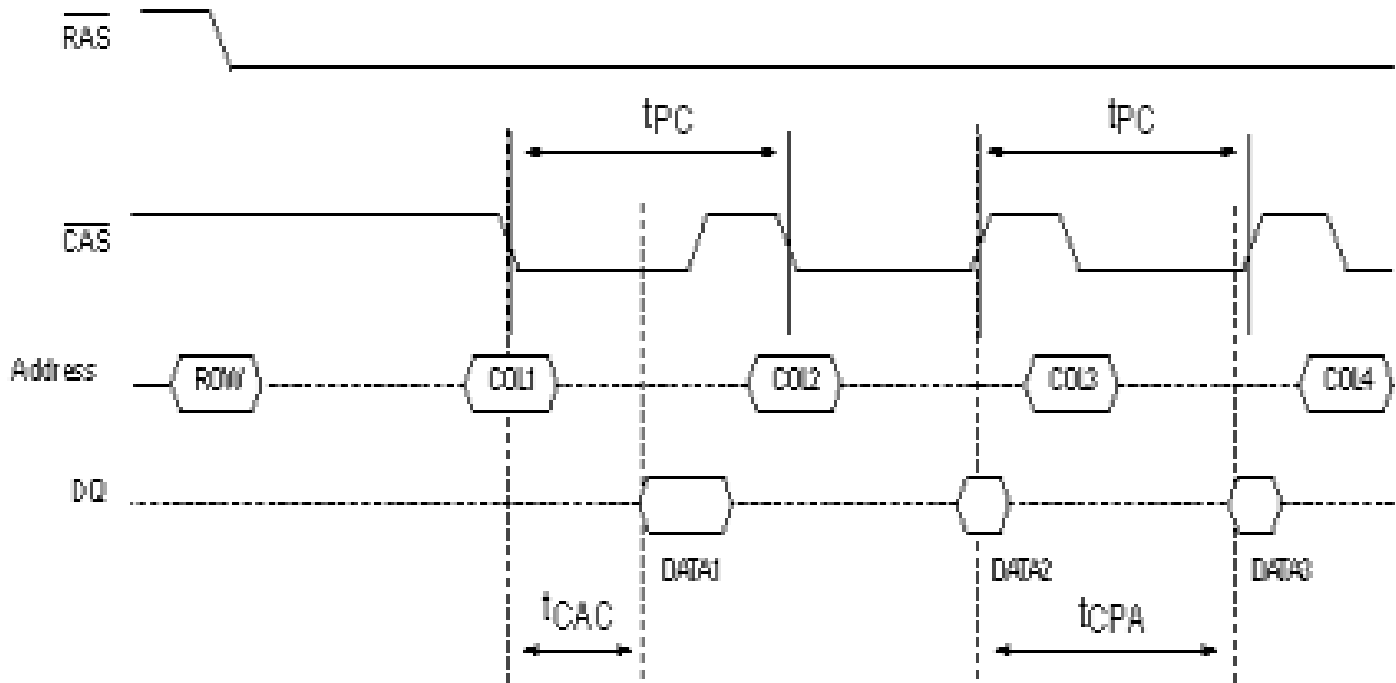
# THE RAM GUIDE

By DEAN KENT

output buffers turn off when CAS goes high. The minimum cycle time is 5ns before the output buffers turn off, which essentially adds at least 5ns to the cycle time.

Today, FPM memory is the least desirable of all available DRAM memory. You should only consider using this if it is either free, or your system does not support any of the later memory types (such as a 486 based system). Typical timings are 6-3-3-3 (initial latency of 3 clocks, with a 3-clock page access). Due to the limited demand, FPM is actually more expensive now than most of the faster memories now available.

## Fast Page Mode:



## HyperPage Mode (EDO)

The last major improvement to asynchronous DRAMs came with the Hyperpage mode, or Extended DataOut. This innovation was simply to no longer turn off the output buffers upon the rising edge of /CAS. In essence, this eliminates the column precharge time while latching the data out. This allows the minimum time for /CAS to be low to be reduced, and the rising edge can come earlier.

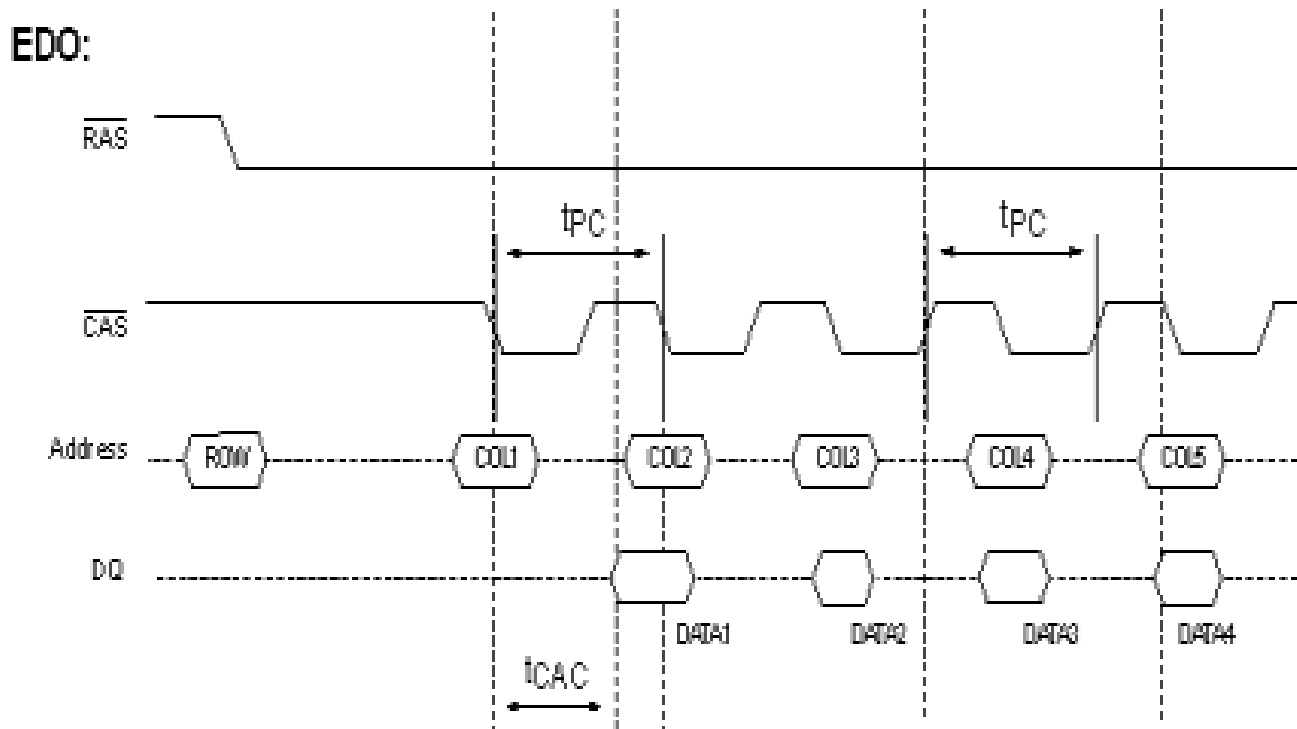
In addition to a 40% or greater improvement in access times, EDO uses the same amount of silicon and the same package size. EDO has been shown to work well with memory bus speeds up to 83MHz with little or no performance penalty. If the chips are sufficiently fast (55ns or faster), EDO can be used even with a 100MHz memory bus. One of the best reasons to use EDO is that all of the current motherboard chipsets support it with no compatibility problems, unlike much of the synchronous memory now being used.

# THE RAM GUIDE

By DEAN KENT

Even with all the stated advantages, EDO is no longer considered mainstream. Most manufacturers no longer produce it, or have limited production. It is only a matter of time before the prices begin to rise, and the equivalent size SDRAM module will be less expensive.

If you already own EDO memory, there is no real reason to jump to SDRAM unless you require bus speeds above 83MHz. With a typical EDO timing of 5-2-2-2 at 66MHz, there is almost no noticeable improvement with SDRAM over EDO, and at 83MHz it is still negligible. If you require 100MHz bus operation, EDO will lag far behind current SDRAM in performance even if it does operate at that speed due to the need for 6-3-3-3 timings. On the other hand, with EDO being phased out, you will likely find SDRAM to be equal to or even lower in price.



## Burst EDO (BEDO)

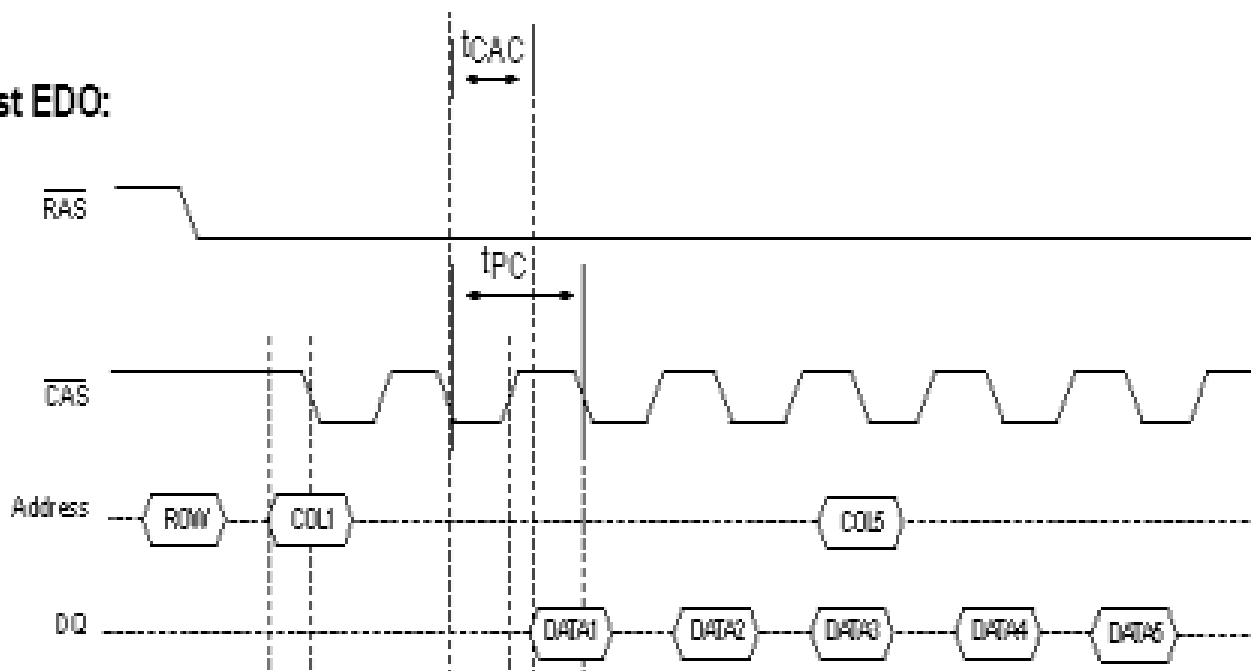
Burst EDO, while a good idea, was dead before it ever was born. The addition of a burst mode, along with a dual bank architecture would have provided the 4-1-1-1 access times at 66MHz that many expected with SDRAM. Burst mode is an advancement over page mode, in that after the first address input, the next 3 addresses are generated internally, thereby eliminating the time necessary to input a new column address. Unfortunately, Intel decided that EDO was no longer viable, and SDRAM was their preferred memory architecture so they did not implement support of BEDO into their chipsets. In fact, several large memory manufacturers had put considerable time and money into the development of SDRAM over the past decade, and were not very happy with the BEDO design.

# THE RAM GUIDE

By DEAN KENT

Except for support of bus speeds of 100MHz and faster, BEDO would probably have been a much faster and more stable memory than SDRAM. Essentially, BEDO lost support as much for political and economic reasons as for technical ones, it seems.

## Burst EDO:



## Synchronous Operation

Once it became apparent that bus speeds would need to run faster than 66MHz, DRAM designers needed to find a way to overcome the significant latency issues that still existed. By implementing a synchronous interface, they were able to do this and gain some additional advantages as well.

With an asynchronous interface, the processor must wait idly for the DRAM to complete its internal operations, which typically takes about 60ns. With synchronous control, the DRAM latches information from the processor under control of the system clock. These latches store the addresses, data and control signals, which allows the processor to handle other tasks. After a specific number of clock cycles the data becomes available and the processor can read it from the output lines.

Another advantage of a synchronous interface is that the system clock is the only timing edge that needs to be provided to the DRAM. This eliminates the need for multiple timing strobes to be propagated. The inputs are simplified as well, since the control signals, addresses and data can all be latched in without the processor monitoring setup and hold timings. Similar benefits are realized for output operations as well.

## JEDEC SDRAM

All DRAMs that have a synchronous interface are known generically as SDRAM. This includes CDRAM (Cache DRAM), RDRAM (Rambus DRAM), ESDRAM (Enhanced SDRAM)

# THE RAM GUIDE

## By DEAN KENT

and others, however the type that most often is called SDRAM is the JEDEC standard synchronous DRAM.

JEDEC SDRAM not only has a synchronous interface controlled by the system clock, it also includes a dual-bank architecture and burst mode (1-bit, 2-bit, 4-bit, 8-bit and full page). A 'mode register' that can be set at power-on and changed during operation controls the burst mode, burst type (sequential or interleave), burst length and CAS latency (1, 2 or 3).

CAS Latency is one of several performance related timings for SDRAM. This measurement is the time it takes to strobe in the Row Address, and to activate the bank. When a burst read cycle is initiated, the addresses are set up and RAS\ and CS\ (chip select) are held low on the next clock cycle (rising edge of CLK), thereby activating the sense amplifiers on the bank. A period of time equal to tRCD (RAS\ to CAS\ delay) must pass after which CAS\ and CS\ are held low (again, at the next clock cycle). After the time period for tCAC (column access time) has passed the first bit of data is on the output line and can be retrieved (at the next clock cycle). The basic rule is that CAS latency times the clock speed (tCLK) must be equal or greater than tCAC (or  $CL \times tCLK \geq tCAC$ ). This means that the column access time is the limiting factor for CAS Latency.

SDRAM was initially introduced as the answer to all performance problems, however it quickly became apparent that there was little performance benefit and a lot of compatibility problems. The first SDRAM modules contained only two clock lines, but it was soon determined that this was insufficient. This created two different module designs (2-clock and 4-clock), and you needed to know which your motherboard required. Though the timings were theoretically supposed to be 5-1-1-1 @ 66MHz, many of the original SDRAM would only run at 6-2-2-2 when run in pairs, mostly because the chipsets (i430VX, SiS5571) had trouble with the speed and coordinating the accesses between modules. The i430TX chipset and later non-Intel chipsets improved upon this, and the SPD chip (serial presence detect) was added to the standard so chipsets could read the timings from the module. Unfortunately, for quite some time the SPD EEPROM was either not included on many modules, or not read by the motherboards.

SDRAM chips are officially rated in MHz, rather than nanoseconds (ns) so that there is a common denominator between the bus speed and the chip speed. This speed is determined by dividing 1 second (1 billion ns) by the output speed of the chip. For example a 67MHz SDRAM chip is rated as 15ns. Note that this nanosecond rating is not measuring the same timing as an asynchronous DRAM chip. Remember, internally all DRAM operates in a very similar manner, and most performance gains are achieved by 'hiding' the internal operations in various ways.

The original SDRAM modules either used 83MHz chips (12ns) or 100MHz chips (10ns), however these were only rated for 66MHz bus operation. Due to some of the delays introduced when having to deal with the various synchronization of signals, the 100MHz chips will produce a module that operates reliably at about 83MHz, in many cases. These SDRAM

# THE RAM GUIDE

By DEAN KENT

modules are now called PC66, to differentiate them from those conforming to Intel's PC100 specification.

## PC100 SDRAM

When Intel decided to officially implement a 100MHz system bus speed, they understood that most of the SDRAM modules available at that time would not operate properly above 83MHz. In order to bring some semblance of order to the marketplace, Intel introduced the PC100 specification as a guideline to manufacturers for building modules that would function properly on their upcoming i440BX. With the PC100 specification, Intel laid out a number of guidelines for trace lengths, trace widths and spacing, number of PCB layers, EEPROM programming specs, etc.

There is still quite a bit of confusion regarding what a 'true' PC100 module actually consists of. Unfortunately, there are quite a few modules being sold today as PC100, yet do not operate reliably at 100MHz. While the chip speed rating is used most often to determine the overall performance of the chip, a number of other timings are very important. tRCD (RAS to CAS Delay), tRP (RAS precharge time) and CAS Latency all play a role in determining the fastest bus speed the module will operate on to still achieve a 4-1-1-1 timing.

PC100 SDRAM on a 100MHz (or faster) system bus will provide a performance boost for Socket 7 systems of between 10% and 15%, since the L2 cache is running at system bus speed. Pentium II systems will not see as big a boost, because the L2 cache is running at ½ processor speed anyway, with the exception of the cacheless Celeron chips of course.

## DDR SDRAM

One limitation of JEDEC SDRAM is that the theoretical limitation of the design is 125MHz, though technology advances may allow up to 133MHz operation. It is obvious that bus speeds will need to increase well beyond that in order for memory bandwidth to keep up with future processors. There are several competing new standards on the horizon that are very promising, however most of them require special pinouts, smaller bus widths, or other design considerations.

In the short term, Double Data Rate SDRAM looks very appealing. Essentially, this design allows the activation of output operations on the chip to occur on both the rising and falling edge of the clock. Currently, only the rising edge signals an event to occur, so the DDR SDRAM design can effectively double the speed of operation up to at least 200MHz.

There is already one Socket 7 chipset that has support for DDR SDRAM, and more will certainly follow if manufacturers decide to make this memory available. In this industry, many times it is the first to market that gains the support, rather than the best technology.

## Enhanced SDRAM (ESDRAM)

In order to overcome some of the inherent latency problems with standard DRAM memory modules, several manufacturers have included a small amount of SRAM directly into the

# THE RAM GUIDE

By DEAN KENT

chip, effectively creating an on-chip cache. One such design that is gaining some attention is ESDRAM from Ramtron International Corporation.

ESDRAM is essentially SDRAM, plus a small amount of SRAM cache which allows for lower latency times and burst operations up to 200MHz. Just as with external cache memory, the goal of a cache DRAM is to hold the most frequently used data in the SRAM cache to minimize accesses to the slower DRAM. One advantage to the on-chip SRAM is that a wider bus can be used between the SRAM and DRAM, effectively increasing the bandwidth and increasing the speed of the DRAM even when there is a cache miss.

As with DDR SDRAM, there is currently at least one Socket 7 chipset with support for ESDRAM. The deciding factor in determining which of these solutions will succeed will likely be the initial cost of the modules. Current estimates show the cost of ESDRAM at about 4 times that of existing DRAM solutions, which will likely not go over well with most users.

## **Protocol Based DRAM**

All of the previously discussed DRAM have separate address, data and control lines which limits the speed at which the device can operate with current technology. In order to overcome this limitation, several designs implement all of these signals on the same bus. The two protocol based designs currently getting the most attention are SyncLink DRAM (now called SLDRAM due to trademark issues) and Direct Rambus DRAM (DRDRAM) licensed by Rambus, Inc. DRDRAM

Intel has placed their money on the proprietary memory design developed by Rambus, Inc. On the surface, this looks to be a very fast solution for system memory due to its fast operation (up to 800MHz). The reality is, however, that the design is only up to twice as fast as current SDRAM operation due to the smaller bus width (16 bits vs. 64 bits).

Despite the claims from Intel and Rambus, Inc., there are some potentially serious issues which need to be addressed with this technology. The higher speeds require short wire lengths and additional shielding to prevent problems with EMI. In addition, latency times are actually worse than currently available fast SDRAM. Since most of today's applications do not actually utilize the full bandwidth of the memory bus even today, simply increasing the bandwidth while ignoring latency issues will likely not provide any real performance improvements. In addition, processors operating with 800MHz bus speeds will certainly require more than double the current memory bandwidth.

While these issues are serious enough, the biggest drawback to the technology is that it is proprietary technology. Manufacturers wishing to implement a solution with DRDRAM will be required to pay a royalty to Intel and Rambus, Inc., and will also have no real control over the technology. This is not an attractive outlook for most memory manufacturers who have no desire to essentially become chip foundries.

# THE RAM GUIDE

By DEAN KENT

## SLDRAM

Many memory manufacturers are putting their support behind SLDRAM as the long-term solution for system performance. While SLDRAM is a protocol-based design, just as RDRAM is, it is an open-industry-standard, which requires no royalty payments. This alone should allow for lower cost. Another cost advantage for the SLDRAM design is that it does not require a redesign of the RAM chips.

Due to the use of packets for address, data and control signals, SLDRAM can operate on a faster bus than standard SDRAM – up to at least 200MHz. Just as DDR SDRAM operates the output signal at twice the clock rate, so can SLDRAM. This puts the output operation as high as 400MHz, with some engineers claiming it can reach 800MHz in the near future.

Compared to DRDRAM, it seems that SLDRAM is a much better solution due to the lower actual clock speed (reducing signal problems), lower latency timings and lower cost due to the royalty-free design and operation on current bus designs. It appears that even the bandwidth of SLDRAM is much higher than DRDRAM at 3.2GB/s vs. 1.6GB/s

Though Intel initially intended to support only DRDRAM in future chipsets, competing chipset manufacturers, memory manufacturers and pressure from end users may force them to include support for SLDRAM as well. If the marketplace can successfully influence Intel to provide this support, we may actually see a situation where the best technology wins over marketing hype.

## DIRECT RAMBUS DRAM (From INTEL)

Working with industry vendors, Intel has adopted a solution that will enable high-performance memory access without adopting exotic and expensive memory technologies. Direct Rambus\* DRAM, also known as Direct RDRAM, is a fast, serial memory technology developed by Rambus Technology. This technology, which will appear in upcoming Intel® Architecture-based workstations, provides an increase in memory bandwidth from 800MBps for SDRAM to 3.2GBps for dual-channel RDRAM.

## The RDRAM Difference

RDRAM differs from today's prevalent synchronous DRAM (SDRAM) in that it uses a thin-and-fast serial connection to move data between the system and memory. Measuring 16 bits wide and ticking away at 800MHz, a single channel of RDRAM memory can provide 1.6GBps of throughput to main system memory. Today's PC100 SDRAM, by contrast, produces 800MBps over a wide, 64-bit connection running at 100MHz.

The advantages go beyond bus widths and clock ticks. The efficient RDRAM interface means that the effective sustained bandwidth—as opposed to the mathematical optimum or "peak bandwidth"—is proportionally higher than that of SDRAM. The protocol has been optimized so that sustained bandwidth can be up to 90 percent of the peak bandwidth. If you take SDRAM designs and run the same application, your sustained bandwidth from the memory is about 65 percent.

# THE RAM GUIDE

By DEAN KENT

Not only does RDRAM provide headroom for intensive streaming media transactions, it also reduces latency by lessening the amount of time it takes to complete memory transactions. What's more, the channeled architecture of the RDRAM interface means that throughput increases as more channels are added. Intel®-based, workstation-class motherboards will employ two channels of RDRAM to deliver 3.2GBps of bandwidth.

By contrast, signaling and timing issues limit clock speeds on the 64-bit SDRAM interface to about 133MHz. Above that level, the varying lengths and electrical loads of SDRAM bus pins make it difficult for system chipsets and memory to reliably distinguish signals. Widening the SDRAM bus is not an attractive option either, as a 128-bit interface would double existing pin counts on the memory bus to over 200 pins, requiring more power as a result.

## Managing Transitions

Direct RDRAM will later be introduced on workstation motherboards featuring the Intel® Pentium® III Xeon™ processor. These motherboards will run at 133MHz and support up to 4GB of RDRAM. The memory itself is packaged much like the dual inline memory modules (DIMMs) used with SDRAM, in a format called RDRAM inline memory modules (or RIMMs). The RIMMs and RIMM sockets will operate within the mechanical and thermal envelope of today's SDRAM designs, easing motherboard and system design.

In order to ease the transition to RDRAM, Intel will provide compatibility with SDRAM. For those more concerned about standardizing on an existing memory platform, SDRAM compatibility will help you buy into the most advanced technology now and make the move to RDRAM later.