New DRAMs Improve Bandwidth (Part 1)

SDRAMs, CDRAMs, and EDRAMs Illustrate Evolutionary Approach

By Steven Przybylski, Consultant, San Jose, CA

This is the first of three parts describing and evaluating alternative DRAM approaches. Steven Przybylski is a consultant on system architecture and product planning and the author of the book "Cache and Memory Hierarchy Design: A Performance-Directed Approach."

The standard multiplexed, asynchronous DRAM interface has been with us since 1974, when Mostek introduced the MK4096 4-Kbit DRAM. Though this interface has served its purpose well, it is increasingly difficult to build high-performance memory systems using a small number of memory chips. Dramatic increases in DRAM density have meant that the available bandwidth per bit of memory has dramatically declined even as processor bandwidth requirements have increased.

Over the past year, several alternative DRAM interfaces have been proposed or demonstrated that address the growing gap between processor needs and the readily available DRAM bandwidth. This article covers the cached DRAM, synchronous DRAM, and enhanced DRAM. Part 2 will discuss the Rambus and RamLink approaches *(see **070304.PDF**)*, and a summary and comparison of these approaches will be presented in Part 3 *(see **070405.PDF**)*.

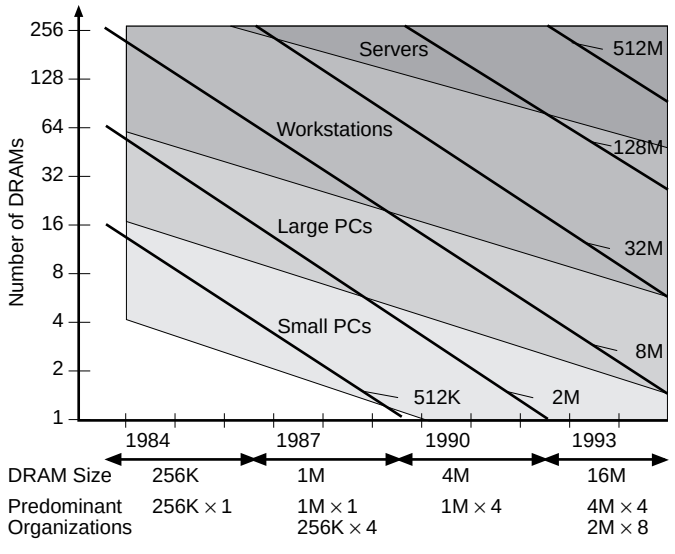

Figure 1. As DRAM sizes grow, the number of parts needed for a fixed-size memory decreases (solid black lines). For a particular type of system (gray bands), main memory size slowly increases, but the total number of parts declines.

Problems With Conventional DRAM

Figure 1 illustrates the trends at the heart of the growing problem with the conventional, narrow DRAM interface. Over time, DRAM sizes have increased at an average rate of about 4× every three years. Meanwhile, the main memory sizes of typical computer systems have grown more slowly, at around 2× every three years. The result is that with each successive generation of DRAMs and systems, the average number of DRAMs in a system has decreased significantly. Consequently, even as the industry has moved from ×1 to ×4 and ×8 parts, the total memory bandwidth has declined over the past decade. In contrast, processor speeds have increased by close to 100× over the same period.

Despite the universal use of caches to reduce the number of references reaching main memory, memory bandwidth requirements have grown significantly over the decade. The end result is a growing gap between the capabilities of memory systems and the needs of the processors to which they are connected. This trend is especially pronounced at the low end of the spectrum where low cost, simplicity, and small size are crucial.

Another important consequence of these trends is the increasing granularity of main memory. Granularity is the smallest amount of memory that can be added at a time. If the DRAM size grows without a corresponding increase in width, then the depth must increase, affecting the memory system's granularity. For example, in a 32-bit-wide memory system built of 1M × 4 RAMs, the granularity is 4 Mbytes.

This problem is compounded in higher-bandwidth memory systems that use interleaving, which increases the effective width of the memory system by using multiple banks of memory. If the previous example used a four-way interleaved structure, it would have a 128-bit effective width and a granularity of 16 Mbytes—unacceptably high for the PC market. A lower granularity can be obtained by using smaller DRAMs, but this can limit the maximum size of main memory unless the physical design can accommodate larger numbers of parts.

A number of effects make it difficult to design small, high-bandwidth memory systems using generic DRAMs. Most significantly, the asynchronous nature of the DRAM interface makes high-speed operation difficult. Timing margin, signal settling time, large voltage swings, and bus dead-time all add significant overhead to the raw DRAM access time. Some high-performance systems minimize this cycle-time overhead using ASIC buffers to partition the memory system and reduce the loading on any heavily-loaded bidirectional buses. These

	Conventional		Evolutionary			Revolutionary	
	Generic DRAM	Wide DRAM	SDRAM	CDRAM	EDRAM	RDRAM	RamLink
Proponent		Various	JEDEC JC42.3 DRAM Committee	Mitsubishi	Ramtron	Rambus	IEEE CS P1596.4 Working Group
Interface	Asynchronous RAS/CAS	Asynchronous RAS/CAS	Synchronous RAS/CAS	Synchronous RAS/CAS + SRAM access	Enhanced asynchronous RAS/CAS	Custom high speed sync. 9-bit bus	Custom synchronous Token Ring
Currently Available Sizes and Organizations	4 or 16 Mbit	4 or 16 Mbit	None	4 Mbit	4 Mbit	4.5 Mbit	None
	1M × 4, 4M × 1, 4M × 4, 16M × 1	×8, ×9, ×16, and ×18		1M × 4	1M × 4, 4M × 1		
Future Organizations		×32 and ×36	4M × 4, 2M × 8@ 66 & 100 MHz	256K × 16, 4M × 4	512K × 8, 4M × 4, 2M × 8	18 Mbit	First availability likely at 64 Mbits
Electrical Interfaces	TTL	TTL	LVTTL and GTL/CTT @ 100 MHz	TTL (4M), GTL or LVTTL (16M)	TTL	Single-ended, low voltage swing	Differential low voltage swing

Table 1. A comparison of conventional and high-speed DRAMs.

ASICs, especially if included on the memory SIMMs, add cost and limit after-market distribution channels for expansion memory.

As DRAMs grow from 4 Mbits to 16 Mbits and beyond, the standard narrow, asynchronous interface will become increasingly inadequate. Six main alternatives have surfaced in the past year. These alternatives span a wide range from the conservative to the radical. They can be partitioned into three groups: conventional, evolutionary, and revolutionary alternatives. Table 1 lists their key characteristics along with those of the narrow generic DRAMs. Each of these new interfaces addresses one or more of the limitations of the existing narrow DRAM interface.

The most conventional approach is to widen the data width without changing the electrical or logical interface at all. The evolutionary designs—synchronous DRAMs (SDRAMs), cached DRAMs (CDRAMs) from Mitsubishi, and enhanced DRAMs (EDRAMs) from Ramtron—modify the logical interface and/or the electrical interface to decrease average latency and/or to increase peak bandwidth. These devices must still be used in parallel to construct a complete memory system.

The revolutionary alternatives—Rambus and RamLink—replace the conventional separate address and data paths with a single byte-wide path on which address, control, and data are communicated at high speed. They are also radical in that a high-performance memory system can be constructed out of a single chip.

Wider DRAMs

The most straightforward solution to the bandwidth problem is to increase the width, or number of data pins per DRAM. Increasing the width from ×1 to ×4 at the 256K and 1M levels painlessly prolonged the life of the interface by a generation. Further increases in width become increasingly costly, however, since the growth in DRAM density that must be matched is exponential. With each generation, the number of bits per DRAM goes up by a factor of four. To maintain the same bandwidth per bit, the width of the DRAMs would also have to quadruple with each new generation. For example, 4M devices in ×16 organizations provide roughly equivalent bandwidth to 1M parts in ×4 widths. At the 16M level, ×64 organizations would be required just to keep pace.

Though such wide organizations are certainly feasible and probably will be offered, they introduce as many problems as they solve. Even at the ×32 level, the large number of signals can create severe ground-bounce problems, and package size and cost add a significant premium over the generic narrow DRAM. Furthermore, any future increase in bandwidth would come at the cost of even wider devices. Thus, increasing DRAM width is a viable temporary solution, but at some point further increases will be unacceptable.

Synchronous DRAM (SDRAM)

Although four of the five new DRAM designs are synchronous in that they include a clock signal, the term "synchronous DRAM" generally refers to one particular style of DRAM interface that uses a registered, multiplexed address bus and a registered data bus (see Figure 2). The control signals are also latched. Although there is a JEDEC committee (EIA/JEDEC JC42.3 DRAM Standards Committee) working on a functional, electrical, and physical specification of a 2M × 8 SDRAM, there will be SDRAMs that, for at least the near term, will not conform to all aspects of that standard. Conforming and nonconforming SDRAMs will probably be available from a variety of vendors and in a variety of organizations in

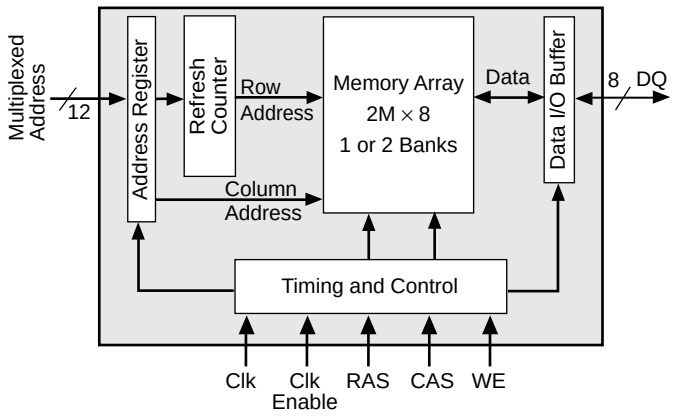

Figure 2. SDRAM block diagram.

the coming years, with the first devices available by the end of this year.

Standard SDRAMs address the problems of the existing interface in an evolutionary way on two separate fronts. First, the asynchronous RAS/CAS interface is replaced with a synchronous interface that permits pipelined accesses at 66 MHz and higher frequencies. Second, the increase in device width from 4 to 8 bits further doubles the per-DRAM bandwidth.

The important features that are being proposed for 16M SDRAMs are:

- Registers on the address, data, and control signals
- Programmable access latencies
- Pipelined row and column accesses
- Overlapped row accesses and data transfers
- Automatic wraparound for burst transfers
- Sub-block ordered burst transfers
- A clock-enable feature (for power-down mode)
- On-chip refresh control

In addition, some of the proposed SDRAM parts have a dual-bank internal architecture that improves opportunities for on-chip parallelism. Many of the new features are designed to allow precharging and row accesses to occur during block data transfers, keeping the data bus fully utilized even in the face of frequent row accesses.

Though the JEDEC committee is working toward a physical and functional standard for SDRAMs, it appears that the first round of synchronous offerings by DRAM vendors will contain some significant differences, potentially delaying their market acceptance. Another inhibitor is the confusion over the electrical interface. Though most manufacturers are designing 3.3V LVTTL (low-voltage TTL) compatible parts at 66 MHz, LVTTL signal quality will be inadequate at 100 MHz for most memory systems. At some frequency, a transition will have to be made to a terminated electrical interface with a low signal swing, such as GTL (Gunning transceiver logic) or CTT (center-tap termination). This pending change in signal levels will complicate the migration of systems designs from initial SDRAMs to subsequent, higher-speed offerings.

It is clear that the SDRAM proposals offer significantly more bandwidth and ease of use than the traditional asynchronous interface. The adoption of the nonproprietary JEDEC standard by many manufacturers will speed the acceptance of this interface. The question that the marketplace will ultimately answer is whether SDRAMs provide adequate bandwidth and design flexibility to serve as a long-term replacement for the current interface, or whether a more revolutionary alternative will also be needed by mainstream DRAM applications.

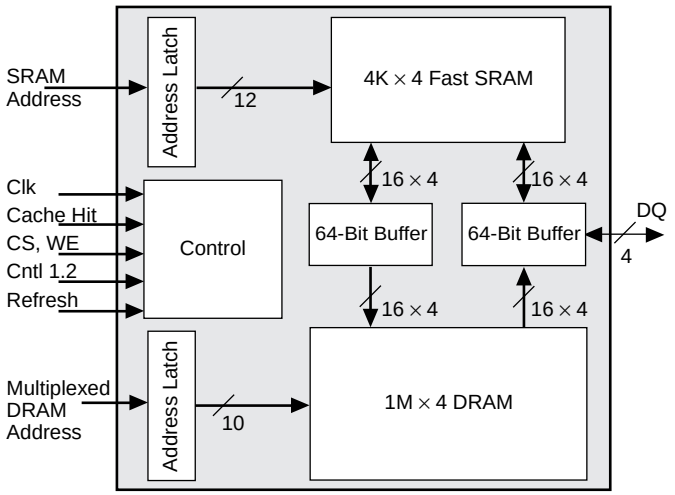

Figure 3. 4M CDRAM block diagram.

Cached DRAM (CDRAM)

The CDRAM is a unique DRAM variant offered by Mitsubishi in 4M and, later this year, 16M sizes. In addition to a DRAM array, the CDRAM includes a small SRAM that acts as a cache (see Figure 3). At the 4M level, the SRAM access time varies between 10 ns and 20 ns, while the row access time of the DRAM ranges between 70 ns and 80 ns, depending on the speed grade. As with SDRAMs, the interfaces are synchronous, with off-chip bandwidths of up to 50 Mbytes/s. Multiple 64-bit datapaths within the CDRAM permit fast transfers between the SRAM and DRAM.

For example, a memory system using eight 4M CDRAMs would have a total SRAM cache of 16 Kbytes. By building this cache into the DRAMs, 64 bytes can be transferred between the main memory and the cache in a single cycle. This high-bandwidth path improves the effectiveness of the relatively small cache. An interesting side effect of this organization is that the cache size, block size, and cache-to-main-memory bandwidth all increase as the width of the memory system increases.

Since the CDRAM contains only the data portion of cache, the tags must be maintained elsewhere within the memory system controller. Although the tag array for a

bank of CDRAMs is not very large (256 tags of 9 bits each), one such array is needed for each bank of eight CDRAMs (4 Mbytes with the 1M × 4 CDRAMs). Either the memory controller must provide enough tag memory for the maximum possible memory size, or the controller must be modular to allow additional tag capacity to be added as the memory size increases.

Of the new DRAM architectures, the CDRAM is unique in two important ways. First, although all modern DRAMs contain fast storage for caching of data, only the CDRAM allows the system designer to control the organization of the fast storage and how it is used. Second, the separate address buses for the DRAM array and the SRAM provide the designer with more control over transfers from DRAM to SRAM and from SRAM to the processor. Together these characteristics make for a flexible part capable of high sustained bandwidths.

Enhanced DRAM (EDRAM)

The Enhanced DRAM is a new DRAM organization developed by Ramtron International (Colorado Springs, CO). The EDRAM has both evolutionary and revolutionary aspects. It is evolutionary in that the basic protocol is very similar to the current static-column or page-mode DRAM interface. It is an asynchronous part with a multiplexed address bus and is available in 1M × 4 and 4M × 1 organizations, with 512K × 8 in development.

As shown in Figure 4, the primary extension to the interface is the addition of a path that allows writes to occur to the DRAM array without disturbing the page cache, which contains the last row read. By latching the write address and data, writes can be retired quickly, before the actual write to the DRAM array completes. Subsequent read references that hit the page cache can thus be satisfied in parallel with the completion of a write operation.

In addition to the usual signals, Ramtron has added a refresh control pin (F*) and a second write-control pin (Write Enable, WE*). The total pin count is 28, just two more than a generic 1M × 4 DRAM.

The revolutionary aspect of the EDRAM is its speed. Ramtron has put a significant emphasis on the speed of the DRAM array and overlapping the precharge time with data transfers. Most commercially-available 4M parts have DRAM array (RAS) access times between 60 and 100 ns. In contrast, the DRAM access time of the Ramtron part is a scant 35 ns. Accesses to the page cache complete in 15 ns, significantly faster than the CAS access times of generic page-mode DRAMs (30–50 ns) and comparable to the peak latency of the initial 66-MHz (15-ns) SDRAMs.

Thus, the EDRAM combines background write completion with fast access times, both for accesses that hit the page cache and for those that miss. In the short term, these parts can provide performance superior to a typical DRAM memory system with SRAM cache, especially if the processor-to-memory interface is running at 50 MHz or below. But in the longer term, the EDRAM interface suffers from many of the same deficiencies of the existing DRAM interface that limit granularity and minimum main-memory size. Once processors with greater than 50-MHz interfaces come into common use, these disadvantages could be overcome by wrapping a more revolutionary interface around Ramtron's uniquely fast DRAM core.

Conclusions

The three evolutionary architectures significantly increase the per-DRAM and per-bit bandwidth without greatly changing the generic DRAM interface. They preserve the multiplexed address path and are all being offered primarily in relatively narrow widths (×1 to ×8). These characteristics require multiple DRAMs to be used in parallel to form 32-bit (or wider) memory systems. Given the continuing exponential growth in the density of DRAMs to 64M and beyond, the question remains whether these evolutionary approaches provide enough added bandwidth to meet system designers' needs for many years to come, or if there will be a need for more radical DRAM interfaces. ♦

*In Part 2, we will look at the more revolutionary Rambus and RamLink designs (see **070304.PDF**), while Part 3 will offer a detailed comparison of the various approaches (see **070405.PDF**).*

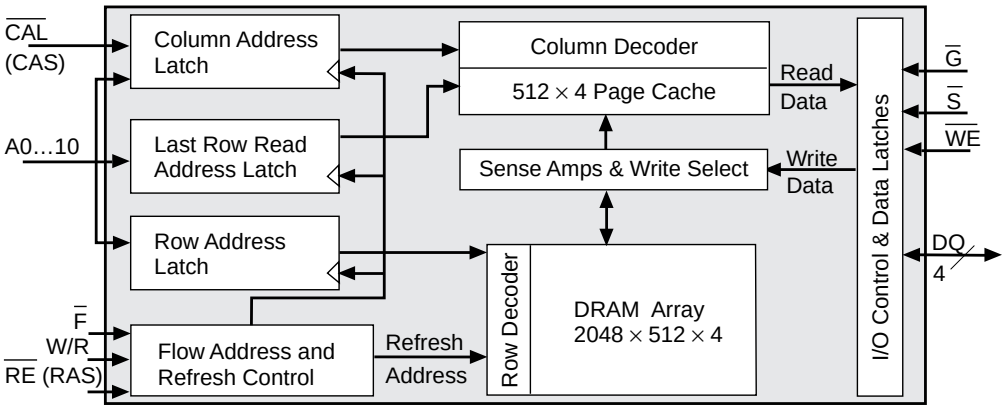

Figure 4. EDRAM block diagram.

