Synchronous DRAMs: The DRAM of the Future

Of all DRAMs manufactured today, approximately 70% are used in desktop and notebook PCs, where they are used to provide two different functions: main storage and frame buffers.

Main storage: Most PCs are offered with an L2 cache to bridge the processor/ memory performance gap. That makes the speed of the DRAM memory used for main storage an important but secondary consideration to price. However, as multitasking increases with large pro- grams, the frequency of L2 cache hits decreases. This degrades overall system performance, since the processor must wait for the DRAM to supply the requested data. To recover lost system performance, larger L2 caches or faster DRAM main memory are required.

Frame buffer: DRAMs used for frame buffering are wide I/O DRAMs, usually 256Kb x 16 DRAMs or 256Kb x 16 dual- ported DRAMs, commonly referred to as Video RAMs (VRAM). The access and cycle time of these devices are at par with commodity DRAMs, so a wide I/O interface is presented to the graphics controller to facilitate parallel processing of video data.

Due to increased chip size and the limited participation of memory suppliers, wide I/O DRAMs and VRAMs command a price premium. However, due to the rigid requirements of frame buffer size and screen refresh rate, these higher-priced DRAMs are still the best price-for- performance solution today.

To reduce system cost, chipset designers and PC manufacturers are investigating a unified memory architecture (UMA), under which main memory DRAM signals will be shared by the system memory controller and the graphics controller (see Figure 1). With UMA as a potential PC design objective, there is a need to develop DRAMs that provide high bandwidth and yet still remain inexpensive.

DRAM Evolution

The work horse of the DRAM industry has been the Fast Page Mode DRAM (FPM). The speed of FPM DRAMs (-60 sort) when performing random accesses within a given page of the memory reaches 28.5MHz— a significant mismatch to microprocessors that require data at rates up to 66MHz. A hierarchical memory subsystem employing an L2 cache plays a critical role in bridging this performance gap, but adds measurably to cost and complexity. A high-performance DRAM would be desirable.

DRAM speed improvements have historically come from process and photolithography advances. More recent improvements in DRAM performance, however, have resulted from making changes to the base DRAM architecture that require little or no increase in die size. The most recent example has been the introduction of Extended Data Output (EDO) DRAMs.

In EDO DRAMs, a minor change was made to the Fast Page Mode architecture to improve the DRAM's page mode cycle time. This, however, did nothing to improve any of the access parameters of the FPM DRAM.

EDO operation: Improvement in page mode cycle time was achieved by convert- ing the normal sequential fast page mode

Figure 1. UMA Block Diagram

operation into a two-stage pipeline. A page address is presented to the DRAM, and the data at that selected address is amplified and latched at the data output drivers. While the output buffers are driving this data off-chip, the address decode and data path circuitry is reset and able to initiate access to the next page address.

A comparison shows that a page mode cycle time of 60ns-sorted DRAM de- creases from 35ns (28.5MHz) for an FPM device to 25ns (40MHz) for an EDO device—a 40% increase in page mode performance. Since the implementation of the EDO concept was a modest variation to the FPM architecture, early suppliers have been able to support both functions (EDO and FPM) on the same silicon die, and memory controllers have been modified to accommodate both FPM and EDO DRAMs.

True random access DRAMs operating at 66MHz do not exist. Designers, however, can come close to the 15ns desired cycle by giving up some of the randomness in memory accessing.

Less Random Access

PC applications commonly access DRAM data in 4-bit burst lengths in either sequential or interleave fashion. Optimum system performance is achieved when data is fed to the processor on par with the system clock. Data accessed from the L2 cache comes close to meeting this requirement since only one wait state is inserted while accessing the first bit of data. However, the remaining three bits are delivered at par with the processor bus speed, denoted as a 2/1/1/1 burst rate.

When, however, the requested data must be accessed directly from the DRAM, the burst rate is significantly degraded. Assuming the burst is from a DRAM page that is already open (page hit), the achievable burst rate at 66MHz using 60ns DRAMs is 5/3/3/3 cycles for FPM DRAMs; for EDO DRAMs it is 5/2/2/2. This clearly illustrates the DRAM bandwidth bottle- neck. Even though EDO DRAMs improve page cycle times by 40%, the overall benefit to system performance is a
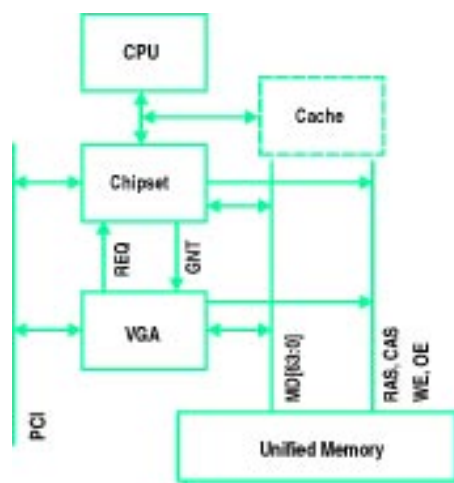
reduction of three wait states, with no improvement at all in accessing the first and most critical bit.

At the other extreme, graphics applications typically burst long streams of data to and from the DRAM. While performing operations to refresh the video display, entire pages of DRAM memory are read out in sequential and methodical fashion. Writeback or update operations to the frame buffer are more sporadic, but often performed in long sequential strings within a page—or even over multiple pages—of the DRAM.

The two scenarios above suggest that the DRAM itself should be rearchitected to deliver sequential data streams yielding faster page cycle times.

To accomplish this, a bursting technique has been incorporated into the DRAM data path. Once the first page address has been accessed, the DRAM itself provides the address of the next memory location to be accessed. This address prediction eliminates the delay associated with detecting and latching an externally provided address to the DRAM.

To implement this bursting feature, a few provisions to the DRAM architecture are required. First, a burst length and a burst type must be defined to the DRAM. In addition to the starting address, the burst length allows the internal address counter to properly generate the next memory location to be accessed. The burst type defines whether the address counter will provide sequential ascending page addresses or interleaved (scrambled) page addresses within the defined burst length. Secondly, a clock is used to increment the address counter and strobe the data off-chip.

The Arrival of Burst DRAMs
Two new DRAMs have been developed that employ this bursting technique: Synchronous DRAMs (SDRAM) were first developed, and were then followed by

Burst EDO (BEDO) DRAMs. Virtually every DRAM supplier is testing the market for—or developing—SDRAMs, with several suppliers already ramping up production.

By employing this bursting technique, the SDRAM architecture with an industry-standard LVTTL I/O interface is capable of delivering data off-chip at burst rates of up to 100MHz. Once the burst has started, all remaining bits of the burst length are delivered at a 10ns rate. Note, however, that the device's random access timing parameters are no better than FPM or EDO DRAMs.

SDRAM was developed to serve not only the commodity DRAM market (i.e. PCs), but was also designed to capture a share of the markets that require extremely high memory bandwidth. To allow users to configure the device for their specific applications, the DRAM's complexity grew in the following ways:

- To best optimize the device for a particular application, a mode register was added to the chip to enable the user to specify burst type, burst lengths, and CAS\ Latency.

- In addition, to support on-chip interleaving of data, the SDRAM architecture allowed two pages of the DRAM to be opened simultaneously.

- Finally, to ease system and setup/hold requirements on addresses and control pins—and to minimize data output skew from the DRAM—the DRAM interface was converted to a clocked or synchronous interface.

For 66MHz PC applications and higher, SDRAMs can reduce burst rates to 5/1/1/1 for L2 cache misses. However, until bus speeds exceed 66MHz, the premium that SDRAMs can command from this application is minimal, since no reduction in first access is gained and the benefit of 100MHz page cycle time cannot be fully exploited. Additionally, DRAM suppliers were not motivated to price SDRAMs at par with EDO devices, due to increased die size and manufacturing test times associated with SDRAMs.

As a result, memory designers returned to the EDO DRAM and incorporated burst features, naming this product Burst EDO DRAM (BEDO), and targeted it explicitly at the PC market. This device supports a burst length of four bits, and the burst can be delivered in sequential or interleave fashion. Again, no improvements were made in accessing the first page address, but subsequent bits are delivered at the required 66MHz rate to match the 5/1/1/1 burst rate of the SDRAM.

Synchronous DRAMs: Commodity DRAM of the Future?
Contention for market share is taking place between these two DRAM camps to determine which DRAM will take the lead in filling the 66MHz niche. More memory builders have supported SDRAM than BEDO to date. SDRAM suppliers—who, for over two years, have invested re-sources to develop and standardize the device at JEDEC and have educated the user community in how these devices operate—remain committed. Differences in functional specifications among suppliers are finally settling out.

Some major SDRAM suppliers are offering a less costly version of the SDRAM, reintroduced as SDRAM Lite. SDRAM Lite supports only a subset of the full SDRAM function set and is offered at 66 and 75MHz, but with improvements to several timing parameters to better compete with BEDO performance. In general, the 66 and 75MHz speed sorts will come from the slow end of the SDRAM performance distribution. This—coupled with a reduction in test cost due to the reduced function—will allow SDRAM Lite devices to be priced competitively with BEDO DRAMs.

Given a choice between BEDO and SDRAM at the same price for PC applications, SDRAM devices will most likely emerge as a migration path to the higher bus speeds already in place. BEDO will make a presence in 1996, but will probably fade in the following years.

The IBM SDRAM
The IBM SDRAM device meets or

exceeds all of JEDEC and industry standards. Its single-chip design supports three I/O organizations: 4M x 4, 2M x 8, and 1M x 16. The x4 and x8 devices are housed in a 400 mil wide, 44 pin plastic TSOP-type II package with 0.8mm lead pitch. The x16 device is packaged in a plastic 400 mil 50 pin TSOP-II package with 0.8mm lead pitch. IBM is sampling 100MHz devices of all three I/O organizations. Early user feedback has been enthusiastic about its performance, AC and DC characteristics, and power dissipation. Customer samples are available through IBM marketing representatives. IBM believes the SDRAM architecture is extendable across a variety of applications and therefore equips its device with the functions and features shown in Table 1.

Two Banks

The SDRAM architecture provides for two row addresses of the DRAM to be opened simultaneously. Memory accesses between two opened banks can be interleaved to hide row precharge and first access delays. In doing so, a seamless data rate of 100MHz can be achieved to read or write the entire device. Additionally, the IBM SDRAM can be optimized for specific applications, including set-top boxes.

Controlled Precharge / Auto Precharge

A row address in a given bank must be properly closed before a new access can begin to a different row address in the same bank. FPM and EDO DRAMs use controlled precharge, where the user must issue the necessary control signals to precharge column and row decoders.

The IBM SDRAM incorporates an auto precharge technique, which automatically closes the bank at the end of the burst operation. An extra address (A10) is used to define whether to invoke Auto precharge, depending on whether a bank is open or closed. Auto precharge is a convenience option for the user community, but in some instances may actually improve system performance. Suppose the device is programmed for a burst length of 4 bits and a CAS latency of 3 clock cycles. By interleaving accesses between banks A and B and by issuing READ commands with auto precharge, the device is able to sustain seamless data rates of 100MHz (10ns per bit) until the entire device is read or written. Furthermore, each burst of 4 bits can be obtained from different pages of the DRAM. By comparison, BEDO DRAMs can sustain a seamless data rate of 66MHz only if data is continuously accessed from the same page of the DRAM. This simple example

- Two banks

- User-programmable mode register
  - Programmable burst length: 1,2,4,8 or full page
  - Programmable wrap sequence: sequential or interleave
  - CAS Latency: 1,2,3
  - Multiple burst read with single burst write operation to support write through cache operation
    - Burst termination via a burst stop command or precharge command
    - Read and write command with or without auto precharge upon completion of the burst length
    - Auto refresh (CBR) and self refresh
    - Byte read and write control via a data-masking pin (DQM)
    - Standard LVTTL I/O interface

Table 1. Functions and Features of IBM's SDRAM

demonstrates the flexibility of the SDRAM device.

Other Features

Features that make the SDRAM common to existing DRAMs include auto refresh (referred to as "CAS\ before RAS\ refresh" on FPM and EDO DRAMs) and self refresh. Self refresh allows the device itself to generate the control signals necessary to refresh the storage cells within the allotted retention interval.

Summary

Both memory manufacturers and system designers are challenged to provide main storage and frame buffer memory whose access speed is sufficient to meet the capability of the processor. At the same time, the industry is investigating UMA as a means to lower the cost of a PC. Of the approaches to creating faster memory, IBM is convinced that SDRAM will eclipse BEDO in a short time. Thus, IBM has developed highly flexible SDRAM devices that meet both JEDEC and industry standards. ■

Summary of Memory Data Rate for a 66MHz Processor Cycletime with a Memory Page Hit, Ranked Fastest to Slowest

Memory Type	Latency (processor cycles per bit delivered)			
	bit 1	bit 2	bit 3	bit 4
L2 (SRAM - Static Random Access Mem.)	2	1	1	1
SDRAM (Synchronous Dynamic Random Access Mem.)	5	1	1	1
BEDO DRAM (Burst Extended Data Out DRAM)	5	1	1	1
EDO DRAM (Extended Data Out DRAM)	5	2	2	2
FPM DRAM (Fast Page Mode DRAM)	5	3	3	3