AN1223

A Zero Wait State Secondary Cache for Intel's Pentium™

Prepared by: Michael Peters, FSRAM Applications Engineer

Due to the increased complexity and sheer memory size requirements of new and forthcoming operating systems (OS), graphical user interfaces (GUI) and application programs, the demand for ever-increasing performance from the desktop machine continues. Next generation machines require more and faster memory. Microsoft's Windows NT™, for instance, will most likely need 12 to 16 MBytes of main memory. Cache size requirements follow accordingly. And Intel's new Pentium CPU has been introduced with external bus speeds of 60 MHz and 66 MHz.

High performance memory is essential in achieving Pentium's full potential. First level (L1), on-chip cache memory hit rates will suffer as a result of users' migration away from DOS to Windows to Windows NT. It has been shown that L1 cache hit rates decrease mainly due to the increased number and types of references demanded by the newer OS.[1] The CPU designer can only afford relatively small increases in L1 cache size in an effort to keep chip size down. So, second level (L2) cache must make up for the lack of an appropriately sized cache and significantly help to avoid time consuming DRAM accesses. In addition, at 60/66 MHz bus speeds, the L2 cache must be capable of reading and writing data fast enough for Pentium's superscalar design.

Motorola's new families of 64Kx18 and 32Kx18 Fast SRAMs establish a new standard in providing a big enough and fast enough data cache for Pentium designs. These families include five synchronous and two asynchronous devices in each family. All x18 SRAMs feature byte-write capability, 3.3 V I/O compatibility, and asynchronous output enable control. A zero wait state solution is possible using four MCM67B618 (or four MCM67B518) BurstRAMs™. The objective of this note is to explain some of the system level, electrical, and timing issues associated with the design of a zero wait state secondary cache.

BurstRAMs vs. ASYNCHRONOUS SRAMs

Although the i486™ and Pentium CPUs support a burst cache line fill protocol, in most cases building a zero wait state bursting cache with a single bank of ordinary SRAMs is simply not practical. Virtually all cache controllers/chipsets designed to work with the i486 accommodate the burst protocol by using an interleaved scheme of two banks of standard asynchronous SRAMs. The speed requirements for this type of caching arrangement allow the use of 20 ns through 35 ns SRAMs. These speeds accommodate 20 through 33 MHz i486 machines, the bulk of today's IBM-compatible PC market. For the i486's 32-bit bus speeds less than 50 MHz, this hook-up is technically feasible, but somewhat expensive and physically large, and it consumes a good deal of power since as many as eight SRAMs are required. However, Pentium's 64-bit bus and bus cycle rates of 60 MHz and faster only exacerbate the difficulties with single and double bank caches using ordinary asynchronous SRAMs. Most chipset vendors will find that the use of synchronous burstable SRAMs will be the only practical zero wait state solution for Pentium.

A single bank scheme must use either extremely fast RAMs (< 7 ns for a 60 MHz bus) or add wait states. With the added wait states, a single bank 3-2-2-2 (three lead-off clock cycles and two clock cycles for each subsequent read) design might still require 12 ns standard SRAMs.

A double bank scheme can be designed with wait states or for high speed with no wait states. Figure 1 shows the timing for a 3-2-2-2 design using sixteen 15 ns 32Kx8 (or x9) SRAMs in a two bank design.

The cache can be expected to consume about 8.6 W. Two banks of 12 ns standard 32Kx8 (or x9) BiCMOS SRAMs might achieve 3-1-1-1 burst, but at an even greater power premium — nearly 12 W. In two bank schemes, even when one bank is de-selected, it will still draw about 65% of the full operating current.

Double bank designs present other issues that must be considered, including address and data bus loading, physical layout, and socketing devices. Two banks of 32Kx8s will present an 80 pF load (plus routing) to the cache controller's address bus. These heavily loaded lines represent additional signal delay and power dissipation compared to a BurstRAM design. And, one cannot afford a 5 ns buffer delay in the address path. When comparing the BurstRAM's 52-lead PLCC package with a standard 32Kx9 SOJ, direct mounting of these devices on a board will yield roughly four square inches versus eight square inches, respectively. Socketing the SRAMs is ill advised since access time will be pushed out, and signal integrity may be compromised.

Although designing caches with asynchronous SRAMs can be done, the control signal timing is far from easy. Of all timing concerns, write pulse generation may be the biggest issue. Burst writes may be next to impossible to perform since both edges of the write pulse must be positioned precisely to accommodate address set-up and data hold times. One can expect 10 ns minimum write pulse widths for 12 ns asynchronous SRAMs; this does not leave much time for the 15 ns cycle processor bus.

Freescale has developed a series of 256Kbit, 512Kbit, and 1Mbit SRAMs, known collectively as BurstRAMs, to solve these problems.[2]

BurstRAM is a trademark of Freescale Semiconductor, Inc.
i486 and Pentium are trademarks of Intel Corp.

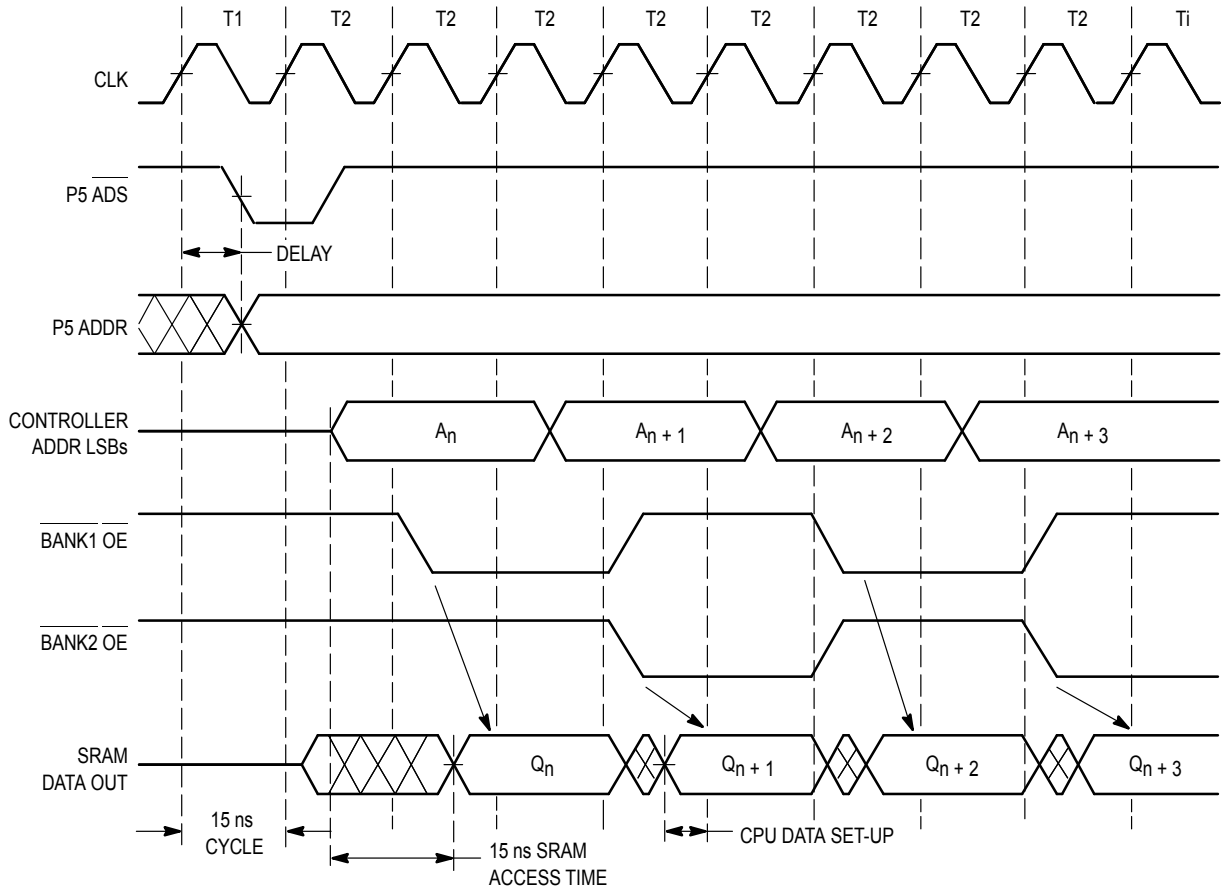


Figure 1. Two Bank Asynchronous SRAMs Performing 3-2-2-2 Burst READ

The MCM62486, a 32Kx9 BurstRAM, was developed for i486 systems. These BurstRAMs are being used in many of the 50 MHz i486 systems built today. The MCM67518, a 32Kx18 device, and the MCM67618, a 64Kx18 device, are the best suited for Pentium-based designs. Key to the success of a zero wait solution is the SRAM's support of Intel's burst protocol. A 2-1-1-1 (zero wait state) burst read cycle can be performed at cycle times of 20 ns and less. Pipelined addressing can further reduce a burst cycle to a 1-1-1-1 count. The MCM67B618 and MCM67B518 are synchronous BiCMOS SRAMs that feature wide x18 data paths, burst reading and writing, byte-write capability, 3.3 V I/O compatibility, and asynchronous output enable control. Note that all BurstRAM operations occur on the rising edge of clock (CLK).

Four (4) MCM67618 devices provide a single bank of 512K byte L2 cache. The interface to the Pentium chip is a direct connection for address and data paths. These new BurstRAMs (MCM67B518, MCM67B618) have been designed to operate at clock rates of up to 66 MHz (15 ns cycle time). They are available in access times of 9/12/18 ns with cycle times of 15/20/30 ns, respectively. The term "access time" is used loosely for synchronous SRAMs and is more accurately, CLK-to-VALID DATA time.

WHAT IS A BurstRAM™ ?

BurstRAMs are synchronous SRAMs that contain input registers for address, write, and enable signals and have an on-chip burst counter that imitates the i486 and Pentium's lower order address burst count. These control signals are registered into the BurstRAM on the rising edge of the CLK input. Three (3) control pins allow complete control of the burst function. ADSP (ADS Processor), ADSC (ADS Controller), and ADV (ADVance) control the burst read/write functions as well as single read/writes. A self-timed write is also provided for the purpose of simpler (and relaxed) write timing. Byte-write capability is provided with the UW and LW (Upper/Lower byte Write) signals. Note that all control signals are active low. See Figure 2.

THE BURST CYCLE

A burst read cycle is performed as follows (see Figure 3):

1. During the first cycle (T1), the CPU generates ADS and a valid address, and the BurstRAMs register the external address A<18:3> and enable on the rising edge of the system clock (CLK). This address can be considered the base address from which the BurstRAM begins its address counting,

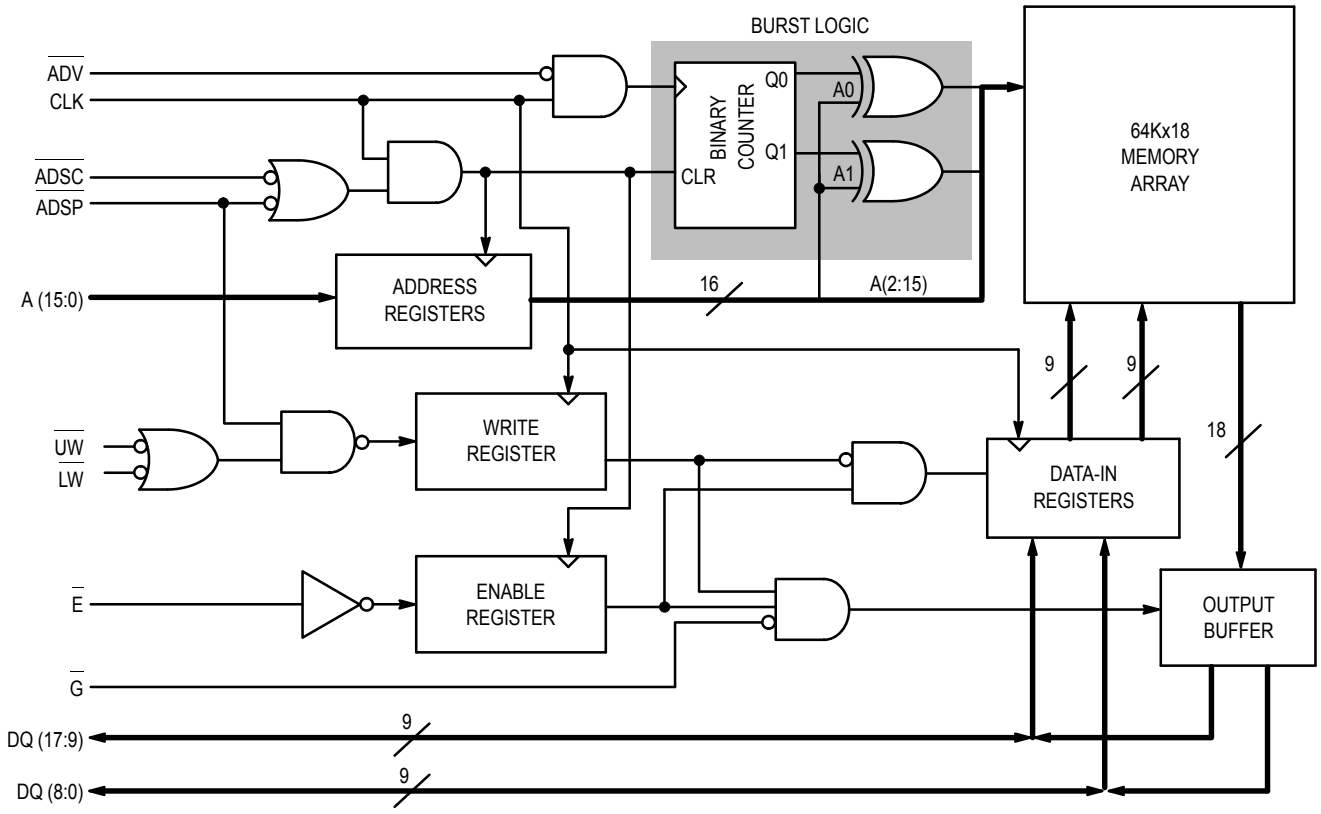


Figure 2. Block Diagram of 64Kx18 BurstRAM

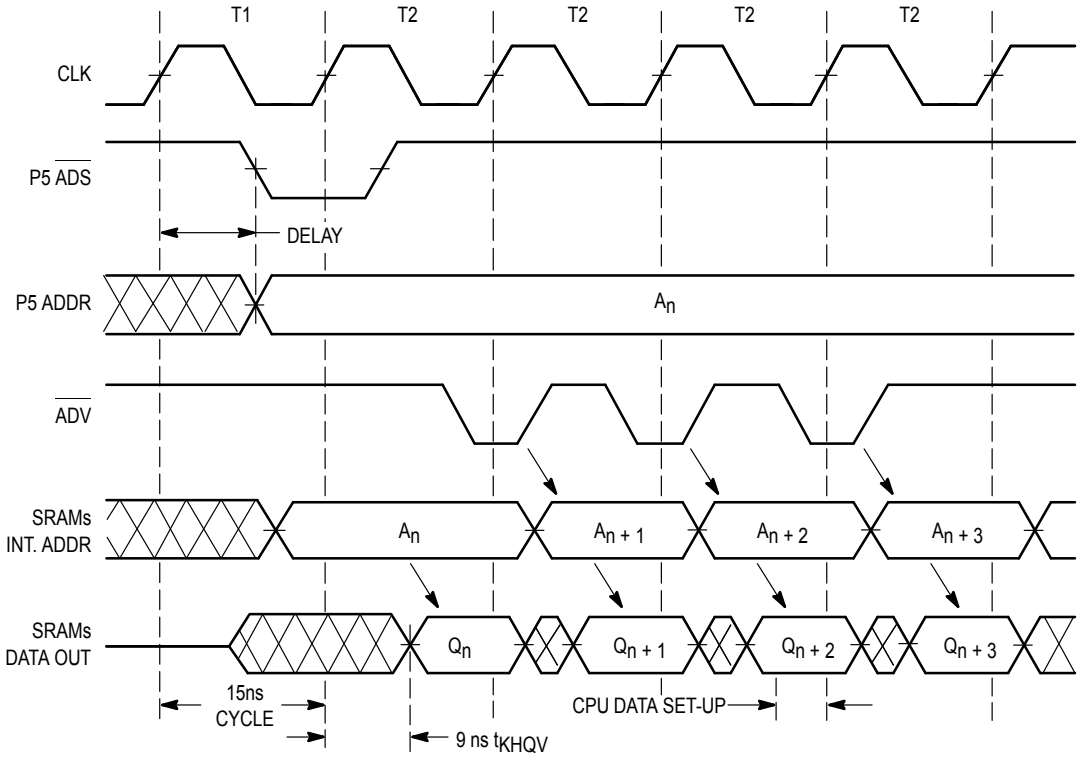


Figure 3. 64Kx18 BurstRAM Performing 2-1-1-1 Burst READ

2. Assuming the cache controller has determined that the cycle is a cache hit, the first 8 bytes of valid data are driven onto the data bus 9 ns after the second rising clock edge,
3. Subsequent cycles present valid data upon the negation of ADS and the assertion of ADV. An entire 32 byte cache line can be supplied to the CPU in just five cycles. The BurstRAM's output enable (G) can be asserted well into the 2nd cycle since it is asynchronous and represents only 5 ns delay.

Pentium operates with external bus speeds of 60 MHz and 66 MHz. This corresponds to 16.6 ns and 15 ns cycle times, respectively. Standard asynchronous SRAMs are hard pressed for a zero-wait state application. A look at the timing reveals that sub-12 ns SRAMs would be required since Pentium's data set-up time is about 3 to 4 ns. The inclusion of on-chip logic allows the BurstRAM to be directly connected to the CPU, and avoids the timing penalty associated with glue logic.

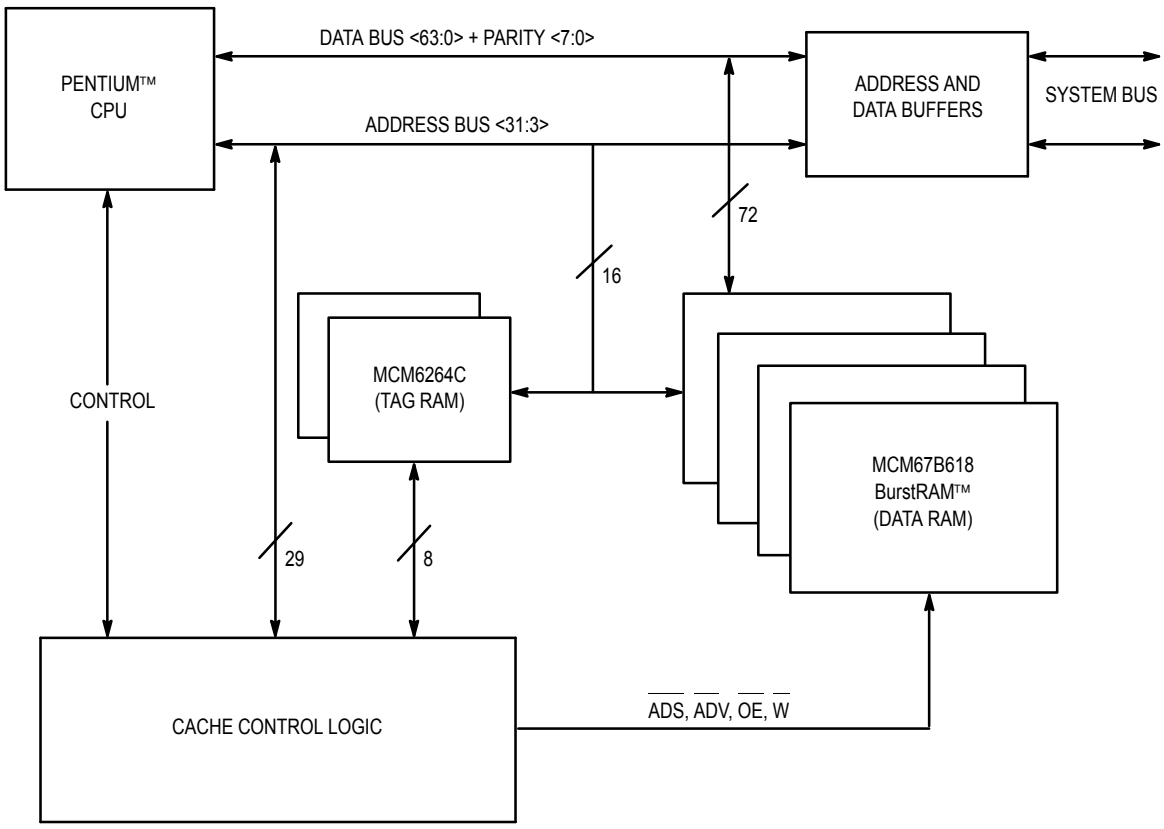
Using the BurstRAM, a zero wait state burst write cycle can be performed as well. Upon the CPU's assertion of ADS, the BurstRAM begins and completes a burst write cycle with the assertion of E, LW, UW, and ADV signals. A burst write cycle can be started using either ADSP or ADSC. If ADSC is sampled low (while ADSP is high), data can be written immediately to the BurstRAM while ADV is asserted on subsequent cycles for the completion of the burst cycle. If ADSP is sampled low (while $\overline{\text{ADSC}}$ is high), the write register is blocked inside the BurstRAM and consequently only allows A<15:0> and E to be registered. On the following cycle (ADSP and ADSC negated), the burst write operation begins assuming LW and UW have been asserted. Again, ADV must be asserted on subsequent cycles to complete the burst cycle.

The use of a synchronous SRAM makes a design simpler in the sense that address and control signals can have looser timing constraints since they are registered in, and the SRAM does the rest. As long as DQ<17:0>, LW, and UW signals comply with the required set-up (2.5 ns) and hold (0.5 ns) times, complex off-chip write pulse generation can be eliminated. An undue burden will be placed on the controller to provide proper write pulse width and write timing edges relative to address and the CPU's valid data.

SYSTEM CONFIGURATIONS

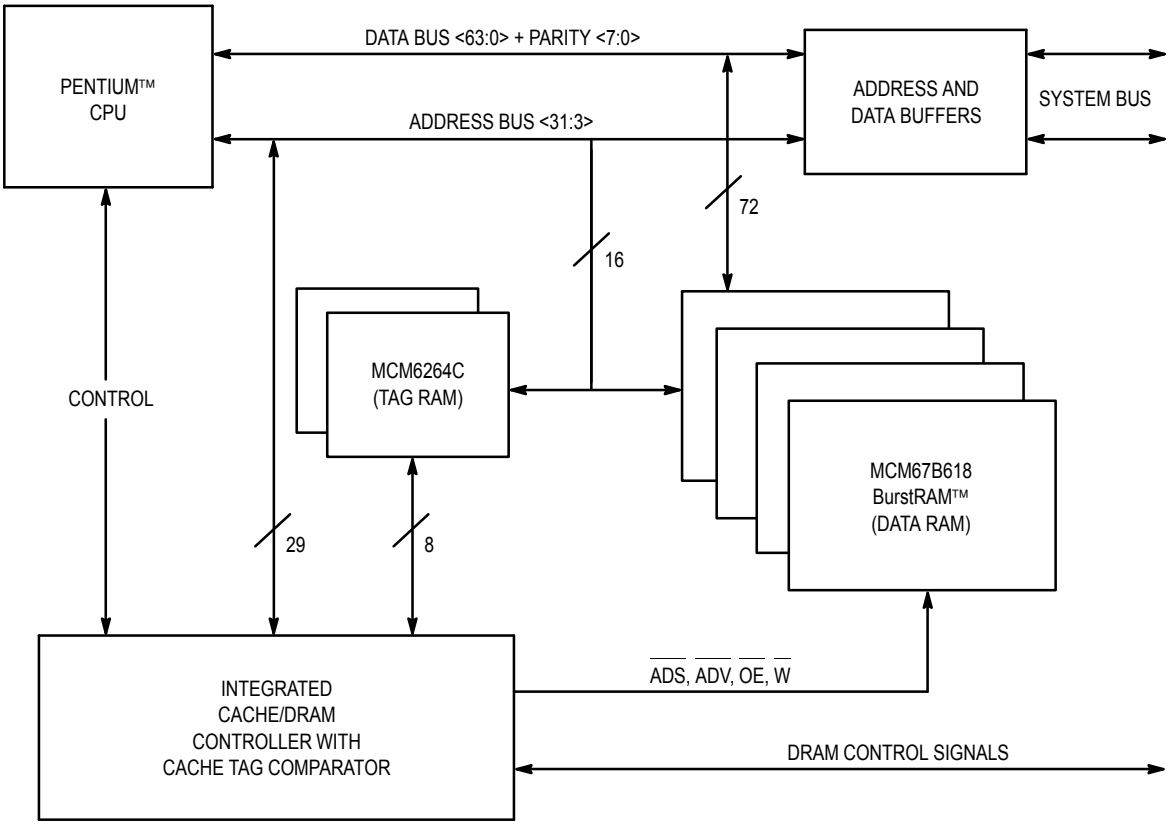
Pentium's 64-bit data path will require four (4) MCM67B618s (or MCM67B518s) to provide a single bank 512K (256K) byte L2 cache. The interface to the Pentium chip is a direct connection for address and data paths. Control signals must come from the cache controller. See Configurations A/B/C of the System Block Diagrams.

Configuration A is the least integrated solution, one that uses external tag RAM and a PAL or ASIC for the cache controller. The DRAM controller would be yet another component in the system.



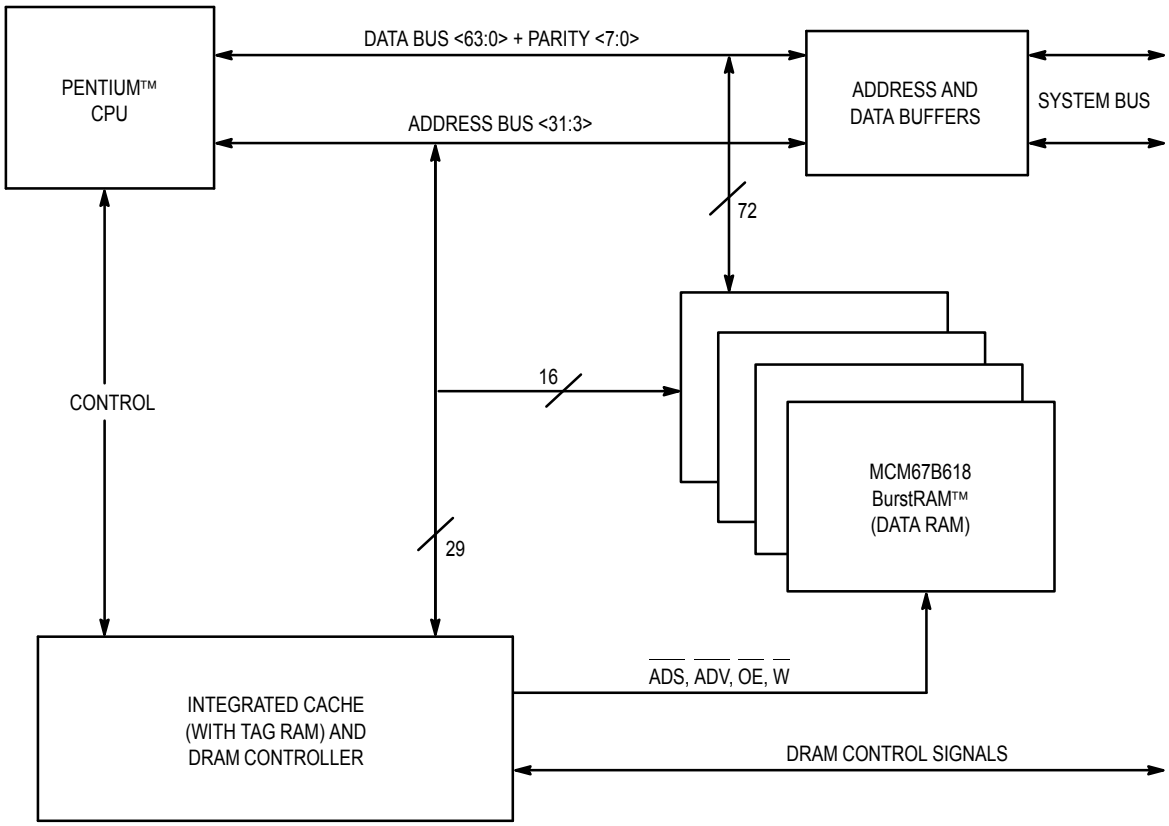
Configuration A
Secondary Cache Solution for Pentium — 512KByte

Configurations B and C are the most likely approaches taken by chipset vendors in which the tag RAM may or may not be integrated, but will probably integrate the DRAM control. For direct-mapped caches such as these, tag RAM size depends on the controller's mapping of tags (or sectors) to cache lines. Each sector may consist of 1, 2, 4, or more cache lines. Tag RAM depth is then 16K, 8K, 4K, or so, respectively.



Configuration B
Secondary Cache Solution for Pentium — 512KByte

The tag RAM must be at least 10 ns for zero wait state performance; otherwise, a lead-off wait state must be added (3-1-1-1). This is determined by the speed of the controller's tag comparison as well. If the cache line size is 32 bytes and the data RAM depth is 64K, the tag RAM will have to be a 16Kx8/10 or 4Kx8/10 organization. The tag RAM's width (data path) is a function of the system's main memory size. An 8-bit tag will allow a cache size of 512KB to cache 128MB of main memory.



Configuration C
Secondary Cache Solution for Pentium — 512KByte

FEATURES OF 64Kx18

The 64Kx18 SRAMs are fabricated on a BiCMOS process and exhibit less dependence on output loading compared to CMOS devices. These SRAMs are powered on a single 5 V supply (± 5%) and are 3.3 V I/O compatible — no additional power supplies are required. The output buffer is composed of an NPN pull-up and an N-channel MOS pull-down. The pull-up circuitry has been carefully designed to limit the NPN's base drive such that the output pulls up to approximately 3.3 V even under high supply conditions (e.g., 5.25 V). These 3.3 V "friendly" output buffers have controlled 3.3 V output swing and will not overdrive a future 3.3 V controller or processor. This important feature allows one to easily migrate from an all 5 V system to a mixed 5 V – 3.3 V system upon the availability of 3.3 V Pentium and controller chips.

SYSTEM CONSIDERATIONS

The entire 64Kx18 SRAM family makes use of multiple power and ground pins on the 52-lead PLCC package. Five (5) power and five (5) ground pins (6 pairs for the asynchronous devices) have been provided to allow adequate supply decoupling and return current paths for such a fast device. Multiple power and ground pins reduce the effective inductance of theses connections. Since the output buffers swing 3.3 V in 1 to 2 ns (t_r/t_f), significant di/dt currents flow in the V_{CC} and V_{SS} pins. Separate power and ground planes on the printed circuit board are highly recommended and will help improve signal integrity, ground bounce, and in turn the SRAM's access time. The use of a 0.001 µF or 0.01 µF chip capacitor or similar leadless (surface mount) capacitor connected within 0.5 inch or so of each pair of V_{CC}/V_{SS} pins will provide a low impedance path for the fastest transients. A single 1 to 4.7 µF chip or ceramic capacitor per device should be sufficient for dc stability.

The use of standard (asynchronous) SRAMs may prove to be very difficult to use in 50+ MHz systems due to the requirements of carefully controlling the signal integrity, maintaining good noise margins, keeping component count down, and reducing board space. Because the BurstRAM, a synchronous device, registers address and control signals during a very brief moment during the system cycle, noise occurring throughout most of the cycle in the system can be tolerated by the BurstRAM. Component count, and therefore board space, is reduced since these SRAMs integrate the burst counter logic and self-timed write circuitry onto the chip and, in addition, have a wide (x18) data path. Because of the on-chip logic, cache control logic can be simplified and some control signal timing can be relaxed.

In cases that demand detailed timing analysis and a close look at the analog effects of your board design, it is recommended that a board-level (Quad Design/Viewlogic) or SPICE simulator is used. Particularly when PCB routing lengths are about 4 inches or more, transmission line effects become dominant over the lumped circuit equivalent. Since interconnect time-of-flight is approximately 175 to 190 ps/inch, a 4 inch route adds about 0.75 ns to a memory access.

When analyzing the cache data read path, the DQ<17:0> are in their active state and drive the data bus. The characteristics of these output pins are important to know when completing a board's physical layout. Use the information in Table 1 (output buffer I-V data), Table 2 (input I-V data), and Table 3 (package parasitics) to help verify your timing and loading effects. This tabular data may be used directly as input to board level simulators, such as those offered by Quad Design, Integrity Engineering, Quantic Labs, etc. Figure 4 shows how to connect the parasitic package components between the chip (output buffer or input) and package pin. An input pin on the 64Kx18 can be modeled as C die = 4 pF.

Table 1. I-V Characteristics of the 64Kx18 I/O Buffers

V_{OL} (V)	I_{OL} (min) (mA)	I_{OL}(max) (mA)	V_{OH} (V)	I_{OH} (min) (mA)	I_{OH} (max) (mA)
0	0	0	0	– 110	– 145
0.5	38	60	0.5	– 106	– 136
1.0	68	107	1.0	– 96	– 124
1.5	90	137	1.5	– 78	– 102
2.0	104	154	2.0	– 55	– 77
2.5	110	160	2.5	– 29	– 45
3.0	112	162	3.0	– 7	– 13
3.5	113	163	3.5	0.3	0.2
4.0	114	164	4.0	0.7	0.6
4.5	115	164	4.5	1.4	1.3
5.0	115	164	5.0	2.0	2.0

Table 2. I-V Characteristics of the 64Kx18 Inputs (Address and Control)

Diode to GND		Diode to V_{CC}	
V_{in} (V)	I_{in} (mA)	V_{in} (V)	I_{in} (mA)
0	0	5.0	0
– 0.4	0	5.4	0
– 0.5	0	5.5	0
– 0.6	0	5.6	0
– 0.7	– 0.1	5.7	0.1
– 0.8	– 2.0	5.8	2.1
– 0.9	– 25	5.9	20
– 1.0	– 70	6.0	50

Table 3. Packaging Characteristics

	Min	Max	Unit
R package	50	200	mΩ
L package	3	6	nH
C package	0.5	1.0	pF
C die	2	7	pF

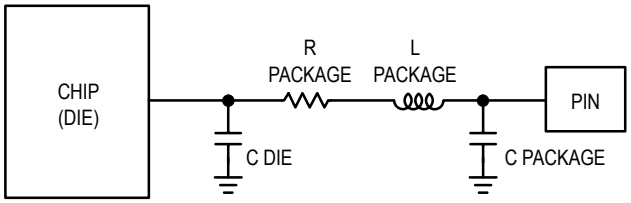


Figure 4. Package Parasitics Schematic

OUTPUT BUFFER CHARACTERISTICS

The access times guaranteed in the datasheet are based on a 50Ω test load and should be derated for unterminated CMOS loads. Refer to the derating curve (Figure 5) for your application. This curve relates the difference in access time between a 50Ω test environment and a lumped capacitive load (no dc load) condition typically found in most applications. The curve is based on worst case conditions, i.e., V_{CC} = 4.75 V and T_A = 70°C. Note that the 50Ω test condition is equivalent to a lumped 10 pF load. For instance, if the BurstRAM outputs see a 30 pF load, derate the access time by about 0.4 ns. So, for a Pentium design that uses the MCM67B618 – 9 ns device, one can expect a worst case access time of 9.4 ns under these conditions.

SUMMARY

For high performance Pentium systems, the use of Motorola's 64Kx18 BurstRAMs provides a straightforward solution to Pentium's secondary cache requirements. Four BiCMOS BurstRAMs support the size and speed required by zero wait state Pentium systems. For equivalent cache size and performance, standard SRAM solutions warrant two bank interleaved approaches that utilize more board space, require more power, and demand a higher performance cache controller.

REFERENCES

1. AP-469: "Cache and Memory Design Considerations for the Intel 486DX2 Microprocessor", Intel Corp.
2. DL156/D: *Fast Static RAM BiCMOS, CMOS, and Module Data*, Freescale Semiconductor, Inc.

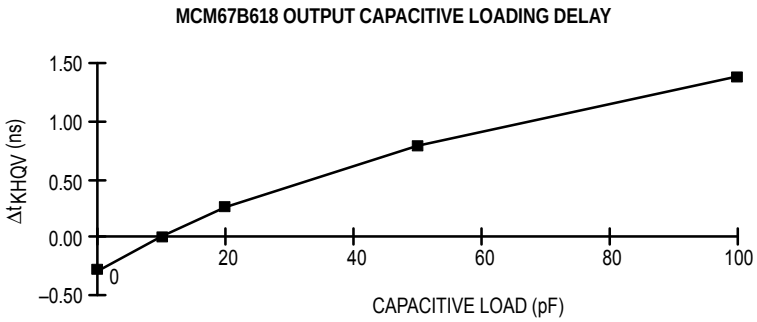


Figure 5. Access Time Derating Curve