High-Speed Memory Design Techniques

AN4010

Introduction

Over the last 20 years, microprocessor clock speeds have increased at an exponential rate. Clock speeds have migrated from 1 MHz with the Intel® 4004 to 2 GHz with the recent version of the Pentium® processor. A parallel increase in bus speeds has occurred with current systems having cycle rates of 60 MHz and 66 MHz and moving toward 2.5 GHz.

Design techniques which were used at slower frequencies (33 MHz and lower) are no longer appropriate for these higher bus frequencies. More attention must be paid to board layout, bus loading and termination to ensure that short clock cycle times can be met without noise, ringing, crosstalk or ground bounce. This application note discusses these issues and the choices a designer faces in high-speed memory system design.

The Memory Hierarchy

Figure 1 shows the memory hierarchy conventionally used in a computer system. High-speed cache memory integrated with the microprocessor is used to store frequently accessed instructions and data and to avoid the time penalties associated with off-chip accesses. However, only a limited amount of cache can be included directly on the chip in the level one (L1) cache (sizes vary from 8 KB to 128 KB). High-speed secondary or level two (L2) cache is included in systems to increase system performance. Sizes for the L2 cache vary depending on the application software. The largest portion of data, stored in the DRAM bulk memory array, is significantly slower and has a large access time penalty. If a cache miss occurs, retrieving data from the DRAM array could take up to six or more processor clock cycles, drastically reducing system performance.

High-speed techniques must be used in evaluating data transfers between the cache SRAM and microprocessor. Timing between the cache and microprocessor is especially critical because of the short cycle and access time required. High bus frequencies require careful design to attain zero-wait-state performance. At these frequencies designers increasingly use synchronous SRAMs to help alleviate timing problems, but even synchronous SRAMs require a thorough timing analysis.

A Cache Timing Example

The following example demonstrates how little timing margin is available for these higher bus speeds and why it is important to analyze bus timing carefully when designing memory subsystems at these frequencies. For this example, we assume that the microprocessor operates at 60 MHz and is using a synchronous cache array organized as 64K x 72, with two 64K x 36 Standard Synchronous SRAMs as shown in Figure 2.

Figure 1. Typical Microprocessor Memory Configuration

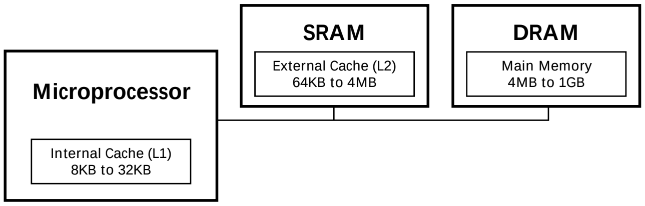
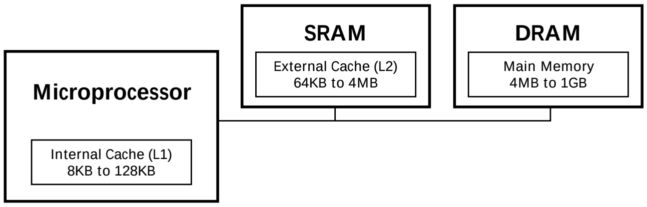


Figure 2. Synchronous Cache Memory Design Using Two Cypress 64K x 36 SRAMs



The timing of a READ cycle for this system is shown in Figure 3. The equation below shows how to calculate the amount of timing margin available in any design. The variable ^{t}CLK represents the clock cycle time of the external bus. For a 60-MHz system this represents a 16.7-ns cycle time.

$t_{margin}= t_{clk} – t_{flight} – t_{setup} – t_{access}$

$= 16.7 – 1.4 – 4 – 10$

$= 1.3$ ns.

In this example a READ cycle is being performed which sends data from the SRAM cache memory to the microprocessor. This example assumes that the address and control signals are valid during the positive edge of the clock pulse and exceed the setup time of the SRAM. In a synchronous system the memory clock cycle begins with the rising edge of the clock which signals the SRAM to use the address on the bus, find the data stored at this address and send it to the outputs. The data appears at the output t_{access}ns later (10 ns for this example). Once data appears at the output, it must travel from the SRAM to the microprocessor through signal traces on the circuit board. This transfer time is called t_{flight} and can vary greatly. Lastly, the microprocessor must latch the data and it must be available to meet the processor setup time (t_{setup}). The hold time must also be met, but it occurs after the rising edge of CLK and does not have to be subtracted as part of the timing calculation.

Altogether the total time is 15.4 ns and the requirement for 60-MHz operation is anything less than 16.7 ns. The margin for error is 1.3 ns, and board layout or other factors can easily exceed this. In the next section we will discuss how t_{flight} can vary. Even with careful design, a t_{flight} of less than 2 ns may be very difficult to obtain. Typical times in some designs could be 3 ns to 5 ns or more. It is no longer sufficient just to connect components without considering the timing impact to the system.

Calculating t_{flight}

t_{flight} consists of the components shown in the equation below:

$t_{flight} = t_{clock\ skew} + t_{propagation\ delay} + t_{rise\ time}$.

The first component, $t_{clock\ skew}$, can be defined as the skew between rising and falling edges of the clock signal for different components on the board. If a clock rises at time t = 0 on the microprocessor clock input, the clock input to the first SRAM might rise at time t = 0.25 ns and t = 0.45 ns on the second. This skew in timing can be due to uneven line lengths or varying load capacitances on the different lines. If a series of buffers is used to distribute the clock signal, delay times through these buffers will also vary and add to the skew. These types of skew are frequently ignored for slower systems but must be accounted for in high-performance ones.

The second component, $t_{propagation\ delay}$, is determined through the characteristics of the transmission line and line load. The propagation delay now consumes a considerable portion of the cycle time of a high-speed system and can no longer be ignored. Designers cannot assume that outputs drive purely capacitive loads and must determine if interconnects should be treated as transmission lines. A purely capacitive load assumes an RC time constant delay consisting of trace resistance, output driver resistance and total lumped capacitance. Transmission line analysis, although more difficult, more accurately reflects actual conditions. Determining propagation delay is discussed in more detail in the next section.

The next component, $t_{rise\ time}$, is determined by the speed of the component driving the line. A faster rise time can help speed the cycle time of a system but may require a huge output driver with a large current dissipation. Rise times can also vary from component to component and worst-case times should be used for design analysis.

A component that should not be ignored is circuit loading. The external capacitive loading is usually accounted for in the access time of the device (t_{access}). A device will have an access time rating that is valid up to a given loading. For example, high-speed synchronous SRAMs are usually rated with an AC loading as shown in Figure 4. Designers can modify their timing margin if the capacitive loading is less than or greater than the specified rating.

Circuit Termination

Unterminated Lines

Because electrical signals travel at a finite velocity through a circuit board, it is necessary to determine how long they take to propagate from driver to receiver. This length of time determines if the output circuit requires termination. As an example, assume that a circuit board uses a polyimide dielectric with a relative dielectric constant (e_r) of 3.5. Common dielectric constants are shown in Table 1.

Figure 3. READ Cycle to Cache Memory

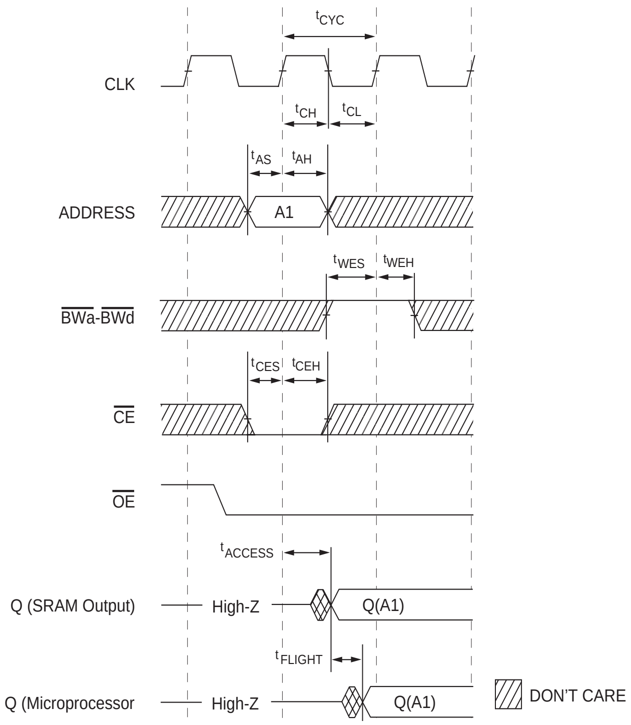
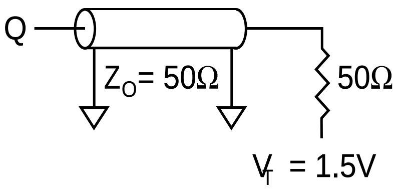


Figure 4. Output Load Equivalent



If the circuit board has a strip conductor and a ground plane separated by a dielectric medium as shown in Figure 5*a* (microstrip line), we could use the following equation to calculate the signal speed:

$$T_d = 0.004\sqrt{0.45\varepsilon_r + 0.67}\ \text{ns per mm}$$

For this circuit board the equation gives us a signal velocity of 6 ps/mm. If the signal conductor were instead sandwiched between two power planes (Figure 5*b*), we could use a stripline equation to calculate the signal velocity.

$$T_d = 0.004\sqrt{\varepsilon_r}\ \text{ns per mm}$$

For the same circuit board we now have a signal velocity of 7.5 ps/mm. The equations above give valid results for reasonable values of trace widths, dielectric constants and dielectric thicknesses. Books are available on transmission line theory for a detailed analysis of propagation delay for various structures.

The purpose of calculating a delay time for signals is to determine if the circuit delay can be treated as an RC time constant. This can be done if the maximum trace length meets the following inequality:

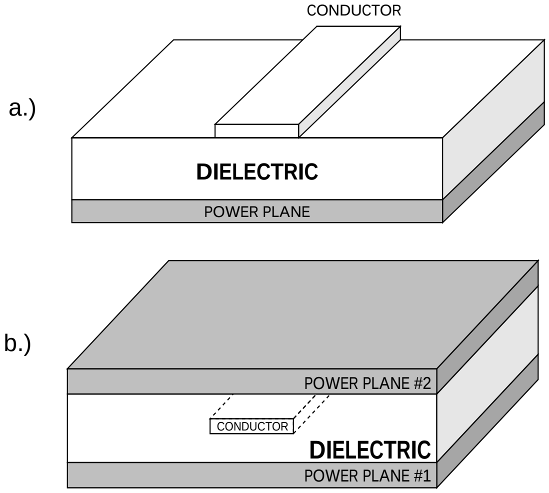
$$L_{MAX} < \frac{t_r}{2T_d}$$

t_r = Rise time or fall time of the output driver.

Table 1. Various Dielectric Constants

Material	Relative Dielectric Constant (e_r)
Ceramic – Alumina	9.4
FR–4	4.7
Glass Epoxy	4.0
Polyimide	3.5
Silicon	11.8

Figure 5. a) Microstripline, b) Stripline



For our example we assume a rise or fall time of 2 ns. This means that as long as the maximum line length is less than 133 mm or 5.2 inches (for stripline), the circuit delay can be treated as an RC time constant. For a given capacitive load we can calculate the actual signal propagation time using the equation:

$$T_{ld} = T_d\sqrt{1+\frac{C_L}{C_o}}\ \text{ns per mm}$$

C_L = Load Capacitance

C_O = Transmission Line Capacitance

where:

$$C_o < \frac{T_d}{Z_o}$$

Z_o = Characteristic impedance of the signal trace.

For our stripline example, C_O = 1.5 pf/cm and the load is C_L = 50 pf with a 5-cm transmission trace length. The actual signal velocity is 22 ps/mm or 1.1 ns for 5 cm of line length. For this example we assumed a value of Z_O of 50 Ohms. This value can be calculated using the equations below, or supplied from a board vendor. It is recommended that a designer use computer software to determine Z_O instead of these equations which are only approximate.

For stripline:

where:

$$Z_o = \frac{30\pi\,(1-t/b)}{\sqrt{\varepsilon_r}\,(W_e\,/\,b+C_f/\pi)}$$

$$C_f = 2\ln\left(\frac{1}{1-t\,/\,b}+1\right)-\frac{t}{b}\ln\left\{\frac{1}{(1-t\,/\,b)^2}-1\right\}$$

$$W_e = W - \left\{\frac{(0.35 - W/b)^2}{1 + 12t/b}\right\}$$

W = width of strip conductor

t = thickness of strip conductor

b = dielectric thickness.

These stripline equations are relatively accurate if the following limitations are met:

$$0.05 \le W/(b\text{-}t) \le 0.35$$
$$t/b \le 25$$

Terminated Lines

If line lengths are greater than L_{MAX}, the above equations can no longer be used and terminations should be considered. For these situations users should use simulation tools to define and analyze their distributed element circuit accurately. Several considerations for terminated lines are discussed in this section.

When transmission line analysis is used, the designer must determine if the design will use incidence-wave or reflected-wave switching. Incidence-wave switching is potentially the quickest way to drive external devices because it does not depend upon the reflected signal to exceed V_{IH} or V_{IL}.

Incidence-wave switching has the drawback that large amounts of power can be generated in the output driver of a chip. Let's calculate what the power of one I/O signal can be using this method. First, the effective characteristic impedance must be calculated using the equation below:

$$Z_L = \frac{Z_O}{\sqrt{1+\frac{C_L}{C_O}}}$$

For our example in the previous section, using a value of Z_o of 50 Ohms and a transmission line of 10 cm, Z_L is equal to 24 Ohms. We can use this number and the output impedance of the driver to determine the instantaneous switching current of the outputs. If the output driver impedance is 25 Ohms, we will see an instantaneous current of 100 mA. A circuit implementation using wide devices could generate a large noise spike that would be very difficult to decouple. The reflection at the end of the line must also be accounted for.

Reflected-wave switching can cut power and noise dramatically because the driving circuit needs to generate only half the output voltage upon switching as the incidence-wave solution. The signal initially propagates at half the required voltage level until it hits the end of the transmission line. Then the reflection causes the voltage level to double. Because the reflected-wave is used, time must be allotted for the reflected-wave propagation. This method was adopted for use in the Peripheral Component Interconnect (PCI) bus.

Types of Termination

Several types of termination are commonly used in designs. The first (shown in Figure 6*a*) is series-resistance termination. A resistor is connected between the output of a driver and the driven elements. A reflected wave from the load which reaches the output can again be reflected, generating noise in the output signal. The series-termination is used to prevent this type of reflection. This is done by making the output resistance (R_{device}) of the driver plus the series resistor (R_S) equal to the line impedance or:

$R_S = Z_o - R_{device}$.

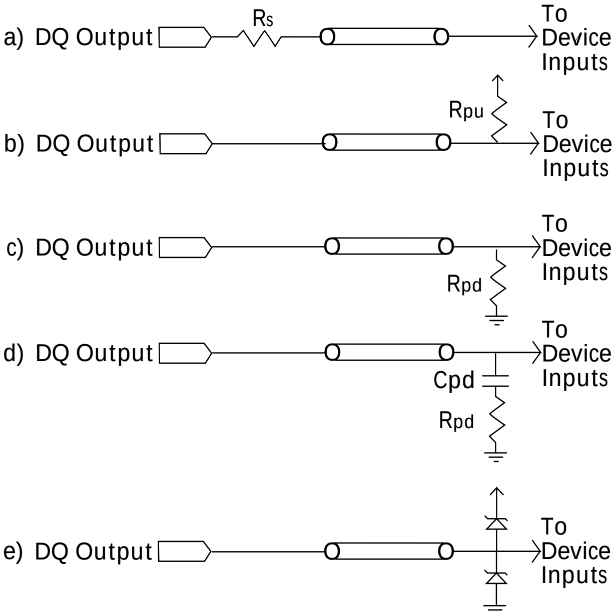
There is a potential disadvantage to the series termination resistor because of the associated voltage drop. This voltage drop could cause problems with noise margin to VOH and VOL and in a bidirectional signal with VIH and VIL. CMOS device inputs have a high impedance and only initial AC power is needed to charge/discharge capacitance. Once the output reaches the final level, current dissipation (and voltage drop) across R_S is minimal in CMOS circuits.

In addition to preventing reflection problems, series termination is commonly used in mixed-voltage systems to prevent high currents. In Figure 7, a 5V device is driving a 3.3V input. Many 3.3V devices contain a protection diode which is connected to 3.3V. If the input was driven to 5V, the diode would become forward-biased and would generate a low resistance path to VDD which could result in potentially damaging currents. In this case, the series termination resistor has the added advantage of limiting the current in these mixed-voltage systems.

Figure 6*b* and Figure 6*c* show pull-up and pull-down termination. A pull-up and pull-down resistor could be also used simultaneously. The main disadvantage of this type of design is that there is a DC power dissipation associated with the devices. For instance, for the pull-up case, there is a DC current path when the output driver is low. This extra current is unacceptable in portable or notebook applications. Mixed 5V and 3.3V designs are recommended to use a pull-up to a 3.3V supply versus a 5V one wherever possible to reduce power consumption.

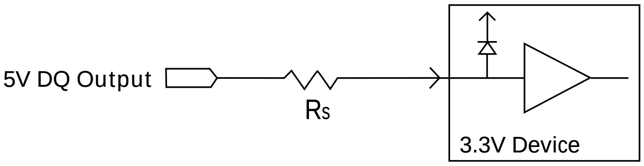
Another advantage of pull-ups is in connecting TTL-level outputs to CMOS-level inputs. TTL outputs have a VOH specification of 2.4V versus a CMOS VIH of 3.15V. CMOS input levels occur in low-power microprocessors and microcontrollers used in portable and hand-held applications. The pull-up to 5V will allow fast SRAMs with TTL-I/O to drive these CMOS circuits.

Figure 6. a) Series-Resistance Termination, b) Pull-Up Termination, c) Pull-Down Termination, d) RC Termination, e) Schottky Diode Termination



RC termination (Figure 6*d*) allows for proper termination without an associated DC current component. The disadvantage of this approach is that it requires an extra component, the capacitor. An advantage is that this circuit acts as a low-pass filter and can absorb unwanted glitches. Care must be taken in the choice of a capacitor since it must be large enough to absorb glitches that may occur in the system but small enough so it does not slow down the system

Figure 7. Diode Protection in 3.3V Circuits Can Cause Problems in Mixed 5V–3.3V Systems.



Lastly, Schottky diode termination (Figure 6*e*) can be used to save power over resistance termination techniques. The advantage of this method is that it prevents overshoot and undershoot problems. The diode turns on if the voltage goes above VDD + V_{diode} or below VSS – V_{diode}. The main advantage of this technique is in prototypes constructed with wire wrap or breadboards where line impedance may not be constant. These diodes provide termination without requiring detailed impedance matching calculations.

Current Trends

The computer industry always moves toward faster architectures. To meet the bus speeds of 10-ns bus cycle times (or faster), I/O standards such as 2.5V I/O, HSTL (high-speed transceiver logic) are now available to minimize output voltage swings and drive transmission lines. Even with these new standards, designers need to analyze their circuit thoroughly to ensure that cycle/access times are met.

Conclusion

As we have seen in this article, high-speed bus design requires detailed analysis to ensure that the system will work properly. A proper analysis of clock skew and propagation delay is essential to ensure that the system will work properly and with the required performance. This detailed analysis is essential for bus speeds over 50 MHz.

In March of 2007, Cypress recataloged all of its Application Notes using a new documentation number and revision code. This new documentation number and revision code (001-xxxxx, beginning with rev. **), located in the footer of the document, will be used in all subsequent revisions.

Intel and Pentium are registered trademarks of Intel Corporation. All other product and company names mentioned in this document are the trademarks of their respective holders.

Cypress Semiconductor
198 Champion Court
San Jose, CA 95134-1709
Phone: 408-943-2600
Fax: 408-943-4730
http://www.cypress.com

Disclaimer: CYPRESS MAKES NO WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, WITH REGARD TO THIS MATERIAL, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. Cypress reserves the right to make changes without further notice to the materials described herein. Cypress does not assume any liability arising out of the application or use of any product or circuit described herein. Cypress does not authorize its products for use as critical components in life-support systems where a malfunction or failure may reasonably be expected to result in significant injury to the user. The inclusion of Cypress' product in a life-support systems application implies that the manufacturer assumes all risk of such use and in doing so indemnifies Cypress against all charges.