NoBL SRAMs and Bus Contention

Introduction

Faster systems demand higher performance especially from the memory. To accommodate the upsurge of networking applications requiring faster memory, various new SRAMs are becoming available. NoBL (No Bus Latency) SRAMs were designed to improve the performance of networking applications. As the demands on these networking systems have increased, so also have the demands of the memory system. The frequency of operation of NoBL SRAMs have gone up over time. Operating NoBL SRAMs at higher frequencies adds several challenges including the possibility of bus contention. This paper delves headlong into bus contention, describing the effects and ways to avoid it.

Operation of a NoBL SRAM

To better understand Bus Contention, it is important that we understand how a NoBL SRAM operates.

NoBL SRAMs are designed to improve performance of systems by eliminating wait states which are required when using standard synchronous memories and interleaving read/write operations. Networking systems access the SRAM memories in an interleaved read-write fashion, which means that every other operation, the controller is either requesting a read from the memory or writing data to the memory.

Synchronous SRAMs introduce idle cycles when there is a transition from a write to a read. This reduces the bandwidth of the system thereby affecting the efficiency on system bus utilization. This is accepted in normal applications where the alternating writes/reads are infrequent. In networking, where this is a common occurrence, these wait states can dramatically affect the system performance.

Cypress's NoBL SRAMs address this need by eliminating the idle cycles and enabling 100% bus utilization. *Figure 1* and *Figure 2* show reads/writes for a Synchronous SRAM and a NoBL SRAM respectively

NoBL SRAMs are designed for high-performance networking applications. In networking applications, data is written into the SRAM and read back very frequently. As a result, the data bus is switching frequently between reads/writes. Therefore, in one clock cycle, control of the bus goes from one device to the another. When the data bus switches, the possibility of the two devices both being on the bus (driving the data) at the same time exists. This usually happens when an ASIC is slower in turning its output buffers OFF, while the SRAM is quicker in turning its output buffers ON and vice versa. This issue is known as Bus Contention and when it occurs, it is usually for a very brief time (0.5–1.0 ns).

Bus Contention can be the cause of many system-level issues. The next section describes bus contention in more detail.

Bus Contention

Bus Contention occurs when two or more devices drive a common bus at the same time. In *Figure 4*, device A is driving a logic level '1' while Device B is driving a logic level '0' onto the common bus. This creates a current path between the contending devices from the power supply of Device A to the ground of Device B. The current path ceases to exist if both the devices try to drive the bus to the same logic level.

Bus Contention can cause excess current to flow through the two devices and thereby increases the power consumption resulting in system noise..

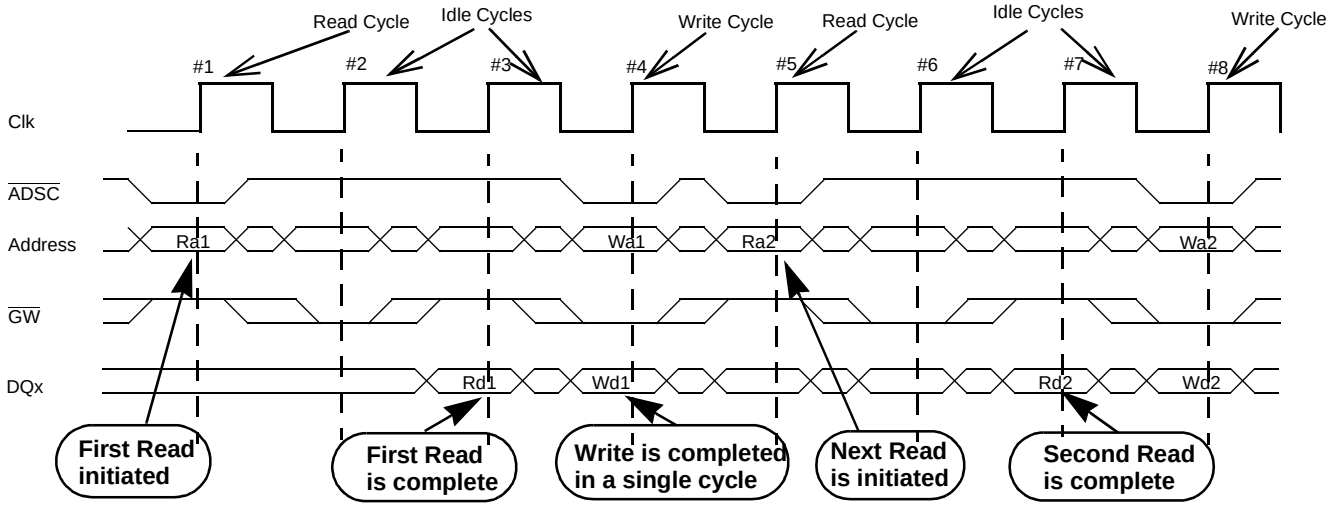

Figure 1. Synchronous Pipelined Burst SRAM Timings for a R-W-R-W Operation

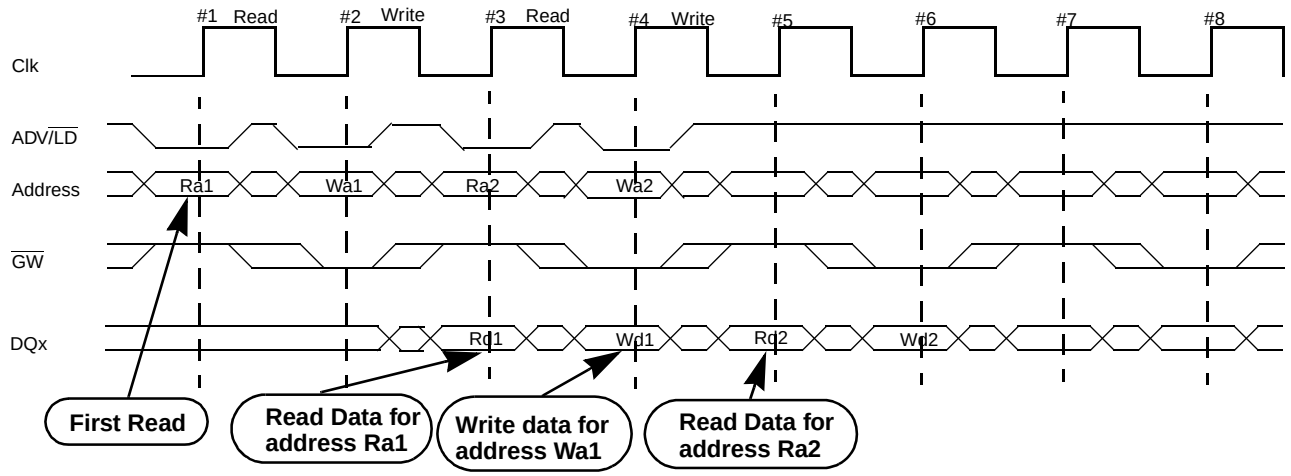

Figure 2. NoBL Timings for a R-W-R-W Operation

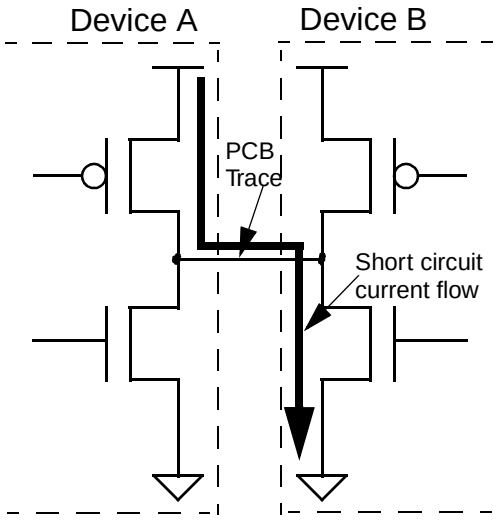

Figure 3. Devices in Contention

Power Analysis

Bus Contention occurs when a device doesn't release control of a data bus before the other device starts driving the same data bus. This draws extra power off the system and also tends to increase the temperature of the devices involved in contention which might lead to failure of the devices.

If contention is unavoidable, as in some designs, a thorough analysis of power dissipation during contention has to be done to ensure that this value is always lower than the maximum power which the device can handle.

First of all, the maximum power which the device can handle should be calculated. Let us assume the following conditions under which you are operating a device:

Max. operating ambient temperature (Ta): 70°C

Max. allowed junction temperature (Tj): 150°C

Theta-ja(θ_{ja}): 25 °C/W

From these values, Safe maximum power the device can handle = $(Tj - Ta)/\theta_{ja} = (150 - 70)/25 = 3.2W$

Normal power drawn from the SRAM without any contention is the sum of core power drawn and the I/O switching power.

The following are assumed to be the conditions under which the device is operated:

I/O supply voltage(V_{DDQ}): 3.3V

Frequency(f): 166 MHz

Active operating current(Icc): 480 mA

No. of Data I/Os(I/O): 36

Capacitive loading on each pin: 20 pF

On an average, let the drivers that are switching to logic '1' be 18. Only logic level '1' need be considered since this is the only time the device consumes power. Let the duration of device read be 50% as in the case of a networking application.

Base power drawn by the SRAM core, i.e. the logic and the memory array circuits excluding I/O circuitry is V_{DDQ} x I_{CC} = 3.3 x 0.48 = 1.584W.

Switching I/O power draw is nothing but the dynamic power consumption which is equal to ½ x f x C x V^2.

Substituting the above values gives

$0.5 \times 166 \times 10^6 \times 3.3 \times 3.3 \times 20 \times 10^{-12} \times 18 \times 50/100$

= 0.163 W.

Total power draw from the device without any contention is the sum of the core power drawn and the switching I/O power drawn.

Total power drawn without any contention =
1.584 + 0.163 = 1.749W

The allowable extra power that could be drawn due to the bus contention: maximum power the device can handle – total power drawn without any contention is
3.2 – 1.749 = 1.451W

Bus Contention Power

The bus contention power can be calculated based on the total contention current drawn by the system. This can be compared against the extra allowable contention power.

Let us assume that the SRAM and the contending device have 50Ω impedances each during contention. Since each I/O of both the SRAM and the contending device can have a '0' or a '1,' the probability of contention is reduced by half or the number of pins in contention could be averaged out to be eighteen.

Current during contention = 3.3/(50 + 50) x 18 = 0.594A.

If the impedance of the system can be thought of as matched, then the combination of the SRAM and the contending device form a voltage divider which terminates the output voltage to 3.3/2 = 1.65V.

Therefore, Bus contention power = 1.65 x 0.594 = 0.98W

Note that this assumes that the contention occurs 100% of the time, which is not the case in reality.

From the above value, we see that it is less than the allowable bus contention power.

Data sheet Specifications

Voltage, temperature and process variation affects the device parameters. The actual value that the device delivers changes depending upon the above factors. In order to ensure a worst case value, all the timing parameters on the datasheet which are crucial for every design are tested at extreme conditions. Not all the AC timing parameters have worst case values under the same conditions.

Two timing parameters of crucial importance regarding bus contention are t_{CHZ} and t_{CLZ}. t_{CHZ} is the SRAM turn-off time, i.e., the time it takes for the SRAM to turn its outputs off or to stop driving the data bus and go into a high-impedance (High-Z) state. t_{CLZ} is the SRAM turn-on time, i.e., the time it takes for the SRAM to turn its outputs on or to start begin driving the data bus and go into a low-impedance mode. During High-Z mode, both the pull-up and pull-down structures are disabled. The output is disabled and the pin behaves as an input.

If two NoBL SRAMs were to be connected onto the same bus and if the bus had to be turned around, then from a preliminary look at these values in the data sheet, it seems impossible to eliminate contention as the t_{CHZ} (maximum) is higher than t_{CLZ} (minimum). However, it is important to note that these two parameters are tested at opposite extreme conditions, t_{CHZ} being at low voltage, slow process corner and high temperature and t_{CLZ} being at high voltage, fast process corner, and low temperature.

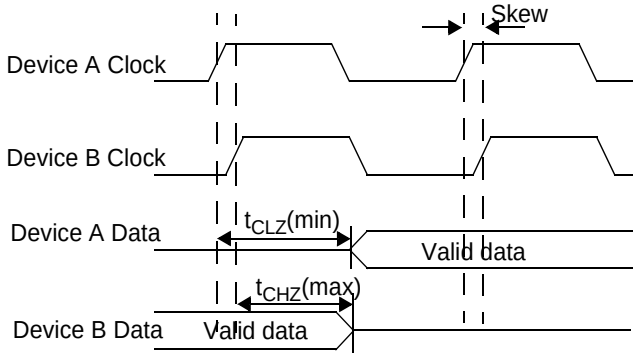

Figure 4. t_{CLZ} and t_{CHZ} timing waveforms

Since the interfacing components are usually placed on the same board, it can be safely assumed that they would be operating at similar voltage and temperature at any given time. Cypress's NoBL SRAMs are designed to ensure that at same operating voltage and temperature, the turn-off time (t_{CHZ}) is faster than the turn-on time (t_{CLZ}) by approximately 1 ns. This enables us to heave a sigh of relief from bus contention.

The above discussion holds equally good for an SRAM and an ASIC interface, that the output buffers are faster at one extreme and the data hold times are worst at another extreme and that two devices on the same board will have very similar conditions (voltage & temperature).

Factors Affecting Bus Contention

Bus Contention is mainly affected by the t_{CHZ} and the t_{CLZ} timings, but there are other factors that affect bus contention. These include clock skew, inductance and capacitive loadings.

Clock Skew

Clock skew is the difference in the arrival of the clock at two different devices that are on the same bus. Possible reasons for clock skew are package parasitic differences and trace lengths on the board. Clock skew might move the turn-off time closer to the turn-on time thus increasing the chances of bus contention as shown in *Figure 4*. As much as possible, clock skew should be minimized to reduce bus contention by careful board design.

Lead Inductance

Bus contention is minimized to some extent by the power supply inductance and the inherent inductance in the leads on the package. Inductance resists the instantaneous change in current thereby reducing the effects of the bus contention.

Capacitive Loading

Capacitive loading on the output also tends to minimize the effects of the bus contention. Let's assume, for the sake of discussion, that a system bus is switching from a logic level '1' to a logic level '0' and also the turn-off time of the bus is higher than the turn-on time of the bus. The capacitor on the bus would have an output of logic '1' at this time. When the pull-down circuit of one device turns on, the pull-up circuit of the other device is still on. The Capacitor resists the instantaneous change in voltage initially and also, since the voltage across the pull-up circuit is zero volts at this time, there would be a zero current flow through the pull-up circuit. Now, the capacitor starts to discharge through the pull-down circuit depending upon the RC constant. This current through the pull-down circuit flows irrespective of whether there was a bus contention or not. By initially preventing any contention current, the capacitor reduces the bus contention. If the capacitor is large, the RC constant would be large and the time for which there will be a logic '1' at the output would be prolonged which further decreases the bus contention. Therefore, a larger output capacitance results in less reduced contention current.

Reliability Issues

Bus Contention can cause high drain to source voltage (V_{DS}) levels across transistors. This might lead to strong electric fields, which will cause hot electrons to flow. Hot electrons, which are high energy carriers in the channel, penetrate the gate oxide and cause gate current to flow. Eventually, this can lead to degradation of the CMOS device threshold voltage leading to the failure of device. Hence long term reliability could be affected by bus contention.

Conclusion

NoBL SRAMs maximize the system performance by eliminating idle cycles and by increasing the bandwidth of the system. This brings in stringent timing specifications which could lead to bus contention. The power analysis detailed in this note illustrates that bus contention in reduced amounts can be acceptable. This paper discussed some of the issues which directly affect the bus contention and how it can be reduced. With careful design, bus contention can be kept to a minimum acceptable value.