3

THE DIGITAL LOGIC LEVEL

At the bottom of the hierarchy of Fig. 1-2 we find the digital logic level, the computer's real hardware. In this chapter, we will examine many aspects of digital logic, as a building block for the study of higher levels in subsequent chapters. This subject is on the boundary of computer science and electrical engineering, but the material is self-contained, so no previous hardware or engineering experience is needed to follow it.

The basic elements from which all digital computers are constructed are amazingly simple. We will begin our study by looking at these basic elements and also at the special two-valued algebra (Boolean algebra) used to analyze them. Next we will examine some fundamental circuits that can be built using gates in simple combinations, including circuits for doing arithmetic. The following topic is how gates can be combined to store information, that is, how memories are organized. After that, we come to the subject of CPUs and especially how single-chip CPUs interface with memory and peripheral devices. Numerous examples from industry will be discussed later in this chapter.

3.1 GATES AND BOOLEAN ALGEBRA

Digital circuits can be constructed from a small number of primitive elements by combining them in innumerable ways. In the following sections we will describe these primitive elements, show how they can be combined, and introduce a powerful mathematical technique that can be used to analyze their behavior.

3.1.1 Gates

A digital circuit is one in which only two logical values are present. Typically, a signal between 0 and 0.5 volt represents one value (e.g., binary 0) and a signal between 1 and 1.5 volts represents the other value (e.g., binary 1). Voltages outside these two ranges are not permitted. Tiny electronic devices, called **gates**, can compute various functions of these two-valued signals. These gates form the hardware basis on which all digital computers are built.

The details of how gates work inside is beyond the scope of this book, belonging to the **device level**, which is below our level 0. Nevertheless, we will now digress ever so briefly to take a quick look at the basic idea, which is not difficult. All modern digital logic ultimately rests on the fact that a transistor can be made to operate as a very fast binary switch. In Fig. 3-1(a) we have shown a bipolar transistor (the circle) embedded in a simple circuit. This transistor has three connections to the outside world: the **collector**, the **base**, and the **emitter**. When the input voltage, V_{in}, is below a certain critical value, the transistor turns off and acts like an infinite resistance. This causes the output of the circuit, V_{out}, to take on a value close to V_{cc}, an externally regulated voltage, typically +1.5 volts for this type of transistor. When V_{in} exceeds the critical value, the transistor switches on and acts like a wire, causing V_{out} to be pulled down to ground (by convention, 0 volts).

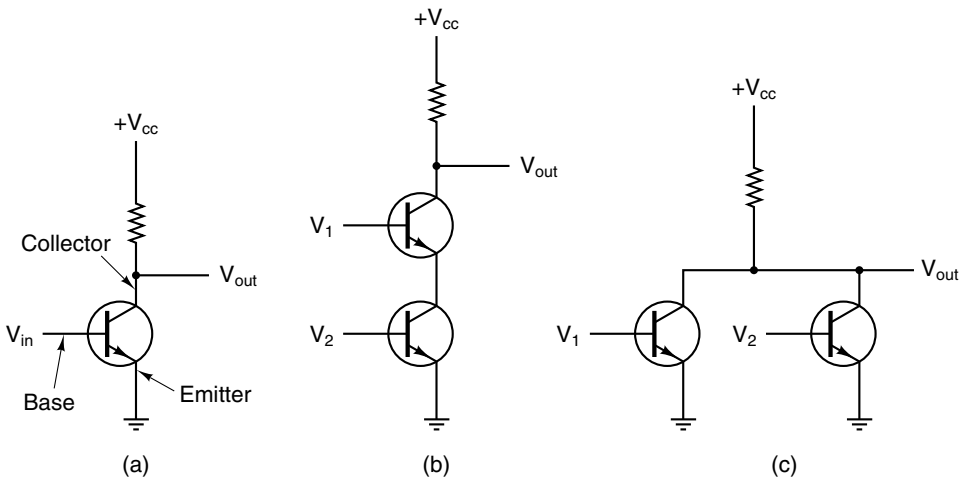


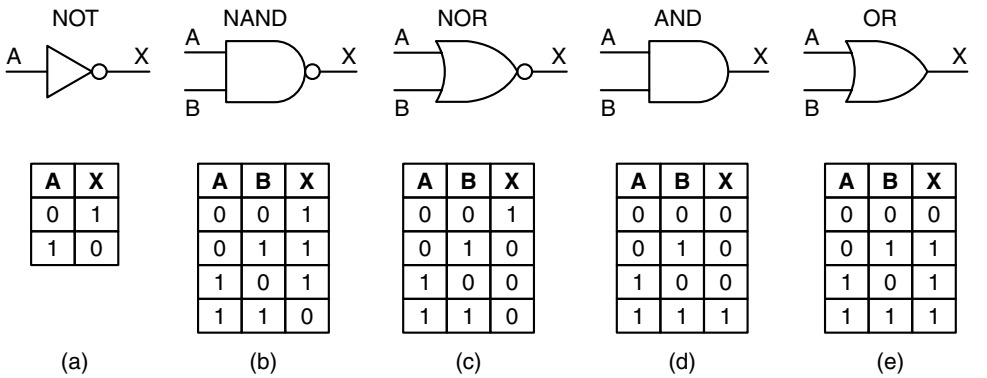
Figure 3-1. (a) A transistor inverter. (b) A NAND gate. (c) A NOR gate.

The important thing to notice is that when V_{in} is low, V_{out} is high, and vice versa. This circuit is thus an inverter, converting a logical 0 to a logical 1, and a logical 1 to a logical 0. The resistor (the jagged line) is needed to limit the amount of current drawn by the transistor so it does not burn out. The time required to switch from one state to the other is typically a nanosecond or less.

In Fig. 3-1(b) two transistors are cascaded in series. If both V_1 and V_2 are high, both transistors will conduct and V_{out} will be pulled low. If either input is low, the corresponding transistor will turn off, and the output will be high. In other words, V_{out} will be low if and only if both V_1 and V_2 are high.

In Fig. 3-1(c) the two transistors are wired in parallel instead of in series. In this configuration, if either input is high, the corresponding transistor will turn on and pull the output down to ground. If both inputs are low, the output will remain high.

These three circuits, or their equivalents, form the three simplest gates. They are called NOT, NAND, and NOR gates, respectively. NOT gates are often called **inverters**; we will use the two terms interchangeably. If we now adopt the convention that "high" (V_{cc} volts) is a logical 1, and that "low" (ground) is a logical 0, we can express the output value as a function of the input values. The symbols used to depict these three gates are shown in Fig. 3-2(a)–(c), along with the functional behavior for each circuit. In these figures, *A* and *B* are inputs and *X* is the output. Each row specifies the output for a different combination of the inputs.



A	X
0	1
1	0

(a)

A	B	X
0	0	1
0	1	1
1	0	1
1	1	0

(b)

A	B	X
0	0	1
0	1	0
1	0	0
1	1	0

(c)

A	B	X
0	0	0
0	1	0
1	0	0
1	1	1

(d)

A	B	X
0	0	0
0	1	1
1	0	1
1	1	1

(e)

Figure 3-2. The symbols and functional behavior for the five basic gates.

If the output signal of Fig. 3-1(b) is fed into an inverter circuit, we get another circuit with precisely the inverse of the NAND gate—namely, a circuit whose output is 1 if and only if both inputs are 1. Such a circuit is called an AND gate; its symbol and functional description are given in Fig. 3-2(d). Similarly, the NOR gate can be connected to an inverter to yield a circuit whose output is 1 if either or both inputs are 1 but 0 if both inputs are 0. The symbol and functional description of this circuit, called an OR gate, are given in Fig. 3-2(e). The small circles used as part of the symbols for the inverter, NAND gate, and NOR gate are called **inversion bubbles**. They are often used in other contexts as well to indicate an inverted signal.

The five gates of Fig. 3-2 are the principal building blocks of the digital logic level. From the foregoing discussion, it should be clear that NAND and NOR gates require two transistors each, whereas the AND and OR gates require three each. For

this reason, many computers are based on NAND and NOR gates rather than the more familiar AND and OR gates. (In practice, all the gates are implemented somewhat differently, but NAND and NOR are still simpler than AND and OR.) In passing it is worth noting that gates may well have more than two inputs. In principle, a NAND gate, for example, may have arbitrarily many inputs, but in practice more than eight inputs is unusual.

Although the subject of how gates are constructed belongs to the device level, we would like to mention the major families of manufacturing technology because they are referred to frequently. The two major technologies are **bipolar** and **MOS** (Metal Oxide Semiconductor). The major bipolar types are **TTL** (Transistor-Transistor Logic), which had been the workhorse of digital electronics for years, and **ECL** (Emitter-Coupled Logic), which was used when very high-speed operation was required. For computer circuits, MOS has now largely taken over.

MOS gates are slower than TTL and ECL but require much less power and take up much less space, so large numbers of them can be packed together tightly. MOS comes in many varieties, including PMOS, NMOS, and CMOS. While MOS transistors are constructed differently from bipolar transistors, their ability to function as electronic switches is the same. Most modern CPUs and memories use CMOS technology, which runs on a voltage in the neighborhood of +1.5 volts. This is all we will say about the device level. Readers interested in pursuing their study of this level should consult the readings suggested on the book's Website.

3.1.2 Boolean Algebra

To describe the circuits that can be built by combining gates, a new type of algebra is needed, one in which variables and functions can take on only the values 0 and 1. Such an algebra is called a **Boolean algebra**, after its discoverer, the English mathematician George Boole (1815–1864). Strictly speaking, we are really referring to a specific type of Boolean algebra, a **switching algebra**, but the term "Boolean algebra" is so widely used to mean "switching algebra" that we will not make the distinction.

Just as there are functions in "ordinary" (i.e., high school) algebra, so are there functions in Boolean algebra. A Boolean function has one or more input variables and yields a result that depends only on the values of these variables. A simple function, f, can be defined by saying that $f(A)$ is 1 if A is 0 and $f(A)$ is 0 if A is 1. This function is the NOT function of Fig. 3-2(a).

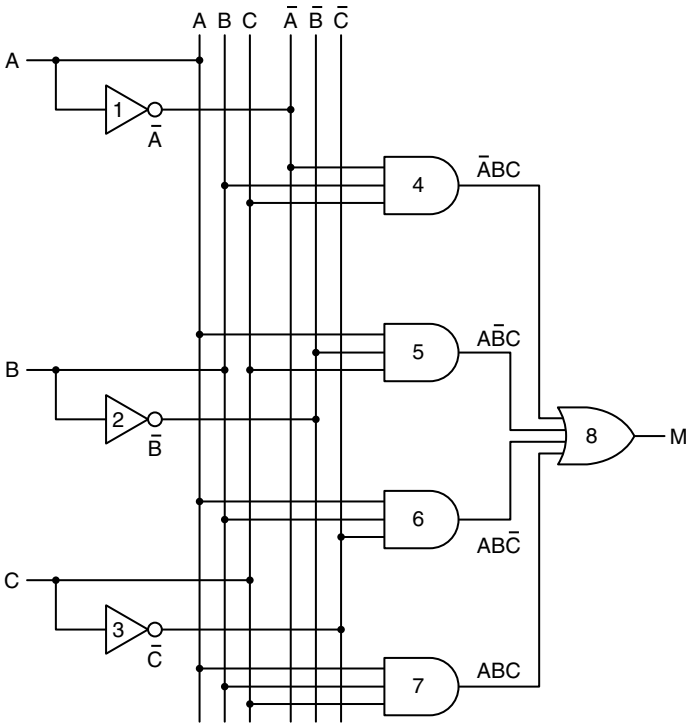
Because a Boolean function of n variables has only 2^n possible combinations of input values, the function can be completely described by giving a table with 2^n rows, each row telling the value of the function for a different combination of input values. Such a table is called a **truth table**. The tables of Fig. 3-2 are all examples of truth tables. If we agree to always list the rows of a truth table in numerical order (base 2), that is, for two variables in the order 00, 01, 10, and 11, the function can be completely described by the 2^n-bit binary number obtained by reading the

result column of the truth table vertically. Thus, NAND is 1110, NOR is 1000, AND is 0001, and OR is 0111. Obviously, only 16 Boolean functions of two variables exist, corresponding to the 16 possible 4-bit result strings. In contrast, ordinary algebra has an infinite number of functions of two variables, none of which can be described by giving a table of outputs for all possible inputs because each variable can take on any one of an infinite number of possible values.

Figure 3-3(a) shows the truth table for a Boolean function of three variables: $M = f(A, B, C)$. This function is the majority logic function, that is, it is 0 if a majority of its inputs are 0 and 1 if a majority of its inputs are 1. Although any Boolean function can be fully specified by giving its truth table, as the number of variables increases, this notation becomes increasingly cumbersome. Instead, another notation is frequently used.

A	B	C	M
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

(a)



(b)

Figure 3-3. (a) The truth table for the majority function of three variables. (b) A circuit for (a).

To see how this other notation comes about, note that any Boolean function can be specified by telling which combinations of input variables give an output value of 1. For the function of Fig. 3-3(a) there are four combinations of input variables that make M equal to 1. By convention, we will place a bar over an input

variable to indicate that its value is inverted. The absence of a bar means that it is not inverted. Furthermore, we will use implied multiplication or a dot to mean the Boolean AND function and + to mean the Boolean OR function. Thus, for example, $A\bar{B}C$ takes the value 1 only when $A = 1$ and $B = 0$ and $C = 1$. Also, $A\bar{B} + B\bar{C}$ is 1 only when ($A = 1$ and $B = 0$) or ($B = 1$ and $C = 0$). The four rows of Fig. 3-3(a) producing 1 bits in the output are: $\bar{A}BC$, $A\bar{B}C$, $AB\bar{C}$, and ABC. The function, M, is true (i.e., 1) if any of these four conditions is true, so we can write

$$M = \bar{A}BC + A\bar{B}C + AB\bar{C} + ABC$$

as a compact way of giving the truth table. A function of n variables can thus be described by giving a "sum" of at most 2^n n-variable "product" terms. This formulation is especially important, as we will see shortly, because it leads directly to an implementation of the function using standard gates.

It is important to keep in mind the distinction between an abstract Boolean function and its implementation by an electronic circuit. A Boolean function consists of variables, such as A, B, and C, and Boolean operators such as AND, OR, and NOT. A Boolean function is described by giving a truth table or a Boolean function such as

$$F = A\bar{B}C + AB\bar{C}$$

A Boolean function can be implemented by an electronic circuit (often in many different ways) using signals that represent the input and output variables and gates such as AND, OR, and NOT. We will generally use the notation AND, OR, and NOT when referring to the Boolean operators and AND, OR, and NOT when referring to the gates, even though it is sometimes ambiguous as to whether we mean the functions or the gates.

3.1.3 Implementation of Boolean Functions

As mentioned above, the formulation of a Boolean function as a sum of up to 2^n product terms leads directly to a possible implementation. Using Fig. 3-3 as an example, we can see how this implementation is accomplished. In Fig. 3-3(b), the inputs, A, B, and C, are shown at the left edge and the output function, M, is shown at the right edge. Because complements (inverses) of the input variables are needed, they are generated by tapping the inputs and passing them through the inverters labeled 1, 2, and 3. To keep the figure from becoming cluttered, we have drawn in six vertical lines, of which three are connected to the input variables, and three connected to their complements. These lines provide a convenient source for the inputs to subsequent gates. For example, gates 5, 6, and 7 all use A as an input. In an actual circuit these gates would probably be wired directly to A without using any intermediate "vertical" wires.

The circuit contains four AND gates, one for each term in the equation for M (i.e., one for each row in the truth table having a 1 bit in the result column). Each

AND gate computes one row of the truth table, as indicated. Finally, all the product terms are ORed together to get the final result.

The circuit of Fig. 3-3(b) uses a convention that we will use repeatedly throughout this book: when two lines cross, no connection is implied unless a heavy dot is present at the intersection. For example, the output of gate 3 crosses all six vertical lines but it is connected only to $\bar{C}$. Be warned that some authors use other conventions.

From the example of Fig. 3-3 it should be clear how we can derive a general method to implement a circuit for any Boolean function:

1. Write down the truth table for the function.
2. Provide inverters to generate the complement of each input.
3. Draw an AND gate for each term with a 1 in the result column.
4. Wire the AND gates to the appropriate inputs.
5. Feed the output of all the AND gates into an OR gate.

Although we have shown how any Boolean function can be implemented using NOT, AND, and OR gates, it is often convenient to implement circuits using only a single type of gate. Fortunately, it is straightforward to convert circuits generated by the preceding algorithm to pure NAND or pure NOR form. All we need is a way to implement NOT, AND, and OR using a single gate type. The top row of Fig. 3-4 shows how all three of these can be implemented using only NAND gates; the bottom row shows how it can be done using only NOR gates. (These are straightforward, but there are other ways, too.)

One way to implement a Boolean function using only NAND or only NOR gates is first follow the procedure given above for constructing it with NOT, AND, and OR. Then replace the multi-input gates with equivalent circuits using two-input gates. For example, $A + B + C + D$ can be computed as $(A + B) + (C + D)$, using three two-input OR gates. Finally, the NOT, AND, and OR gates are replaced by the circuits of Fig. 3-4.

Although this procedure does not lead to the optimal circuits, in the sense of the minimum number of gates, it does show that a solution is always feasible. Both NAND and NOR gates are said to be **complete**, because any Boolean function can be computed using either of them. No other gate has this property, which is another reason they are often preferred for the building blocks of circuits.

3.1.4 Circuit Equivalence

Circuit designers often try to reduce the number of gates in their products to reduce the chip area needed to implement them, minimize power consumption, and increase speed. To reduce the complexity of a circuit, the designer must find another circuit that computes the same function as the original but does so with fewer

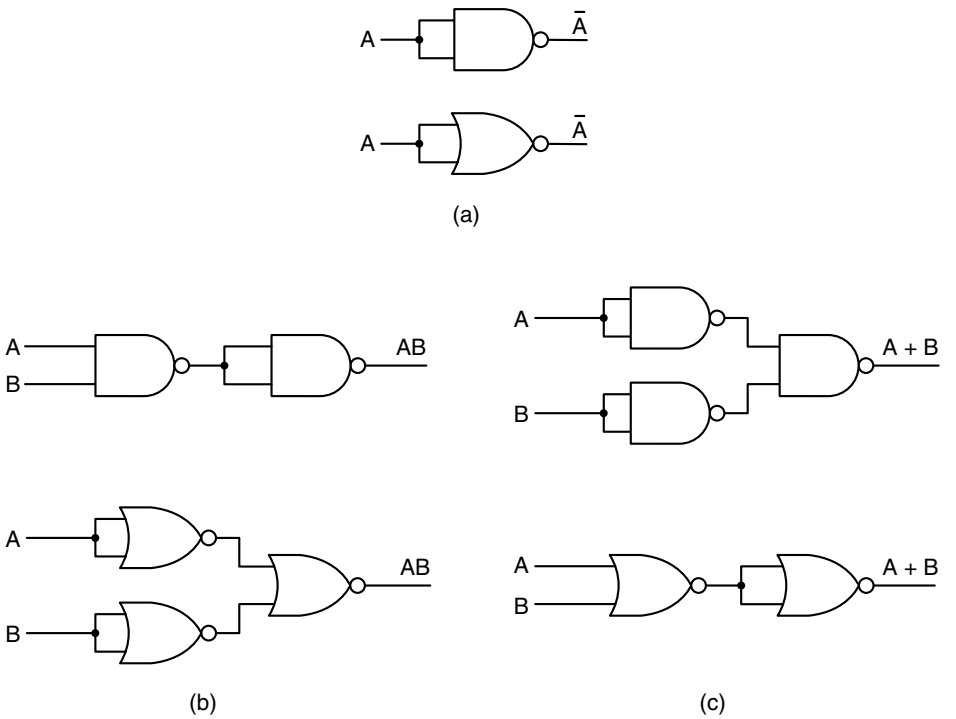


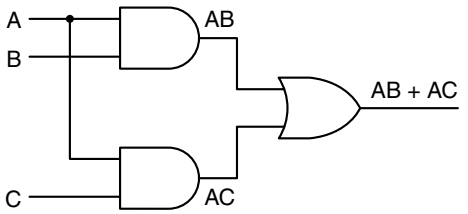
Figure 3-4. Construction of (a) NOT, (b) AND, and (c) OR gates using only NAND gates or only NOR gates.

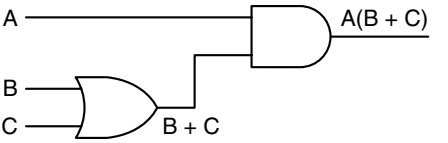
gates (or perhaps with simpler gates, for example, two-input gates instead of four-input gates). In the search for equivalent circuits, Boolean algebra can be a valuable tool.

As an example of how Boolean algebra can be used, consider the circuit and truth table for $AB + AC$ shown in Fig. 3-5(a). Although we have not discussed them yet, many of the rules of ordinary algebra also hold for Boolean algebra. In particular, $AB + AC$ can be factored into $A(B + C)$ using the distributive law. Figure 3-5(b) shows the circuit and truth table for $A(B + C)$. Because two functions are equivalent if and only if they have the same output for all possible inputs, it is easy to see from the truth tables of Fig. 3-5 that $A(B + C)$ is equivalent to $AB + AC$. Despite this equivalence, the circuit of Fig. 3-5(b) is clearly better than that of Fig. 3-5(a) because it contains fewer gates.

In general, a circuit designer starts with a Boolean function and then applies the laws of Boolean algebra to it in an attempt to find a simpler but equivalent one. From the final function, a circuit can be constructed.

To use this approach, we need some identities from Boolean algebra. Figure 3-6 shows some of the major ones. It is interesting to note that each law has two





A	B	C	AB	AC	AB + AC
0	0	0	0	0	0
0	0	1	0	0	0
0	1	0	0	0	0
0	1	1	0	0	0
1	0	0	0	0	0
1	0	1	0	1	1
1	1	0	1	0	1
1	1	1	1	1	1

(a)

A	B	C	A	B + C	A(B + C)
0	0	0	0	0	0
0	0	1	0	1	0
0	1	0	0	1	0
0	1	1	0	1	0
1	0	0	1	0	0
1	0	1	1	1	1
1	1	0	1	1	1
1	1	1	1	1	1

(b)

Figure 3-5. Two equivalent functions. (a) $AB + AC$. (b) $A(B + C)$.

forms that are **duals** of each other. By interchanging AND and OR and also 0 and 1, either form can be produced from the other one. All the laws can be easily proven by constructing their truth tables. Except for De Morgan's law, the absorption law, and the AND form of the distributive law, the results should be understandable with some study. De Morgan's law can be extended to more than two variables, for example, $\overline{ABC} = \bar{A} + \bar{B} + \bar{C}$.

De Morgan's law suggests an alternative notation. In Fig. 3-7(a) the AND form is shown with negation indicated by inversion bubbles, both for input and output. Thus, an OR gate with inverted inputs is equivalent to a NAND gate. From Fig. 3-7(b), the dual form of De Morgan's law, it should be clear that a NOR gate can be drawn as an AND gate with inverted inputs. By negating both forms of De Morgan's law, we arrive at Fig. 3-7(c) and (d), which show equivalent representations of the AND and OR gates. Analogous symbols exist for the multiple-variable forms of De Morgan's law (e.g., an n input NAND gate becomes an OR gate with n inverted inputs).

Using the identities of Fig. 3-7 and the analogous ones for multi-input gates, it is easy to convert the sum-of-products representation of a truth table to pure NAND or pure NOR form. As an example, consider the EXCLUSIVE OR function of Fig. 3-8(a). The standard sum-of-products circuit is shown in Fig. 3-8(b). To

Name	AND form	OR form
Identity law	$1A = A$	$0 + A = A$
Null law	$0A = 0$	$1 + A = 1$
Idempotent law	$AA = A$	$A + A = A$
Inverse law	$A\bar{A} = 0$	$A + \bar{A} = 1$
Commutative law	$AB = BA$	$A + B = B + A$
Associative law	$(AB)C = A(BC)$	$(A + B) + C = A + (B + C)$
Distributive law	$A + BC = (A + B)(A + C)$	$A(B + C) = AB + AC$
Absorption law	$A(A + B) = A$	$A + AB = A$
De Morgan's law	$\overline{AB} = \bar{A} + \bar{B}$	$\overline{A + B} = \bar{A}\bar{B}$

Figure 3-6. Some identities of Boolean algebra.

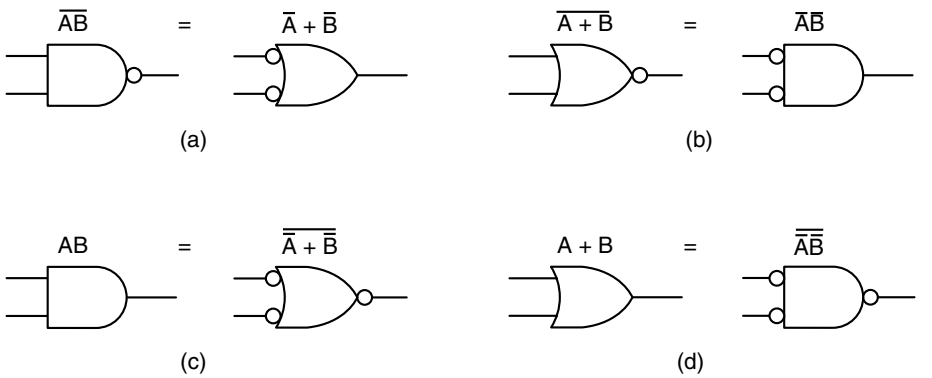


Figure 3-7. Alternative symbols for some gates: (a) NAND (b) NOR (c) AND (d) OR.

convert to NAND form, the lines connecting the output of the AND gates to the input of the OR gate should be redrawn with two inversion bubbles, as shown in Fig. 3-8(c). Finally, using Fig. 3-7(a), we arrive at Fig. 3-8(d). The variables $\bar{A}$ and $\bar{B}$ can be generated from A and B using NAND or NOR gates with their inputs tied together. Note that inversion bubbles can be moved along a line at will, for example, from the outputs of the input gates in Fig. 3-8(d) to the inputs of the output gate.

As a final note on circuit equivalence, we will now demonstrate the surprising result that the same physical gate can compute different functions, depending on the conventions used. In Fig. 3-9(a) we show the output of a certain gate, F, for different input combinations. Both inputs and outputs are shown in volts. If we

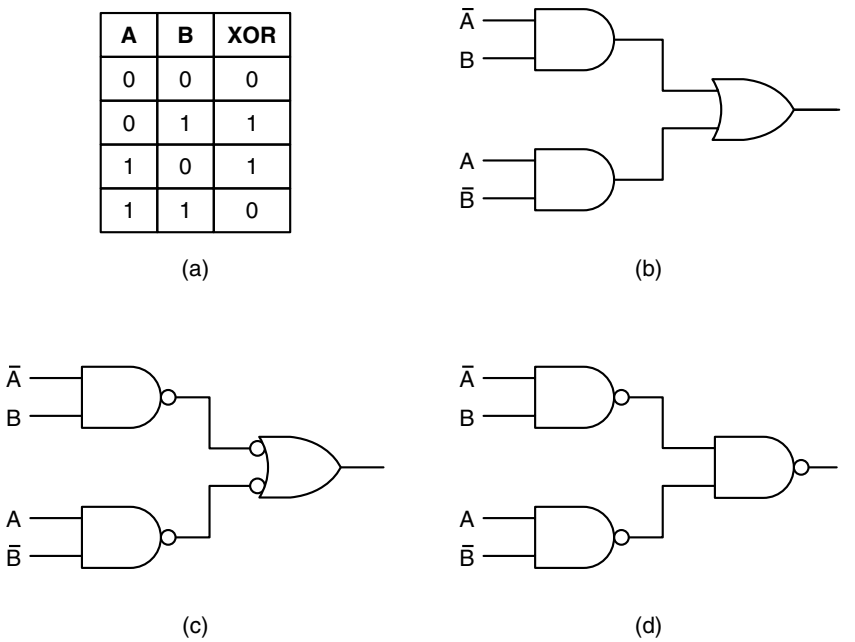


Figure 3-8. (a) The truth table for the XOR function. (b)–(d) Three circuits for computing it.

adopt the convention that 0 volts is logical 0 and 1.5 volts is logical 1, called **positive logic**, we get the truth table of Fig. 3-9(b), the AND function. If, however, we adopt **negative logic**, which has 0 volts as logical 1 and 1.5 volts as logical 0, we get the truth table of Fig. 3-9(c), the OR function.

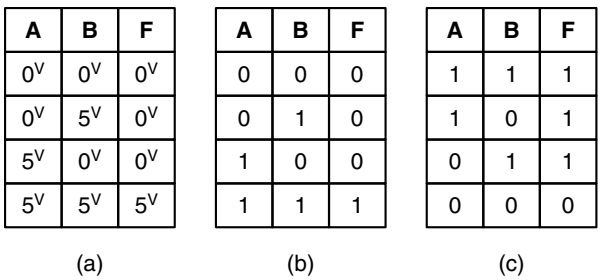


Figure 3-9. (a) Electrical characteristics of a device. (b) Positive logic. (c) Negative logic.

Thus, the convention chosen to map voltages onto logical values is critical. Except where otherwise specified, we will henceforth use positive logic, so the terms logical 1, true, and high are synonyms, as are logical 0, false, and low.

3.2 BASIC DIGITAL LOGIC CIRCUITS

In the previous sections we saw how to implement truth tables and other simple circuits using individual gates. In practice, few circuits are actually constructed gate-by-gate anymore, although this once was common. Nowadays, the usual building blocks are modules containing a number of gates. In the following sections we will examine these building blocks more closely and see how they are used and how they can be constructed from individual gates.

3.2.1 Integrated Circuits

Gates are not manufactured or sold individually but rather in units called **Integrated Circuits**, often called **ICs** or **chips**. An IC is a rectangular piece of silicon of varied size depending on how many gates are required to implement the chip's components. Small dies will measure about 2 mm $\times$ 2 mm, while larger dies can be as large as 18 mm $\times$ 18 mm. ICs are mounted into plastic or ceramic packages that can be much larger than the dies they house, if many pins are required to connect the chip to the outside world. Each pin connects to the input or output of some gate on the chip or to power or to ground.

Figure 3-10 shows a number of common IC packages used for chips today. Smaller chips, such as those used to house microcontrollers or RAM chips, will use **Dual Inline Packages** or **DIPs**. A DIP is a package with two rows of pins that fit into a matching socket on the motherboard. The most common DIP packages have 14, 16, 18, 20, 22, 24, 28, 40, 64, or 68 pins. For large chips, square packages with pins on all four sides or on the bottom are often used. Two common packages for larger chips are **Pin Grid Arrays** or **PGAs** and **Land Grid Arrays** or **LGAs**. PGA have pins on the bottom of the package, which fit into a matching socket on the motherboard. PGA sockets often utilize a zero-insertion-force mechanism in which the PGA can be placed into the socket without force, then a lever can be thrown which will apply lateral pressure to all of the PGA's pins, holding it firmly in the PGA socket. LGAs, on the other hand, have small flat pads on the bottom of the chip, and an LGA socket will have a cover that fits over the LGA and applies a downward force on the chip, ensuring that all of the LGA pads make contact with the LGA socket pads.

Because many IC packages are symmetric in shape, figuring out which orientation is correct is a perennial problem with IC installation. DIPs typically have a notch in one end which matches a corresponding mark on the DIP socket. PGAs typically have one pin missing, so if you attempt to insert the PGA into the socket incorrectly, the PGA will not insert. Because LGAs do not have pins, correct installation is enforced by placing a notch on one or two sides of the LGA, which matches a notch in the LGA socket. The LGA will not enter the socket unless the two notches match.

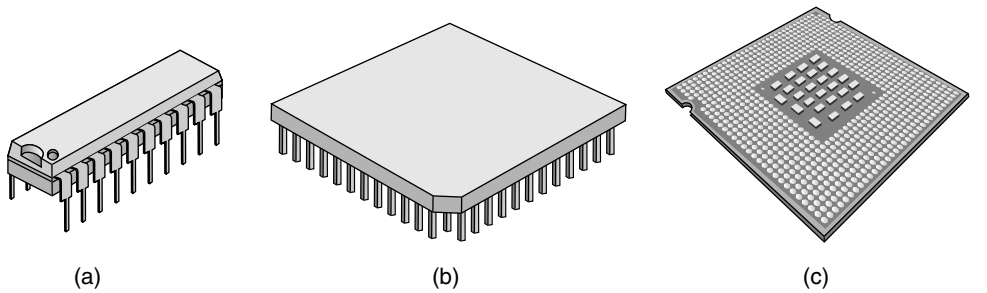


Figure 3-10. Common types of integrated-circuit packages, including a dual-in-line package (a), pin grid array (b), and land grid array (c).

For our purposes, all gates are ideal in the sense that the output appears as soon as the input is applied. In reality, chips have a finite **gate delay**, which includes both the signal propagation time through the chip and the switching time. Typical delays are 100s of picoseconds to a few nanoseconds.

It is within the current state of the art to put more than 1 billion transistors on a single chip. Because any circuit can be built up from NAND gates, you might think that a manufacturer could make a very general chip containing 500 million NAND gates. Unfortunately, such a chip would need 1,500,000,002 pins. With the standard pin spacing of 1 millimeter, an LGA would have to be 38 meters on a side to accommodate all of those pins, which might have a negative effect on sales. Clearly, the only way to take advantage of the technology is to design circuits with a high gate/pin ratio. In the following sections we will look at simple circuits that combine a number of gates internally to provide a useful function requiring only a limited number of external connections (pins).

3.2.2 Combinational Circuits

Many applications of digital logic require a circuit with multiple inputs and outputs in which the outputs are uniquely determined by the current input values. Such a circuit is called a **combinational circuit**. Not all circuits have this property. For example, a circuit containing memory elements may generate outputs that depend on the stored values as well as the input variables. A circuit implementing a truth table, such as that of Fig. 3-3(a), is a typical example of a combinational circuit. In this section we will examine some frequently used combinational circuits.

Multiplexers

At the digital logic level, a **multiplexer** is a circuit with 2^n data inputs, one data output, and n control inputs that select one of the data inputs. The selected data input is "gated" (i.e., sent) to the output. Figure 3-11 is a schematic diagram

for an eight-input multiplexer. The three control lines, *A*, *B*, and *C*, encode a 3-bit number that specifies which of the eight input lines is gated to the OR gate and thence to the output. No matter what value is on the control lines, seven of the AND gates will always output 0; the other one may output either 0 or 1, depending on the value of the selected input line. Each AND gate is enabled by a different combination of the control inputs. The multiplexer circuit is shown in Fig. 3-11.

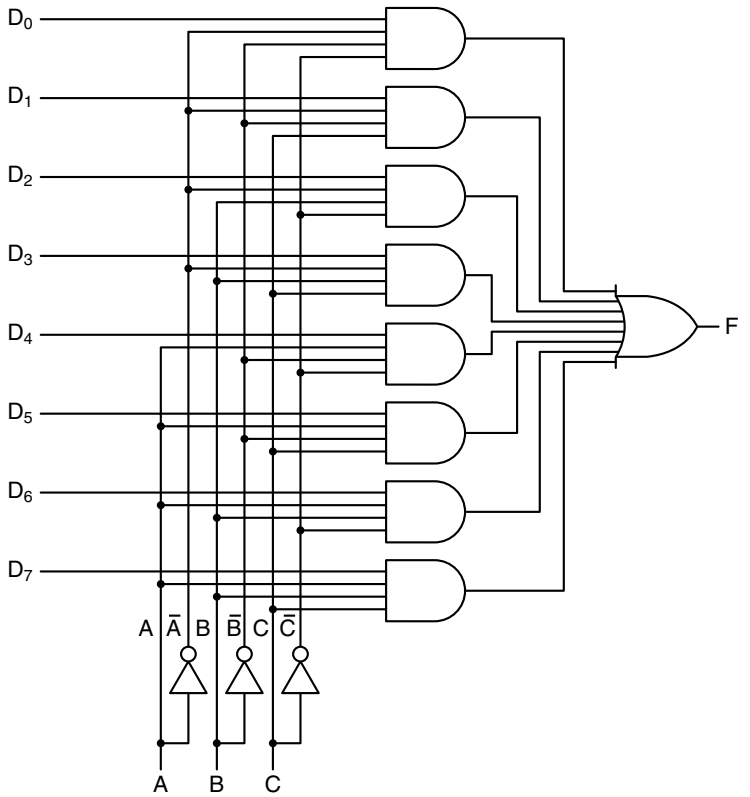


Figure 3-11. An eight-input multiplexer circuit.

Using the multiplexer, we can implement the majority function of Fig. 3-3(a), as shown in Fig. 3-12(b). For each combination of *A*, *B*, and *C*, one of the data input lines is selected. Each input is wired to either V_{cc} (logical 1) or ground (logical 0). The algorithm for wiring the inputs is simple: input D_i is the same as the value in row *i* of the truth table. In Fig. 3-3(a), rows 0, 1, 2, and 4 are 0, so the corresponding inputs are grounded; the remaining rows are 1, so they are wired to logical 1. In this manner any truth table of three variables can be implemented using the chip of Fig. 3-12(a).

We just saw how a multiplexer chip can be used to select one of several inputs and how it can implement a truth table. Another of its many applications is as a

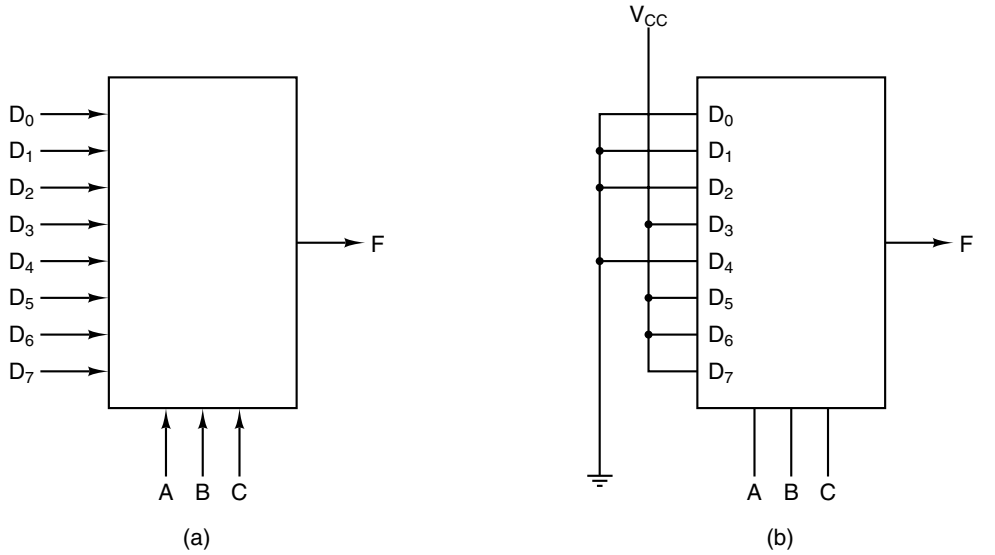


Figure 3-12. (a) An eight-input multiplexer. (b) The same multiplexer wired to compute the majority function.

parallel-to-serial data converter. By putting 8 bits of data on the input lines and then stepping the control lines sequentially from 000 to 111 (binary), the 8 bits are put onto the output line in series. A typical use for parallel-to-serial conversion is in a keyboard, where each keystroke implicitly defines a 7- or 8-bit number that must be output over a serial link, such as USB.

The inverse of a multiplexer is a **demultiplexer**, which routes its single input signal to one of 2^n outputs, depending on the values of the n control lines. If the binary value on the control lines is k, output k is selected.

Decoders

As a second example, we will now look at a circuit that takes an n-bit number as input and uses it to select (i.e., set to 1) exactly one of the 2^n output lines. Such a circuit, illustrated for $n = 3$ in Fig. 3-13, is called a **decoder**.

To see where a decoder might be useful, imagine a small memory consisting of eight chips, each containing 256 MB. Chip 0 has addresses 0 to 256 MB, chip 1 has addresses 256 MB to 512 MB, and so on. When an address is presented to the memory, the high-order 3 bits are used to select one of the eight chips. Using the circuit of Fig. 3-13, these 3 bits are the three inputs, A, B, and C. Depending on the inputs, exactly one of the eight output lines, $D_0, \ldots, D_7$, is 1; the rest are 0. Each output line enables one of the eight memory chips. Because only one output line is set to 1, only one chip is enabled.

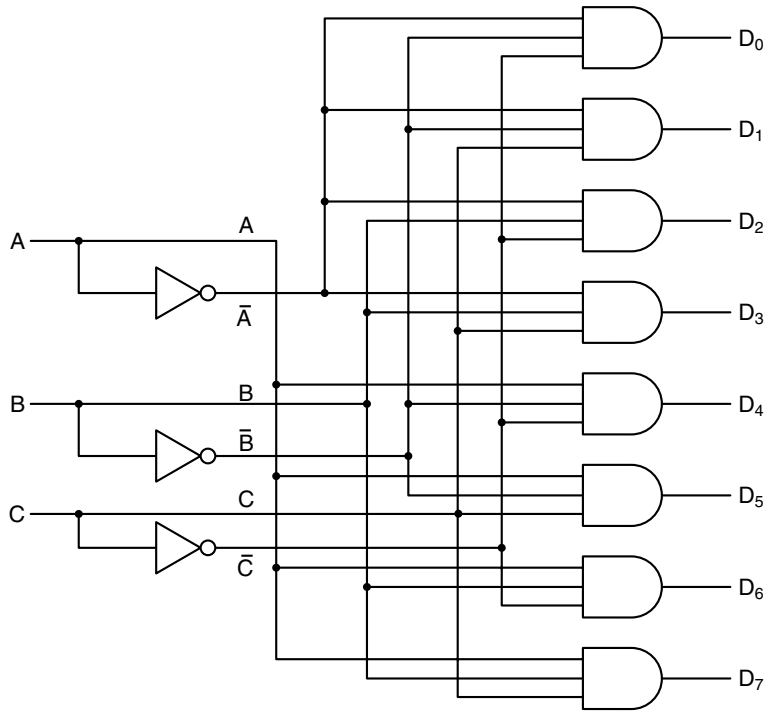


Figure 3-13. A three-to-eight decoder circuit.

The operation of the circuit of Fig. 3-13 is straightforward. Each AND gate has three inputs, of which the first is either A or $\bar{A}$, the second is either B or $\bar{B}$, and the third is either C or $\bar{C}$. Each gate is enabled by a different combination of inputs: D_0 by $\bar{A}\ \bar{B}\ \bar{C}$, D_1 by $\bar{A}\ \bar{B}\ C$, and so on.

Comparators

Another useful circuit is the **comparator**, which compares two input words. The simple comparator of Fig. 3-14 takes two inputs, A and B, each of length 4 bits, and produces a 1 if they are equal and a 0 otherwise. The circuit is based on the XOR (EXCLUSIVE OR) gate, which puts out a 0 if its inputs are equal and a 1 otherwise. If the two input words are equal, all four of the XOR gates must output 0. These four signals can then be ORed together; if the result is 0, the input words are equal, otherwise not. In our example we have used a NOR gate as the final stage to reverse the sense of the test: 1 means equal, 0 means unequal.

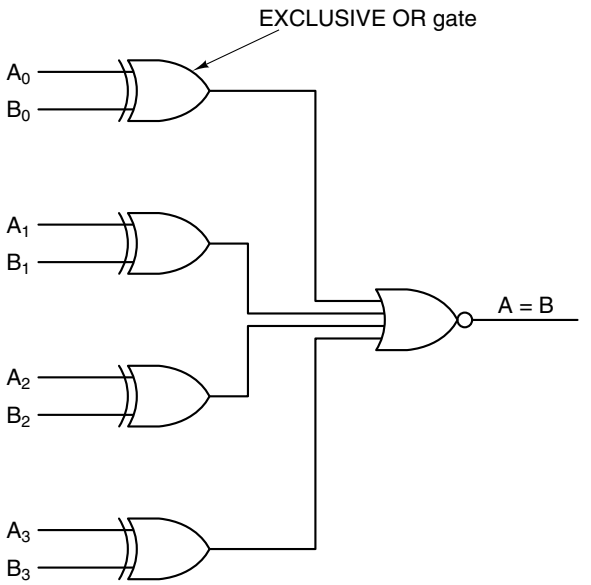


Figure 3-14. A simple 4-bit comparator.

3.2.3 Arithmetic Circuits

It is now time to move on from the general-purpose circuits discussed above to combinational circuits. As a reminder, combinational circuits have outs that are functions of their inputs, but circuits used for doing arithmetic do not have this property. We will begin with a simple 8-bit shifter, then look at how adders are constructed, and finally examine arithmetic logic units, which play a central role in any computer.

Shifters

Our first arithmetic circuit is an eight-input, eight-output shifter (see Fig. 3-15). Eight bits of input are presented on lines $D_0, \ldots, D_7$. The output, which is just the input shifted 1 bit, is available on lines $S_0, \ldots, S_7$. The control line, C, determines the direction of the shift, 0 for left and 1 for right. On a left shift, a 0 is inserted into bit 7. Similarly, on a right shift, a 1 is inserted into bit 0.

To see how the circuit works, notice the pairs of AND gates for all the bits except the gates on the end. When $C = 1$, the right member of each pair is turned on, passing the corresponding input bit to output. Because the right AND gate is wired to the input of the OR gate to its right, a right shift is performed. When $C = 0$, it is the left member of the AND gate pair that turns on, doing a left shift.

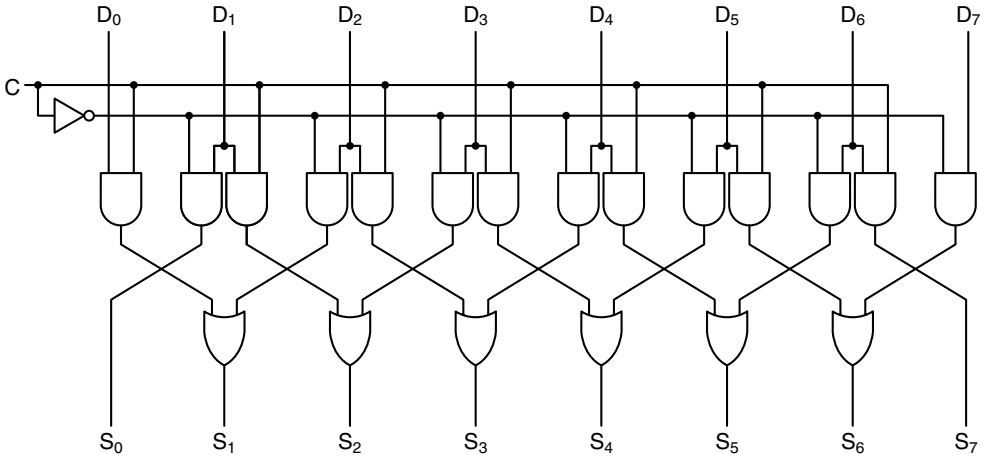


Figure 3-15. A 1-bit left/right shifter.

Adders

A computer that cannot add integers is almost unthinkable. Consequently, a hardware circuit for performing addition is an essential part of every CPU. The truth table for addition of 1-bit integers is shown in Fig. 3-16(a). Two outputs are present: the sum of the inputs, *A* and *B*, and the carry to the next (leftward) position. A circuit for computing both the sum bit and the carry bit is illustrated in Fig. 3-16(b). This simple circuit is generally known as a **half adder**.

A	B	Sum	Carry
0	0	0	0
0	1	1	0
1	0	1	0
1	1	0	1

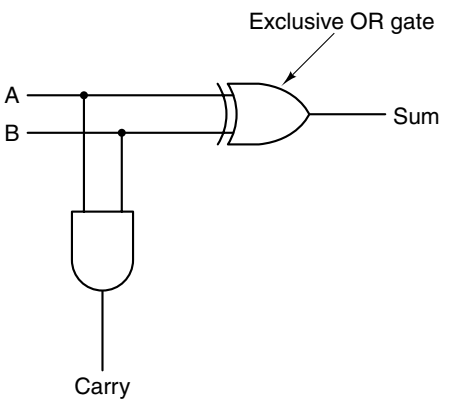
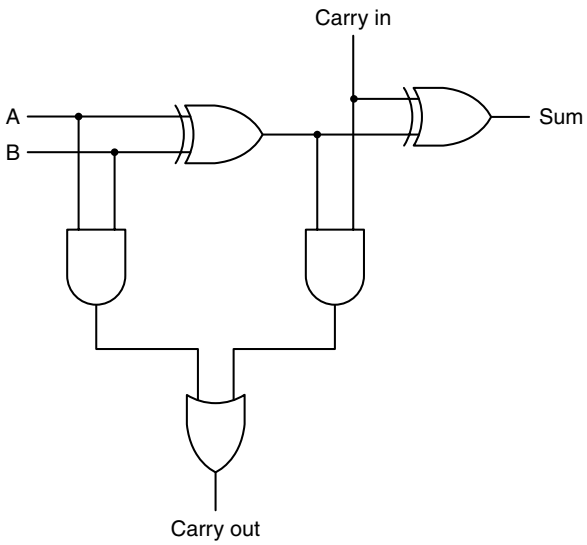


Figure 3-16. (a) Truth table for 1-bit addition. (b) A circuit for a half adder.

Although a half adder is adequate for summing the low-order bits of two multibit input words, it will not do for a bit position in the middle of the word because it does not handle the carry into the position from the right. Instead, the **full adder** of Fig. 3-17 is needed. From inspection of the circuit it should be clear that a full adder is built up from two half adders. The *Sum* output line is 1 if an odd number of *A*, *B*, and the *Carry in* are 1. The *Carry out* is 1 if either *A* and *B* are both 1 (left input to the OR gate) or exactly one of them is 1 and the *Carry in* bit is also 1. Together the two half adders generate both the sum and the carry bits.

A	B	Carry in	Sum	Carry out
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

(a)



(b)

Figure 3-17. (a) Truth table for full adder. (b) Circuit for a full adder.

To build an adder for, say, two 16-bit words, one just replicates the circuit of Fig. 3-17(b) 16 times. The carry out of a bit is used as the carry into its left neighbor. The carry into the rightmost bit is wired to 0. This type of adder is called a **ripple carry adder**, because in the worst case, adding 1 to 111...111 (binary), the addition cannot complete until the carry has rippled all the way from the rightmost bit to the leftmost bit. Adders that do not have this delay, and hence are faster, also exist and are usually preferred.

As a simple example of a faster adder, consider breaking up a 32-bit adder into a 16-bit lower half and a 16-bit upper half. When the addition starts, the upper adder cannot yet get to work because it will not know the carry into it for 16 addition times.

However, consider this modification to the circuit. Instead of having a single upper half, give the adder two upper halves in parallel by duplicating the upper

half's hardware. Thus, the circuit now consists of three 16-bit adders: a lower half and two upper halves, *U0* and *U1* that run in parallel. A 0 is fed into *U0* as a carry; a 1 is fed into *U1* as a carry. Now both of these can start at the same time the lower half starts, but only one will be correct. After 16 bit-addition times, it will be known what the carry into the upper half is, so the correct upper half can now be selected from the two available answers. This trick reduces the addition time by a factor of two. Such an adder is called a **carry select adder**. This trick can then be repeated to build each 16-bit adder out of replicated 8-bit adders, and so on.

Arithmetic Logic Units

Most computers contain a single circuit for performing the AND, OR, and sum of two machine words. Typically, such a circuit for n-bit words is built up of n identical circuits for the individual bit positions. Figure 3-18 is a simple example of such a circuit, called an **Arithmetic Logic Unit** or **ALU**. It can compute any one of four functions—namely, A AND B, A OR B, $\bar{B}$, or $A + B$, depending on whether the function-select input lines F_0 and F_1 contain 00, 01, 10, or 11 (binary). Note that here $A + B$ means the arithmetic sum of A and B, not the Boolean OR.

The lower left-hand corner of our ALU contains a 2-bit decoder to generate enable signals for the four operations, based on the control signals F_0 and F_1. Depending on the values of F_0 and F_1, exactly one of the four enable lines is selected. Setting this line allows the output for the selected function to pass through to the final OR gate for output.

The upper left-hand corner has the logic to compute A AND B, A OR B, and $\bar{B}$, but at most one of these results is passed onto the final OR gate, depending on the enable lines coming out of the decoder. Because exactly one of the decoder outputs will be 1, exactly one of the four AND gates driving the OR gate will be enabled; the other three will output 0, independent of A and B.

In addition to being able to use A and B as inputs for logical or arithmetic operations, it is also possible to force either one to 0 by negating ENA or ENB, respectively. It is also possible to get $\bar{A}$, by setting INVA. We will see uses for INVA, ENA, and ENB in Chap. 4. Under normal conditions, ENA and ENB are both 1 to enable both inputs and INVA is 0. In this case, A and B are just fed into the logic unit unmodified.

The lower right-hand corner of the ALU contains a full adder for computing the sum of A and B, including handling the carries, because it is likely that several of these circuits will eventually be wired together to perform full-word operations. Circuits like Fig. 3-18 are actually available and are known as **bit slices**. They allow the computer designer to build an ALU of any desired width. Figure 3-19 shows an 8-bit ALU built up of eight 1-bit ALU slices. The INC signal is useful only for addition operations. When present, it increments (i.e., adds 1 to) the result, making it possible to compute sums like $A + 1$ and $A + B + 1$.

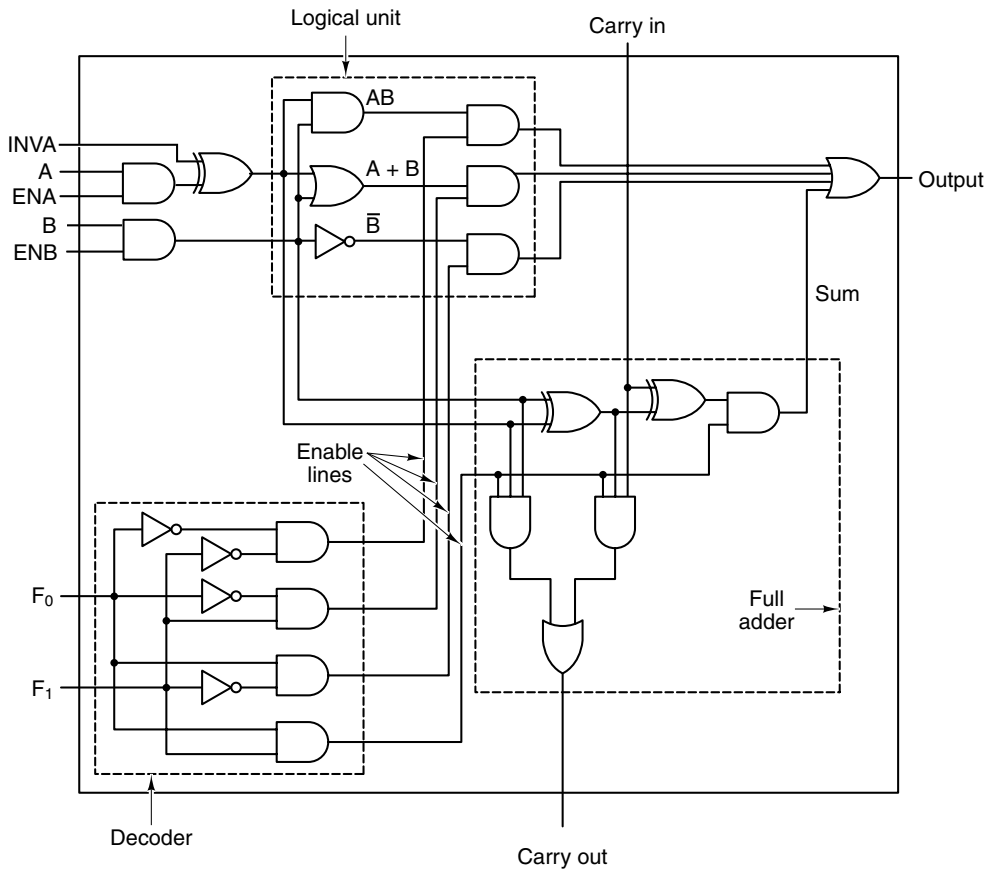


Figure 3-18. A 1-bit ALU.

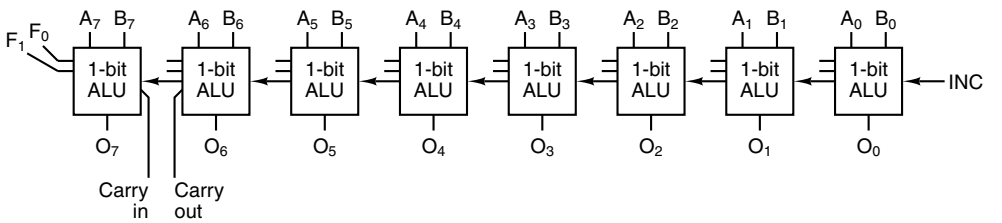


Figure 3-19. Eight 1-bit ALU slices connected to make an 8-bit ALU. The enables and invert signals are not shown for simplicity.

Years ago, a bit slice was an actual chip you could buy. Nowadays, a bit slice is more likely to be a library a chip designer can replicate the desired number of times in a computer-aided-design program that produces an output file that drives the chip-production machines. But the idea is essentially the same.

3.2.4 Clocks

In many digital circuits the order in which events happen is critical. Sometimes one event must precede another, sometimes two events must occur simultaneously. To allow designers to achieve the required timing relations, many digital circuits use clocks to provide synchronization. A **clock** in this context is a circuit that emits a series of pulses with a precise pulse width and precise interval between consecutive pulses. The time interval between the corresponding edges of two consecutive pulses is known as the **clock cycle time**. Pulse frequencies are commonly between 100 MHz and 4 GHz, corresponding to clock cycles of 10 nsec to 250 psec. To achieve high accuracy, the clock frequency is usually controlled by a crystal oscillator.

In a computer, many events may happen during a single clock cycle. If these events must occur in a specific order, the clock cycle must be divided into subcycles. A common way of providing finer resolution than the basic clock is to tap the primary clock line and insert a circuit with a known delay in it, thus generating a secondary clock signal that is phase-shifted from the primary, as shown in Fig. 3-20(a). The timing diagram of Fig. 3-20(b) provides four time references for discrete events:

1. Rising edge of C1.
2. Falling edge of C1.
3. Rising edge of C2.
4. Falling edge of C2.

By tying different events to the various edges, the required sequencing can be achieved. If more than four time references are needed within a clock cycle, more secondary lines can be tapped from the primary, with one with a different delay if necessary.

In some circuits, one is interested in time intervals rather than discrete instants of time. For example, some event may be allowed to happen whenever C1 is high, rather than precisely at the rising edge. Another event may happen only when C2 is high. If more than two different intervals are needed, more clock lines can be provided or the high states of the two clocks can be made to overlap partially in time. In the latter case four distinct intervals can be distinguished: $\overline{\text{C1}}$ AND $\overline{\text{C2}}$, $\overline{\text{C1}}$ AND C2, C1 AND $\overline{\text{C2}}$, and C1 AND C2.

As an aside, clocks are symmetric, with time spent in the high state equal to the time spent in the low state, as shown in Fig. 3-20(b). To generate an asymmetric pulse train, the basic clock is shifted using a delay circuit and ANDed with the original signal, as shown in Fig. 3-20(c) as C.

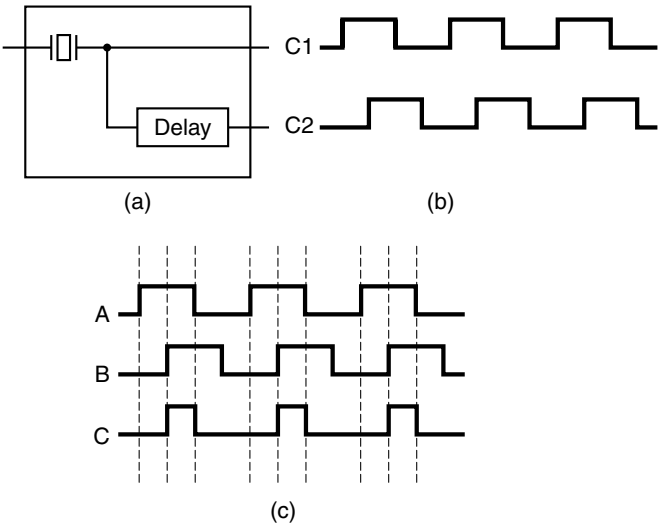


Figure 3-20. (a) A clock. (b) The timing diagram for the clock. (c) Generation of an asymmetric clock.

3.3 MEMORY

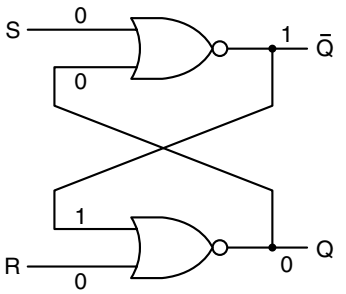
An essential component of every computer is its memory. Without memory there could be no computers as we now know them. Memory is used for storing both instructions to be executed and data. In the following sections we will examine the basic components of a memory system starting at the gate level to see how they work and how they are combined to produce large memories.

3.3.1 Latches

To create a 1-bit memory, we need a circuit that somehow "remembers" previous input values. Such a circuit can be constructed from two NOR gates, as illustrated in Fig. 3-21(a). Analogous circuits can be built from NAND gates. We will not mention these further, however, because they are conceptually identical to the NOR versions.

The circuit of Fig. 3-21(a) is called an **SR latch**. It has two inputs, S, for Setting the latch, and R, for Resetting (i.e., clearing) it. It also has two outputs, Q and $\bar{Q}$, which are complementary, as we will see shortly. Unlike a combinational circuit, the outputs of the latch are not uniquely determined by the current inputs.

To see how this comes about, let us assume that both S and R are 0, which they are most of the time. For argument's sake, let us further assume that $Q = 0$. Because Q is fed back into the upper NOR gate, both of its inputs are 0, so its output,



A	B	NOR
0	0	1
0	1	0
1	0	0
1	1	0

(c)

Figure 3-21. (a) NOR latch in state 0. (b) NOR latch in state 1. (c) Truth table for NOR.

$\bar{Q}$, is 1. The 1 is fed back into the lower gate, which then has inputs 1 and 0, yielding $Q = 0$. This state is at least consistent and is depicted in Fig. 3-21(a).

Now let us imagine that Q is not 0 but 1, with R and S still 0. The upper gate has inputs of 0 and 1, and an output, $\bar{Q}$, of 0, which is fed back to the lower gate. This state, shown in Fig. 3-21(b), is also consistent. A state with both outputs equal to 0 is inconsistent, because it forces both gates to have two 0s as input, which, if true, would produce 1, not 0, as output. Similarly, it is impossible to have both outputs equal to 1, because that would force the inputs to 0 and 1, which yields 0, not 1. Our conclusion is simple: for $R = S = 0$, the latch has two stable states, which we will refer to as 0 and 1, depending on Q.

Now let us examine the effect of the inputs on the state of the latch. Suppose that S becomes 1 while $Q = 0$. The inputs to the upper gate are then 1 and 0, forcing the $\bar{Q}$ output to 0. This change makes both inputs to the lower gate 0, forcing the output to 1. Thus, setting S (i.e., making it 1) switches the state from 0 to 1. Setting R to 1 when the latch is in state 0 has no effect because the output of the lower NOR gate is 0 for inputs of 10 and inputs of 11.

Using similar reasoning, it is easy to see that setting S to 1 when in state $Q = 1$ has no effect but that setting R drives the latch to state $Q = 0$. In summary, when S is set to 1 momentarily, the latch ends up in state $Q = 1$, regardless of what state it was previously in. Likewise, setting R to 1 momentarily forces the latch to state $Q = 0$. The circuit "remembers" whether S or R was last on. Using this property, we can build computer memories.

Clocked SR Latches

It is often convenient to prevent the latch from changing state except at certain specified times. To achieve this goal, we modify the basic circuit slightly, as shown in Fig. 3-22, to get a **clocked SR latch**.

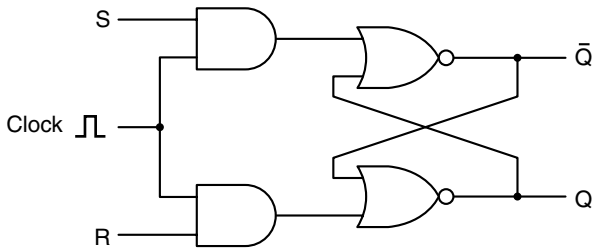


Figure 3-22. A clocked SR latch.

This circuit has an additional input, the clock, which is normally 0. With the clock 0, both AND gates output 0, independent of S and R, and the latch does not change state. When the clock is 1, the effect of the AND gates vanishes and the latch becomes sensitive to S and R. Despite its name, the clock signal need not be driven by a clock. The terms **enable** and **strobe** are also widely used to mean that the clock input is 1; that is, the circuit is sensitive to the state of S and R.

Up until now we have carefully swept under the rug the problem of what happens when both S and R are 1. And for good reason: the circuit becomes nondeterministic when both R and S finally return to 0. The only consistent state for $S = R = 1$ is $Q = \bar{Q} = 0$, but as soon as both inputs return to 0, the latch must jump to one of its two stable states. If either input drops back to 0 before the other, the one remaining 1 longest wins, because when just one input is 1, it forces the state. If both inputs return to 0 simultaneously (which is very unlikely), the latch jumps to one of its stable states at random.

Clocked D Latches

A good way to resolve the SR latch's instability (caused when $S = R = 1$) is to prevent it from occurring. Figure 3-23 gives a latch circuit with only one input, D. Because the input to the lower AND gate is always the complement of the input to the upper one, the problem of both inputs being 1 never arises. When $D = 1$ and the clock is 1, the latch is driven into state $Q = 1$. When $D = 0$ and the clock is 1, it is driven into state $Q = 0$. In other words, when the clock is 1, the current value of D is sampled and stored in the latch. This circuit, called a **clocked D latch**, is a true 1-bit memory. The value stored is always available at Q. To load the current value of D into the memory, a positive pulse is put on the clock line.

This circuit requires 11 transistors. More sophisticated (but less obvious) circuits can store 1 bit with as few as six transistors. In practice, such designs are normally used. This circuit can remain stable indefinitely as long as power (not shown) is applied. Later we will see memory circuits that quickly forget what state they are in unless constantly "reminded" somehow.

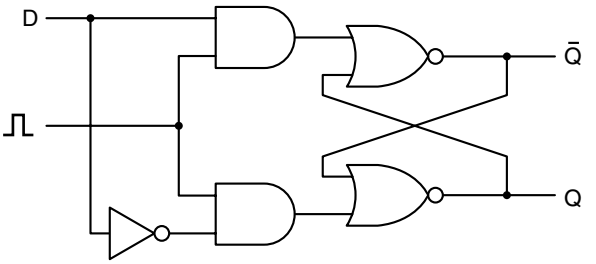


Figure 3-23. A clocked D latch.

3.3.2 Flip-Flops

In many circuits it is necessary to sample the value on a certain line at a particular instant in time and store it. In this variant, called a **flip-flop**, the state transition occurs not when the clock is 1 but during the clock transition from 0 to 1 (rising edge) or from 1 to 0 (falling edge) instead. Thus, the length of the clock pulse is unimportant, as long as the transitions occur fast.

For emphasis, we will repeat the difference between a flip-flop and a latch. A flip-flop is **edge triggered**, whereas a latch is **level triggered**. Be warned, however, that in the literature these terms are often confused. Many authors use "flip-flop" when they are referring to a latch, and vice versa.

There are various approaches to designing a flip-flop. For example, if there were some way to generate a very short pulse on the rising edge of the clock signal, that pulse could be fed into a D latch. There is, in fact, such a way, and the circuit for it is shown in Fig. 3-24(a).

At first glance, it might appear that the output of the AND gate would always be zero, since the AND of any signal with its inverse is zero, but the situation is a bit more subtle than that. The inverter has a small, but nonzero, propagation delay through it, and that delay is what makes the circuit work. Suppose that we measure the voltage at the four measuring points *a*, *b*, *c*, and *d*. The input signal, measured at *a*, is a long clock pulse, as shown in Fig. 3-24(b) on the bottom. The signal at *b* is shown above it. Notice that it is both inverted and delayed slightly, typically hundreds of picoseconds, depending on the kind of inverter used.

The signal at *c* is delayed, too, but only by the signal propagation time (at the speed of light). If the physical distance between *a* and *c* is, for example, 20 microns, then the propagation delay is 0.0001 nsec, which is certainly negligible compared to the time for the signal to propagate through the inverter. Thus, for all intents and purposes, the signal at *c* is as good as identical to the signal at *a*.

When the inputs to the AND gate, *b* and *c*, are ANDed together, the result is a short pulse, as shown in Fig. 3-24(b), where the width of the pulse, Δ, is equal to the gate delay of the inverter, typically 5 nsec or less. The output of the AND gate is just this pulse shifted by the delay of the AND gate, as shown at the top of

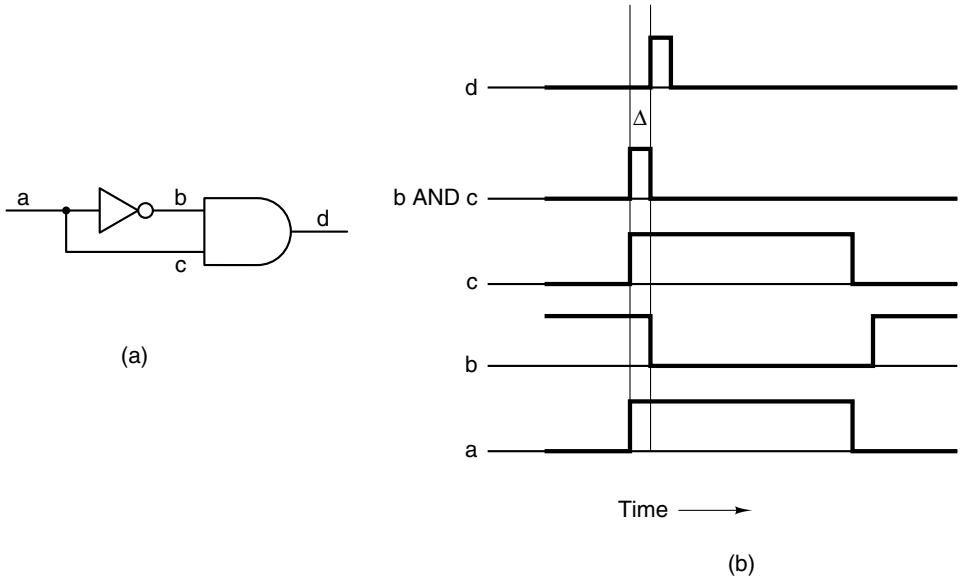


Figure 3-24. (a) A pulse generator. (b) Timing at four points in the circuit.

Fig. 3-24(b). This time shifting just means that the D latch will be activated at a fixed delay after the rising edge of the clock, but it has no effect on the pulse width. In a memory with a 10-nsec cycle time, a 1-nsec pulse telling it when to sample the *D* line may be short enough, in which case the full circuit can be that of Fig. 3-25. It is worth noting that this flip-flop design is nice because it is easy to understand, but in practice more sophisticated flip-flops are normally used.

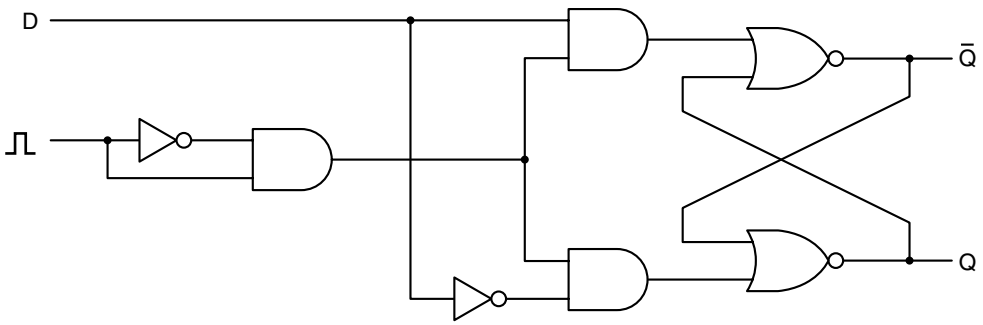


Figure 3-25. A D flip-flop.

The standard symbols for latches and flip-flops are shown in Fig. 3-26. Figure 3-26(a) is a latch whose state is loaded when the clock, *CK*, is 1. It is in contrast to Fig. 3-26(b) which is a latch whose clock is normally 1 but which drops to 0

momentarily to load the state from *D*. Figure 3-26(c) and (d) are flip-flops rather than latches, which is indicated by the pointy symbol on the clock inputs. Figure 3-26(c) changes state on the rising edge of the clock pulse (0-to-1 transition), whereas Fig. 3-26(d) changes state on the falling edge (1-to-0 transition). Many, but not all, latches and flip-flops also have $\bar{Q}$ as an output, and some have two additional inputs *Set* or *Preset* (force state to $Q = 1$) and *Reset* or *Clear* (force state to $Q = 0$).

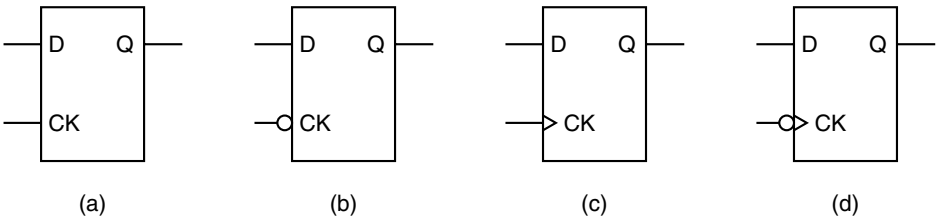


Figure 3-26. D latches and flip-flops.

3.3.3 Registers

Flip-flops can be combined in groups to create registers, which hold data types larger than 1 bit in length. The register in Fig. 3-27 shows how eight flip-flops can be ganged together to form an 8-bit storage register. The register accepts an 8-bit input value (*I*0 to *I*7) when the clock *CK* transitions. To implement a register, all the clock lines are connected to the same input signal *CK*, such that when the clock transitions, each register will accept the new 8-bit data value on the input bus. The flip-flops themselves are of the Fig. 3-26(d) type, but the inversion bubbles on the flip-flops are canceled by the inverter tied to the clock signal *CK*, such that the flip-flops are loaded on the rising transition of the clock. All eight clear signals are also ganged, so when the clear signal *CLR* goes to 0, all the flip-flops are forced to their 0 state. In case you are wondering why the clock signal *CK* is inverted at the input and then inverted again at each flip-flop, an input signal may not have enough current to drive all eight flip-flops; the input inverter is really being used as an amplifier.

Once we have designed an 8-bit register, we can use it as a building block to create larger registers. For example, a 32-bit register could be created by combining two 16-bit registers by tying their clock signals *CK* and clear signals *CLR*. We will look at registers and their uses more closely in Chap. 4.

3.3.4 Memory Organization

Although we have now progressed from the simple 1-bit memory of Fig. 3-23 to the 8-bit memory of Fig. 3-27, to build large memories a fairly different organization is required, one in which individual words can be addressed. A widely used

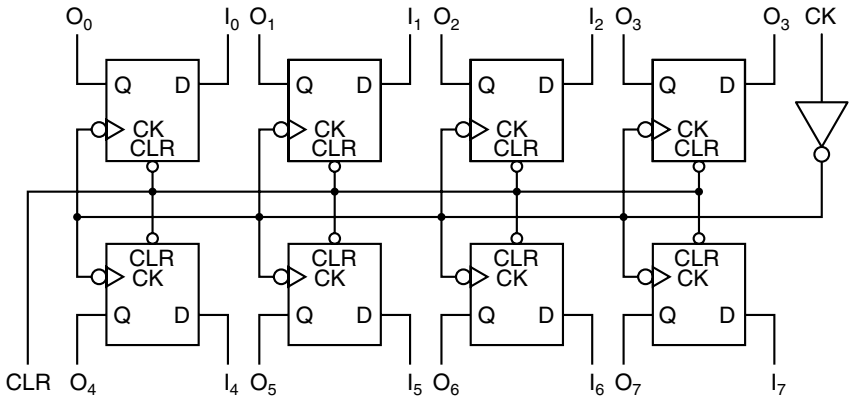


Figure 3-27. An 8-bit register constructed from single-bit flip-flops.

memory organization that meets this criterion is shown in Fig. 3-28. This example illustrates a memory with four 3-bit words. Each operation reads or writes a full 3-bit word. While the total memory capacity of 12 bits is hardly more than our octal flip-flop, it requires fewer pins and, most important, the design extends easily to large memories. Note the number of words is always a power of 2.

While the memory of Fig. 3-28 may look complicated at first, it is really quite simple due to its regular structure. It has eight input lines and three output lines. Three inputs are data: I_0, I_1, and I_2; two are for the address: A_0 and A_1; and three are for control: CS for Chip Select, RD for distinguishing between read and write, and OE for Output Enable. The three outputs are for data: O_0, O_1, and O_2. It is interesting to note that this 12-bit memory requires fewer signals than the previous 8-bit register. The 8-bit register requires 20 signals, including power and ground, while the 12-bit memory requires only 13 signals. The memory block requires fewer signals because, unlike the register, memory bits share an output signal. In this memory, 4 memory bits each share one output signal. The value of the address lines determine which of the 4 memory bits is allowed to input or output a value.

To select this memory block, external logic must set CS high and also set RD high (logical 1) for read and low (logical 0) for write. The two address lines must be set to indicate which of the four 3-bit words is to be read or written. For a read operation, the data input lines are not used, but the word selected is placed on the data output lines. For a write operation, the bits present on the data input lines are loaded into the selected memory word; the data output lines are not used.

Now let us look at Fig. 3-28 closely to see how it works. The four word-select AND gates at the left of the memory form a decoder. The input inverters have been placed so that each gate is enabled (output is high) by a different address. Each gate drives a word select line, from top to bottom, for words 0, 1, 2, and 3. When the chip has been selected for a write, the vertical line labeled CS $\cdot$ $\overline{\text{RD}}$ will be high,

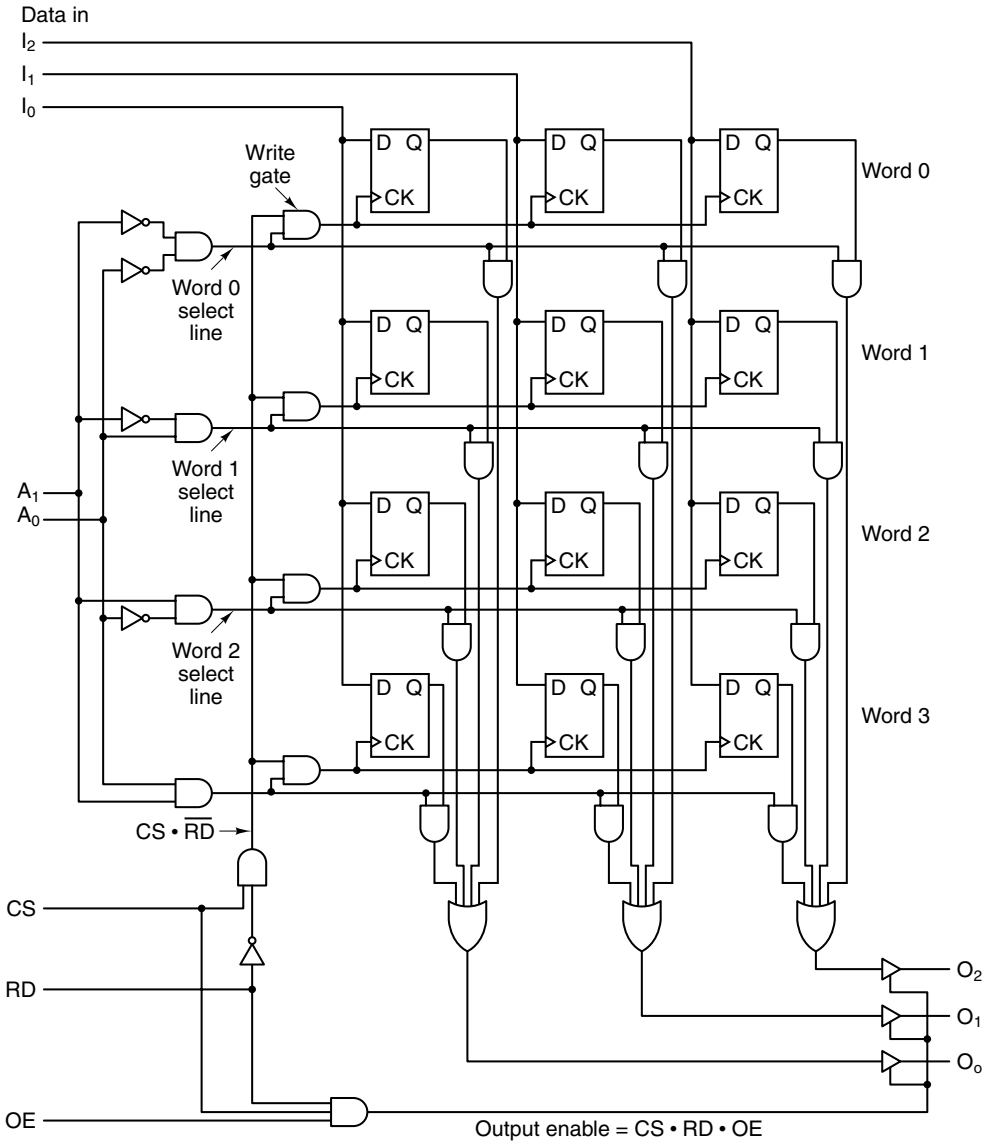


Figure 3-28. Logic diagram for a 4 × 3 memory. Each row is one of the four 3-bit words. A read or write operation always reads or writes a complete word.

enabling one of the four write gates, depending on which word select line is high. The output of the write gate drives all the CK signals for the selected word, loading the input data into the flip-flops for that word. A write is done only if CS is high and RD is low, and even then only the word selected by A_0 and A_1 is written; the other words are not changed at all.

Read is similar to write. The address decoding is exactly the same as for write. But now the CS · $\overline{\text{RD}}$ line is low, so all the write gates are disabled and none of the flip-flops is modified. Instead, the word select line that is chosen enables the AND gates tied to the Q bits of the selected word. Thus, the selected word outputs its data into the four-input OR gates at the bottom of the figure, while the other three words output 0s. Consequently, the output of the OR gates is identical to the value stored in the word selected. The three words not selected make no contribution to the output.

Although we could have designed a circuit in which the three OR gates were just fed into the three output data lines, doing so sometimes causes problems. In particular, we have shown the data input lines and the data output lines as being different, but in actual memories the same lines are used. If we had tied the OR gates to the data output lines, the chip would try to output data, that is, force each line to a specific value, even on writes, thus interfering with the input data. For this reason, it is desirable to have a way to connect the OR gates to the data output lines on reads but disconnect them completely on writes. What we need is an electronic switch that can make or break a connection in a fraction of a nanosecond.

Fortunately, such switches exist. Figure 3-29(a) shows the symbol for what is called a **noninverting buffer**. It has a data input, a data output, and a control input. When the control input is high, the buffer acts like a wire, as shown in Fig. 3-29(b). When the control input is low, the buffer acts like an open circuit, as shown in Fig. 3-29(c); it is as though someone detached the data output from the rest of the circuit with a wirecutter. However, in contrast to the wirecutter analogy the connection can be subsequently restored in a fraction of a nanosecond by just making the control signal high again.

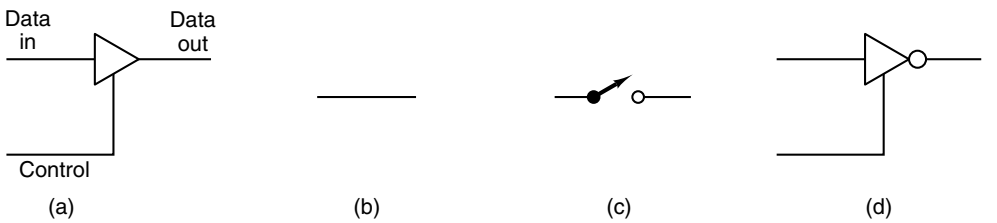


Figure 3-29. (a) A noninverting buffer. (b) Effect of (a) when control is high. (c) Effect of (a) when control is low. (d) An inverting buffer.

Figure 3-29(d) shows an **inverting buffer**, which acts like a normal inverter when control is high and disconnects the output from the circuit when control is low. Both kinds of buffers are **tri-state devices**, because they can output 0, 1, or none of the above (open circuit). Buffers also amplify signals, so they can drive many inputs simultaneously. They are sometimes used in circuits for this reason, even when their switching properties are not needed.

Getting back to the memory circuit, it should now be clear what the three noninverting buffers on the data output lines are for. When CS, RD, and OE are all high, the output enable signal is also high, enabling the buffers and putting a word onto the output lines. When any one of CS, RD, or OE is low, the data outputs are disconnected from the rest of the circuit.

3.3.5 Memory Chips

The nice thing about the memory of Fig. 3-28 is that it extends easily to larger sizes. As we drew it, the memory is 4×3, that is, four words of 3 bits each. To extend it to 4×8 we need only add five more columns of four flip-flops each, as well as five more input lines and five more output lines. To go from 4×3 to 8×3 we must add four more rows of three flip-flops each, as well as an address line A_2. With this kind of structure, the number of words in the memory should be a power of 2 for maximum efficiency, but the number of bits in a word can be anything.

Because integrated-circuit technology is well suited to making chips whose internal structure is a repetitive two-dimensional pattern, memory chips are an ideal application for it. As the technology improves, the number of bits that can be put on a chip keeps increasing, typically by a factor of two every 18 months (Moore's law). The larger chips do not always render the smaller ones obsolete due to different trade-offs in capacity, speed, power, price, and interfacing convenience. Commonly, the largest chips currently available sell at a premium and thus are more expensive per bit than older, smaller ones.

For any given memory size, there are various ways of organizing the chip. Figure 3-30 shows two possible organizations for an older memory chip of size 4 Mbit: $512\text{K} \times 8$ and $4096\text{K} \times 1$. (As an aside, memory-chip sizes are usually quoted in bits, rather than in bytes, so we will stick to that convention here.) In Fig. 3-30(a), 19 address lines are needed to address one of the 2^{19} bytes, and eight data lines are needed for loading or storing the byte selected.

A note on terminology is in order here. On some pins, the high voltage causes an action to happen. On others, the low voltage causes the action. To avoid confusion, we will consistently say that a signal is **asserted** (rather than saying it goes high or goes low) to mean that it is set to cause some action. Thus, for some pins, asserting it means setting it high. For others, it means setting the pin low. Pins that are asserted low are given signal names containing an overbar. Thus, a signal named CS is asserted high, but one named $\overline{\text{CS}}$ is asserted low. The opposite of asserted is **negated**. When nothing special is happening, pins are negated.

Now let us get back to our memory chip. Since a computer normally has many memory chips, a signal is needed to select the chip that is currently needed so that it responds and all the others do not. The $\overline{\text{CS}}$ (Chip Select) signal is provided for this purpose. It is asserted to enable the chip. Also, a way is needed to distinguish reads from writes. The $\overline{\text{WE}}$ signal (Write Enable) is used to indicate that data are

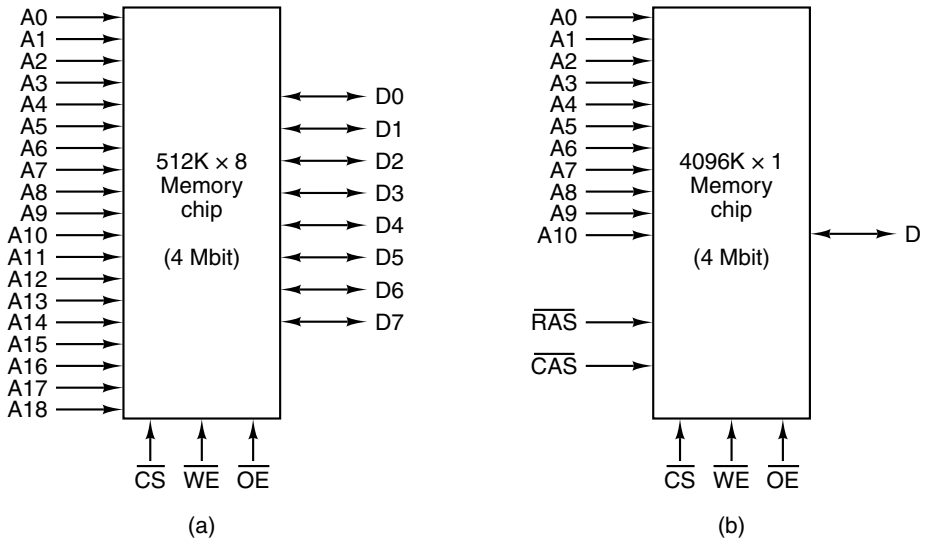


Figure 3-30. Two ways of organizing a 4-Mbit memory chip.

being written rather than being read. Finally, the $\overline{\text{OE}}$ (Output Enable) signal is asserted to drive the output signals. When it is not asserted, the chip output is disconnected from the circuit.

In Fig. 3-30(b), a different addressing scheme is used. Internally, this chip is organized as a 2048×2048 matrix of 1-bit cells, which gives 4 Mbits. To address the chip, first a row is selected by putting its 11-bit number on the address pins. Then the $\overline{\text{RAS}}$ (Row Address Strobe) is asserted. After that, a column number is put on the address pins and $\overline{\text{CAS}}$ (Column Address Strobe) is asserted. The chip responds by accepting or outputting one data bit.

Large memory chips are often constructed as $n \times n$ matrices that are addressed by row and column. This organization reduces the number of pins required but also makes addressing the chip slower, since two addressing cycles are needed, one for the row and one for the column. To win back some of the speed lost by this design, some memory chips can be given a row address followed by a sequence of column addresses to access consecutive bits in a row.

Years ago, the largest memory chips were often organized like Fig. 3-30(b). As memory words have grown from 8 bits to 32 bits and beyond, 1-bit-wide chips began to be inconvenient. To build a memory with a 32-bit word from $4096\text{K} \times 1$ chips requires 32 chips in parallel. These 32 chips have a total capacity of at least 16 MB, whereas using $512\text{K} \times 8$ chips requires only four chips in parallel and allows memories as small as 2 MB. To avoid having 32 chips for memory, most chip manufacturers now have chip families with 4-, 8-, and 16-bit widths. And the situation with 64-bit words is even worse, of course.

Two examples of 512-Mbit chips are given in Fig. 3-31. These chips have four internal memory banks of 128 Mbit each, requiring two bank select lines to choose a bank. The design of Fig. 3-31(a) is a 32M × 16 design, with 13 lines for the $\overline{\text{RAS}}$ signal, 10 lines for the $\overline{\text{CAS}}$ signal, and 2 lines for the bank select. Together, these 25 signals allow each of the 2^{25} internal 16-bit cells to be addressed. In contrast, Fig. 3-31(b) is a 128M × 4 design, with 13 lines for the $\overline{\text{RAS}}$ signal, 12 lines for the $\overline{\text{CAS}}$ signal, and 2 lines for the bank select. Here, 27 signals can select any of the 2^{27} internal 4-bit cells to be addressed. The decision about how many rows and how many columns a chip has is made for engineering reasons. The matrix need not be square.

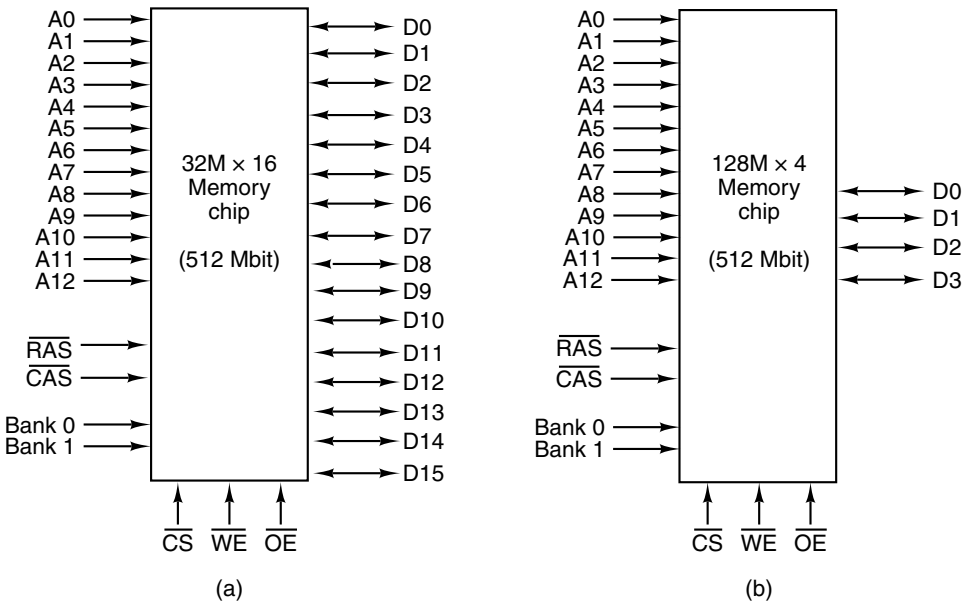


Figure 3-31. Two ways of organizing a 512-Mbit memory chip.

These examples demonstrate two separate and independent issues for memory-chip design. First is the output width (in bits): does the chip deliver 1, 4, 8, 16, or some other number of bits at once? Second, are all the address bits presented on separate pins at once or are the rows and columns presented sequentially as in the examples of Fig. 3-31? A memory-chip designer has to answer both questions before starting the chip design.

3.3.6 RAMs and ROMs

The memories we have studied so far can all be read and written. Such memories are called **RAM**s (Random Access Memories), which is a misnomer because all memory chips are randomly accessible, but the term is too well established to

get rid of now. RAMs come in two varieties, static and dynamic. **Static RAM**s (**SRAM**s) are constructed internally using circuits similar to our basic D flip-flop. These memories have the property that their contents are retained as long as the power is kept on: seconds, minutes, hours, even days. Static RAMs are very fast. A typical access time is on the order of a nanosecond or less. For this reason, static RAMS are popular as cache memory.

Dynamic RAMs (**DRAM**s), in contrast, do not use flip-flops. Instead, a dynamic RAM is an array of cells, each cell containing one transistor and a tiny capacitor. The capacitors can be charged or discharged, allowing 0s and 1s to be stored. Because the electric charge tends to leak out, each bit in a dynamic RAM must be **refreshed** (reloaded) every few milliseconds to prevent the data from leaking away. Because external logic must take care of the refreshing, dynamic RAMs require more complex interfacing than static ones, although in many applications this disadvantage is compensated for by their larger capacities.

Since dynamic RAMs need only one transistor and one capacitor per bit (vs. six transistors per bit for the best static RAM), dynamic RAMs have a very high density (many bits per chip). For this reason, main memories are nearly always built out of dynamic RAMs. However, this large capacity has a price: dynamic RAMs are slow (tens of nanoseconds). Thus, the combination of a static RAM cache and a dynamic RAM main memory attempts to combine the good properties of each.

Several types of dynamic RAM chips exist. The oldest type still around (in elderly computers) is **FPM** (**Fast Page Mode**) DRAM. Internally it is organized as a matrix of bits and it works by having the hardware present a row address and then step through the column addresses, as we described with $\overline{\text{RAS}}$ and $\overline{\text{CAS}}$ in the context of Fig. 3-30. Explicit signals tell the memory when it is time to respond, so the memory runs asynchronously from the main system clock.

FPM DRAM was replaced with **EDO** (**Extended Data Output**) DRAM, which allows a second memory reference to begin before the previous memory reference has been completed. This simple pipelining did not make a single memory reference go faster but did improve the memory bandwidth, giving more words per second.

FPM and EDO worked reasonably well when memory chips had cycle times of 12 nsec and slower. When processors got so fast that faster memories were really needed, FPM and EDO were replaced by **SDRAM** (**Synchronous DRAM**), which is a hybrid of static and dynamic RAM and is driven by the main system clock. The big advantage of SDRAM is that the clock eliminates the need for control signals to tell the memory chip when to respond. Instead, the CPU tells the memory how many cycles it should run, then starts it. On each subsequent cycle, the memory outputs 4, 8, or 16 bits, depending on how many output lines it has. Eliminating the need for control signals increases the data rate between CPU and memory.

The next improvement over SDRAM was **DDR** (**Double Data Rate**) SDRAM. With this kind of memory, the memory chip produces output on both the rising

edge of the clock and the falling edge, doubling the data rate. Thus, an 8-bit-wide DDR chip running at 200 MHz outputs two 8-bit values 200 million times a second (for a short interval, of course), giving a theoretical burst rate of 3.2 Gbps. The DDR2 and DDR3 memory interfaces provide additional performance over DDR by increasing the memory-bus speeds to 533 MHz and 1067 MHz, respectively. At the time this book went to press, the fastest DDR3 chips could output data at 17.067 GB/sec.

Nonvolatile Memory Chips

RAMs are not the only kind of memory chips. In many applications, such as toys, appliances, and cars, the program and some of the data must remain stored even when the power is turned off. Furthermore, once installed, neither the program nor the data are ever changed. These requirements have led to the development of **ROM**s (Read-Only Memories), which cannot be changed or erased, intentionally or otherwise. The data in a ROM are inserted during its manufacture, essentially by exposing a photosensitive material through a mask containing the desired bit pattern and then etching away the exposed (or unexposed) surface. The only way to change the program in a ROM is to replace the entire chip.

ROMs are much cheaper than RAMs when ordered in large enough volumes to defray the cost of making the mask. However, they are inflexible, because they cannot be changed after manufacture, and the turnaround time between placing an order and receiving the ROMs may be weeks. To make it easier for companies to develop new ROM-based products, the **PROM** (Programmable ROM) was invented. A PROM is like a ROM, except that it can be programmed (once) in the field, eliminating the turnaround time. Many PROMs contain an array of tiny fuses inside. A specific fuse can be blown out by selecting its row and column and then applying a high voltage to a special pin on the chip.

The next development in this line was the **EPROM** (Erasable PROM), which can be not only field-programmed but also field-erased. When the quartz window in an EPROM is exposed to a strong ultraviolet light for 15 minutes, all the bits are set to 1. If many changes are expected during the design cycle, EPROMs are far more economical than PROMs because they can be reused. EPROMs usually have the same organization as static RAMs. The 4-Mbit 27C040 EPROM, for example, uses the organization of Fig. 3-31(a), which is typical of a static RAM. What is interesting is that ancient chips like this one do not die off. They just become cheaper and find their way into lower-end products that are highly cost sensitive. A 27C040 can now be bought retail for under $3 and much less in large volumes.

Even better than the EPROM is the **EEPROM** which can be erased by applying pulses to it instead of putting it in a special chamber for exposure to ultraviolet light. In addition, an EEPROM can be reprogrammed in place, whereas an EPROM has to be inserted in a special EPROM programming device to be programmed. On the minus side, the biggest EEPROMs are typically only 1/64 as

large as common EPROMs and they are only half as fast. EEPROMs cannot compete with DRAMs or SRAMs because they are 10 times slower, 100 times smaller in capacity, and much more expensive. They are used only in situations where their nonvolatility is crucial.

A more recent kind of EEPROM is **flash memory**. Unlike EPROM, which is erased by exposure to ultraviolet light, and EEPROM, which is byte erasable, flash memory is block erasable and rewritable. Like EEPROM, flash memory can be erased without removing it from the circuit. Various manufacturers produce small printed-circuit cards with up to 64 GB of flash memory on them for use as "film" for storing pictures in digital cameras and many other purposes. As we discussed in Chap. 2, flash memory is now starting to replace mechanical disks. As a disk, flash memory provides faster access times at lower power, but with a much higher cost per bit. A summary of the various kinds of memory is given in Fig. 3-32.

Type	Category	Erasure	Byte alterable	Volatile	Typical use
SRAM	Read/write	Electrical	Yes	Yes	Level 2 cache
DRAM	Read/write	Electrical	Yes	Yes	Main memory (old)
SDRAM	Read/write	Electrical	Yes	Yes	Main memory (new)
ROM	Read-only	Not possible	No	No	Large-volume appliances
PROM	Read-only	Not possible	No	No	Small-volume equipment
EPROM	Read-mostly	UV light	No	No	Device prototyping
EEPROM	Read-mostly	Electrical	Yes	No	Device prototyping
Flash	Read/write	Electrical	No	No	Film for digital camera

Figure 3-32. A comparison of various memory types.

Field-Programmable Gate Arrays

As we saw in Chap. 1, **field-programmable gate arrays** (**FPGA**s) are chips which contain programmable logic such that we can form arbitrary logic circuit by simply loading the FPGA with appropriate configuration data. The main advantage of FPGAs is that new hardware circuits can be built in hours, rather than the months it takes to fabricate ICs. Integrated circuits are not going the way of the dodo, however, as they still hold a significant cost advantage over FPGAs for high-volume applications, and they run faster and use much less power. Because of their design-time advantages, however, FPGAs are often used for design prototyping and low-volume applications.

Let's now look inside an FPGA and understand how it can be used to implement a wide range of logic circuits. The FPGA chip contains two primary components that are replicated many times: **LUT**s (**LookUp Tables**) and **programmable interconnects**. Let us now examine how they are used.

A LUT, shown in Fig. 3-33(a), is a small programmable memory that produces a signal output optionally to a register, which is then output to the programmable interconnect. The programmable memory is used to create an arbitrary logic function. The LUT in the figure has a 16×4 memory, which can emulate any logic circuit with 4 bits of input and 4 bits of output. Programming the LUT requires loading the memory with the appropriate responses of the combinational logic being emulated. In other words, if the combinational logic produces the value Y when given the input X, the value Y would be written into the LUT at index X.

The example design in Fig. 3-32(b) shows how a single 4-input LUT could implement a 3-bit counter with reset. The example counter continually counts up by adding one (modulo 4) to the current value of the counter, unless the reset signal *CLR* is asserted, in which case the counter resets its value to zero.

To implement the example counter, the upper four entries of the LUT are all zero. These entries output the value zero when the counter is reset. Thus, the most significant bit of the LUT input (I_3) represents the reset input (*CLR*) which is asserted with a logic 1. For the remaining LUT entries, the value at index $I_{0..3}$ of the LUT contains the value $(I + 1)$ modulo 4. To complete the design, the output signal $O_{0..3}$ must be connected, using the programmable interconnect to the internal input signal $I_{0..3}$.

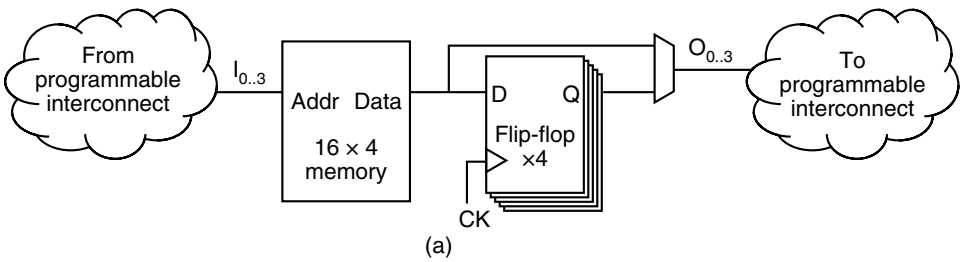


Figure 3-33. (a) A field-programmable logic array lookup table (LUT). (b) The LUT configuration to create a 3-bit clearable counter.

To better understand the FPGA-based counter with reset, let's consider its operation. If, for example, the current state of the counter is 2 and the reset (*CLR*) signal is not asserted, the input address to the LUT will be 2, which will produce an output to the flip-flops of 3. If the reset signal (*CLR*) were asserted for the same state, the input to the LUT would be 6, which would produce the next state of 0.

All in all, this may seem like an arcane way to build a counter with reset, and in fact a fully custom design with an incrementer circuit and reset signals to the flip-flops would be smaller, faster, and use less power. The main advantage of the FPGA-based design is that you can craft it in an hour at home, whereas the more efficient fully custom design must be fabricated from silicon, which could take a month or more.

To use an FPGA, the design must be described using a circuit description or a hardware description language (i.e., a programming language used to describe hardware structures). The design is then processed by a synthesizer, which maps the circuit to a specific FPGA architecture. One challenge of using FPGAs is that the design you want to map never seems to fit. FPGAs are manufactured with varying number of LUTs, with larger quantities costing more. In general, if your design does not fit, you need to simplify or throw away some functionality, or purchase a larger (and more expensive) FPGA. Very large designs may not fit into the largest FPGAs, which will require the designer to map the design into multiple FPGAs; this task is definitely more difficult, but still a walk in the park compared to designing a complete custom integrated circuit.

3.4 CPU CHIPS AND BUSES

Armed with information about integrated circuits, clocks, and memory chips, we can now start to put all the pieces together to look at complete systems. In this section, we will first look at some general aspects of CPUs as viewed from the digital logic level, including **pinout** (what the signals on the various pins mean). Since CPUs are so closely intertwined with the design of the buses they use, we will also provide an introduction to bus design in this section. In subsequent sections we will give detailed examples of both CPUs and their buses and how they are interfaced.

3.4.1 CPU Chips

All modern CPUs are contained on a single chip. This makes their interaction with the rest of the system well defined. Each CPU chip has a set of pins, through which all its communication with the outside world must take place. Some pins output signals from the CPU to the outside world; others accept signals from the outside world; some can do both. By understanding the function of all the pins, we can learn how the CPU interacts with the memory and I/O devices at the digital logic level.

The pins on a CPU chip can be divided into three types: address, data, and control. These pins are connected to similar pins on the memory and I/O chips via a collection of parallel wires called a bus. To fetch an instruction, the CPU first puts the memory address of that instruction on its address pins. Then it asserts one or

more control lines to inform the memory that it wants to read (for example) a word. The memory replies by putting the requested word on the CPU's data pins and asserting a signal saying that it is done. When the CPU sees this signal, it accepts the word and carries out the instruction.

The instruction may require reading or writing data words, in which case the whole process is repeated for each additional word. We will go into the detail of how reading and writing works below. For the time being, the important thing to understand is that the CPU communicates with the memory and I/O devices by presenting signals on its pins and accepting signals on its pins. No other communication is possible.

Two of the key parameters that determine the performance of a CPU are the number of address pins and the number of data pins. A chip with m address pins can address up to 2^m memory locations. Common values of m are 16, 32, and 64. Similarly, a chip with n data pins can read or write an n-bit word in a single operation. Common values of n are 8, 32, and 64. A CPU with 8 data pins will take four operations to read a 32-bit word, whereas one with 32 data pins can do the same job in one operation. Thus, the chip with 32 data pins is much faster but is invariably more expensive as well.

In addition to address and data pins, each CPU has some control pins. They regulate the flow and timing of data to and from the CPU and have other miscellaneous uses. All CPUs have pins for power (usually +1.2 to +1.5 volts), ground, and a clock signal (a square wave at some well-defined frequency), but the other pins vary greatly from chip to chip. Nevertheless, the control pins can be roughly grouped into the following major categories:

1. Bus control.
2. Interrupts.
3. Bus arbitration.
4. Coprocessor signaling.
5. Status.
6. Miscellaneous.

We will briefly describe each of these categories below. When we look at the Intel Core i7, TI OMAP4430, and Atmel ATmega168 chips later, we will provide more detail. A generic CPU chip using these signal groups is shown in Fig. 3-34.

The bus control pins are mostly outputs from the CPU to the bus (thus inputs to the memory and I/O chips) telling whether the CPU wants to read or write memory or do something else. The CPU uses these pins to control the rest of the system and tell it what it wants to do.

The interrupt pins are inputs from I/O devices to the CPU. In most systems, the CPU can tell an I/O device to start an operation and then go off and do some

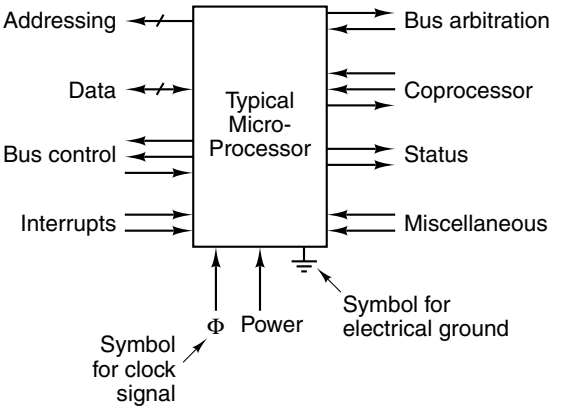


Figure 3-34. The logical pinout of a generic CPU. The arrows indicate input signals and output signals. The short diagonal lines indicate that multiple pins are used. For a specific CPU, a number will be given to tell how many.

other activity, while the I/O device is doing its work. When the I/O has been completed, the I/O controller chip asserts a signal on one of these pins to interrupt the CPU and have it service the I/O device, for example to check whether if I/O errors occurred. Some CPUs have an output pin to acknowledge the interrupt signal.

The bus arbitration pins are needed to regulate traffic on the bus, in order to prevent two devices from trying to use it at the same time. For arbitration purposes, the CPU counts as a device and has to request the bus like any other device.

Some CPU chips are designed to operate with coprocessors such as floating-point chips, but sometimes graphics or other chips as well. To facilitate communication between CPU and coprocessor, special pins are provided for making and granting various requests.

In addition to these signals, there are various miscellaneous pins that some CPUs have. Some of these provide or accept status information, others are useful for debugging or resetting the computer, and still others are present to assure compatibility with older I/O chips.

3.4.2 Computer Buses

A **bus** is a common electrical pathway between multiple devices. Buses can be categorized by their function. They can be used internal to the CPU to transport data to and from the ALU, or external to the CPU to connect it to memory or to I/O devices. Each type of bus has its own requirements and properties. In this section and the following ones, we will focus on buses that connect the CPU to the memory and I/O devices. In the next chapter we will examine more closely the buses inside the CPU.

Early personal computers had a single external bus or **system bus**. It consisted of 50 to 100 parallel copper wires etched onto the motherboard, with connectors spaced at regular intervals for plugging in memory and I/O boards. Modern personal computers generally have a special-purpose bus between the CPU and memory and (at least) one other bus for the I/O devices. A minimal system, with one memory bus and one I/O bus, is illustrated in Fig. 3-35.

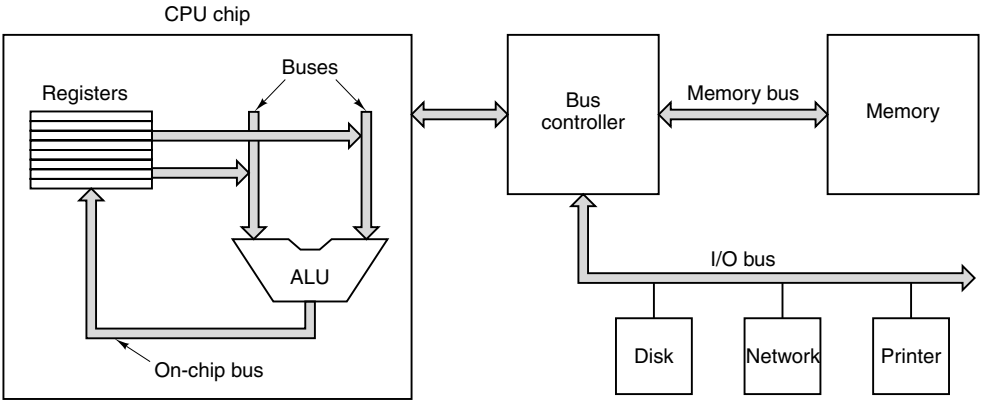


Figure 3-35. A computer system with multiple buses.

In the literature, buses are sometimes drawn as "fat" arrows, as in this figure. The distinction between a fat arrow and a single line with a diagonal line through it and a bit count next to it is subtle. When all the bits are of the same type, say, all address bits or all data bits, then the short-diagonal-line approach is commonly used. When address, data, and control lines are involved, a fat arrow is more common.

While the designers of the CPU are free to use any kind of bus they want inside the chip, in order to make it possible for boards designed by third parties to attach to the system bus, there must be well-defined rules about how the external bus works, which all devices attached to it must obey. These rules are called the **bus protocol**. In addition, there must be mechanical and electrical specifications, so that third-party boards will fit in the card cage and have connectors that match those on the motherboard mechanically and in terms of voltages, timing, etc. Still other buses do not have mechanical specifications because they are designed to be used only within an integrated circuit, for example, to connect components together within a system-on-a-chip (SoC).

A number of buses are in widespread use in the computer world. A few of the better-known ones, historical and current (with examples), are the Omnibus (PDP-8), Unibus (PDP-11), Multibus (8086), VME bus (physics lab equipment), IBM PC bus (PC/XT), ISA bus (PC/AT), EISA bus (80386), Microchannel (PS/2), Nubus (Macintosh), PCI bus (many PCs), SCSI bus (many PCs and workstations),

Universal Serial Bus (modern PCs), and FireWire (consumer electronics). The world would probably be a better place if all but one would suddenly vanish from the face of the earth (well, all right, how about all but two?). Unfortunately, standardization in this area seems very unlikely, as there is already too much invested in all these incompatible systems.

As an aside, there is another interconnect, PCI Express, that is widely referred to as a bus but is not a bus at all. We will study it later in this chapter.

Let us now begin our study of how buses work. Some devices that attach to a bus are active and can initiate bus transfers, whereas others are passive and wait for requests. The active ones are called **masters**; the passive ones are called **slaves**. When the CPU orders a disk controller to read or write a block, the CPU is acting as a master and the disk controller is acting as a slave. However, later on, the disk controller may act as a master when it commands the memory to accept the words it is reading from the disk drive. Several typical combinations of master and slave are listed in Fig. 3-36. Under no circumstances can memory ever be a master.

Master	Slave	Example
CPU	Memory	Fetching instructions and data
CPU	I/O device	Initiating data transfer
CPU	Coprocessor	CPU handing instruction off to coprocessor
I/O device	Memory	DMA (Direct Memory Access)
Coprocessor	CPU	Coprocessor fetching operands from CPU

Figure 3-36. Examples of bus masters and slaves.

The binary signals that computer devices output are frequently too weak to power a bus, especially if it is relatively long or has many devices on it. For this reason, most bus masters are connected to the bus by circuitry called a **bus driver**, which is essentially a digital amplifier. Similarly, most slaves are connected to the bus by a **bus receiver**. For devices that can act as both master and slave, a combined circuit called a **bus transceiver** is used. These bus interfaces are often tri-state devices, to allow them to float (disconnect) when they are not needed, or are hooked up in a somewhat different way, called **open collector**, that achieves a similar effect. When two or more devices on an open-collector line assert the line at the same time, the result is the Boolean OR of all the signals. This arrangement is often called **wired-OR**. On most buses, some of the lines are tri-state and others, which need the wired-OR property, are open collector.

Like a CPU, a bus also has address, data, and control lines. However, there is not necessarily a one-to-one mapping between the CPU pins and the bus signals. For example, some CPUs have three pins that encode whether the CPU is doing a memory read, memory write, I/O read, I/O write, or some other operation. A typical bus might have one line for memory read, a second for memory write, a third for I/O read, a fourth for I/O write, and so on. A decoder circuit would then be

needed between the CPU and such a bus to match the two sides up, that is, to convert the 3-bit encoded signal into separate signals that can drive the bus lines.

Bus design and operation are sufficiently complex subjects that a number of entire books have been written about them (Anderson et al., 2004, Solari and Willse, 2004). The principal bus design issues are bus width, bus clocking, bus arbitration, and bus operations. Each of these issues has a substantial impact on the speed and bandwidth of the bus. We will now examine each of these in the next four sections.

3.4.3 Bus Width

Bus width is the most obvious design parameter. The more address lines a bus has, the more memory the CPU can address directly. If a bus has n address lines, then a CPU can use it to address 2^n different memory locations. To allow large memories, buses need many address lines. That sounds simple enough.

The problem is that wide buses need more wires than narrow ones. They also take up more physical space (e.g., on the motherboard) and need bigger connectors. All of these factors make the bus more expensive. Thus, there is a tradeoff between maximum memory size and system cost. A system with a 64-line address bus and 2^{32} bytes of memory will cost more than one with 32 address lines and the same 2^{32} bytes of memory. The possibility of expansion later is not free.

The result of this observation is that many system designers tend to be shortsighted, with unfortunate consequences later. The original IBM PC contained an 8088 CPU and a 20-bit address bus, as shown in Fig. 3-37(a). Having 20 bits allowed the PC to address 1 MB of memory.

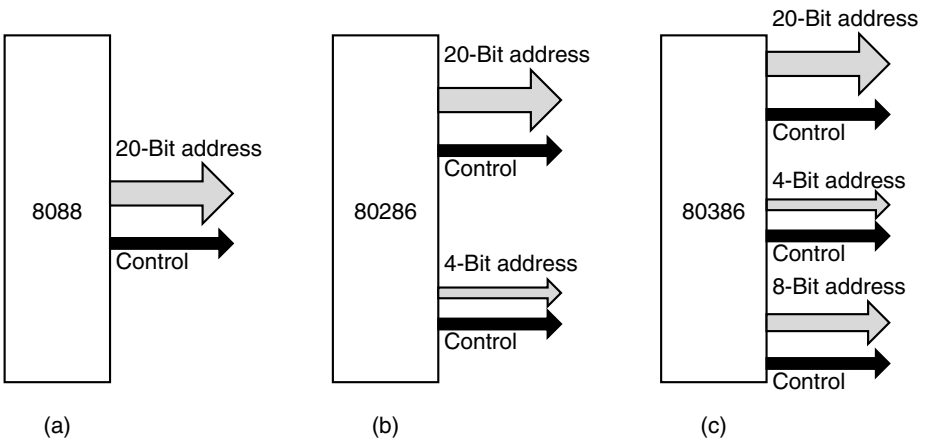


Figure 3-37. Growth of an address bus over time.

When the next CPU chip (the 80286) came out, Intel flet it had to increase the address space to 16 MB, so four more bus lines were added (without disturbing the

original 20, for reasons of backward compatibility), as illustrated in Fig. 3-37(b). Unfortunately, more control lines had to be added to deal with the new address lines. When the 80386 came out, another eight address lines were added, along with still more control lines, as shown in Fig. 3-37(c). The resulting design (the EISA bus) is much messier than it would have been had the bus been given 32 lines at the start.

Not only does the number of address lines tend to grow over time, but so does the number of data lines, albeit for a different reason. There are two ways to increase the bandwidth of a bus: decrease the bus cycle time (more transfers/sec) or increase the data bus width (more bits/transfer). Speeding the bus up is possible (but difficult) because the signals on different lines travel at slightly different speeds, a problem known as **bus skew**. The faster the bus, the more the skew.

Another problem with speeding up the bus is it will not be backward compatible. Old boards designed for the slower bus will not work with the new one. Invalidating old boards makes both the owners and manufacturers of the old boards unhappy. Therefore the usual approach to improving performance is to add more data lines, analogous to Fig. 3-37. As you might expect, however, this incremental growth does not lead to a clean design in the end. The IBM PC and its successors, for example, went from 8 data lines to 16 and then 32 on essentially the same bus.

To get around the problem of very wide buses, sometimes designers opt for a **multiplexed bus**. In this design, instead of the address and data lines being separate, there are, say, 32 lines for address and data together. At the start of a bus operation, the lines are used for the address. Later on, they are used for data. For a write to memory, for example, this means that the address lines must be set up and propagated to the memory before the data can be put on the bus. With separate lines, the address and data can be put on together. Multiplexing the lines reduces bus width (and cost) but results in a slower system. Bus designers have to carefully weigh all these options when making choices.

3.4.4 Bus Clocking

Buses can be divided into two distinct categories depending on their clocking. A **synchronous bus** has a line driven by a crystal oscillator. The signal on this line consists of a square wave with a frequency generally between 5 and 133 MHz. All bus activities take an integral number of these cycles, called **bus cycles**. The other kind of bus, the **asynchronous bus**, does not have a master clock. Bus cycles can be of any length required and need not be the same between all pairs of devices. Below we will examine each bus type.

Synchronous Buses

As an example of how a synchronous bus works, consider the timing of Fig. 3-38(a). In this example, we will use a 100-MHz clock, which gives a bus cycle of 10 nsec. While this may seem a bit slow compared to CPU speeds of 3

GHz and more, few existing PC buses are much faster. For example, the popular PCI bus usually runs at either 33 or 66 MHz, and the upgraded (but now defunct) PCI-X bus ran at a speed of up to 133 MHz. The reasons current buses are slow were given above: technical design problems such as bus skew and the need for backward compatibility.

In our example, we will further assume that reading from memory takes 15 nsec from the time the address is stable. As we will see shortly, with these parameters, it will take three bus cycles to read a word. The first cycle starts at the rising edge of T_1 and the third one ends at the rising edge of T_4, as shown in the figure. Note that none of the rising or falling edges has been drawn vertically, because no electrical signal can change its value in zero time. In this example we will assume that it takes 1 nsec for a signal to change. The clock, ADDRESS, DATA, $\overline{\text{MREQ}}$, $\overline{\text{RD}}$, and $\overline{\text{WAIT}}$ lines are all shown on the same time scale.

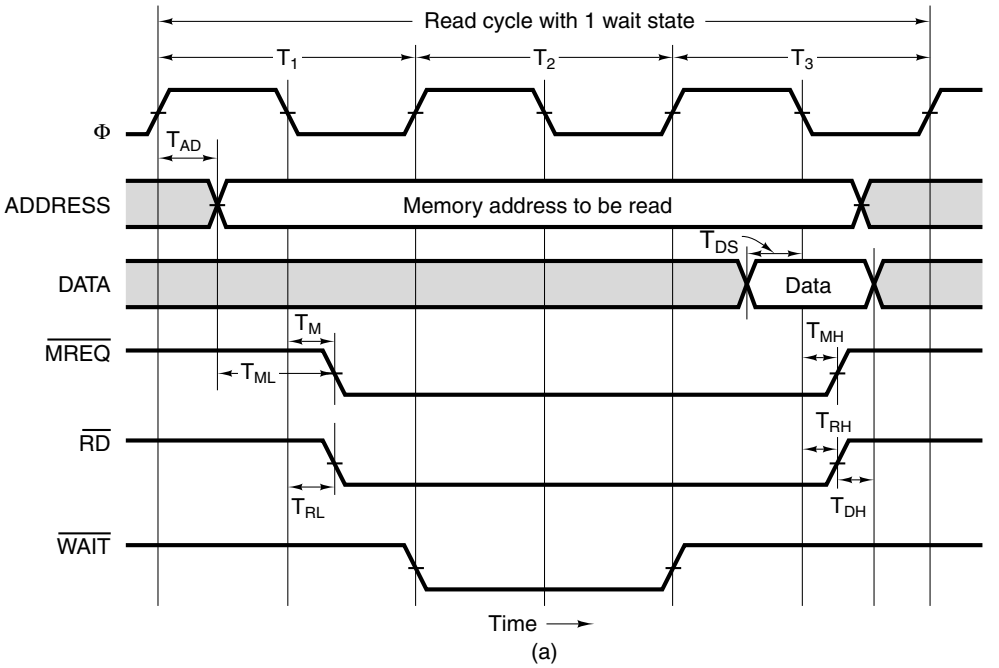
The start of T_1 is defined by the rising edge of the clock. Partway through T_1 the CPU puts the address of the word it wants on the address lines. Because the address is not a single value, like the clock, we cannot show it as a single line in the figure; instead, it is shown as two lines, with a crossing at the time that the address changes. Furthermore, the shading prior to the crossing indicates that the shaded value is not important. Using the same shading convention, we see that the contents of the data lines are not significant until well into T_3.

After the address lines have had a chance to settle down to their new values, $\overline{\text{MREQ}}$ and $\overline{\text{RD}}$ are asserted. The former indicates that memory (as opposed to an I/O device) is being accessed, and the latter is asserted for reads and negated for writes. Since the memory takes 15 nsec after the address is stable (partway into the first clock cycle), it cannot provide the requested data during T_2. To tell the CPU not to expect it, the memory asserts the $\overline{\text{WAIT}}$ line at the start of T_2. This action will insert **wait states** (extra bus cycles) until the memory is finished and negates $\overline{\text{WAIT}}$. In our example, one wait state (T_2) has been inserted because the memory is too slow. At the start of T_3, when it is sure it will have the data during the current cycle, the memory negates $\overline{\text{WAIT}}$.

During the first half of T_3, the memory puts the data onto the data lines. At the falling edge of T_3 the CPU strobes (i.e., reads) the data lines, latching (i.e., storing) the value in an internal register. Having read the data, the CPU negates $\overline{\text{MREQ}}$ and $\overline{\text{RD}}$. If need be, another memory cycle can begin at the next rising edge of the clock. This sequence can be repeatedly indefinitely.

In the timing specification of Fig. 3-38(b), eight symbols that occur in the timing diagram are further clarified. T_{AD}, for example, is the time interval between the rising edge of the T_1 clock and the address lines being set. According to the timing specification, $T_{\text{AD}} \leq 4$ nsec. This means that the CPU manufacturer guarantees that during any read cycle, the CPU will output the address to be read within 4 nsec of the midpoint of the rising edge of T_1.

The timing specifications also require that the data be available on the data lines at least T_{DS} (2 nsec) before the falling edge of T_3, to give it time to settle



Symbol	Parameter	Min	Max	Unit
T_{AD}	Address output delay		4	nsec
T_{ML}	Address stable prior to $\overline{\text{MREQ}}$	2		nsec
T_M	$\overline{\text{MREQ}}$ delay from falling edge of Φ in T_1		3	nsec
T_{RL}	RD delay from falling edge of Φ in T_1		3	nsec
T_{DS}	Data setup time prior to falling edge of Φ	2		nsec
T_{MH}	$\overline{\text{MREQ}}$ delay from falling edge of Φ in T_3		3	nsec
T_{RH}	$\overline{\text{RD}}$ delay from falling edge of Φ in T_3		3	nsec
T_{DH}	Data hold time from negation of $\overline{\text{RD}}$	0		nsec

(b)

Figure 3-38. (a) Read timing on a synchronous bus. (b) Specification of some critical times.

down before the CPU strobes it in. The combination of constraints on T_{AD} and T_{DS} means that, in the worst case, the memory will have only $25 - 4 - 2 = 19$ nsec from the time the address appears until it must produce the data. Because 10 nsec is enough, even in the worst case, a 10-nsec memory can always respond during T_3. A 20-nsec memory, however, would just miss and have to insert a second wait state and respond during T_4.

The timing specification further guarantees that the address will be set up at least 2 nsec prior to $\overline{\text{MREQ}}$ being asserted. This time can be important if $\overline{\text{MREQ}}$ drives chip select on the memory chip because some memories require an address setup time prior to chip select. Clearly, the system designer should not choose a memory chip that needs a 3-nsec setup time.

The constraints on T_M and T_{RL} mean that $\overline{\text{MREQ}}$ and $\overline{\text{RD}}$ will both be asserted within 3 nsec from the T_1 falling clock. In the worst case, the memory chip will have only 10 + 10 − 3 − 2 = 15 nsec after the assertion of $\overline{\text{MREQ}}$ and $\overline{\text{RD}}$ to get its data onto the bus. This constraint is in addition to (and independent of) the 15-nsec interval needed after the address is stable.

T_{MH} and T_{RH} tell how long it takes $\overline{\text{MREQ}}$ and $\overline{\text{RD}}$ to be negated after the data have been strobed in. Finally, T_{DH} tells how long the memory must hold the data on the bus after $\overline{\text{RD}}$ has been negated. As far as our example CPU is concerned, the memory can remove the data from the bus as soon as $\overline{\text{RD}}$ has been negated. On some actual CPUs, however, the data must be kept stable a little longer.

We would like to point out that Fig. 3-38 is a highly simplified version of real timing constraints. In reality, many more critical times are always specified. Nevertheless, it gives a good flavor for how a synchronous bus works.

A last point worth making is that control signals can be asserted high or low. It is up to the bus designers to determine which is more convenient, but the choice is essentially arbitrary. One can regard it as the hardware equivalent of a programmer's choice to represent free disk blocks in a bit map as 0s vs. 1s.

Asynchronous Buses

Although synchronous buses are easy to work with due to their discrete time intervals, they also have some problems. For example, everything works in multiples of the bus clock. If a CPU and memory are able to complete a transfer in 3.1 cycles, they have to stretch it to 4.0 because fractional cycles are forbidden.

Worse yet, once a bus cycle has been chosen, and memory and I/O cards have been built for it, it is difficult to take advantage of future improvements in technology. For example, suppose a few years after the system of Fig. 3-38 was built, new memories became available with access times of 8 nsec instead of 15 nsec. These would get rid of the wait state, speeding up the machine. Then suppose 4-nsec memories became available. There would be no further gain in performance because the minimum time for a read is two cycles with this design.

Putting this in slightly different terms, if a synchronous bus has a heterogeneous collection of devices, some fast and some slow, the bus has to be geared to the slowest one and the fast ones cannot use their full potential.

Mixed technology can be handled by going to an asynchronous bus, that is, one with no master clock, as shown in Fig. 3-39. Instead of tying everything to the clock, when the bus master has asserted the address, $\overline{\text{MREQ}}$, $\overline{\text{RD}}$, and anything else

it needs to, it then asserts a special signal that we will call $\overline{\text{MSYN}}$ (Master SYNchronization). When the slave sees this, it performs the work as fast as it can. When it is done, it asserts $\overline{\text{SSYN}}$ (Slave SYNchronization).

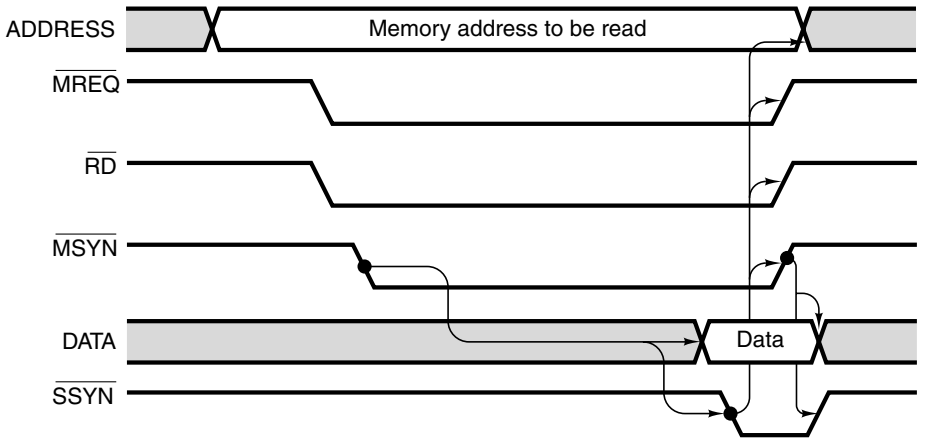


Figure 3-39. Operation of an asynchronous bus.

As soon as the master sees $\overline{\text{SSYN}}$ asserted, it knows that the data are available, so it latches them and then negates the address lines, along with $\overline{\text{MREQ}}$, $\overline{\text{RD}}$, and $\overline{\text{MSYN}}$. When the slave sees the negation of $\overline{\text{MSYN}}$, it knows that the cycle has been completed, so it negates $\overline{\text{SSYN}}$, and we are back in the original situation, with all signals negated, waiting for the next master.

Timing diagrams of asynchronous buses (and sometimes synchronous buses as well) use arrows to show cause and effect, as in Fig. 3-39. The assertion of $\overline{\text{MSYN}}$ causes the data lines to be asserted and also causes the slave to assert $\overline{\text{SSYN}}$. The assertion of $\overline{\text{SSYN}}$, in turn, causes the negation of the address lines, $\overline{\text{MREQ}}$, $\overline{\text{RD}}$, and $\overline{\text{MSYN}}$. Finally, the negation of $\overline{\text{MSYN}}$ causes the negation of $\overline{\text{SSYN}}$, which ends the read and returns the system to its original state.

A set of signals that interlocks this way is called a **full handshake**. The essential part consists of four events:

1. $\overline{\text{MSYN}}$ is asserted.
2. $\overline{\text{SSYN}}$ is asserted in response to $\overline{\text{MSYN}}$.
3. $\overline{\text{MSYN}}$ is negated in response to $\overline{\text{SSYN}}$.
4. $\overline{\text{SSYN}}$ is negated in response to the negation of $\overline{\text{MSYN}}$.

It should be clear that full handshakes are timing independent. Each event is caused by a prior event, not by a clock pulse. If a particular master/slave pair is slow, that in no way affects a subsequent master/slave pair that is much faster.

The advantage of an asynchronous bus should now be clear, but the fact is that most buses are synchronous. The reason is that it is easier to build a synchronous system. The CPU just asserts its signals, and the memory just reacts. There is no feedback (cause and effect), but if the components have been chosen properly, everything will work without handshaking. Also, there is a lot of investment in synchronous bus technology.

3.4.5 Bus Arbitration

Up until now, we have tacitly assumed that there is only one bus master, the CPU. In reality, I/O chips have to become bus master to read and write memory, and also to cause interrupts. Coprocessors may also need to become bus master. The question then arises: "What happens if two or more devices all want to become bus master at the same time?" The answer is that some **bus arbitration** mechanism is needed to prevent chaos.

Arbitration mechanisms can be centralized or decentralized. Let us first consider centralized arbitration. One particularly simple form of this is shown in Fig. 3-40(a). In this scheme, a single bus arbiter determines who goes next. Many CPUs have the arbiter built into the CPU chip, but sometimes a separate chip is needed. The bus contains a single wired-OR request line that can be asserted by one or more devices at any time. There is no way for the arbiter to tell how many devices have requested the bus. The only categories it can distinguish are some requests and no requests.

When the arbiter sees a bus request, it issues a grant by asserting the bus grant line. This line is wired through all the I/O devices in series, like a cheap string of Christmas tree lamps. When the device physically closest to the arbiter sees the grant, it checks to see if it has made a request. If so, it takes over the bus but does not propagate the grant further down the line. If it has not made a request, it propagates the grant to the next device in line, which behaves the same way, and so on until some device accepts the grant and takes the bus. This scheme is called **daisy chaining**. It has the property that devices are effectively assigned priorities depending on how close to the arbiter they are. The closest device wins.

To get around the implicit priorities based on distance from the arbiter, many buses have multiple priority levels. For each priority level there is a bus request line and a bus grant line. The one of Fig. 3-40(b) has two levels, 1 and 2 (real buses often have 4, 8, or 16 levels). Each device attaches to one of the bus request levels, with more time-critical devices attaching to the higher-priority ones. In Fig. 3-40(b) devices, 1, 2, and 4 use priority 1 while devices 3 and 5 use priority 2.

If multiple priority levels are requested at the same time, the arbiter issues a grant only on the highest-priority one. Among devices of the same priority, daisy chaining is used. In Fig. 3-40(b), in the event of conflicts, device 2 beats device 4, which beats 3. Device 5 has the lowest-priority because it is at the end of the lowest priority daisy chain.

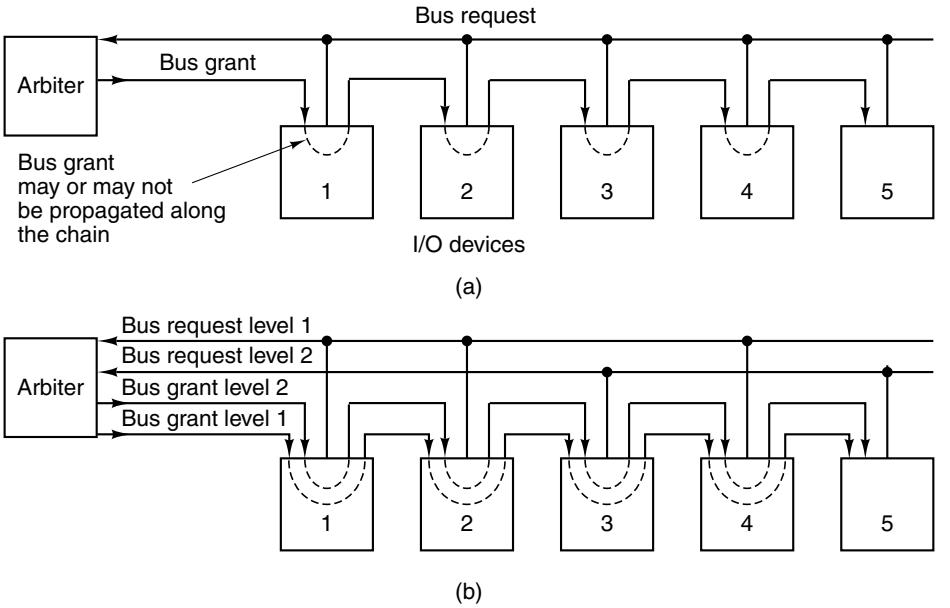


Figure 3-40. (a) A centralized one-level bus arbiter using daisy chaining. (b) The same arbiter, but with two levels.

As an aside, it is not technically necessary to wire the level 2 bus grant line serially through devices 1 and 2, since they cannot make requests on it. However, as an implementation convenience, it is easier to wire all the grant lines through all the devices, rather than making special wiring that depends on which device has which priority.

Some arbiters have a third line that a device asserts when it has accepted a grant and seized the bus. As soon as it has asserted this acknowledgement line, the request and grant lines can be negated. As a result, other devices can request the bus while the first device is using the bus. By the time the current transfer is finished, the next bus master will have already been selected. It can start as soon as the acknowledgement line has been negated, at which time the following round of arbitration can begin. This scheme requires an extra bus line and more logic in each device, but it makes better use of bus cycles.

In systems in which memory is on the main bus, the CPU must compete with all the I/O devices for the bus on nearly every cycle. One common solution for this situation is to give the CPU the lowest priority, so it gets the bus only when nobody else wants it. The idea here is that the CPU can always wait, but I/O devices frequently must acquire the bus quickly or lose incoming data. Disks rotating at high speed cannot wait. This problem is avoided in many modern computer systems by

putting the memory on a separate bus from the I/O devices so they do not have to compete for access to the bus.

Decentralized bus arbitration is also possible. For example, a computer could have 16 prioritized bus request lines. When a device wants to use the bus, it asserts its request line. All devices monitor all the request lines, so at the end of each bus cycle, each device knows whether it was the highest-priority requester, and thus whether it is permitted to use the bus during the next cycle. Compared to centralized arbitration, this arbitration method requires more bus lines but avoids the potential cost of the arbiter. It also limits the number of devices to the number of request lines.

Another kind of decentralized bus arbitration, shown in Fig. 3-41, uses only three lines, no matter how many devices are present. The first bus line is a wired-OR line for requesting the bus. The second bus line is called BUSY and is asserted by the current bus master. The third line is used to arbitrate the bus. It is daisy chained through all the devices. The head of this chain is held asserted by tying it to the power supply.

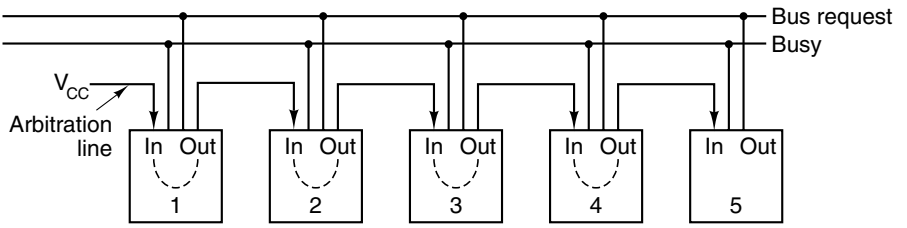


Figure 3-41. Decentralized bus arbitration.

When no device wants the bus, the asserted arbitration line is propagated through to all devices. To acquire the bus, a device first checks to see if the bus is idle and the arbitration signal it is receiving, IN, is asserted. If IN is negated, it may not become bus master, and it negates OUT. If IN is asserted, however, and the device wants the bus, the device negates OUT, which causes its downstream neighbor to see IN negated and to negate its OUT. Then all downstream devices all see IN negated and correspondingly negate OUT. When the dust settles, only one device will have IN asserted and OUT negated. This device becomes bus master, asserts BUSY and OUT, and begins its transfer.

Some thought will reveal that the leftmost device that wants the bus gets it. Thus, this scheme is similar to the original daisy chain arbitration, except without having the arbiter, so it is cheaper, faster, and not subject to arbiter failure.

3.4.6 Bus Operations

Up until now, we have discussed only ordinary bus cycles, with a master (typically the CPU) reading from a slave (typically the memory) or writing to one. In fact, several other kinds of bus cycles exist. We will now look at some of these.

Normally, one word at a time is transferred. However, when caching is used, it is desirable to fetch an entire cache line (e.g., 8 consecutive 64-bit words) at once. Often block transfers can be made more efficient than successive individual transfers. When a block read is started, the bus master tells the slave how many words are to be transferred, for example, by putting the word count on the data lines during T_1. Instead of just returning one word, the slave outputs one word during each cycle until the count has been exhausted. Figure 3-42 shows a modified version of Fig. 3-38(a), but now with an extra signal $\overline{\text{BLOCK}}$ that is asserted to indicate that a block transfer is requested. In this example, a block read of 4 words takes 6 cycles instead of 12.

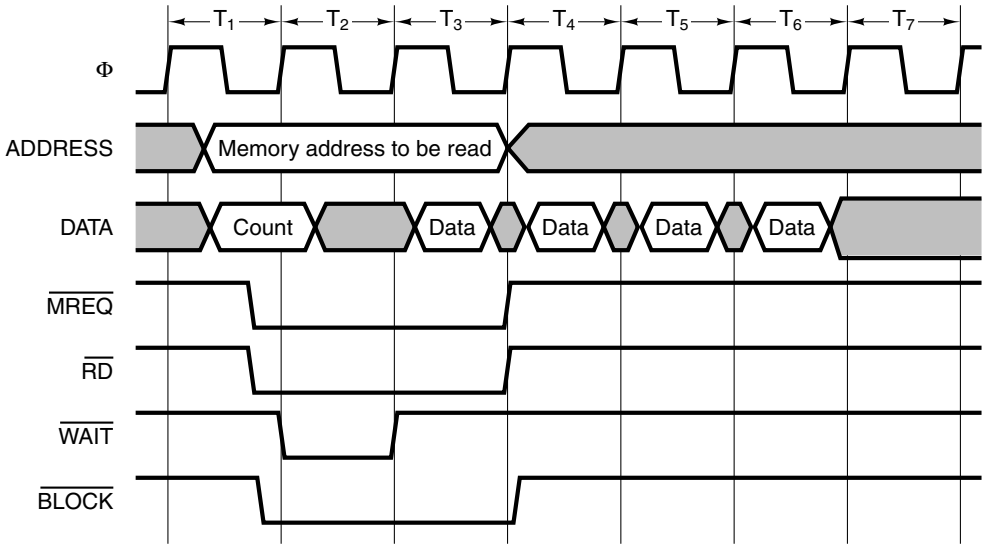


Figure 3-42. A block transfer.

Other kinds of bus cycles also exist. For example, on a multiprocessor system with two or more CPUs on the same bus, it is often necessary to make sure that only one CPU at a time uses some critical data structure in memory. A typical way to arrange this is to have a variable in memory that is 0 when no CPU is using the data structure and 1 when it is in use. If a CPU wants to gain access to the data structure, it must read the variable, and if it is 0, set it to 1. The trouble is, with some bad luck, two CPUs might read it on consecutive bus cycles. If each one sees that the variable is 0, then each one sets it to 1 and thinks that it is the only CPU using the data structure. This sequence of events leads to chaos.

To prevent this situation, multiprocessor systems often have a special read-modify-write bus cycle that allows any CPU to read a word from memory, inspect and modify it, and write it back to memory, all without releasing the bus.

This type of cycle prevents competing CPUs from being able to use the bus and thus interfere with the first CPU's operation.

Another important kind of bus cycle is for handling interrupts. When the CPU commands an I/O device to do something, it usually expects an interrupt when the work is done. The interrupt signaling requires the bus.

Since multiple devices may want to cause an interrupt simultaneously, the same kind of arbitration problems are present here that we had with ordinary bus cycles. The usual solution is to assign priorities to devices and use a centralized arbiter to give priority to the most time-critical devices. Standard interrupt interfaces exist and are widely used. In Intel-processor-based PCs, the chipset incorporates an 8259A interrupt controller, illustrated in Fig. 3-43.

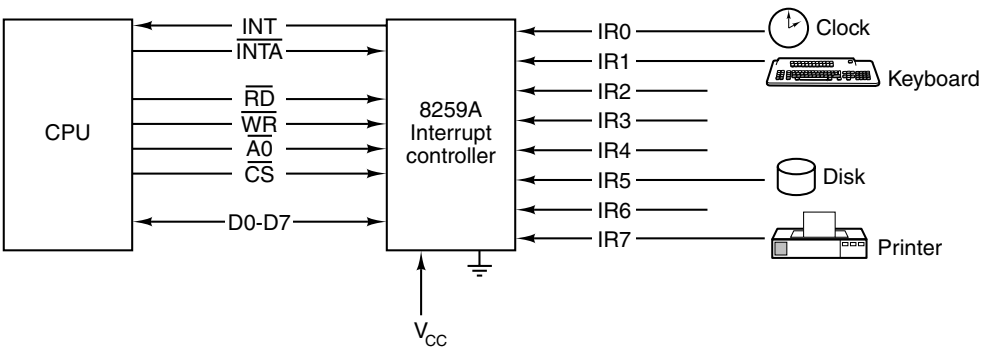


Figure 3-43. Use of the 8259A interrupt controller.

Up to eight 8259A I/O controllers can be directly connected to the eight IRx (Interrupt Request) inputs to the 8259A. When any of these devices wants to cause an interrupt, it asserts its input line. When one or more inputs are asserted, the 8259A asserts INT (INTerrupt), which directly drives the interrupt pin on the CPU. When the CPU is able to handle the interrupt, it sends a pulse back to the 8259A on INTA (INTerrupt Acknowledge). At that point the 8259A must specify which input caused the interrupt by outputting that input's number on the data bus. This operation requires a special bus cycle. The CPU hardware then uses that number to index into a table of pointers, called **interrupt vectors**, to find the address of the procedure to run to service the interrupt.

The 8259A has several registers inside that the CPU can read and write using ordinary bus cycles and the $\overline{\text{RD}}$ (ReaD), $\overline{\text{WR}}$ (WRite), $\overline{\text{CS}}$ (Chip Select), and $\overline{\text{A0}}$ pins. When the software has handled the interrupt and is ready to take the next one, it writes a special code into one of the registers, which causes the 8259A to negate INT, unless it has another interrupt pending. These registers can also be written to put the 8259A in one of several modes, mask out a set of interrupts, and enable other features.

When more than eight I/O devices are present, the 8259As can be cascaded. In the most extreme case, all eight inputs can be connected to the outputs of eight

more 8259As, allowing for up to 64 I/O devices in a two-stage interrupt network. The Intel ICH10 I/O controller hub, one of the of the chips in the Core i7 chipset, incorporates two 8259A interrupt controllers. This gives the ICH10 15 external interrupts, one less than 16 interrupts on the two 8259A controllers because one of the interrupts is used to cascade the second 8259A onto the first one. The 8259A has a few pins devoted to handling this cascading, which we have omitted for the sake of simplicity. Nowadays, the "8259A" is really part of another chip.

While we have by no means exhausted the subject of bus design, the material above should give enough background to understand the essentials of how a bus works, and how CPUs and buses interact. Let us now move from the general to the specific and look at some examples of actual CPUs and their buses.

3.5 EXAMPLE CPU CHIPS

In this section we will examine the Intel Core i7, TI OMAP4430, and Atmel ATmega168 chips in some detail at the hardware level.

3.5.1 The Intel Core i7

The Core i7 is a direct descendant of the 8088 CPU used in the original IBM PC. The first Core i7 was introduced in November 2008 as a four-processor 731-million transistor CPU running up to 3.2 GHz with a line width of 45 nanometers. The line width is how wide the wires between transistors are (as well as being a measure of the size of the transistors themselves). The narrower the line width, the more transistors can fit on the chip. Moore's law is fundamentally about the ability of process engineers to keep reducing the line widths. For comparison purposes, human hairs range from 20,000 to 100,000 nanometers in diameter, with blonde hair being finer than black hair.

The initial release of the Core i7 architecture was based on the "Nahalem" architecture; however, the newest versions of the Core i7 are built on the newer "Sandy Bridge" architecture. The architecture in this context represents the internal organization of the CPU, which is often given a code name. Despite being generally serious people, computer architects will sometimes come up with very clever code names for their projects. One of particular note was the AMD K-series architectures, which were designed to break Intel's seeming invulnerable hold on the desktop CPU market. The K-series processors' code name was "Kryptonite," a reference to the only substance that could hurt Superman, and a clever jab at the dominant Intel.

The new Sandy-Bridge-based Core i7 has evolved to having 1.16 billion transistors and running at speeds up to 3.5 GHz with line widths of 32 nanometers. Although the Core i7 is a far cry from the 29,000-transistor 8088, it is fully backward

compatible with the 8088 and can run unmodified 8088 binary programs (not to mention programs for all the intermediate processors as well).

From a software point of view, the Core i7 is a full 64-bit machine. It has all the same user-level ISA features as the 80386, 80486, Pentium, Pentium II, Pentium Pro, Pentium III, and Pentium 4 including the same registers, same instructions, and a full on-chip implementation of the IEEE 754 floating-point standard. In addition, it has some new instructions intended primarily for cryptographic operations.

The Core i7 processor is a multicore CPU, thus the silicon die contains multiple processors. The CPU is sold with a varying number of processors, ranging from 2 to 6 with more planned for the near future. If programmers write a parallel program, using threads and locks, it is possible to gain significant program speedups by exploiting parallelism on multiple processors. In addition, the individual CPUs are "hyperthreaded" such that multiple hardware threads can be active simultaneously. Hyperthreading (more typically called "simultaneous multithreading" by computer architects) allows very short latencies, such as cache misses, to be tolerated with hardware thread switches. Software-based threading can tolerate only very long latencies, such as page faults, due to the hundreds of cycles needed to implement software-based thread switches.

Internally, at the microarchitecture level, the Core i7 is a very capable design. It is based on the architecture of its predecessors, the Core 2 and Core 2 Duo. The Core i7 processor can carry out up to four instructions at once, making it a 4-wide superscalar machine. We will examine the microarchitecture in Chap. 4.

The Core i7 processors all have three levels of cache. Each processor in a Core i7 processor has a 32-KB level 1 (L1) data cache and a 32-KB level 1 instruction cache. Each core also has its own 256-KB level 2 (L2) cache. The second-level cache is unified, which means that it can hold a mixture of instructions and data. All cores share a single level 3 (L3) unified cache, the size of which varies from 4 to 15 MB depending on the processor model. Having three levels of cache significantly improves processor performance but at a great cost in silicon area, as Core i7 CPUs can have as much as 17 MB total cache on a single silicon die.

Since all Core i7 chips have multiple processors with private data caches, a problem arises when a processor modifies a word in this private cache that is contained in another processor's cache. If the other processor tries to read that word from memory, it will get a stale value, since modified cache words are not written back to memory immediately. To maintain memory consistency, each CPU in a multiprocessor system **snoops** on the memory bus looking for references to words it has cached. When it sees such a reference, it jumps in and supplies the required data before the memory gets a chance to do so. We will study snooping in Chap. 8.

Two primary external buses are used in Core i7 systems, both of them synchronous. A DDR3 memory bus is used to access the main memory DRAM, and a PCI Express bus connects the processor to I/O devices. High-end versions of the Core i7 include multiple memory and PCI Express buses, and they also include a

Quick Path Interconnect (QPI) port. The QPI port connects the processor to an external multiprocessor interconnect, allowing systems with more than six processors to be built. The QPI port sends and receives cache coherency requests, plus a variety of other multiprocessor management messages such as interprocessor interrupts.

A problem with the Core i7 as well as with most other modern desktop-class CPUs, is the power it consumes and the heat it generates. To prevent damaging the silicon, the heat must be moved away from the processor die soon after it is produced. The Core i7 consumes between 17 and 150 watts, depending on the frequency and model. Consequently, Intel is constantly searching for ways to manage the heat produced by its CPU chips. Cooling technologies and heat-conductive packaging are vital to protecting the silicon from burning up.

The Core i7 comes in a square LGA package 37.5 mm on edge. It contains 1155 pads on the bottom, of which 286 are for power and 360 are grounded to reduce noise. The pads are arranged roughly as a 40×40 square, with the middle 17×25 missing. In addition, 20 more pads are missing at the perimeter in an asymmetric pattern, to prevent the chip from being inserted incorrectly in its socket. The physical pinout is shown in Fig. 3-44.

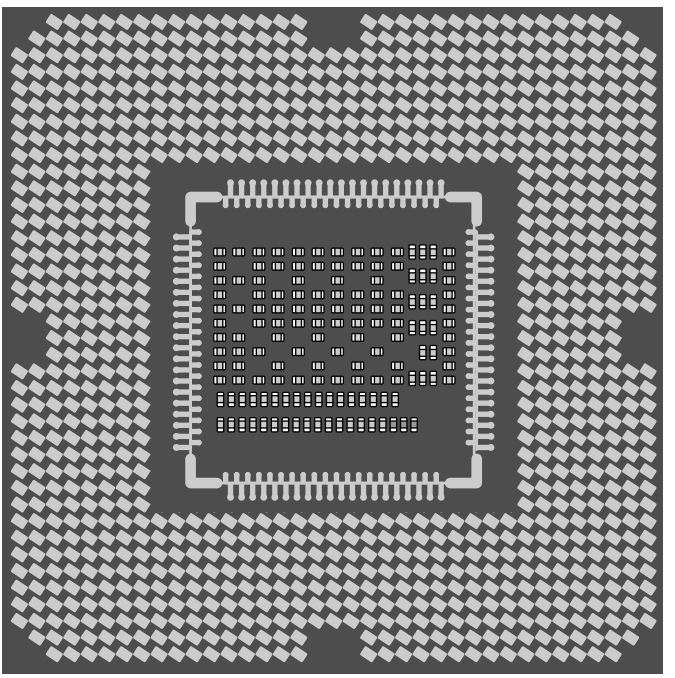


Figure 3-44. The Core i7 physical pinout.

The chip is outfitted with a mounting plate for a heat sink to distribute the heat and a fan to cool it. To get some idea of how large the power problem is, turn on a

150-watt incandescent light bulb, let it warm up, and then put your hands around it (but do not touch it). This amount of heat must be dissipated continuously by a high-end Core i7 processor. Consequently, when a Core i7 has outlived its usefulness as a CPU, it can always be used as a camp stove.

According to the laws of physics, anything that puts out a lot of heat must suck in a lot of energy. In a portable computer with a limited battery charge, using a lot of energy is not desirable because it drains the battery quickly. To address this issue, Intel has provided a way to put the CPU to sleep when it is idle and to put it into a deep sleep when it is likely to be that way for a while. Five states are provided, ranging from fully active to deep sleep. In the intermediate states, some functionality (such as cache snooping and interrupt handling) is enabled, but other functions are disabled. When in deep sleep state, the register values are preserved, but the caches are flushed and turned off. When in deep sleep, a hardware signal is required to wake it up. It is not known whether a Core i7 can dream when it is in deep sleep.

The Core i7's Logical Pinout

The 1155 pins on the Core i7 are used for 447 signals, 286 power connections (at several different voltages), 360 grounds, and 62 spares for future use. Some of the logical signals use two or more pins (such as the memory-address requested), so there are only 131 different signals. A somewhat simplified logical pinout is given in Fig. 3-45. On the left side of the figure are five major groups of bus signals; on the right side are various miscellaneous signals.

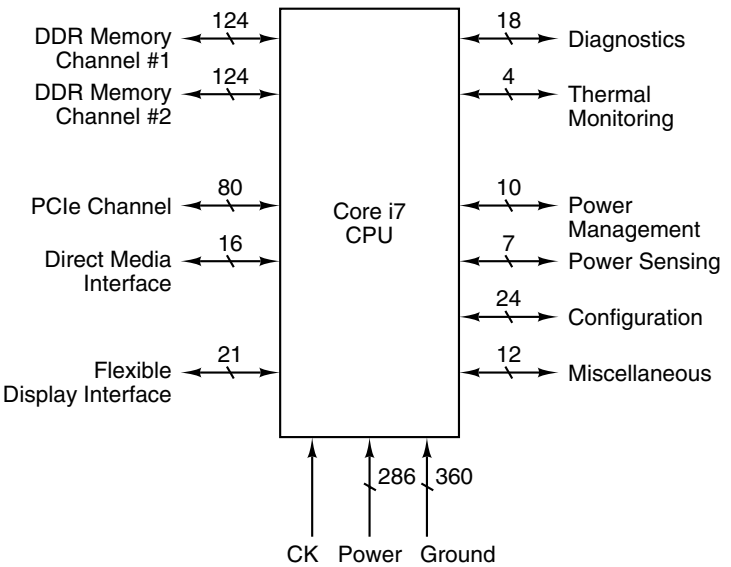


Figure 3-45. Logical pinout of the Core i7.

Let us examine the signals, starting with the bus signals. The first two bus signals are used to interface to DDR3 compatible DRAM. This group of signals provides address, data, control and clock to the DRAM bank. The Core i7 supports two independent DDR3 DRAM channels, running with a 666-MHz bus clock that transfers on both edges to allow 1333 million transactions per second. The DDR3 interface is 64 bits wide, thus, the two DDR3 interfaces work in tandem to provide memory-hungry programs up to 20 gigabytes of data each second.

The third bus group is the PCI Express interface, which is used to connect peripherals directly to the Core i7 CPU. The PCI Express interface is a high-speed serial interface, with each single serial link forming a "lane" of communication with peripherals. The Core i7 link is an x16 interface, which means that it can utilize 16 lanes simultaneously for an aggregate bandwidth of 16 GB/sec. Despite its being a serial channel, a rich set of commands travel over PCI Express links, including device reads, writes, interrupts, and configuration setup commands.

The next bus group is the Direct Media Interface (DMI), which is used to connect the Core i7 CPU to its companion **chipset**. The DMI interface is similar to the PCI Express interface, although it runs at about half the speed since four lanes can provide only up to 2.5-GB-per-second data transfer rates. A CPU's chipset contains a rich set of additional peripheral interface support, typically required for higher-end system with many I/O devices. The Core i7 chipset is composed of the P67 and ICH10 chips. The P67 chip is the Swiss Army knife of chips, providing SATA, USB, Audio, PCIe, and Flash memory interfaces. The ICH10 chip provides legacy interface support, including a PCI interface and the 8259A interrupt control functionality. Additionally, the ICH10 contains a handful of other circuits, such as real-time clocks, event timers, and direct memory access (DMA) controllers. Having chips like these greatly simplifies construction of a full PC.

The Core i7 can be configured to use interrupts the same way as on the 8088 (for purposes of backward compatibility), or it can also use a new interrupt system using a device called an **APIC** (**Advanced Programmable Interrupt Controller**).

The Core i7 can run at any one of several predefined voltages, but it has to know which one. The power-management signals are used for automatic power-supply voltage selection, telling the CPU that power is stable, and other power-related matters. Managing the various sleep states is also done here since sleeping is done for reasons of power management.

Despite sophisticated power management, the Core i7 can get very hot. To protect the silicon, each Core i7 processor contains multiple internal heat sensors that detect when the chip is about to overheat. The thermal monitoring group deals with thermal management, allowing the CPU to indicate to its environment that it is in danger of overheating. One of the pins is asserted by the CPU if its internal temperature reaches 130°C (266°F). If a CPU ever hits this temperature, it is probably dreaming about retirement and becoming a camp stove.

Even at camp-stove temperatures, you need not worry about the safety of the Core i7. If the internal sensors detect that the processor is about to overheat, it will

initiate **thermal throttling**, which is a technique that quickly reduces heat generation by running the processor only every Nth clock cycle. The larger the value of N, the more the processor is throttled down, and the more quickly it will cool. Of course, the cost of this throttling is a decrease in system performance. Prior to the invention of thermal throttling, CPUs would burn up if their cooling system failed. Evidence of these dark times of CPU thermal management can be found by searching for "exploding CPU" on YouTube. The video is fake but the problem is not.

The Clock signal provides the system clock to the processor, which internally is used to generate variety of clocks based on a multiple or fraction of the system clock. Yes, it is possible to generate a multiple of the system clock frequency, using a very clever device called a delay-locked loop, or DLL.

The Diagnostics group contains signals for testing and debugging systems in conformance with the IEEE 1149.1 JTAG (Joint Test Action Group) test standard. Finally, the miscellaneous group is a hodge-podge of other signals that have various special purposes.

Pipelining on the Core i7's DDR3 Memory Bus

Modern CPUs like the Core i7 place heavy demands on DRAM memories. Individual processors can produce access requests much faster than a slow DRAM can produce values, and this problem is compounded when multiple processors are making simultaneous requests. To keep the CPUs from starving for lack of data, it is essential to get the maximum possible throughput from the memory. For this reason, the Core i7 DDR3 memory bus can be operated in a pipelined manner, with as many as four simultaneous memory transactions going on at the same time. We saw the concept of pipelining in Chap. 2 in the context of a pipelined CPU (see Fig. 2-4), but memories can also be pipelined.

To allow pipelining, Core i7 memory requests have three steps:

1. The memory ACTIVATE phase, which "opens" a DRAM memory row, making it ready for subsequent memory accesses.
2. The memory READ or WRITE phase, where multiple accesses can be made to individual words within the currently open DRAM row or to multiple sequential words within the current DRAM row using a burst mode.
3. The PRECHARGE phase, which "closes" the current DRAM memory row, and prepares the DRAM memory for the next ACTIVATE command.

The secret to the Core i7's pipelined memory accesses is that DDR3 DRAMs are organized with multiple **banks** within the DRAM chip. A bank is a block of DRAM memory, which may be accessed in parallel with other DRAM memory

banks, even if they are contained in the same chip. A typical DDR3 DRAM chip will have as many as 8 banks of DRAM. However, the DDR3 interface specification allows only up to four concurrent accesses on a single DDR3 channel. The timing diagram in Fig. 3-46 illustrates the Core i7 making 4 memory accesses to three distinct DRAM banks. The accesses are fully overlapped, such that the DRAM reads occur in parallel within the DRAM chip. The figure shows which commands lead to later operations through the use of arrows in the timing diagram.

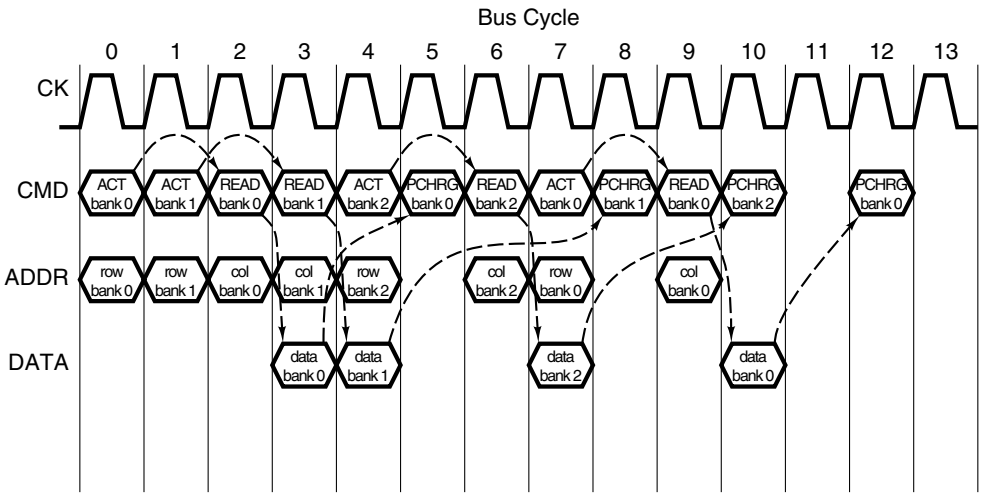


Figure 3-46. Pipelining memory requests on the Core i7's DDR3 interface.

As shown in Fig. 3-46, the DDR3 memory interface has four primary signal paths: bus clock (CK), bus command (CMD), address (ADDR), and data (DATA). The bus clock signal CK orchestrates all bus activity. The bus command CMD indicates what activity is requested of the connect DRAM. The ACTIVATE command specifies the address of the DRAM row to open via the ADDR signal. When a READ is executed, the DRAM column address is given via the ADDR signals, and the DRAM produces the read value a fixed time later on the DATA signals. Finally, the PRECHARGE command indicates the bank to precharge via the ADDR signals. For the purpose of the example, the ACTIVATE command must precede the first READ to the same bank by two DDR3 bus cycles, and data are produced one bus cycle after the READ command. Additionally, the PRECHARGE operation must occur at least two bus cycles after the last READ operation to the same DRAM bank.

The parallelism in the memory requests can be seen in the overlapping of the READ requests to the different DRAM banks. The first two READ accesses to banks 0 and 1 are completely overlapped, producing results in bus cycles 3 and 4, respectively. The access to bank 2 partially overlaps with the first access of bank 1, and finally the second read of bank 0 partially overlaps with the access to bank 2.

You might be wondering how the Core i7 knows when to expect its READ command data to return, and when it can make a new memory request. The answer is that it knows when to receive and initiate requests because it fully models the internal activities of each attached DDR3 DRAM. Thus, it will anticipate the return of data in the correct cycle, and it will know to avoid initiating a precharge operation until two cycles after its last read operation. The Core i7 can anticipate all of these activities because the DDR3 memory interface is a **synchronous memory interface**. Thus, all activities take a well-known number of DDR3 bus cycles. Even with all of this knowledge, building a high-performance fully pipelined DDR3 memory interface is a nontrivial task, requiring many internal timers and conflict detectors to implement efficient DRAM request handling.

3.5.2 The Texas Instruments OMAP4430 System-on-a-Chip

As our second example of a CPU chip, we will now examine the Texas Instruments (TI) OMAP4430 **system-on-a-chip** (**SoC**). The OMAP4430 implements the ARM instruction set, and it is targeted at mobile and embedded applications such as smartphones, tablets, and Internet appliances. Aptly named, a system-on-a-chip incorporates a wide range of devices such that the SoC combined with physical peripherals (touchscreen, flash memory, etc.) implements a complete computing device.

The OMAP4430 SoC includes two ARM A9 cores, additional accelerators, and a wide range of peripheral interfaces. The internal organization of the OMAP4430 is shown in Fig. 3-47. The ARM A9 cores are 2-wide superscalar microarchitectures. In addition, there are three more accelerator processors on the OMAP4430 die: a POWERVR SGX540 graphics processor, an image signal processor (ISP), and an IVA3 video processor. The SGX540 provides efficient programmable 3D rendering, similar to the GPUs found in desktop PCs, albeit smaller and slower. The ISP is a programmable processor designed for efficient image manipulation, for the type of operations that would be required in a high-end digital camera. The IVA3 implements efficient video encoding and decoding, with enough performance to support 3D applications like those found in handheld game consoles. Also included in the OMAP4430 SoC is a wide range of peripheral interfaces, including a touchscreen and keypad controllers, DRAM and flash interfaces, USB, and HDMI. Texas Instruments has detailed the roadmap for the OMAP series of CPUs. Future designs will have more of everything—more ARM cores, more GPUs, and more diverse peripherals.

The OMAP4430 SoC was introduced in early 2011 with two ARM A9 cores running at 1 GHz using a 45-nanometer silicon implementation. A key aspect of the OMAP4430 design is that it performs significant amounts of computation with very little power, since it is targeted to mobile applications that are powered by a battery. In battery-powered mobile applications, the more efficiently the design operates, the longer the user can go between battery charges.

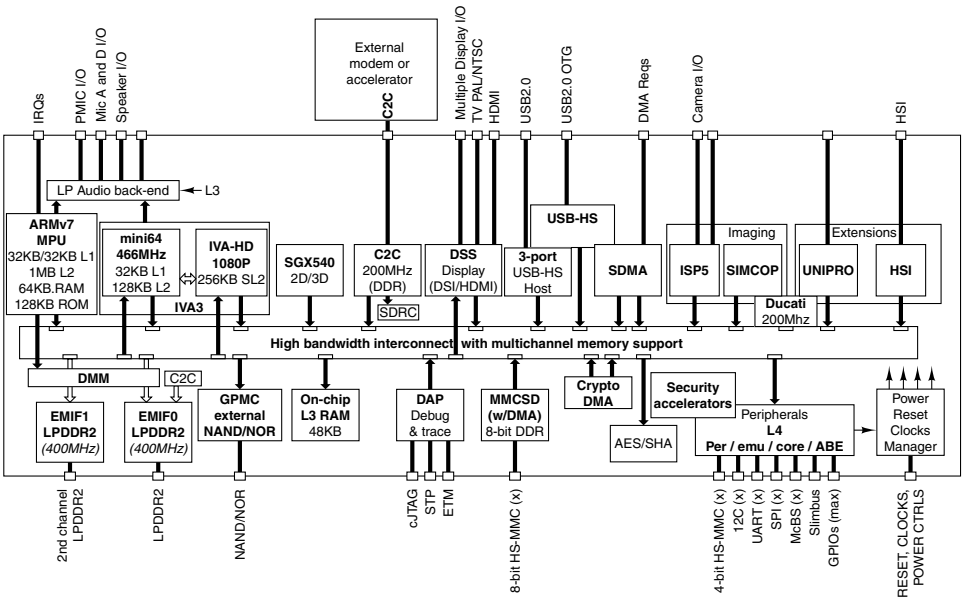


Figure 3-47. The internal organization of the OMAP4430 system-on-a-chip.

The many processors of the OMAP4430 are incorporated specifically to support its mission of low-power operation. The graphics processor, ISP, and IVA3 are all programmable accelerators that provide efficient computation capabilities at significantly less energy compared to the same tasks running on the ARM A9 CPUs alone. Fully powered, the OMAP4430 SoC draws only 600 mW of power. Compared to a high-end Core i7, the OMAP4430 uses about 1/250 as much power. The OMAP4430 also implements a very efficient sleep mode; when all components are asleep, the design draws only 100 μW. Efficient sleep modes are crucial to mobile applications with long periods of standby time, such as a cell phone. The less energy used in sleep mode, the longer the cell phone can stay in standby mode.

To further reduce power demands of the OMAP4430, the design incorporates a variety of power-management facilities, including **dynamic voltage scaling** and **power gating**. Dynamic voltage scaling allows components to run slower at a lower voltage, which significantly reduces power requirements. If you do not need the CPU's most blazing speed for computation, the voltage of the design can be lowered to run the CPU at a slower speed and much energy will be saved. Power gating is an even more aggressive power-management technique where a component is powered down completely when it is not in use, thereby eliminating its power draw. For example in a tablet application, if the user is not watching a movie, the IVA3 video processor is completely powered down and draws no power. Conversely, when the user is watching a movie, the IVA3 video processor is speeding through its video decoding tasks, while the two ARM A9 CPUs are asleep.

Despite its bent for joule-frugal operation, the ARM A9 cores utilize a very capable microarchitecture. They can decode and execute up to two instructions each cycle. As we will learn in Chap. 4, this execution rate represents the maximum throughput of the microarchitecture. But do not expect it to execute this many instructions each cycle. Rather, think of this rate as the manufacturer's guaranteed maximum performance, a level that the processor will never exceed, no matter what. In many cycles, fewer than two instructions will execute due to the myriad of "hazards" that can stall instructions, leading to lower execution throughput. To address many of these throughput limiters, the ARM A9 incorporates a powerful branch predictor, out-of-order instruction scheduling, and a highly optimized memory system.

The OMAP4430's memory system has two main internal L1 caches for each ARM A9 processor: a 32-KB cache for instructions and a 32-KB cache for data. Like the Core i7, it also uses an on-chip level 2 (L2) cache, but unlike the Core i7, it is a relatively tiny 1 MB in size, and it is shared by both ARM A9 cores. The caches are fed with dual LPDDR2 low-power DRAM channels. LPDDR2 is derived from the DDR2 memory interface standard, but changed to require fewer wires and to operate at more power-efficient voltages. Additionally, the memory controller incorporates a number of memory-access optimizations, such as tiled memory prefetching and in-memory rotation support.

While we will discuss caching in detail in Chap. 4, a few words about it here will be useful. All of main memory is divided up into cache lines (blocks) of 32 bytes. The 1024 most heavily used instruction lines and the 1024 most heavily used data lines are in the level 1 cache. Cache lines that are heavily used but which do not fit in the level 1 cache are kept in the level 2 cache. This cache contains both data lines and instruction lines from both ARM A9 CPUs mixed at random. The level 2 cache contains the most recently touched 32,768 lines in main memory.

On a level 1 cache miss, the CPU sends the identifier of the line it is looking for (Tag address) to the level 2 cache. The reply (Tag data) provides the information for the CPU to tell whether the line is in the level 2 cache, and if so, what state it is in. If the line is cached there, the CPU goes and gets it. Getting a value out of the level 2 cache takes 19 cycles. This is a long time to wait for data, so clever programmers will optimize their programs to use less data, making it more likely to find data in the fast level 1 cache.

If the cache line is not in the level 2 cache, it must be fetched from main memory via the LPDDR2 memory interface. The OMAP4430 LPDDR2 interface is implemented on-chip such that LPDDR2 DRAM can be connected directly to the OMAP4430. To access memory, the CPU must first send the upper portion of the DRAM address to the DRAM chip, using the 13 address lines. This operation, called an *ACTIVATE*, loads an entire row of memory within the DRAM into a row buffer. Subsequently, the CPU can issue multiple READ or $\overline{\text{WRITE}}$ commands, sending the remainder of the address on the same 13 address lines, and sending (or receiving) the data for the operation on the 32 data lines.

While waiting for the results, the CPU may well be able to continue with other work. For example, a cache miss while prefetching an instruction does not inhibit the execution of one or more instructions already fetched, each of which may refer to data not in any cache. Thus, multiple transactions on the two LPDDR2 interfaces may be outstanding at once, even for the same processor. It is up to the memory controller to keep track of all this and to make actual memory requests in the most efficient order.

When data finally arrives from the memory, it can come in 4 bytes at a time. A memory operation may utilize a burst mode read or write, which will allow multiple contiguous addresses within the same DRAM row to be read or written. This mode is particularly efficient for reading or writing cache blocks. Just for the record, the description of the OMAP4430 given above, like that of the Core i7 before it, has been highly simplified, but the essence of its operation has been described.

The OMAP4430 comes in a 547-pin **ball grid array** (PBGA), as shown in Fig. 3-48. A ball grid array is similar to a land grid array except that the connections on the chip are small metal balls, rather than the square pads used in the LGA. The two packages are not compatible, providing further evidence that you cannot stick a square peg into a round hole. The OMAP4430's package consists of a rectangular array of 28×26 balls, with two inner rings of balls missing, plus two asymmetric half rows and columns of balls missing to prevent the chip from being inserted incorrectly in the BGA socket.

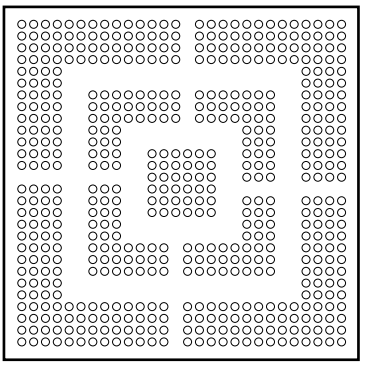


Figure 3-48. The OMAP4430 system-on-a-chip pinout.

It is difficult to compare a CISC chip (like the Core i7) and a RISC chip (like the OMAP4430) based on clock speed alone. For example, the two ARM A9 cores in the OMAP4430 have a peak execution speed of four instructions per clock cycle, giving it almost the same execution rate as that of the Core i7's 4-wide superscalar processors. The Core i7 achieves faster program execution, however, since it has up to six processors running with a clock speed 3.5 times faster (3.5 GHz) than the OMAP4430. The OMAP4430 may seem like a turtle running next

to the Core i7 rabbit, but the turtle uses much less power, and the turtle may finish first, especially if the rabbit's battery is not very big.

3.5.3 The Atmel ATmega168 Microcontroller

Both the Core i7 and the OMAP4430 are examples of high-performance computing platforms designed for building highly capable computing devices, with the Core i7 focusing on desktop applications while the OMAP4430 focuses on mobile applications. When many people think about computers, systems like these come to mind. However, another whole world of computers exists that is actually far more pervasive: embedded systems. In this section we will take a brief look at that world.

It is probably only a slight exaggeration to say that every electrical device costing more than $100 has a computer in it. Certainly televisions, cell phones, electronic personal organizers, microwave ovens, camcorders, VCRs, laser printers, burglar alarms, hearing aids, electronic games, and other devices too numerous to mention are all computer controlled these days. The computers inside these things tend to be optimized for low price rather than for high performance, which leads to different trade-offs than the high-end CPUs we have been studying so far.

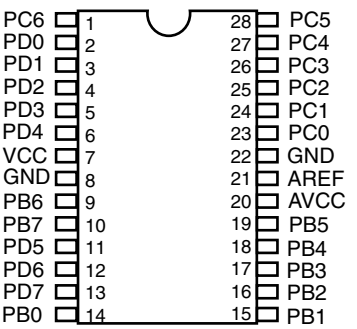


Figure 3-49. Physical pinout of the ATmega168.

As we mentioned in Chap. 1, the Atmel ATmega168 microcontroller is widely used, mostly due to its very low cost (about $1). As we will see shortly, it is also a versatile chip, which makes interfacing to it simple and inexpensive. So let us now examine this chip, whose physical pinout is shown in Fig. 3-49.

As can be seen from the figure, the ATmega168 normally comes in a standard 28-pin package (although other packages are available). At first glance, you probably noticed that the pinout on this chip is a bit strange compared to the previous two designs we examined. In particular, this chip has no address and data lines. This is because the chip is not designed to be connected to memory, only to devices. All of the memory, SRAM and flash, is contained within the processor, obviating the need for any address and data pins as shown in Fig. 3-50.

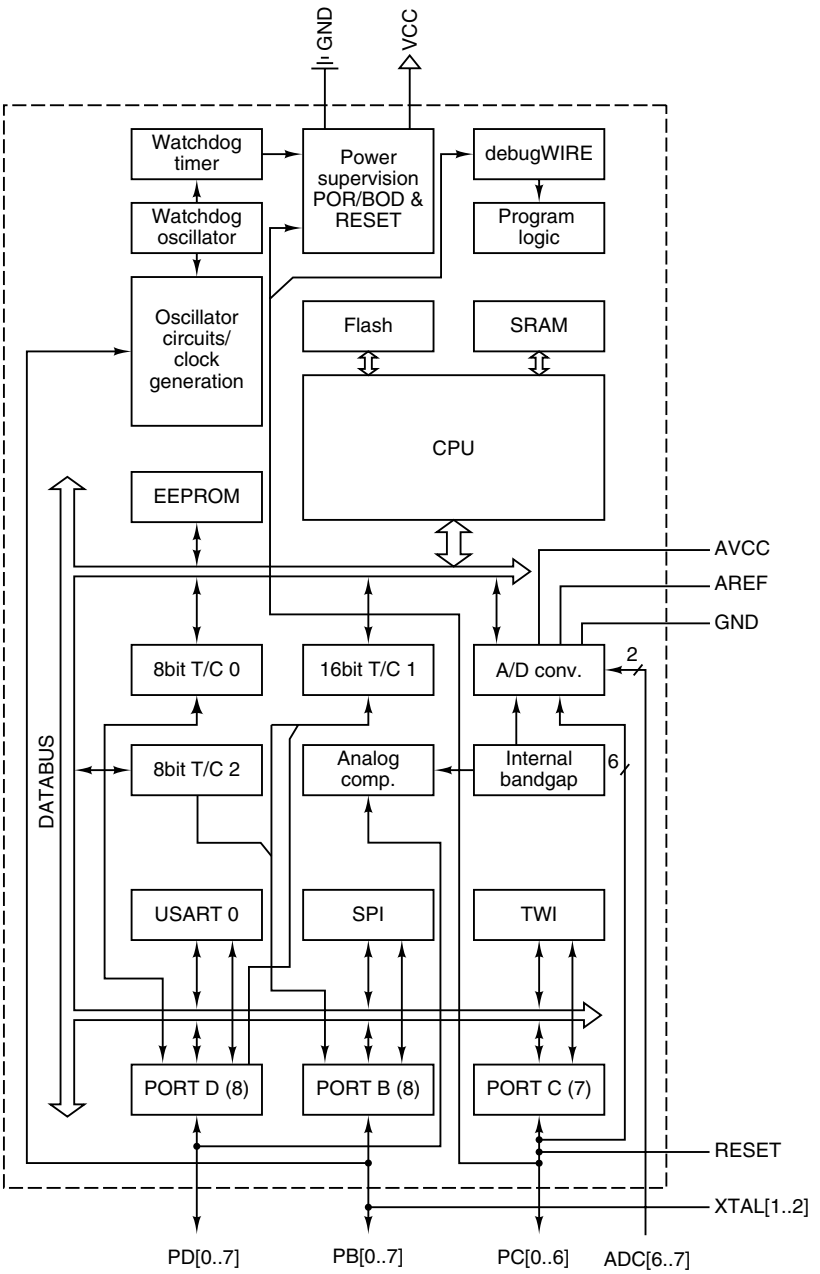


Figure 3-50. The internal architecture and logical pinout of the ATmega168.

Instead of address and data pins, the ATmega168 has 27 digital I/O ports, 8 lines in port B and D, and 7 lines in port C. These digital I/O lines are designed to be connected to I/O peripherals, and each line can be internally configured by startup software to be an input or an output. For example, when used in a microwave oven, one digital I/O line would be an input from the "door open" sensor. Another digital I/O line would be an output used to turn the microwave generator on and off. Software in the ATmega168 would check that the door was closed before turning on the microwave generator. If the door is suddenly opened, the software must kill the power. In practice, hardware interlocks are always present, too.

Optionally, six of the inputs from port C can be configured to be analog I/O. Analog I/O pins can read the voltage level of an input or set the voltage level of an output. Extending our microwave oven example, some ovens have a sensor that allows the user to heat food to a given temperature. The temperature sensor would be connected to a C port input, and software could read the voltage of the sensor and then convert it to a temperature using a sensor-specific translation function. The remaining pins on the ATmega168 are the power input (VCC), two ground pins (GND), and two pins to configure the analog I/O circuitry (AREF, AVCC).

The internal architecture of the ATmega168, like that of the OMAP4430, is a system-on-a-chip with a rich array of internal devices and memory. The ATmega168 comes with up to 16 KB of internal flash memory, for storage of infrequently changing nonvolatile information such as program instructions. It also includes up to 1 KB of EEPROM, which is nonvolatile memory that can be written by software. The EEPROM stores system-configuration data. Again, using our microwave example, the EEPROM would store a bit indicating whether the microwave displayed time in 12- or 24-hour format. The ATmega168 also incorporates up to 1 KB of internal SRAM, where software can store temporary variables.

The internal processor runs the AVR instruction set, which is composed of 131 instructions, each 16 bits in length. The processor is an 8-bit processor, which means that it operates on 8-bit data values, and internally its registers are 8 bits in size. The instruction set incorporates special instructions that allow the 8-bit processor to efficiently operate on larger data types. For example, to perform 16-bit or larger additions, the processor provides the "add-with-carry" instruction, which adds two values plus the carry out of the previous addition. The remaining internal components include the real-time clock and a variety of interface logic, including support for serial links, pulse-width-modulated (PWM) links, I2C (Inter-IC bus) link, and analog and digital I/O controllers.

3.6 EXAMPLE BUSES

Buses are the glue that hold computer systems together. In this section we will take a close look at some popular buses: the PCI bus and the Universal Serial Bus. The PCI bus is the primary I/O peripheral bus used today in PCs. It comes in two

forms, the older PCI bus, and the new and much faster PCI Express (PCIe) bus. The Universal Serial Bus is an increasingly popular I/O bus for low-speed peripherals such as mice and keyboards. A second and third version of the USB bus run at much higher speeds. In the following sections, we will look at each of these buses in turn to see how they work.

3.6.1 The PCI Bus

On the original IBM PC, most applications were text based. Gradually, with the introduction of Windows, graphical user interfaces came into use. None of these applications put much strain on early system buses such as the ISA bus. However, as time went on and many applications, especially multimedia games, began to use computers to display full-screen, full-motion video, the situation changed radically.

Let us make a simple calculation. Consider a 1024×768 color video with 3 bytes/pixel. One frame contains 2.25 MB of data. For smooth motion, at least 30 screens/sec are needed, for a data rate of 67.5 MB/sec. In fact, it is worse than this, since to display a video from a hard disk, CD-ROM, or DVD, the data must pass from the disk drive over the bus to the memory. Then for the display, the data must travel over the bus again to the graphics adapter. Thus, we need a bus bandwidth of 135 MB/sec for the video alone, not counting the bandwidth that the CPU and other devices need.

The PCI bus' predecessor, the ISA bus, ran at a maximum rate of 8.33 MHz and could transfer 2 bytes per cycle, for a maximum bandwidth of 16.7 MB/sec. The enhanced ISA bus, called the EISA bus, could move 4 bytes per cycle, to achieve 33.3 MB/sec. Clearly, neither of these approached what is needed for full-screen video.

With modern full HD video the situation is even worse. It requires 1920×1080 frames at 30 frames/sec for a data rate of 155 MB/sec (or 310 MB/sec if the data have to traverse the bus twice). Clearly, the EISA bus does not even come close to handling this.

In 1990, Intel saw this coming and designed a new bus with a far higher bandwidth than the EISA bus. It was called the **PCI bus** (**Peripheral Component Interconnect bus**). To encourage its use, Intel patented the PCI bus and then put all the patents into the public domain, so any company could build peripherals for it without having to pay royalties. Intel also formed an industry consortium, the PCI Special Interest Group, to manage the future of the PCI bus. As a result, the PCI bus became extremely popular. Virtually every Intel-based computer since the Pentium has a PCI bus, and many other computers do, too. The PCI bus is covered in gory detail in Shanley and Anderson (1999) and Solari and Willse (2004).

The original PCI bus transferred 32 bits per cycle and ran at 33 MHz (30-nsec cycle time) for a total bandwidth of 133 MB/sec. In 1993, PCI 2.0 was introduced,

and in 1995, PCI 2.1 came out. PCI 2.2 has features for mobile computers (mostly for saving battery power). The PCI bus runs at up to 66 MHz and can handle 64-bit transfers, for a total bandwidth of 528 MB/sec. With this kind of capacity, full-screen, full-motion video is doable (assuming the disk and the rest of the system are up to the job). In any event, the PCI bus will not be the bottleneck.

Even though 528 MB/sec sounds pretty fast, the PCI bus still had two problems. First, it was not good enough for a memory bus. Second, it was not compatible with all those old ISA cards out there. The solution Intel thought of was to design computers with three or more buses, as shown in Fig. 3-51. Here we see that the CPU can talk to the main memory on a special memory bus, and that an ISA bus can be connected to the PCI bus. This arrangement met all requirements, and as a consequence it was widely used in the 1990s.

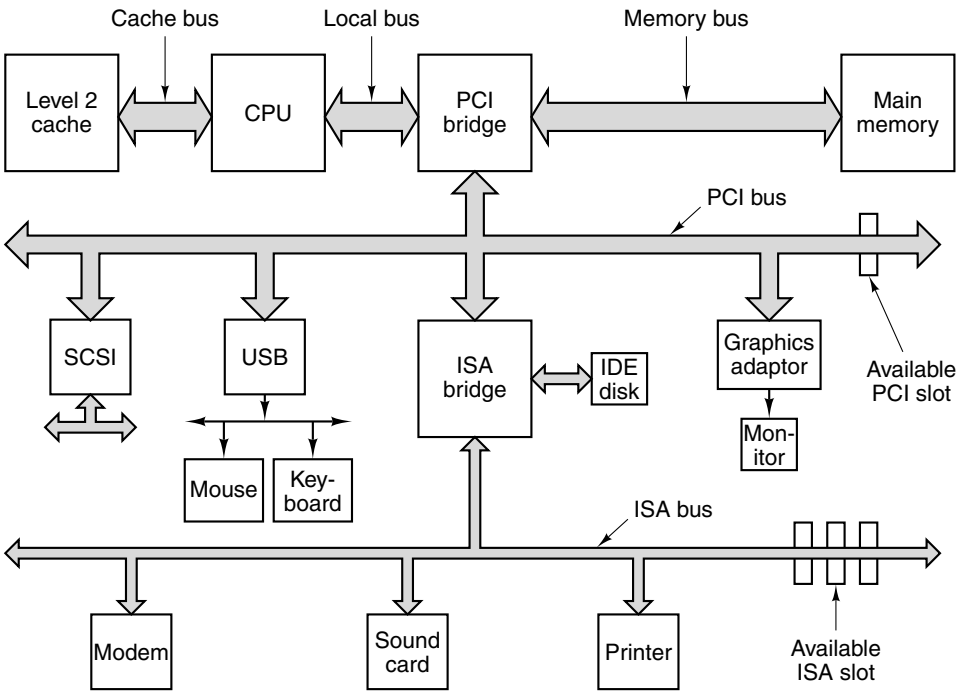


Figure 3-51. Architecture of an early Pentium system. The thicker buses have more bandwidth than the thinner ones but the figure is not to scale.

Two key components in this architecture are the two bridge chips (which Intel manufactures—hence its interest in this whole project). The PCI bridge connects the CPU, memory, and PCI bus. The ISA bridge connects the PCI bus to the ISA bus and also supports one or two IDE disks. Nearly all PC systems using this architecture would have one or more free PCI slots for adding new high-speed peripherals, and one or more ISA slots for adding low-speed peripherals.

The big advantage of the design of Fig. 3-51 is that the CPU has an extremely high bandwidth to memory using a proprietary memory bus; the PCI bus offers high bandwidth for fast peripherals such as SCSI disks, graphics adaptors, etc.; and old ISA cards can still be used. The USB box in the figure refers to the Universal Serial Bus, which will be discussed later in this chapter.

It would have been nice had there been only one kind of PCI card. Unfortunately, such is not the case. Options are provided for voltage, width, and timing. Older computers often use 5 volts and newer ones tend to use 3.3 volts, so the PCI bus supports both. The connectors are the same except for two bits of plastic that are there to prevent people from inserting a 5-volt card in a 3.3-volt PCI bus or vice versa. Fortunately, universal cards exist that support both voltages and can plug into either kind of slot. In addition to the voltage option, cards come in 32-bit and 64-bit versions. The 32-bit cards have 120 pins; the 64-bit cards have the same 120 pins plus an additional 64. A PCI bus system that supports 64-bit cards can also take 32-bit cards, but the reverse is not true. Finally, PCI buses and cards can run at either 33 or 66 MHz. The choice is made by having one pin wired either to the power supply or to ground. The connectors are identical for both speeds.

By the late 1990s, pretty much everyone agreed that the ISA bus was dead, so new designs excluded it. By then, however, monitor resolution had increased in some cases to 1600×1200 and the demand for full-screen full motion video had also increased, especially in the context of highly interactive games, so Intel added yet another bus just to drive the graphics card. This bus was called the **AGP bus** (**Accelerated Graphics Port bus**). The initial version, AGP 1.0, ran at 264 MB/sec, which was defined as 1x. While slower than the PCI bus, it was dedicated to driving the graphics card. Over the years, newer versions came out, with AGP 3.0 running at 2.1 GB/sec (8x). Today, even the high-performance AGP 3.0 bus has been usurped by even faster upstarts, in particular, the PCI Express bus, which can pump an amazing 16 GB/sec of data over high-speed serial bus links. A modern Core i7 system is illustrated in Fig. 3-52.

In a modern Core i7 based system, a number of interfaces have been integrated directly into the CPU chip. The two DDR3 memory channels, running at 1333 transactions/sec, connect to main memory and provide an aggregate bandwidth of 10 GB/sec per channel. Also integrated into the CPU is a 16-lane PCI Express channel, which optimally can be configured into a single 16-bit PCI Express bus or dual independent 8-bit PCI Express buses. The 16 lanes together provide a bandwidth of 16 GB/sec to I/O devices.

The CPU connects to the primary bridge chip, the P67, via the 20-Gb/sec (2.5 GB/sec) serial direct media interface (DMI). The P67 provides interfaces to a number of modern high-performance I/O interfaces. Eight additional PCI Express lanes are provided, plus SATA disk interfaces. The P67 also implements 14 USB 2.0 interfaces, 10G Ethernet and an audio interface.

The ICH10 chip provides legacy interface support for old devices. It is connected to the P67 via a slower DMI interface. The ICH10 implements the PCI bus,

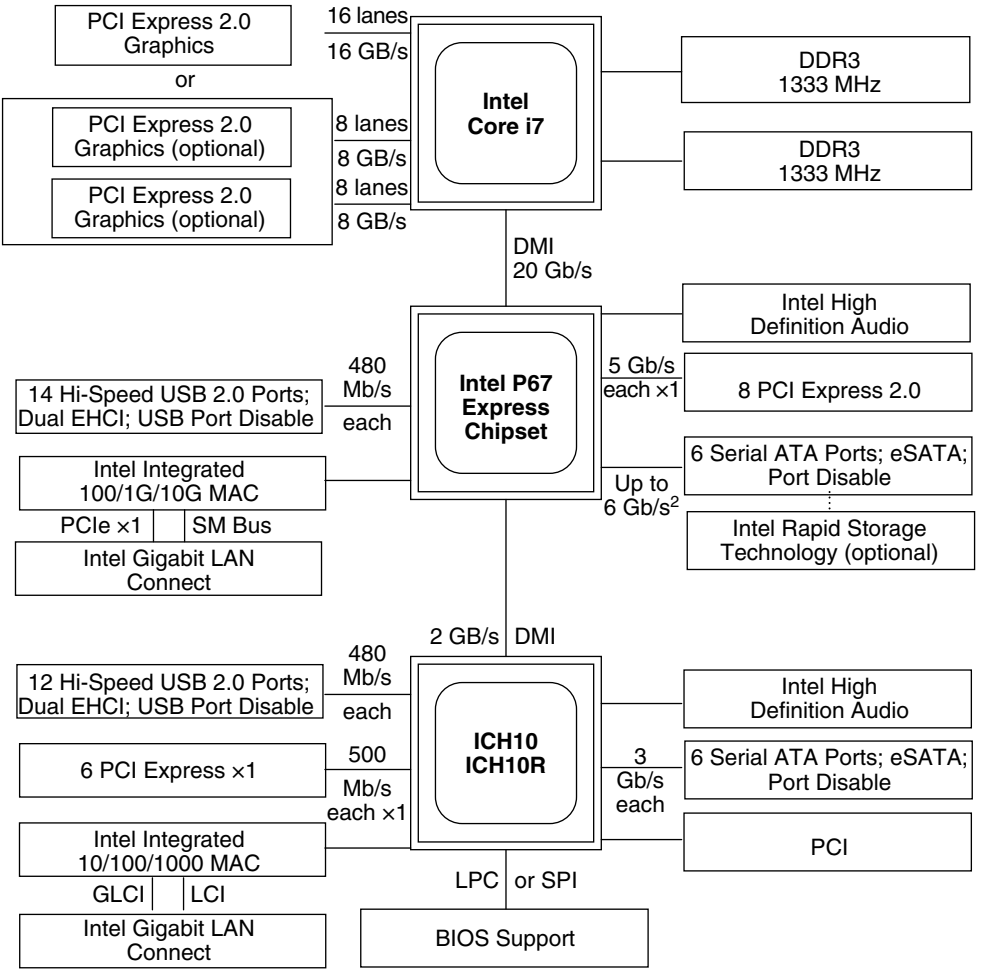


Figure 3-52. The bus structure of a modern Core i7 system.

1G Ethernet, USB ports, and old-style PCI Express and SATA interfaces. Newer systems may not incorporate the ICH10; it is required only if the system needs to support legacy interfaces.

PCI Bus Operation

Like all PC buses going back to the original IBM PC, the PCI bus is synchronous. All transactions on the PCI bus are between a master, officially called the **initiator**, and a slave, officially called the **target**. To keep the PCI pin count

down, the address and data lines are multiplexed. In this way, only 64 pins are needed on PCI cards for address plus data signals, even though PCI supports 64-bit addresses and 64-bit data.

The multiplexed address and data pins work as follows. On a read operation, during cycle 1, the master puts the address onto the bus. On cycle 2, the master removes the address and the bus is turned around so the slave can use it. On cycle 3, the slave outputs the data requested. On write operations, the bus does not have to be turned around because the master puts on both the address and the data. Nevertheless, the minimum transaction is still three cycles. If the slave is not able to respond in three cycles, it can insert wait states. Block transfers of unlimited size are also allowed, as well as several other kinds of bus cycles.

PCI Bus Arbitration

To use the PCI bus, a device must first acquire it. PCI bus arbitration uses a centralized bus arbiter, as shown in Fig. 3-53. In most designs, the bus arbiter is built into one of the bridge chips. Every PCI device has two dedicated lines running from it to the arbiter. One line, REQ#, is used to request the bus. The other line, GNT#, is used to receive bus grants. Note: REQ# is PCI-speak for $\overline{\text{REQ}}$.

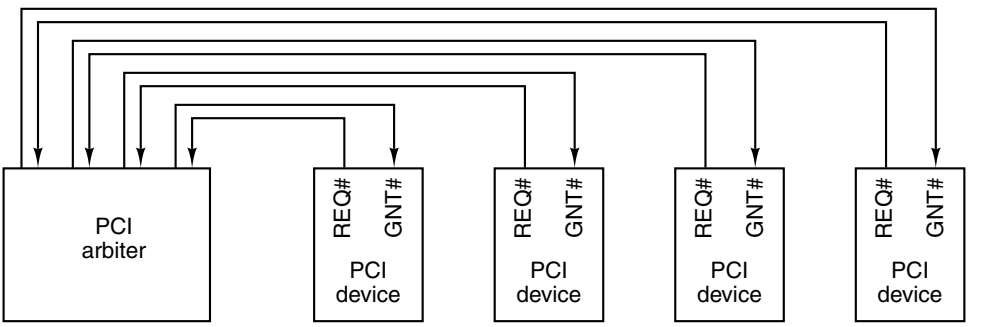


Figure 3-53. The PCI bus uses a centralized bus arbiter.

To request the bus, a PCI device (including the CPU) asserts REQ# and waits until it sees its GNT# line asserted by the arbiter. When that event happens, the device can use the bus on the next cycle. The algorithm used by the arbiter is not defined by the PCI specification. Round-robin arbitration, priority arbitration, and other schemes are all allowed. Clearly, a good arbiter will be fair, so as not to let some devices wait forever.

A bus grant is for only one transaction, although the length of this transaction is theoretically unbounded. If a device wants to run a second transaction and no other device is requesting the bus, it can go again, although often one idle cycle between transactions has to be inserted. However, under special circumstances, in

the absence of competition for the bus, a device can make back-to-back transactions without having to insert an idle cycle. If a bus master is making a very long transfer and some other device has requested the bus, the arbiter can negate the GNT# line. The current bus master is expected to monitor the GNT# line, so when it sees the negation, it must release the bus on the next cycle. This scheme allows very long transfers (which are efficient) when there is only one candidate bus master but still gives fast response to competing devices.

PCI Bus Signals

The PCI bus has a number of mandatory signals, shown in Fig. 3-54(a), and a number of optional signals, shown in Fig. 3-54(b). The remainder of the 120 or 184 pins are used for power, ground, and related miscellaneous functions and are not listed here. The *Master* (initiator) and *Slave* (target) columns tell who asserts the signal on a normal transaction. If the signal is asserted by a different device (e.g., CLK), both columns are left blank.

Let us now look briefly at each of the PCI bus signals. We will start with the mandatory (32-bit) signals, then move on to the optional (64-bit) signals. The CLK signal drives the bus. Most of the other signals are synchronous with it. A PCI bus transaction begins at the falling edge of CLK, which is in the middle of the cycle.

The 32 AD signals are for the address and data (for 32-bit transactions). Generally, during cycle 1 the address is asserted and during cycle 3 the data are asserted. The PAR signal is a parity bit for AD. The C/BE# signal is used for two different things. On cycle 1, it contains the bus command (read 1 word, block read, etc.). On cycle 2 it contains a bit map of 4 bits, telling which bytes of the 32-bit word are valid. Using C/BE# it is possible to read or write any 1, 2, or 3 bytes, as well as an entire word.

The FRAME# signal is asserted by the master to start a bus transaction. It tells the slave that the address and bus commands are now valid. On a read, usually IRDY# is asserted at the same time as FRAME#. It says the master is ready to accept incoming data. On a write, IRDY# is asserted later, when the data are on the bus.

The IDSEL signal relates to the fact that every PCI device must have a 256-byte configuration space that other devices can read (by asserting IDSEL). This configuration space contains properties of the device. The plug-and-play feature of some operating systems uses the configuration space to find out what devices are on the bus.

Now we come to signals asserted by the slave. The first of these, DEVSEL#, announces that the slave has detected its address on the AD lines and is prepared to engage in the transaction. If DEVSEL# is not asserted within a certain time limit, the master times out and assumes the device addressed is either absent or broken.

The second slave signal is TRDY#, which the slave asserts on reads to announce that the data are on the AD lines and on writes to announce that it is prepared to accept data.

Signal	Lines	Master	Slave	Description
CLK	1			Clock (33 or 66 MHz)
AD	32	×	×	Multiplexed address and data lines
PAR	1	×		Address or data parity bit
C/BE	4	×		Bus command/bit map for bytes enabled
FRAME#	1	×		Indicates that AD and C/BE are asserted
IRDY#	1	×		Read: master will accept; write: data present
IDSEL	1	×		Select configuration space instead of memory
DEVSEL#	1		×	Slave has decoded its address and is listening
TRDY#	1		×	Read: data present; write: slave will accept
STOP#	1		×	Slave wants to stop transaction immediately
PERR#	1			Data parity error detected by receiver
SERR#	1			Address parity error or system error detected
REQ#	1			Bus arbitration: request for bus ownership
GNT#	1			Bus arbitration: grant of bus ownership
RST#	1			Reset the system and all devices

(a)

Signal	Lines	Master	Slave	Description
REQ64#	1	×		Request to run a 64-bit transaction
ACK64#	1		×	Permission is granted for a 64-bit transaction
AD	32	×		Additional 32 bits of address or data
PAR64	1	×		Parity for the extra 32 address/data bits
C/BE#	4	×		Additional 4 bits for byte enables
LOCK	1	×		Lock the bus to allow multiple transactions
SBO#	1			Hit on a remote cache (for a multiprocessor)
SDONE	1			Snooping done (for a multiprocessor)
INTx	4			Request an interrupt
JTAG	5			IEEE 1149.1 JTAG test signals
M66EN	1			Wired to power or ground (66 MHz or 33 MHz)

(b)

Figure 3-54. (a) Mandatory PCI bus signals. (b) Optional PCI bus signals.

The next three signals are for error reporting. The first of these is STOP#, which the slave asserts if something disastrous happens and it wants to abort the current transaction. The next one, PERR#, is used to report a data parity error on the previous cycle. For a read, it is asserted by the master; for a write it is asserted by the slave. It is up to the receiver to take the appropriate action. Finally, SERR# is for reporting address errors and system errors.

The REQ# and GNT# signals are for doing bus arbitration. These are not asserted by the current bus master, but rather by a device that wants to become bus master. The last mandatory signal is RST#, used for resetting the system, either due to the user pushing the RESET button or some system device noticing a fatal error. Asserting this signal resets all devices and reboots the computer.

Now we come to the optional signals, most of which relate to the expansion from 32 bits to 64 bits. The REQ64# and ACK64# signals allow the master to ask permission to conduct a 64-bit transaction and allow the slave to accept, respectively. The AD, PAR64, and C/BE# signals are just extensions of the corresponding 32-bit signals.

The next three signals are not related to 32 bits vs. 64 bits, but to multiprocessor systems, something that PCI boards are not required to support. The LOCK signal allows the bus to be locked for multiple transactions. The next two relate to bus snooping to maintain cache coherence.

The INTx signals are for requesting interrupts. A PCI card can have up to four separate logical devices on it, and each one can have its own interrupt request line. The JTAG signals are for the IEEE 1149.1 JTAG testing procedure. Finally, the M66EN signal is either wired high or wired low, to set the clock speed. It must not change during system operation.

PCI Bus Transactions

The PCI bus is really very simple (as buses go). To get a better feel for it, consider the timing diagram of Fig. 3-55. Here we see a read transaction, followed by an idle cycle, followed by a write transaction by the same bus master.

When the falling edge of the clock happens during T_1, the master puts the memory address on AD and the bus command on C/BE#. It then asserts FRAME# to start the bus transaction.

During T_2, the master floats the address bus to let it turn around in preparation for the slave to drive it during T_3. The master also changes C/BE# to indicate which bytes in the word addressed it wants to enable (i.e., read in).

In T_3, the slave asserts DEVSEL# so the master knows it got the address and is planning to respond. It also puts the data on the AD lines and asserts TRDY# to tell the master that it has done so. If the slave were not able to respond so quickly, it would still assert DEVSEL# to announce its presence but keep TRDY# negated until it could get the data out there. This procedure would introduce one or more wait states.

In this example (and often in reality), the next cycle is idle. Starting in T_5 we see the same master initiating a write. It starts out by putting the address and command onto the bus, as usual. Only now, in the second cycle it asserts the data. Since the same device is driving the AD lines, there is no need for a turnaround cycle. In T_7, the memory accepts the data.

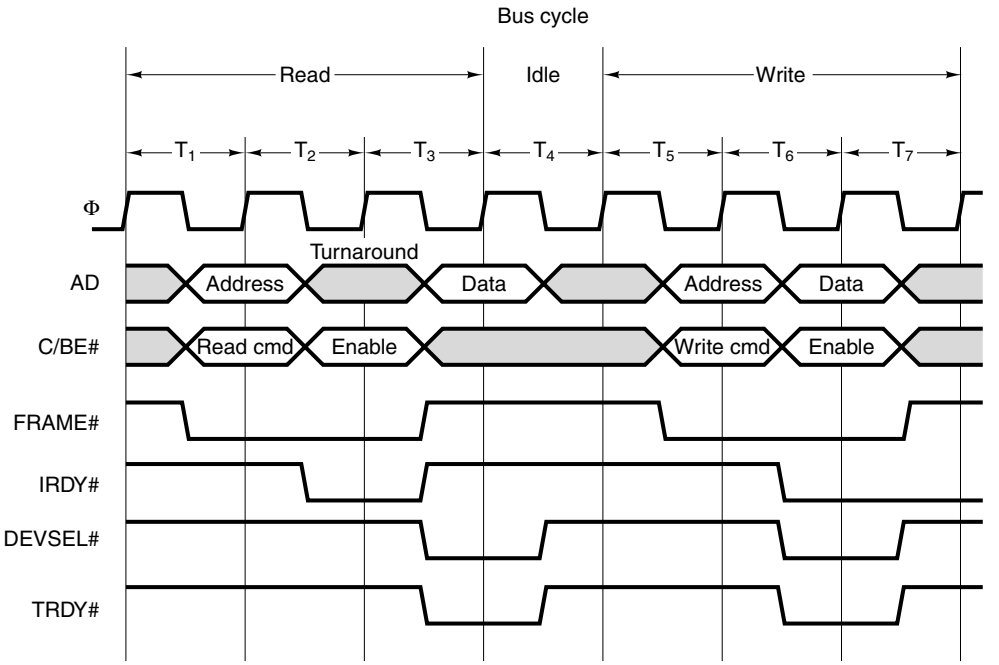


Figure 3-55. Examples of 32-bit PCI bus transactions. The first three cycles are used for a read operation, then an idle cycle, and then three cycles for a write operation.

3.6.2 PCI Express

Although the PCI bus works adequately for most current applications, the need for greater I/O bandwidth is making a mess of the once-clean internal PC architecture. In Fig. 3-52, it is clear that the PCI bus is no longer the central element that holds the parts of the PC together. The bridge chip has taken over part of that role.

The essence of the problem is that increasingly many I/O devices are too fast for the PCI bus. Cranking up the clock frequency on the bus is not a good solution because then problems with bus skew, crosstalk between the wires, and capacitance effects just get worse. Every time an I/O device gets too fast for the PCI bus (like the graphics card, hard disk, network, etc.), Intel adds a new special port to the bridge chip to allow that device to bypass the PCI bus. Clearly, this is not a long-term solution either.

Another problem with the PCI bus is that the cards are quite large. Standard PCI cards are generally 17.5 cm by 10.7 cm and low-profile cards are 12.0 cm by 3.6 cm. Neither of these fit well in laptop computers and and certainly not in mobile devices. Manufacturers would like to produce even smaller devices. Also,

some companies like to repartition the PC, with the CPU and memory in a tiny sealed box and the hard disk inside the monitor. With PCI cards, doing this is impossible.

Several solutions have been proposed, but the one that won a place in all modern PCs today (in no small part because Intel was behind it) is called **PCI Express**. It has little to do with the PCI bus and in fact is not a bus at all, but the marketing folks did not like letting go of the well-known PCI name. PCs containing it are now the standard. Let us now see how it works.

The PCI Express Architecture

The heart of the PCI Express solution (often abbreviated PCIe) is to get rid of the parallel bus with its many masters and slaves and go to a design based on high-speed point-to-point serial connections. This solution represents a radical break with the ISA/EISA/PCI bus tradition, borrowing many ideas from the world of local area networking, especially switched Ethernet. The basic idea comes down to this: deep inside, a PC is a collection of CPU, memory, and I/O controller chips that need to be interconnected. What PCI Express does is provide a general-purpose switch for connecting chips using serial links. A typical configuration is illustrated in Fig. 3-56.

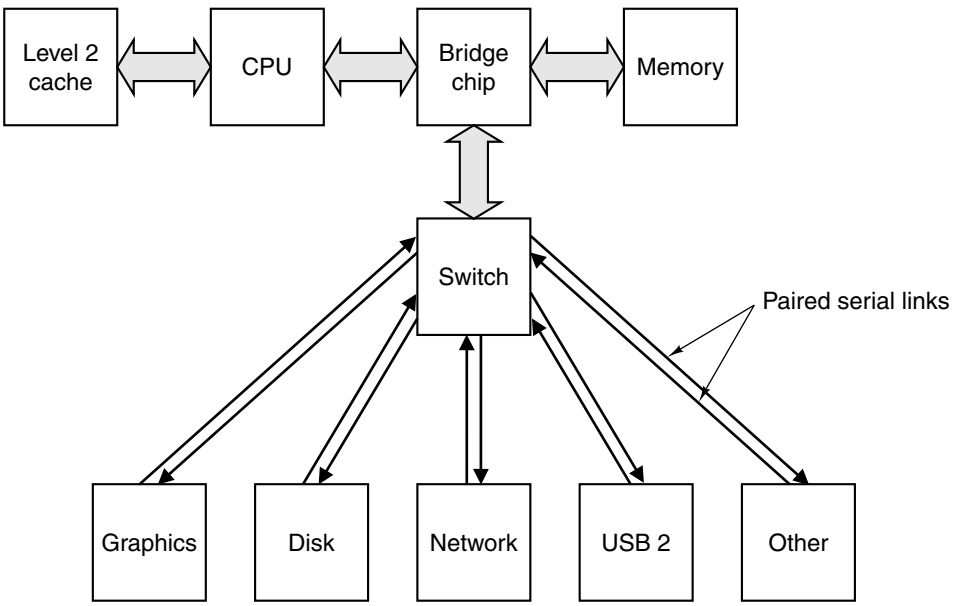


Figure 3-56. A typical PCI Express system.

As illustrated in Fig. 3-56, the CPU, memory, and cache are connected to the bridge chip in the traditional way. What is new here is a switch connected to the

bridge (possibly part of the bridge chip itself or integrated directly into the processor). Each I/O chip has a dedicated point-to-point connection to the switch. Each connection consists of a pair of unidirectional channels, one to the switch and one from it. Each channel is made up of two wires, one for the signal and one for ground, to provide high noise immunity during high-speed transmission. This architecture has replaced the old one with a much more uniform model, in which all devices are treated equally.

The PCI Express architecture differs from the old PCI bus architecture in three key ways. We have already seen two of them: a centralized switch vs. a multidrop bus and a the use of narrow serial point-to-point connections vs. a wide parallel bus. The third difference is more subtle. The conceptual model behind the PCI bus is that of a bus master issuing a command to a slave to read a word or a block of words. The PCI Express model is that of a device sending a data packet to another device. The concept of a **packet**, which consists of a header and a payload, is taken from the networking world. The **header** contains control information, thus eliminating the need for the many control signals present on the PCI bus. The **payload** contains the data to be transferred. In effect, a PC with PCI Express is a miniature packet-switching network.

In addition to these three major breaks with the past, there are also several minor differences as well. Fourth, an error-detecting code is used on the packets, providing a higher degree of reliability than on the PCI bus. Fifth, the connection between a chip and the switch is longer than it was, up to 50 cm, to allow system partitioning. Sixth, the system is expandable because a device may actually be another switch, allowing a tree of switches. Seventh, devices are hot pluggable, meaning that they can be added or removed from the system while it is running. Finally, since the serial connectors are much smaller than the old PCI connectors, devices and computers can be made much smaller. All in all, PCI Express is a major departure from the PCI bus.

The PCI Express Protocol Stack

In keeping with the model of a packet-switching network, the PCI Express system has a layered protocol stack. A **protocol** is a set of rules governing the conversation between two parties. A protocol stack is a hierarchy of protocols that deal with different issues at different layers. For example, consider a business letter. It has certain conventions about the placement and content of the letterhead, the recipient's address, the date, the salutation, the body, the signature, and so on. This might be thought of as the letter protocol. In addition, there is another set of conventions about the envelope, such as its size, where the sender's address goes and its format, where the receiver's address goes and its format, where the stamp goes, and so on. These two layers and their protocols are independent. For example, it is possible to reformat the letter but use the same envelope or vice versa. Layered

protocols make for a modular and flexible design, and have been widely used in the world of network software for decades. What is new here is building them into the "bus" hardware.

The PCI Express protocol stack is shown in Fig. 3-57(a). It is discussed below.

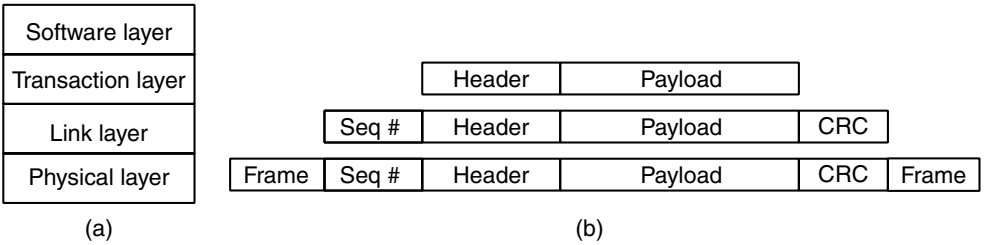


Figure 3-57. (a) The PCI Express protocol stack. (b) The format of a packet.

Let us examine the layers from the bottom up. The lowest is the **physical layer**. It deals with moving bits from a sender to a receiver over a point-to-point connection. Each point-to-point connection consists of one or more pairs of simplex (i.e., unidirectional) links. In the simplest case, there is one pair in each direction, but having 2, 4, 8, 16, or 32 pairs is also allowed. Each link is called a **lane**. The number of lanes in each direction must be the same. First-generation products must support a data rate each way of at least 2.5 Gbps, but the speed is expected to migrate to 10 Gbps each way fairly soon.

Unlike the ISA/EISA/PCI buses, PCI Express does not have a master clock. Devices are free to start transmitting as soon as they have data to send. This freedom makes the system faster but also leads to a problem. Suppose that a 1 bit is encoded as +3 volts and a 0 bit as 0 volts. If the first few bytes are all 0s, how does the receiver know data are being transmitted? After all, a run of 0 bits looks the same as an idle link. The problem is solved using what is called **8b/10b encoding**. In this scheme, 10 bits are used to encode 1 byte of actual data in a 10-bit symbol. Of the 1024 possible 10-bit symbols, the legal ones have been chosen to have enough clock transitions to keep the sender and receiver synchronized on the bit boundaries even without a master clock. A consequence of 8b/10b encoding is that a link with a gross capacity of 2.5 Gbps can carry only 2 Gbps of (net) user data.

Whereas the physical layer deals with bit transmission, the **link layer** deals with packet transmission. It takes the header and payload given to it by the transaction layer and adds to them a sequence number and an error-correcting code called a **CRC** (**Cyclic Redundancy Check**). The CRC is generated by running a certain algorithm on the header and payload data. When a packet is received, the receiver performs the same computation on the header and data and compares the result with the CRC attached to the packet. If they agree, it sends back a short **acknowledgment packet** affirming its correct arrival. If they disagree, the receiver asks for a retransmission. In this manner, data integrity is greatly improved

over the PCI bus system, which does not have any provision for verification and retransmission of data sent over the bus.

To prevent having a fast sender bury a slow receiver in packets it cannot handle, a **flow control** mechanism is used. The mechanism is that the receiver gives the transmitter a certain number of credits, basically corresponding to the amount of buffer space it has available to store incoming packets. When the credits are used up, the transmitter has to stop sending until it is given more credits. This scheme, which is widely used in all networks, prevents losing data due to a mismatch of transmitter and receiver speeds.

The **transaction layer** handles bus actions. Reading a word from memory requires two transactions: one initiated by the CPU or DMA channel requesting some data and one initiated by the target supplying the data. But the transaction layer does more than handle pure reads and writes. It adds value to the raw packet transmission offered by the link layer. To start with, it can divide each lane into up to eight **virtual circuits**, each handling a different class of traffic. The transaction layer can tag packets according to their traffic class, which may include attributes such as high priority, low priority, do not snoop, may be delivered out of order, and more. The switch may use these tags when deciding which packet to handle next.

Each transaction uses one of four address spaces:

1. Memory space (for ordinary reads and writes).
2. I/O space (for addressing device registers).
3. Configuration space (for system initialization, etc.).
4. Message space (for signaling, interrupts, etc.).

The memory and I/O spaces are similar to what current systems have. The configuration space can be used to implement features such as plug-and-play. The message space takes over the role of many of the existing control signals. Something like this space is needed because none of the PCI bus' control lines exist in PCI Express.

The **software layer** interfaces the PCI Express system to the operating system. It can emulate the PCI bus, making it possible to run existing operating systems unmodified on PCI Express systems. Of course, operating like this will not exploit the full power of PCI Express, but backward compatibility is a necessary evil that is needed until operating systems have been modified to fully utilize PCI Express. Experience shows this can take a while.

The flow of information is illustrated in Fig. 3-57(b). When a command is given to the software layer, it hands it to the transaction layer, which formulates it in terms of a header and a payload. These two parts are then passed to the link layer, which attaches a sequence number to the front and a CRC to the back. This enlarged packet is then passed on to the physical layer, which adds framing information on each end to form the physical packet that is actually transmitted. At the

receiving end, the reverse process takes place, with the link header and trailer being stripped and the result being given to the transaction layer.

The concept of each layer adding additional information to the data as it works its way down the protocol has been used for decades in the networking world with great success. The big difference between a network and PCI Express is that in the networking world the code in the various layers is nearly always software that is part of the operating system. With PCI Express it is all part of the device hardware.

PCI Express is a complicated subject. For more information see Mayhew and Krishnan (2003) and Solari and Congdon (2005). It is also still evolving. In 2007, PCIe 2.0 was released. It supports 500 MB/s per lane up to 32 lines, for a total bandwidth of 16 GB/sec. Then came PCIe 3.0 in 2011, which changed the encoding from 8b/10b to 128b/130b and can run at 8 billion transactions/sec, double PCIe 2.0.

3.6.3 The Universal Serial Bus

The PCI bus and PCI Express are fine for attaching high-speed peripherals to a computer, but they are too expensive for low-speed I/O devices such as keyboards and mice. Historically, each standard I/O device was connected to the computer in a special way, with some free ISA and PCI slots for adding new devices. Unfortunately, this scheme has been fraught with problems from the beginning.

For example, each new I/O device often comes with its own ISA or PCI card. The user is often responsible for setting switches and jumpers on the card and making sure the settings do not conflict with other cards. Then the user must open up the case, carefully insert the card, close the case, and reboot the computer. For many users, this process is difficult and error prone. In addition, the number of ISA and PCI slots is very limited (two or three typically). Plug-and-play cards eliminate the jumper settings, but the user still has to open the computer to insert the card and bus slots are still limited.

To deal with this problem, in 1993, representatives from seven companies (Compaq, DEC, IBM, Intel, Microsoft, NEC, and Northern Telecom) got together to design a better way to attach low-speed I/O devices to a computer. Since then, hundreds of other companies have joined them. The resulting standard, officially released in 1998, is called **USB** (**Universal Serial Bus**) and it is now widely implemented in personal computers. It is described further in Anderson (1997) and Tan (1997).

Some of the goals of the companies that originally conceived of the USB and started the project were as follows:

1. Users must not have to set switches or jumpers on boards or devices.
2. Users must not have to open the case to install new I/O devices.
3. There should be only one kind of cable to connect all devices.

4. I/O devices should get their power from the cable.
5. Up to 127 devices should be attachable to a single computer.
6. The system should support real-time devices (e.g., sound, telephone).
7. Devices should be installable while the computer is running.
8. No reboot should be needed after installing a new device.
9. The new bus and its I/O devices should be inexpensive to manufacture.

USB meets all these goals. It is designed for low-speed devices such as keyboards, mice, still cameras, snapshot scanners, digital telephones, and so on. Version 1.0 has a bandwidth of 1.5 Mbps, which is enough for keyboards and mice. Version 1.1 runs at 12 Mbps, which is enough for printers, digital cameras, and many other devices. Version 2.0 supports devices with up to 480 Mbps, which is sufficient to support external disk drives, high-definition webcams, and network interfaces. The recently defined USB version 3.0 pushes speeds up to 5 Gbps; only time will tell what new and bandwidth-hungry applications will spring forth from this ultra-high-bandwidth interface.

A USB system consists of a **root hub** that plugs into the main bus (see Fig. 3-51). This hub has sockets for cables that can connect to I/O devices or to expansion hubs, to provide more sockets, so the topology of a USB system is a tree with its root at the root hub, inside the computer. The cables have different connectors on the hub end and on the device end, to prevent people from accidentally connecting two hub sockets together.

The cable consists of four wires: two for data, one for power (+5 volts), and one for ground. The signaling system transmits a 0 as a voltage transition and a 1 as the absence of a voltage transition, so long runs of 0s generate a regular pulse stream.

When a new I/O device is plugged in, the root hub detects this event and interrupts the operating system. The operating system then queries the device to find out what it is and how much USB bandwidth it needs. If the operating system decides that there is enough bandwidth for the device, it assigns the new device a unique address (1–127) and downloads this address and other information to configuration registers inside the device. In this way, new devices can be added on-the-fly, without requiring any user configuration or the installation of any new ISA or PCI cards. Uninitialized cards start out with address 0, so they can be addressed. To make the cabling simpler, many USB devices contain built-in hubs to accept additional USB devices. For example, a monitor might have two hub sockets to accommodate the left and right speakers.

Logically, the USB system can be viewed as a set of bit pipes from the root hub to the I/O devices. Each device can split its bit pipe up into a maximum of 16 subpipes for different types of data (e.g., audio and video). Within each pipe or

subpipe, data flows from the root hub to the device or the other way. There is no traffic between two I/O devices.

Precisely every 1.00 ± 0.05 msec, the root hub broadcasts a new frame to keep all the devices synchronized in time. A frame is associated with a bit pipe and consists of packets, the first of which is from the root hub to the device. Subsequent packets in the frame may also be in this direction, or they may be back from the device to the root hub. A sequence of four frames is shown in Fig. 3-58.

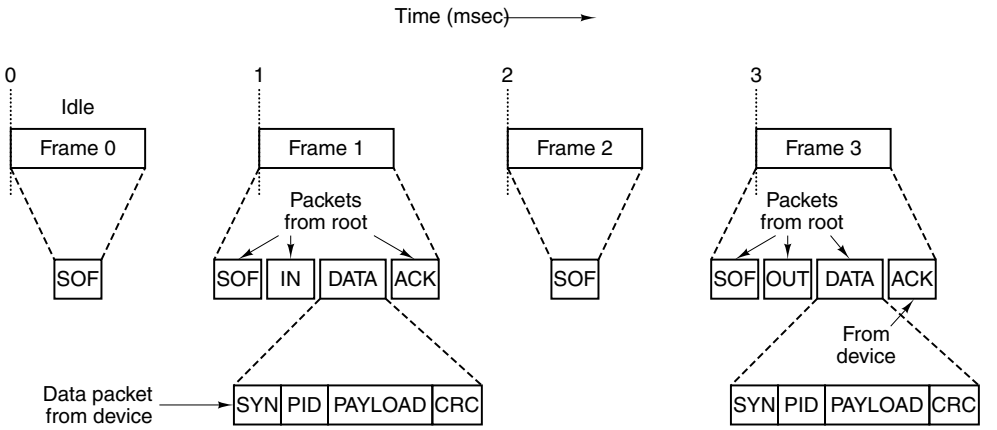


Figure 3-58. The USB root hub sends out frames every 1.00 msec.

In Fig. 3-58 there is no work to be done in frames 0 and 2, so all that is needed is one SOF (Start of Frame) packet. This packet is always broadcast to all devices. Frame 1 is a poll, for example a request to a scanner to return the bits it has found on the image it is scanning. Frame 3 consists of delivering data to some device, for example to a printer.

USB supports four kinds of frames: control, isochronous, bulk, and interrupt. Control frames are used to configure devices, give them commands, and inquire about their status. Isochronous frames are for real-time devices such as microphones, loudspeakers, and telephones that need to send or accept data at precise time intervals. They have a highly predictable delay but provide no retransmissions in the event of errors. Bulk frames are for large transfers to or from devices with no real-time requirements, such as printers. Finally, interrupt frames are needed because USB does not support interrupts. For example, instead of having the keyboard cause an interrupt whenever a key is struck, the operating system can poll it every 50 msec to collect any pending keystrokes.

A frame consists of one or more packets, possibly some in each direction. Four kinds of packets exist: token, data, handshake, and special. Token packets are from the root to a device and are for system control. The SOF, IN, and OUT packets in Fig. 3-58 are token packets. The SOF packet is the first in each frame and marks the beginning of the frame. If there is no work to do, the SOF packet is the

only one in the frame. The IN token packet is a poll, asking the device to return certain data. Fields in the IN packet tell which bit pipe is being polled so the device knows which data to return (if it has multiple streams). The OUT token packet announces that data for the device will follow. A fourth type of token packet, SETUP (not shown in the figure), is used for configuration.

Besides the token packet, three other kinds exist. These are DATA (used to transmit up to 64 bytes of information either way), handshake, and special packets. The format of a DATA packet is shown in Fig. 3-58. It consists of an 8-bit synchronization field, an 8-bit packet type (PID), the payload, and a 16-bit **CRC** (**Cyclic Redundancy Check**) to detect errors. Three kinds of handshake packets are defined: ACK (the previous data packet was correctly received), NAK (a CRC error was detected), and STALL (please wait—I am busy right now).

Now let us look at Fig. 3-58 again. Every 1.00 msec a frame must be sent from the root hub, even if there is no work. Frames 0 and 2 consist of just an SOF packet, indicating that there was no work. Frame 1 is a poll, so it starts out with SOF and IN packets from the computer to the I/O device, followed by a DATA packet from the device to the computer. The ACK packet tells the device that the data were received correctly. In case of an error, a NAK would be sent back to the device and the packet would be retransmitted for bulk data (but not for isochronous data). Frame 3 is similar in structure to frame 1, except that now the data flow is from the computer to the device.

After the USB standard was finalized in 1998, the USB designers had nothing to do so they began working on a new high-speed version of USB, called **USB 2.0**. This standard is similar to the older USB 1.1 and backward compatible with it, except that it adds a third speed, 480 Mbps, to the two existing speeds. There are also some minor differences, such as the interface between the root hub and the controller. With USB 1.1 two new interfaces were available. The first one, **UHCI** (**Universal Host Controller Interface**), was designed by Intel and put most of the burden on the software designers (read: Microsoft). The second one, **OHCI** (**Open Host Controller Interface**), was designed by Microsoft and put most of the burden on the hardware designers (read: Intel). In USB 2.0 everyone agreed to a single new interface called **EHCI** (**Enhanced Host Controller Interface**).

With USB now operating at 480 Mbps, it clearly competes with the IEEE 1394 serial bus popularly called FireWire, which runs at 400 Mbps or 800 Mbps. Because virtually every new Intel-based PC now comes with USB 2.0 or USB 3.0 (see below) 1394 is likely to vanish in due course. This disappearance is not so much due to obsolescence as to turf wars. USB is a product of the computer industry whereas 1394 comes from the consumer electronics industry. When it came to connecting cameras to computers, each industry wanted everyone to use its cable. It looks like the computer folks won this one.

Eight years after the introduction of USB 2.0, the **USB 3.0** interface standard was announced. USB 3.0 supports a whopping 5-Gbps bandwidth over the cable, although the link modulation is adaptive, and one is likely to achieve this top speed

only with professional-grade cabling. USB 3.0 devices are structurally identical to earlier USB devices, and they fully implement the USB 2.0 standard. If plugged into a USB 2.0 socket, they will operate correctly.

3.7 INTERFACING

A typical small- to medium-sized computer system consists of a CPU chip, chipset, memory chips, and some I/O devices, all connected by a bus. Sometimes, all of these devices are integrated into a system-on-a-chip, like the TI OMAP4430 SoC. We have already studied memories, CPUs, and buses in some detail. Now it is time to look at the last part of the puzzle, the I/O interfaces. It is through these I/O ports that the computer communicates with the external world.

3.7.1 I/O Interfaces

Numerous I/O interfaces are already available and new ones are being introduced all the time. Common interfaces include UARTs, USARTs, CRT controllers, disk controllers, and PIOs. A **UART** (**Universal Asynchronous Receiver Transmitter**) is an I/O interface that can read a byte from the data bus and output it a bit at a time on a serial line for a terminal, or input data from a terminal. UARTs usually allow various speeds from 50 to 19,200 bps; character widths from 5 to 8 bits; 1, 1.5, or 2 stop bits; and provide even, odd, or no parity, all under program control. **USART**s (**Universal Synchronous Asynchronous Receiver Transmitters**) can handle synchronous transmission using a variety of protocols as well as performing all the UART functions. Since UARTs have become less important as telephone modems are vanishing, let us now study the parallel interface as an example of an I/O chip.

PIO Interfaces

A typical **PIO** (**Parallel Input/Output**) interface (based on the classic Intel 8255A PIO design) is illustrated in Fig. 3-59. It has a collection of I/O lines (e.g., 24 I/O lines in the example in the figure) that can interface to any digital logic device interface, for example, keyboards, switches, lights, or printers. In a nutshell, the CPU program can write a 0 or 1 to any line, or read the input status of any line, providing great flexibility. A small CPU-based system using a PIO interface can control a variety of physical devices, such as a robot, toaster, or electron microscope. Typically, PIO interfaces are found in embedded systems.

The PIO interface is configured with a 3-bit configuration register, which specifies if the three independent 8-bit ports are to be used for digital signal input (0) or output (1). Setting the appropriate value in the configuration register will allow any combination of input and output for the three ports. Associated with

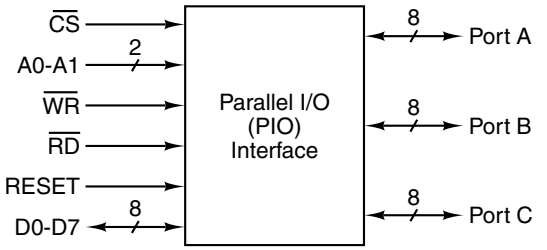


Figure 3-59. A 24-bit PIO Interface.

each port is an 8-bit latch register. To set the lines on an output port, the CPU just writes an 8-bit number into the corresponding register, and the 8-bit number appears on the output lines and stays there until the register is rewritten. To use a port configured for input, the CPU just reads the corresponding 8-bit latch register.

It is possible to build more sophisticated PIO interfaces. For example, one popular operating mode provides for handshaking with external devices. For example, to output to a device that is not always ready to accept data, the PIO can present data on an output port and wait for the device to send a pulse back saying that it has accepted the data and wants more. The necessary logic for latching such pulses and making them available to the CPU includes a ready signal plus an 8-bit register queue for each output port.

From the functional diagram of the PIO we can see that in addition to 24 pins for the three ports, it has eight lines that connect directly to the data bus, a chip select line, read and write lines, two address lines, and a line for resetting the chip. The two address lines select one of the four internal registers, corresponding to ports A, B, C, and the port configuration register. Normally, the two address lines are connected to the low-order bits of the address bus. The chip select line allows the 24-bit PIO to be combined to form larger PIO interfaces, by adding additional address lines and using them to select the proper PIO interface by asserting its chip select line.

3.7.2 Address Decoding

Up until now we have been deliberately vague about how chip select is asserted on the memory and I/O chips we have looked at. It is now time to look more carefully at how this is done. Let us consider a simple 16-bit embedded computer consisting of a CPU, a 2KB $\times$ 8 byte EPROM for the program, a $2-$KB $\times$ 8 byte RAM for the data, and a PIO interface. This small system might be used as a prototype for the brain of a cheap toy or simple appliance. Once in production, the EPROM might be replaced by a ROM.

The PIO interface can be selected in one of two ways: as a true I/O device or as part of memory. If we choose to use it as an I/O device, then we must select it

using an explicit bus line that indicates that an I/O device is being referenced, and not memory. If we use the other approach, **memory-mapped I/O**, then we must assign it 4 bytes of the memory space for the three ports and the control register. The choice is somewhat arbitrary. We will choose memory-mapped I/O because it illustrates some interesting issues in I/O interfacing.

The EPROM needs 2 KB of address space, the RAM also needs 2K of address space, and the PIO needs 4 bytes. Because our example address space is 64K, we must make a choice about where to put the three devices. One possible choice is shown in Fig. 3-60. The EPROM occupies addresses to 2K, the RAM occupies addresses 32 KB to 34 KB, and the PIO occupies the highest 4 bytes of the address space, 65532 to 65535. From the programmer's point of view, it makes no difference which addresses are used; however, for interfacing it does matter. If we had chosen to address the PIO via the I/O space, it would not need any memory addresses (but it would need four I/O space addresses).

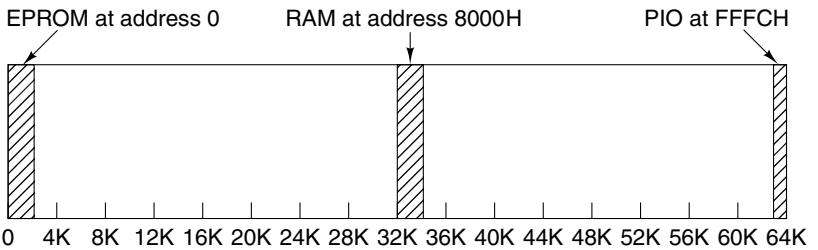


Figure 3-60. Location of the EPROM, RAM, and PIO in our 64-KB address space.

With the address assignments of Fig. 3-60, the EPROM should be selected by any 16-bit memory address of the form 00000xxxxxxxxxxx (binary). In other words, any address whose 5 high-order bits are all 0s falls in the bottom 2 KB of memory, hence in the EPROM. Thus, the EPROM's chip select could be wired to a 5-bit comparator, one of whose inputs was permanently wired to 00000.

A better way to achieve the same effect is to use a five-input OR gate, with the five inputs attached to address lines A11 to A15. If and only if all five lines are 0 will the output be 0, thus asserting $\overline{\text{CS}}$ (which is asserted low). This addressing approach is illustrated in Fig. 3-60(a) and is called full-address decoding.

The same principle can be used for the RAM. However, the RAM should respond to binary addresses of the form 10000xxxxxxxxxxx, so an additional inverter is needed as shown in the figure. The PIO address decoding is somewhat more complicated, because it is selected by the four addresses of the form 11111111111111xx. A possible circuit that asserts $\overline{\text{CS}}$ only when the correct address appears on the address bus is shown in the figure. It uses two eight-input NAND gates to feed an OR gate.

However, if the computer really consists of only the CPU, two memory chips, and the PIO, we can use a trick to greatly simplify the address decoding. The trick

is based on the fact that all EPROM addresses, and only EPROM addresses, have a 0 in the high-order bit, A15. Therefore, we can just wire $\overline{\text{CS}}$ to A15 directly, as shown in Fig. 3-61(b).

At this point the decision to put the RAM at 8000H may seem much less arbitrary. The RAM decoding can be done by noting that the only valid addresses of the form 10xxxxxxxxxxxxx are in the RAM, so 2 bits of decoding are sufficient. Similarly, any address starting with 11 must be a PIO address. The complete decoding logic is now two NAND gates and an inverter.

The address decoding logic of Fig. 3-61(b) is called **partial address decoding**, because the full addresses are not used. It has the property that a read from addresses 0001000000000000, 0001100000000000, or 0010000000000000 will give the same result. In fact, every address in the bottom half of the address space will select the EPROM. Because the extra addresses are not used, no harm is done, but if one is designing a computer that may be expanded in the future (an unlikely occurrence in a toy), partial decoding should be avoided because it ties up too much address space.

Another common address-decoding technique is to use a decoder, such as that shown in Fig. 3-13. By connecting the three inputs to the three high-order address lines, we get eight outputs, corresponding to addresses in the first 8K, second 8K, and so on. For a computer with eight RAMs, each 8K × 8, one such chip provides the complete decoding. For a computer with eight 2K × 8 memory chips, a single decoder is also sufficient, provided that the memory chips are each located in distinct 8-KB chunks of address space. (Remember our earlier remark that the position of the memory and I/O chips within the address space matters.)

3.8 SUMMARY

Computers are constructed from integrated circuit chips containing tiny switching elements called gates. The most common gates are AND, OR, NAND, NOR, and NOT. Simple circuits can be built up by directly combining individual gates.

More complex circuits are multiplexers, demultiplexers, encoders, decoders, shifters, and ALUs. Arbitrary Boolean functions can be programmed using a FPGA. If many Boolean functions are needed, FPGAs are often more efficient. The laws of Boolean algebra can be used to transform circuits from one form to another. In many cases more economical circuits can be produced this way.

Computer arithmetic is done by adders. A single-bit full adder can be constructed from two half adders. An adder for a multibit word can be built by connecting multiple full adders in such a way as to allow the carry out of each one feed into its left-hand neighbor.

The components of (static) memories are latches and flip-flops, each of which can store one bit of information. These can be combined linearly into latches and

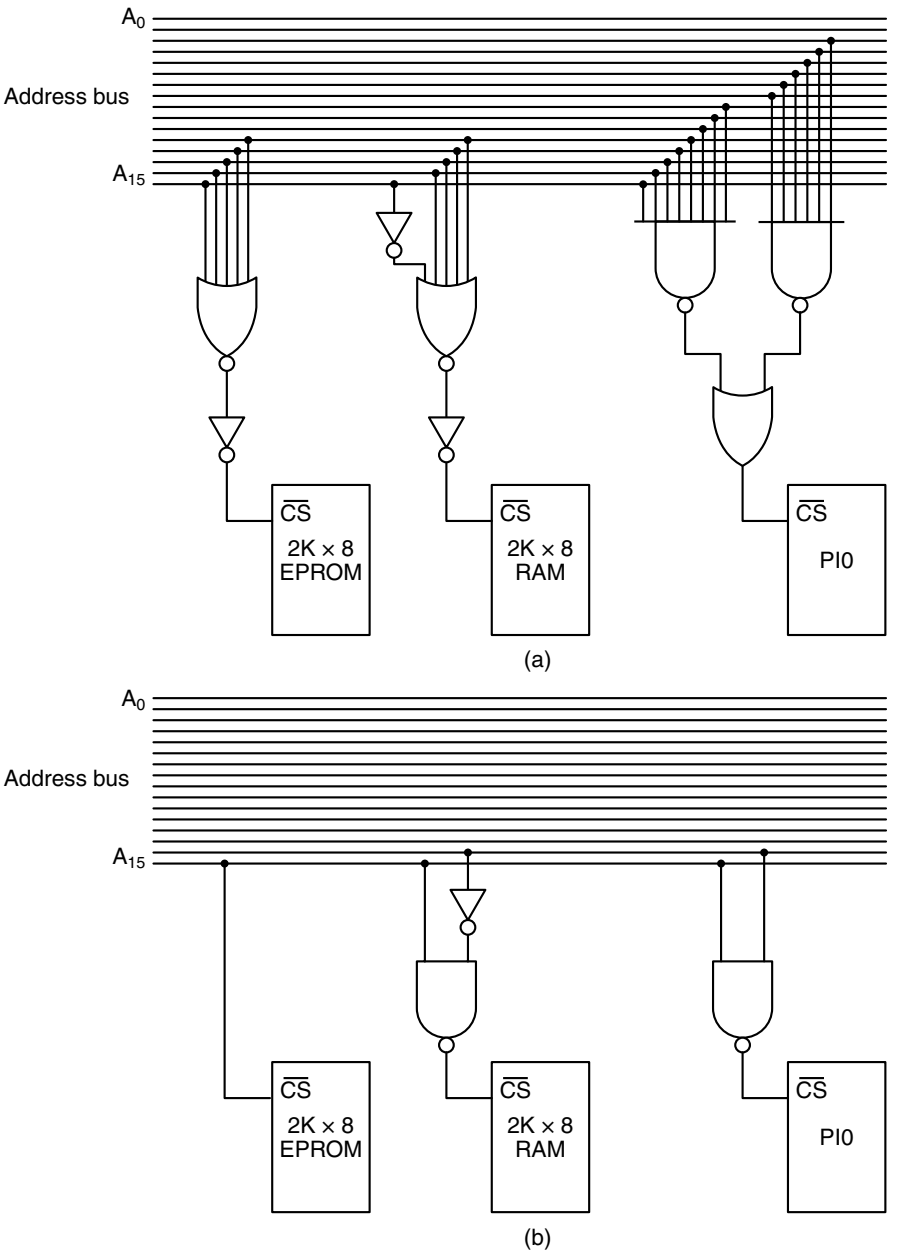


Figure 3-61. (a) Full address decoding. (b) Partial address decoding.

flip-flops for memories with any word size desired. Memories are available as RAM, ROM, PROM, EPROM, EEPROM, and flash. Static RAMs need not be refreshed; they keep their stored values as long as the power remains on. Dynamic RAMs, on the other hand, must be refreshed periodically to compensate for leakage from the little capacitors on the chip.

The components of a computer system are connected by buses. Many, but not all, of the pins on a typical CPU chip directly drive one bus line. The bus lines can be divided into address, data, and control lines. Synchronous buses are driven by a master clock. Asynchronous buses use full handshaking to synchronize the slave to the master.

The Core i7 is an example of a modern CPU. Modern systems using it have a memory bus, a PCIe bus, and a USB bus. The PCIe interconnect is the dominant way to connect the internal parts of a computer at high speeds. The ARM is also a modern high-end CPU but is intended for embedded systems and mobile devices where low power consumption is important. The Atmel ATmega168 is an example of a low-priced chip good for small, inexpensive appliances and many other price-sensitive applications.

Switches, lights, printers, and many other I/O devices can be interfaced to computers using parallel I/O interfaces. These chips can be configured to be part of the I/O space or the memory space, as needed. They can be fully decoded or partially decoded, depending on the application.

PROBLEMS

1. Analog circuits are subject to noise that can distort their output. Are digital circuits immune to noise? Discuss your answer.

2. A logician drives into a drive-in restaurant and says, "I want a hamburger or a hot dog and french fries." Unfortunately, the cook flunked out of sixth grade and does not know (or care) whether "and" has precedence over "or." As far as he is concerned, one interpretation is as good as the other. Which of the following cases are valid interpretations of the order? (Note that in English "or" means "exclusive or.")

a. Just a hamburger.
b. Just a hot dog.
c. Just french fries.
d. A hot dog and french fries.
e. A hamburger and french fries.
f. A hot dog and a hamburger.
g. All three.
h. Nothing—the logician goes hungry for being a wiseguy.