CHAPTER 5
WHEN THE WAFERS ARE DELIVERED

The return of the finished wafers is an exciting moment for the IC designer. Proper preparation for this day can help to prevent frustrating delays in *wafer separation, packaging* and *testing,* thus promoting rapid feedback concerning the success of the designs. Testing complexity increases even faster than design complexity. Thus testing needs suitable preparation as already mentioned in Section 2.7. If "multiple iteration with fast turn-around" is to become a successful IC design methodology, testing must also be efficient and streamlined. Often the designer tends to relax after the masks have been submitted, when instead he should immediately begin preparation of the test set-up for his chip. The following discussion of wafer separation, chip bonding and testing techniques is primarily aimed at those in a research, rather than a production environment.

Testing progresses in a multistep, hierarchical approach. First it has to be determined whether the wafer was properly processed. This should be done on the uncut wafer, since there is no point in wasting the effort of chip separation and mounting on defective wafers. For large designs it may be worthwhile to do some preliminary electrical tests by probing individual chips in order to determine which ones should be mounted.

5.1 Process Testing

Some tests can be performed visually by inspecting the wafers under a microscope. All wafers should be checked for a clean look, for lack of obvious defects such as scratches or missing parts, for sharply defined patterns in all visible mask levels, and for the absence of obvious shorts between adjacent lines or breakage in narrow paths. Some of these visual tests can be facilitated by special test patterns included on the chip. In particular, a series of L-shaped paths (see section 6.1) of various widths and spacings provide quick information about the resolution and cleanliness of the various processing steps.

Special test structures are useful for evaluating the overall process quality or determining process parameters, for example oxide thickness or the conductivity of various levels of interconnects. Some features that have proven useful in our experience are: narrowly spaced interdigited combs to detect bridging between adjacent lines; long, narrow paths to detect breakage; metal paths running across other features such as poly lines to test step coverage; a series connection of many short sections of metal paths and poly or diffusion paths to test for open contact holes.

These and other structures permit one to measure electrical and process parameters. If the geometry of the long paths is known, then the resistivity of the corresponding layer can be derived from a measurement of the end-to-end resistance. A large MOS capacitor is used for capacitance *vs* gate voltage measurements to determine gate oxide thickness, fixed oxide charge and density of interface states. Some isolated MOS transistors give feedback on threshold voltages and saturation currents in enhancement and depletion mode devices. If the transistors are arranged so that they can be readily connected as inverters, then the characteristic of this all-important building block can also be checked. A circular connection of an odd number of inverter stages forms a ring oscillator, and its the fundamental oscillation frequency yields a good first measurement of the basic pair delay of the inverter stages.

It is often useful to include a set of test patterns on each chip, then if the circuit does not perform as expected the designer can probe the test patterns. The results of those tests give a quick yes/no indication of whether devices on the particular chip are functioning, and hence that the chip was properly fabricated. If the test patterns work correctly, a design or bonding problem may well be at fault. Otherwise the processing is likely to be marginal, and a different chip can be tested.

5.2 Wafer Separation

The wafers, as returned from the fabrication line, are disks of silicon that must be broken into chips that fit into the cavity in a dual-in-line package *(DIP).* Wafer separation is accomplished in one of two ways: *scribing* or *sawing.*

Scribing is a simple operation similar to glass cutting. The wafer is held on a vacuum table, which is an integral part of the scribing machine, or *scriber,* and a diamond-tipped scribing tool is dragged across the surface within the confines of the scribe lines. Vertical stress cracks are induced where the diamond tip of the scribing tool slides over bare silicon. The pressure that the scribing tool exerts on the silicon is critical, too little results in random breakage in the fracturing operation, while too much produces stress cracks in the horizontal direction. Such cracks cause splintering of the wafer radially from the scribe lines, probably into active circuit elements. After each scribe line in the grid has been "scratched" in this fashion, the wafer (at this point it is still in one piece) is removed from the scriber and fractured into chips. This may be achieved in a number of ways, for example by sandwiching the wafer in some soft material (rubber sheeting, filter paper), supporting it on a foam rubber block, and rolling a cylindrical bar over it. If the wafer was properly scribed the flexing force is concentrated at the scribe lines and the wafer fractures cleanly along them.

Sawing is an alternative to scribing. In this technique a thin saw blade with an edge containing diamond-dust is used to cut approximately 2/3 of the way through the silicon wafer. Again, the wafer is removed from the saw in one piece and fractured by techniques similar to the one outlined above.

Sawing offers several advantages over scribing. The saw can slice anywhere on the wafer, thus no scribe lines are needed. This allows dense packing of projects on a multi-project chip; when the wafers are returned from fabrication each designer can have a wafer sliced up without regard to the location of other projects on the wafer (i.e. by sacrificing neighboring projects to the saw blade). Most saws can be set up to automatically step across the wafer, making parallel cuts at fixed intervals. This convenient feature provides repeatable die sizing and saves a considerable amount of time, particularly for operators who dice wafers infrequently. Sawing also leaves square edges after fracturing which makes manipulating the chips with tweezers an easy task.

Disadvantages include the necessity of removing the silicon dust (*slurry*) generated in the sawing process — this means an extra cleaning step following wafer separation. Sawing equipment is somewhat more complex and expensive ($12,000-$20,000) than scribers ($2,000-$5,000). Usually there is more maintainance required and more setup overhead involved.

5.3 Chip Packaging

Once the wafers are fractured into chips only *packaging* remains before they are ready to be tested. Packaging encompasses two different operations, *chip attachment* and *wire bonding*. In the first operation the chip is permanently affixed to the IC package; the second involves connecting the aluminum pads on the chip to pads surrounding the package cavity. These package pads are connected through the ceramic to the external pins.

Chip attachment is a straightforward process, especially in a low volume research environment. The chip must be solidly attached to the mounting pad (*header*) in the package cavity. The bond should exhibit low thermal resistance and make good electrical contact with the silicon substrate. Common means of attachment include *solder bonding* (both header and chip must be heated to the melting point of the solder used) and *eutectic bonding* (usually utilizing a gold-silicon alloy, see [Glaser 1977]). By far the most convenient for the researcher is *epoxy bonding*: the backside of the chip or the header is dabbed with a commercially available silver/epoxy mixture and then pressed onto the header. Tweezers suffice for handling the chips. The header and chip are baked at a low temperature for a few hours to cure the epoxy and the assembly is ready for wire bonding.

All three of the manual wire bonding techniques in widespread use require considerable skill on the machine operator's part. *Thermocompression bonding* relies on pressure and heat to produce a strong bond. The header and the chip are maintained at about 350 degrees centigrade. A gold ball (on the end of a fine gold wire) is squashed against the aluminum bonding pad on the chip, forming a bond in a fraction of a second. As the capillary, which guides the wire, is withdrawn from the bonding pad, wire is automatically payed out; the operator maneuvers the capillary over the desired post and the wire is mashed against it, forming the second bond. As the capillary is backed away, the wire is cut by a gas flame that simultaneously forms a gold ball for the next bond. This leaves a short length of wire, from the second bond to the point where the flame made the cut, that must be removed by hand after all wires are attached.

Ultrasonic bonding utilizes aluminum wire and ultrasonic energy to make bonds. The aluminum wire is pressed against the bonding pad and a short burst of ultrasonic energy locally heats the wire/pad interface so that a bond is formed. Similarly, a second bond is made on a post, and the wire is cut, usually by mechanical means. The header may be heated to assist the bonding process.

Ultrasonic bonding offers low materials cost but is less flexible than the thermocompression technique, which allows "daisy-chaining" of connection points. Thermocompression also gives more freedom to choose the angles at which wires leave the bonding pads, enabling some further flexibility which may be needed in a research environment.

Thermosonic bonding effects bonds with a combination of thermal and ultrasonic energy. The header and the capillary are both heated to a somewhat lower temperature than required in thermocompression bonding; ultrasound provides the additional energy to form the bond at the instant of wire/pad contact. This technique has the advantage that the chip and header are cooler than in thermocompression bonding, and so the epoxies used to bond chip to header need not be so heat resistant. At the same time, more wire-positioning freedom is possible than with ultrasonic bonding.

The importance of this wire-positioning freedom should not be overlooked. On a multi-project chip the operator is often faced with less than optimal pad location, and the ability to make bonds at strange angles can mean the difference between being able and not being able to wire bond a particular project. Thus, thermocompression or thermosonic bonding equipment seems most reasonable for those applications.

Chip attachment and wire bonding will doubtless be carried out in a central location, and the packaged chips distributed from there. We have found that the operator of the wire bonding machine can work most efficiently if he or she is allowed to choose the pad/pin mapping as the chip is bonded. For this purpose we provide a number of blank wire bonding maps (see Figure 5.3.1) that the operator fills in as the bonds are made. Once the project is wire bonded, it is packed and distributed with its map. In general, research chips need not be hermetically sealed in their packages, often a piece of tape over the cavity (or no cover at all) will prove adequate. Users should be aware, however, that MOS circuits exhibit very different device characteristics when operated in light (in fact, they may not work correctly).

5.4 Functional Testing

The next step is to functionally test the mounted chip. As mentioned previously, testing really starts at design time. The issues that should be taken into consideration have been discussed in Section 2.7. Here we discuss the actual testing procedure.

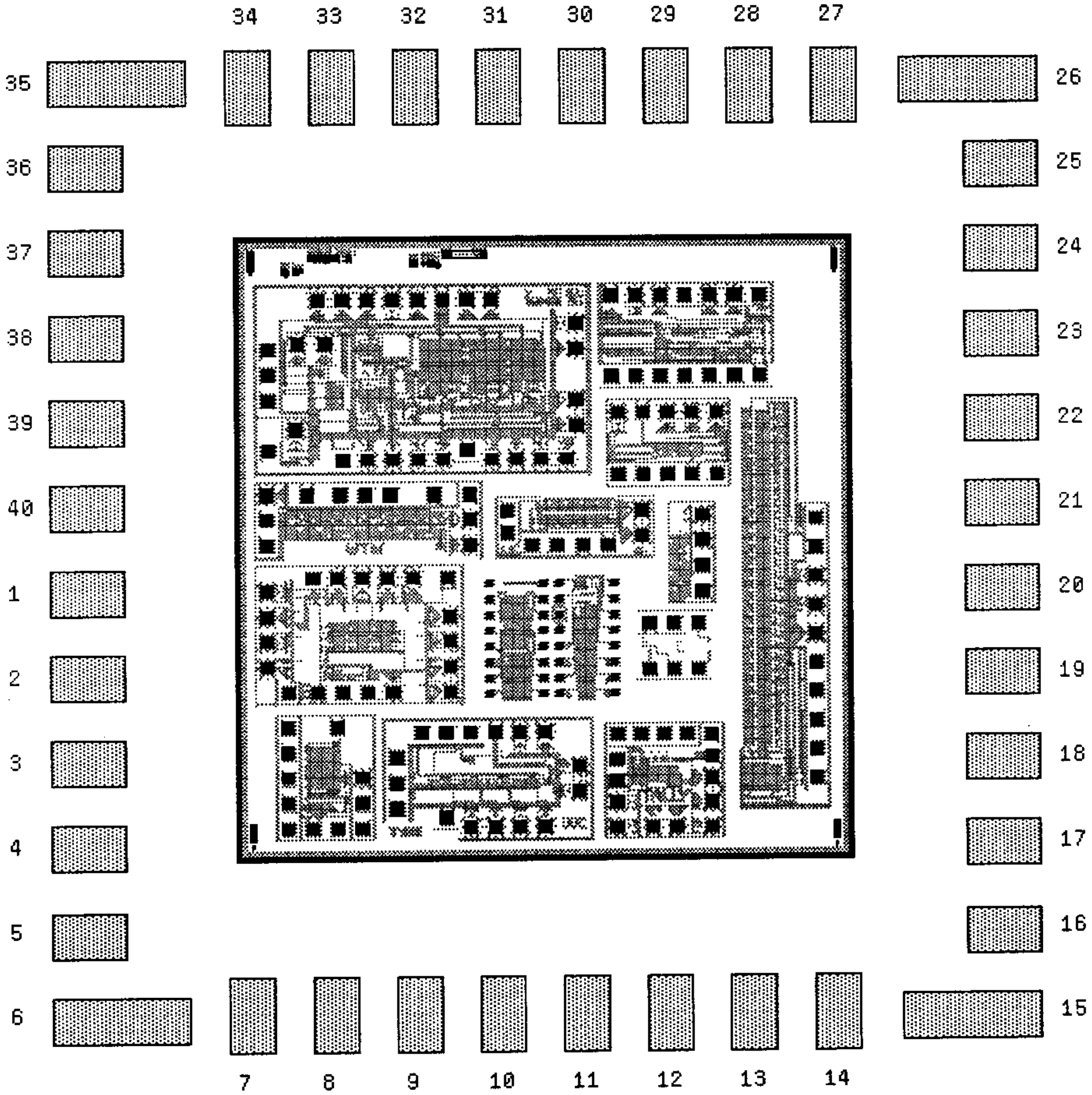


Figure 5.3.1 Wire-Bonding Map

The most burning question to the designer is: *does the chip work?* A preliminary answer to this question should be readily obtainable by plugging the chip into the test set-up that *the designer* should have prepared while waiting for the wafers to be processed. Of course, even if the expected output signals are observed and if it looks like everything is OK, the real testing is only just beginning.

Whether a chip works or not is not just a binary decision. The *quality* of a chip is described by such parameters as maximum operating speed, power consumption and drive capabilities. In addition its *operating margins* must be determined, i.e. what are maximum and minimum supply voltages, maximum clock frequencies and minimum clock pulse widths for which the circuit still works. Even if it is not intended to ever operate under these conditions, the width of these margins will give an indication of how robust the design really is. This ties in with other issues of reliability, for example, is the chip susceptible to external noise pick-up or is the chip subject to an aging process either while sitting on the shelf or during prolonged operation? Even after all these questions have been answered to the designers satisfaction, there may still be some doubts about the *correctness* of a design. Will it really behave properly under *all* allowed conditions — can it be proven? Can the chip be tested exhaustively?

Initial tests may also lead to the agonizing conclusion that *the chip does not work.* The first thought is to blame it on processing. However, since the wafers have been process tested before separation, this is an unlikely excuse. A careful inspection under a microscope often reveals defects which could be responsible for the malfunctioning of the device. Such defects could result from a small local defect in the mask set, from a random defect in fabrication or, perhaps most likely, from mishandling during packaging. Often there are scratches extending several hundred microns from bonding pads into the circuitry.

In most cases there will be at least some signals coming out of the chip, which provide clues about what is going on inside and which can be used to start systematic debugging. Since the logical procedure depends strongly on the particular circuit, debugging is often more an art than a science. If possible a few general rules should be followed. If there are enough intermediate access points, it should be possible to systematically proceed from input to output and to verify the function of subsequent stages until the faulty one is found. It is now that proper partitioning and the inclusion of special testing aids will pay off.

5.5 Simple Test Systems

Except for simple blocks of Boolean logic, typical MOS IC's rarely permit manual testing. Circuit complexity, a large number of inputs, clock signals and outputs, or the multitude of possible states may make manual testing prohibitively time-consuming. Furthermore, if there are any internal nodes which store charge in a dynamic manner, manually exercising such a chip may be too slow to test its operation. In general it will be necessary to use a computer to exercise the chip (see figure 5.5.1). Properly defined *excitation vectors*, provided by the computer's *output port*, are applied to the inputs of the circuit under test. The *response vector* of the device under test can then be read by the computer's *input port*. This response can then be properly formatted, output in hardcopy, or displayed on a screen in textual or graphical form. Alternatively, it could be compared to a stored correct response, with the computer programmed to respond with an error message upon detecting deviation from the expected pattern. In any case, the test vectors must be carefully selected if they are to supply relevant information about the chip.

Typically, only a small amount of hardware is required between the computer and the device under test (for example, see [Mathews 1979]). A minimal system could consist of a general purpose test board with a zero insertion force socket, a tristate driver and read buffer connected to each pin, a couple of latches to store the state of the input and/or output variable at each pin, and a communication interface to the computer. For simple chips, a serial data line, transmitting individual characters for each change of the state of an input pin or for each read request of some outputs, may be sufficient. Obviously the maximum rate at which a chip in such a set-up can be exercised is limited by the bandwidth of the serial link and the number of pins that need to be changed in each phase. This approach has the advantage that a minimal amount of additional hardware is needed. Such a software-based testing system, however, might not be fast enough to capture some quickly changing response vectors, or it may have trouble exercising a complex device which may require some minimum clock rates for proper operation. Testing the maximum operating speed of an IC also requires input rates which often can not be provided by a simple software system.

More performance can be obtained with a parallel link to the computer. If the excitation vector is wider than the typical 8 or 16 bit link to the host computer, the vector may be transmitted in several pieces and assembled in a set of registers on the interface board. Even higher I/O vector speeds can be obtained by moving more functions from software into hardware. An improved tester could use semiconductor memory to store the excitation and response vectors of a whole test sequence. A hardware counter is used to step the memory through the required words. The captured response vectors can later be analyzed at a slower rate. If two memory banks are used each for input and output vectors, the device can be kept running in a continuous test loop, while the other input bank is reloaded with the instructions for a new test cycle or while the second output bank is read by the computer.

The tester is now a *peripheral* device to the host computer (see figure 5.5.2). The advantage of ever more self-contained testing hardware is two-fold; it allows higher testing speeds and frees the computer to perform other tasks while testing is in progress. The ultimate step is to provide a stand-alone processor for testing, flexible enough to test any conceivable digital IC. The excitation vector for the device under test can be understood as a set of *control words* emanating from a computer's *control unit*, with the result vector out of the device acting as a *condition vector* to this

Figure 5.5.1
Software approach to IC testing. A computer is dedicated to exciting the device under test (DUT) and collecting the response. The disadvantage of this method is that it is slow and wasteful of computer time. Note that the computer is directly connected to the device under test.

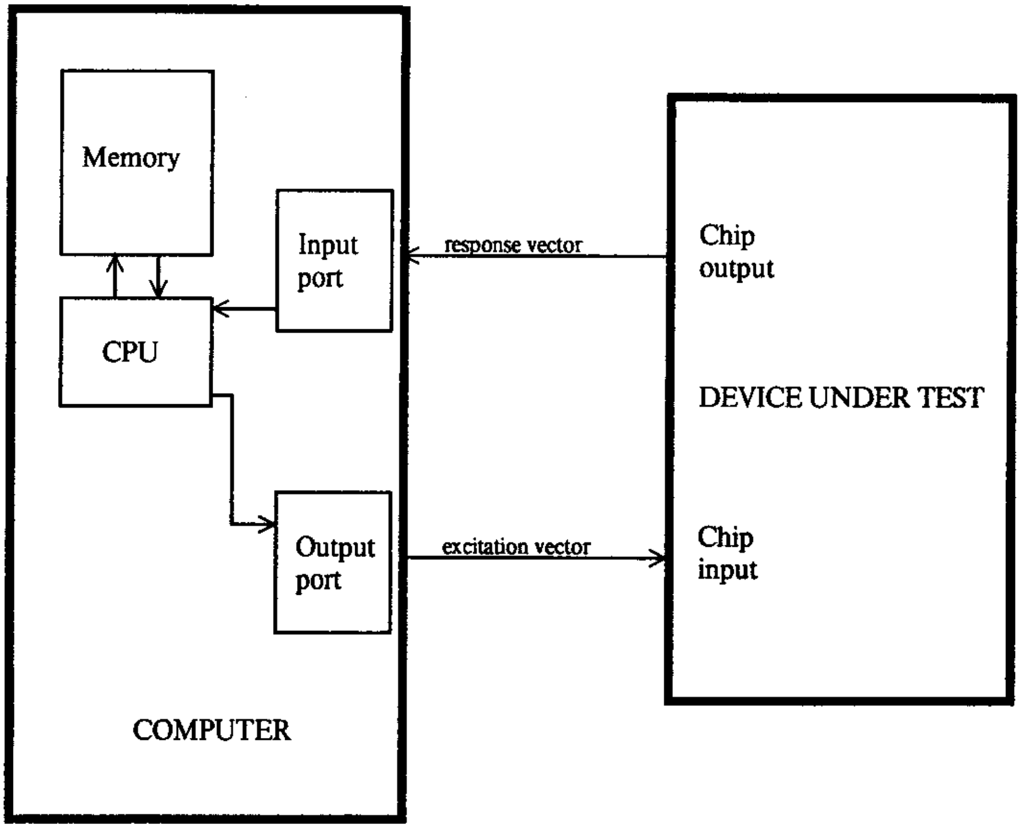
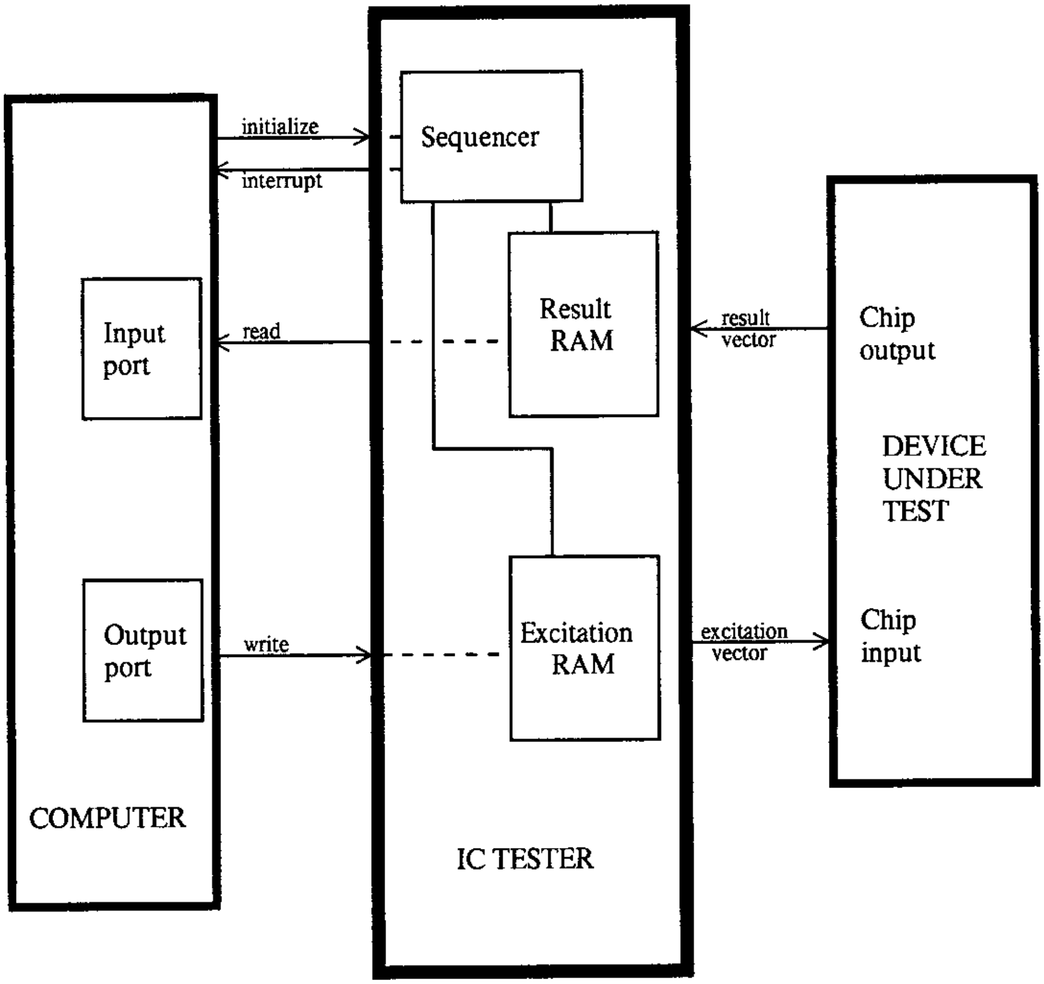


Figure 5.5.2
Hardware approach to IC testing. The computer initializes an IC Tester that is connected as a peripheral device. The sequencer (counter) steps both RAMS, sending an excitation vector and collecting the result vector. When the test is over, the sequencer interrupts the computer. This method is fast and requires little CPU time. Note that the computer is no longer directly connected to the DUT.



control unit. The tester thus takes the form of a *microprogrammed* controller. A tester based on this principle can exercise very complex devices due to its inherent ability to make logical decisions based on some of the results. When the test is concluded, relevant results are as before stored in a result RAM, and the host computer is signaled to fetch them. This kind of a tester can be adapted to new tasks by changing the *microcode* (see figure 5.5.3); it may even do suitable branching dependent on the outcome of a few preliminary tests.

Almost any computer or microcomputer can be used as the host. The only requirement is that it have an accessible I/O port. The control unit for the tester could be built using one of the fast bipolar *bit-slice* microprocessors.

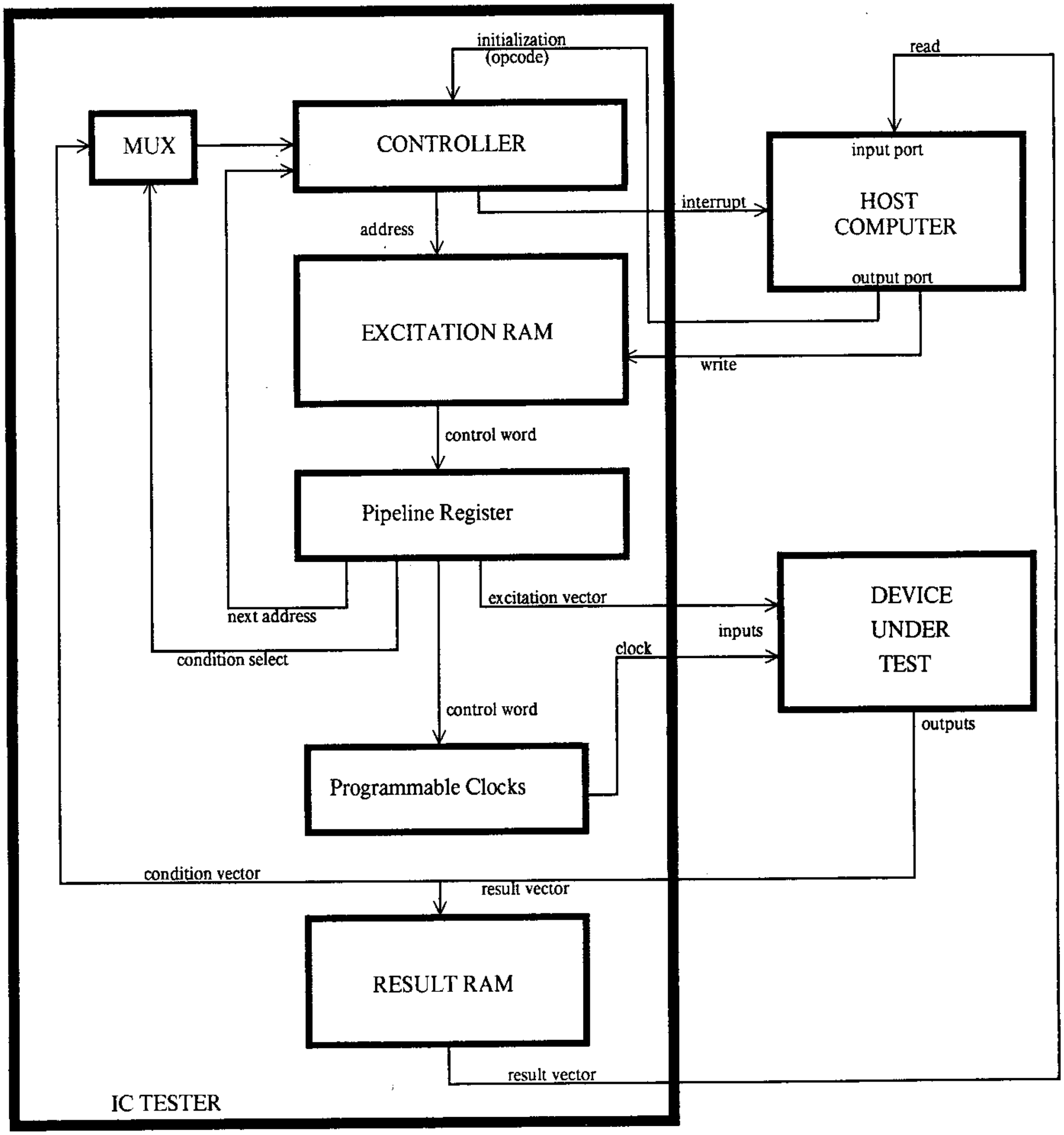
5.6 A Concluding Remark

It should be emphasized that preparing for the day when the wafers come back from the fab line may be as large and complicated a task as the original design. The proper custom made interface board between the chip and the test system has to be built and the test routines have to be written. In preparing for testing, the designer should keep in mind the possibility that the chip does not work at all and plan a strategy to deal with this case. Debugging and testing is the responsibility of the designer and should be kept in mind from the early stages of the design process to the day of the delivery of a finished product to a customer. The availability of quick turnaround IC implementation could permit a new and more experimental way of designing and debugging integrated systems, in the same style that programmers now use to build large software systems. It is clear that this iterative design loop must be closed by the IC designer, who must provide the hooks to extract all relevant information from the silicon chip.

References

[Mathews 1979]
R. Mathews, "A Minimal Tester", Stanford CS Technical Report, November 1979.

Figure 5.5.3
Block diagram of a Microprogrammed IC Tester. This approach allows maximum flexibility by being programmable. The computer initializes the controller which outputs the address of the first microinstruction to the microprogram RAM. The microprogram RAM supplies the next address and enables the controller to sequence through various testing microsubroutines.


initialization
(opcode)
read
MUX
CONTROLLER
input port
HOST
COMPUTER
interrupt
address
output port
EXCITATION RAM
write
control word
Pipeline Register
excitation vector
DEVICE
UNDER
TEST
inputs
next address
condition select
clock
control word
outputs
Programmable Clocks
condition vector
result vector
RESULT RAM
result vector
IC TESTER