Semiconductor Manufacturing Considerations for VLSI Designers

Steve McMinn, American Microsystems, Inc.

In the past five years, semiconductor processing technology has made startling advances. Minimum production line widths have been reduced from 6 microns (1977) to the state-of-the-art 3 microns. At the same time, improvements in photolithography have let die areas nearly double, with no appreciable yield loss. Together, these advances mean that up to four times as much circuitry can be implemented on a single chip than was possible only five years ago.

This capability has led to a resurgence of the truly customized chip. Many of the peripheral circuits previously required to make a standard IC operational in the required configuration can now be included on-chip. In the long run, this inclusion can offset the non-recurring costs of designing a custom chip, and can also reduce the overall space required for the system.

More and more systems-oriented designers are beginning to design custom ICs—especially with the emergence of structured design approaches. Because many designers and systems houses do not have access to an internal fabrication facility, the need for production-oriented customer-owned tooling operations is greater than ever. However, problems still exist with the customer-owned tooling interface; they must be understood by designers so as to optimize throughput and to minimize circuit-development and production costs.

Use Industry-Standard Processes

Whenever possible, designs should be limited to widely available processes. Locking a design into one manufacturer's offbeat process can be dangerous; it can lead to shipping delays if the supplier runs into a fab problem. It is also wise not to push a process to its limit in hopes of achieving peak performance. To expect optimal performance from every wafer run is unrealistic; it can contribute to yield problems if a certain "margin" is not designed into the circuit from the start. On the other hand, being too conservative in layout, or trying to integrate "too much system" can result in large dies that substantially affect yields (see box).

Choosing a Foundry

When selecting a "silicon foundry" one must be sure the manufacturer is in the customer-owned tooling business to stay. Many companies will gear up foundry operations in slack times, when excess capacity is available, only to snuff them out as soon as orders pick up and standard products recapture the capacity. To avoid delays due to insufficient data, a foundry should have a good set of documentation outlining its requirements. Flexibility is also important. Although changes such as additional masking steps require a substantial amount of engineering for a fab line, most foundries are willing to vary threshold voltages for the sake of optimal circuit performance.

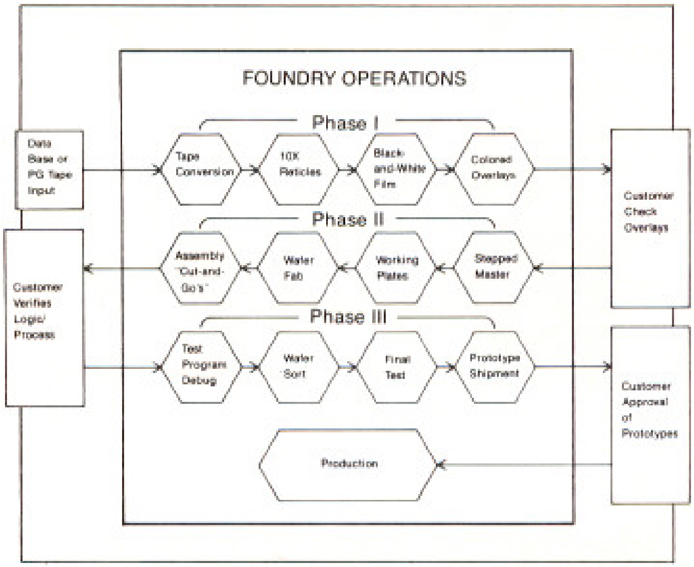


FIGURE 1. Typical operations performed by a semiconductor company that offers customer-owned tooling services.

The Development Cycle

Figure 1 outlines a typical set of operations performed by a customer-owned tooling operation during the development cycle. This chart shows where the lines of responsibility are drawn between the designer and the foundry.

AMI separates the development cycle into three distinct phases:

- Phase I: from receipt of pattern-generator (PG) tape to shipment of colored overlays
- Phase II: from overlay approval to shipment of wafers or untested assemblies
- Phase III: from approval of assemblies to prototype shipment (Parts are tested with a fully debugged test program.)

Each phase ends with some type of verification or approval required from the designer. For Phase I, by far the easiest and most common interface for both the designer and the foundry is a pattern-generator (PG) tape. This is because three different tape formats dominate the industry: Mebes (e-beam standard), David Mann, or Electromask. By interfacing at this point in the design, most foundries can afford to develop and maintain

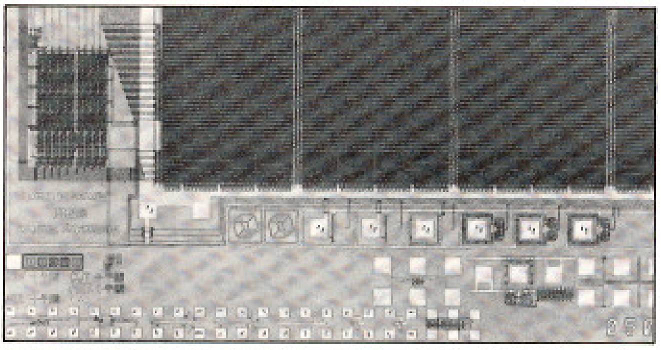


FIGURE 2. A "test strip" which contains the critical dimension should be included on each die in a multi-project wafer.

conversion programs that adapt the input format to the format required by their own pattern generator. In this case, the advantage of a PG tape over a data-base tape (Calma or Applicon, for example) is clear. Although a data base tape must be input if computerized design-rule checks are required, it is generally not preferred because of the many ways in which a circuit can be formatted on the tape. This circumstance prevents a foundry from converting all types of data-base tapes, because the expense of creating and maintaining the required conversion programs is not justified.

Another common interface point into the silicon foundry is the working plate. This input is widely used for the multi-chip approach, because of the complexity of merging several different designs on one wafer. Because most foundry photo-shops are production oriented, they often cannot merge designs effectively. This exercise is usually left to mask houses, or to the "silicon brokers" who specialize in this activity.

The working-plate medium is also an excellent tool for determining process compatibility. For a relatively low price, a single run of perhaps ten wafers can be fabricated from customer-supplied working plates, with virtually no expenditure on engineering resources (which would have to be recovered from a production-volume commitment) by the foundry. Once a circuit is destined for volume production, the working-plate input is no longer preferred, because semiconductor manufacturers usually require internal control over tooling. There are two reasons for this: 1) the foundry will be assuming responsibility for the yield once production begins on a piece-part basis, therefore, it will want control over the quality of the working plate used; and 2) the availability of internal tooling will help avoid disrupting the production cycle when reprints are needed.

Once the PG tape is received, the conversion routine mentioned above is performed, and the tape is sent to the photoshop. Ten-power (10X) reticles are generated, and colored overlays are produced from them. A copy at a previously agreed upon scale is sent to the designer for a level-to-level check, to make sure that the data contained on the tape has been correctly reproduced on the reticles.

During this cycle, the manufacturer must ensure that the geometries present on the reticles are the exact size required to give the designer the final etched dimensions desired on the wafer. Because the photolithography area must develop the reticles in a manner similar to the development of film, the size of these critical dimensions can vary due to the length of the etching cycle. Most foundries require a pictorial representation of the circuit, showing a spot in which the smallest (*i.e.*, most

The Effect of Die Size on Yield

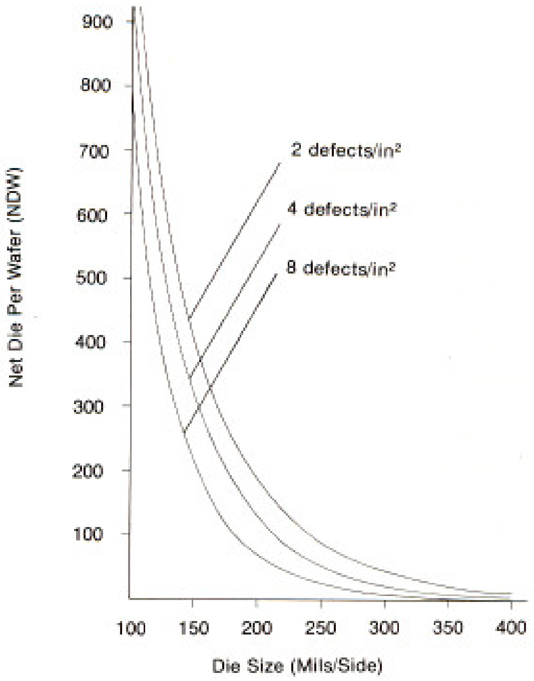
The trade-off between die size and projected yield is an important consideration when partitioning your system for implementation on a custom chip. In their zest to integrate a maximum portion of a system on a chip, many designers overlook the consequences of building a chip too large to be processed and still provide reasonable wafer-sort yields. The accompanying graph shows the effect of increased die sizes on projected wafer sort yields. AMI uses the following equation to calculate project yields:

$$\mathrm{ND/W} = \frac{\pi(r - A)^2}{A(1 + AD)^n}$$

ND/W = net die per wafer
r = radius of the wafer in inches
A = area of the die in square inches
D = defects per square inch per level
n = number of critical layers in the process

NOTE: The term $\frac{\pi(r - A)^2}{A}$ = gross die per wafer

The graph was drawn for a typical 5-micron silicon-gate nMOS process using 4-inch wafers, assuming five critical mask levels (diffusion, buried contact, polysilicon, contact cut, and metal). A family of curves is shown, assuming a defect rate of 2, 4, and 8 defects per square inch. As an example, the projected yield for a die 200 mils on a side (assuming four defects per square inch) is 121 net die per wafer. Using the same criteria, a 300-mil-per-side die would yield only 22 die. In other words, slightly more than doubling the area causes a nearly six-fold decrease in the projected number of die per wafer! In view of the additive effects of defects during working-plate and fabrication operations, most manufacturers would be satisfied with defect density in the range of 6 to 12 defects per square inch. It is clearly in designers' best interests to keep die sizes within reasonable limits. It is not unusual for die sizes in the range of 300 to 400 mils per side to have a projected yield of one die per wafer.



Die size has a dramatic effect on projected yields.

critical) geometry appears on the circuit. This technique aids the photo area in tooling generation, and also helps the fabrication area maintain precise control over geometries. One problem that can arise in multi-chip projects is that the same critical dimension is not present on all of the dies, because the circuits are different. Having a fabrication person search for the die that contains the dimension to be measured is far too cumbersome; therefore, the solution lies in ensuring that the standard "test strip" present on each die contains the correct critical dimension. Figure 2 illustrates this technique. The chip contains a standard set of test structures from which the process parameters and critical dimensions can be obtained.

Tooling and Wafer Fab

Phase II begins the actual tooling and wafer fabrication operations of the development cycle. As shown further in Figure 1, the 10X reticles are optically reduced to 1X, and a mirror-imaged stepped master is produced. This master is then used to print working plates for contact-printing as in normal production.

The technique used for advanced projection printing is different from that used in contact printing in that the actual master is used for the working plate. This more expensive plate can be used for production because the tooling never comes into contact with the wafer during processing; thus, it never wears out. The difference in cost between a master plate and a reprinted working plate is typically one order of magnitude ($500 versus $50). Because a normal chrome working plate is good for only about 75 exposures before high defect counts make it useless, the expense of using masters in a contact-printing process is prohibitively high.

Once the tooling has passed through a critical quality inspection, the wafer fabrication cycle begins. On completion, parametric data is taken from each wafer. This information can be obtained either from a test device located within each die, or from a much more elaborate standard monitor stepped into each wafer. Depending on the level of control desired, two to five of these process-control monitors (PCMs) can be inserted. More control is generally desirable for multi-chip projects, because there may be only 5 to 10 sites for each circuit on the wafer. However, customers must be prepared to pay a premium for wafers in which the manufacturer's acceptance criteria are upgraded. One of the extremely attractive points of semiconductor technology is evident here: the fact that processing is completely independent from design. Data taken from these standard test inserts can determine the "goodness" of a wafer, which in turn helps define the point at which the responsibility for the device is returned to the designer. Before ordering wafers from an outside source, designers should clearly define the exact acceptance and rejection criteria for their wafers, in terms of the parametric limits, the percentage of data that must be good before a wafer will be accepted, and the amount of optical inspection required.

Defining these items early in negotiations can prevent the friction that might occur if a designer's circuit does not work and if the foundry believes that it has met the required process parameters. If it is found that the wafer run has been correctly processed, wafers can be delivered to the customer, or else the foundry can provide assemblies in the form of "cut-and-go's"-—untested, unsealed assemblies consisting of dies (randomly selected from the center of a wafer) that meet the acceptance criteria. These cut-and-go's are ordinarily used to verify correct logic implementation, or to verify that the process does help the design reach its expected performance.

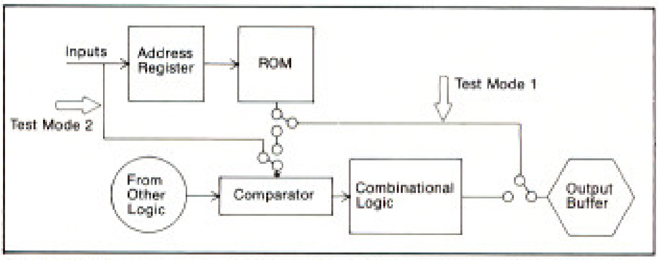


FIGURE 3. The addition of extra pads or extra logic for access to internal nodes can greatly reduce testing time.

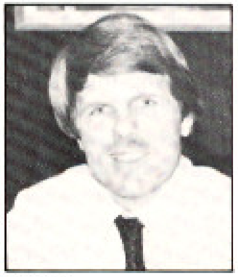
Debugging and Testing

Once the designers are satisfied that the chip performs correctly, they move into Phase III: debugging the test program and proving to the manufacturing facility that the chip is a viable, reproducible product. The test program is probably the most frequently neglected part of the design process and causes more delays in putting the product into production than does any other single factor—undoubtedly because it is not needed to determine that the design works. It is very important to design the chip so that it can be thoroughly tested as fast as possible. Because test time (including overhead) can cost about $.06 per second, a 45-second test time could add nearly $3 to the cost of a part. This cost is unacceptable; therefore, alternatives must be used that reduce testing time.

The conventional approach to reducing test time has been to overdesign device sizes to allow testing at higher speeds. Although this is always a viable approach, the test-time reduction usually does not justify the increased die size. Figure 3 shows a more acceptable approach—one that uses extra logic to provide access to internal nodes. By routing the ROM output to existing output buffers (Test Mode 1), the ROM content can be checked quickly, without going through the combinational logic that would entail a lengthy test time. This combinational logic can then be checked by artifically setting logic levels via access to an input (Test Mode 2). The insertion of extra logic would result in a minimal increase in die size, and a minimal yield loss; but it would drastically reduce the total testing time. This simple example emphasizes the importance of considering the test plan before the chip is laid out.

Depending on the process, Phase I and Phase II together will take from four to twelve weeks (assuming that the foundry has the inputs it needs to eliminate costly misunderstandings). Phase III depends on the soundness of the test program, and on whether yields are adequate to make the part profitable for manufacturers to produce.

About the Author



Stephen E. McMinn received the EE degree from Louisiana State University at Baton Rouge (1977), and the MBA degree from the University of Santa Clara (1981). He spent eighteen months as a product engineer in the discrete bipolar transistor division of Fairchild before joining AMI in 1978 as a development engineer in the Customer-Owned Tooling Department. He is now the department manager for all of the engineering development of the Customer-Owned Tooling Groups.