Chapter 3

Fabrication Process and its Impact on Physical Design

The biggest driving force behind growth in the semiconductor, computer, networking, consumer electronics, and software industries in the last half century has been the continuous scaling, or miniaturization, of the transistor. Computers and other electronic devices have become smaller, more portable, cheaper, easier to use, and more accessible to everyone. As long as we can make the transistor faster and smaller, make the wires that interconnect them less resistive to electrical current, and make each chip denser, the digital revolution will continue.

The manufacture of ICs, like any other high volume manufacturing business, is very cost sensitive. The yield of the fab must be very high to be profitable. So in any given process generation, semiconductor manufacturers use process equipment and critical dimensions that allow them acceptable yields. As more and more chips are manufactured and tested in a process, the process matures and the yield of the process increases. When the yield increases, more aggressive (that is, smaller) critical dimensions can be used to shrink the layout. This process of shrinking the layout, in which every dimension is reduced by a factor is called *scaling*. In general, scaling refers to making the transistors, and the technology that connects them, smaller. As a transistor becomes smaller, it becomes faster, conducts more electricity, and uses less power, the cost of producing each transistor goes down, and more of them can be packed on a chip.

If a chip is scaled, it leads to a smaller die size, increased yield, and increased performance. Suppose a chip in 0.25 micron process generation is x microns wide and x microns high. Assume a shrink factor of 0.7 from 0.25 to 0.18 micron process. Therefore, on 0.18 process, we can essentially produce a $0.7x$ micron wide and $0.7x$ micron high chip. That is, the scaled chip is half the size of the original chip, in terms of area. It will have better yield, due to smaller die size and it will have better performance due to faster transistors and shorter interconnect.

As transistors are scaled, their characteristics (such a delay, leakage current, threshold voltage, etc) also scale but not uniformly. For example, power may not scale with device size. In particular, by middle of next decade, it is expected that the leakage current of a transistor will be of the same value whether the transistor is on or off.

The biggest concern in scaling, is the mismatch in the scaling of devices and interconnect. Interconnect delay is not scaling at the same rate as the device delay. As a result, it has become a more dominant factor in overall delay. This has fundamentally changed the perspective of physical design. Earlier it was possible to lay down the devices and interconnect them and be sure that the design would work. It was possible to ignore the delay in the interconnect, as it was 5 – 10% of the overall delay. As interconnect has now become 50 – 70% of the overall delay, it was necessary to find optimal locations for devices so that interconnect delay is as small as possible. This has led to the consideration of physical design (in particular interconnect planning) at very early stages of design (even at architectural level), as well as throughout the VLSI design cycle.

As a result of the potential side effects, it is important to be aware of process technology and innovations, so as to understand the impact on physical design. The purpose of this chapter is to explain the process scaling, process innovations, process side-effects and their impact on physical design. Our focus is to identify potential future physical design problems due to process. Section 3.1 discusses the scaling methods, Section 3.2 presents the status of fabrication process (circa 1998). Section 3.3 is dedicated to issues with the process such as the parasitics effects, interconnect delay and noise, power dissipation, and yield, among others. In Section 3.4, we discuss the future of the process and innovations that might solve the current and future process related problems. Section 3.5 discusses the solutions and options for interconnect problems. Finally, Section 3.6 discusses CAD tools needed to help in process development.

3.1 Scaling Methods

There are two basic types of scaling to consider. One is *full scaling,* in which all device dimensions, both surface and vertical, and all voltages are reduced by the same factor. Other type is called the *constant-voltage scaling,* wherein only device dimensions are scaled, while maintaining voltage levels. These two methods are compared in Table 3.1, where the scaling factor is specified as S.

In full scaling, devices improve in speed, dissipate less power and can be packed closer together. As a result, high speed chips that integrate large numbers of transistors are made possible. However, devices cannot be scaled down beyond a certain limit. This limit is imposed by a number of second order effects which arise due to simple scaling technique.

In constant voltage scaling, ICs can be used in existing systems without multiple power supplies. In addition, the gate delay is reduced by an additional factor of S. On the other hand, constant voltage scaling leads to higher power

Parameter	Full scaling	CV scaling
Dimensions: width, length, oxide thickness	$1/S$	$1/S$
Voltages: Power, threshold	$1/S$	1
Gate capacitance	$1/S$	$1/S$
Current	$1/S$	S
Propagation delay	$1/S$	$1/S^2$

Table 3.1: Scaling effect on device parameters.

dissipation density, and increments in electric fields, which may lead to oxide breakdown problems.

Except for features related to bonding pads and scribe lines, all other features can be scaled. If the design is not limited by the pad-size, the layout can be scaled and then pads of original size can be placed around the shrunken layout.

3.2 Status of Fabrication Process

In 1998, the standard production fabrication process is the 0.25 micron CMOS process. It has a 1.8V VDD and 4.2 nm oxide. Transistors use complementary doped poly and Shallow Trench Isolation (STI). In terms of interconnect, process supports 5-6 layers of aluminum interconnect using (typically) a Ti/Al-Cu/Ti/TiN stack. Some manufacturers provide a local interconnect layer as well. The key feature of interconnect is the high aspect ratio metal lines for improved resistance and electro-migration. Contact and via layers are filled with tungsten plugs and planarized by CMP. ILD layers are also planarized by CMP. New processes support copper layers for interconnect.

3.2.1 Comparison of Fabrication Processes

In this section, we compare 0.25 micron processes of leading manufacturers. Table 3.2 shows the key features of production fabrication processes of five leading semiconductor manufacturers. The five processes are listed from IBM (International Business Machines), AMD (Advanced Micro Devices), DEC (Digital Equipment Corporation, which is now part of Intel and Compaq), TI (Texas Instruments) and Intel Corporation.

Several interesting observations can be made about the processes.

1. Synergy between Metal Layers: Some manufacturers have syngerized the metal layers. For example, DEC's CMOS-7 process has (1:2) ratio between two lower metal layers and three higher metal layers. Similarly, TI's CO7 has (1:3) ratio between first four layers and the last layer. While Intel's process favors syngergy between M2 and M3 and there is

Company	IBM	AMD	DEC	TI	Intel
Process	CMOS-6x	CS-44	CMOS-7	C07	P856
No. of metal layers	6	5	6	5	5
M0	Yes	Yes	No	No	No
Stacked Vias	Yes	Yes	Yes	Yes	Yes
Voltage	1.8V	2.5V	1.8V	1.8V	1.8V
M1	0.7	0.88	0.84	0.85	0.64
M2	0.9	0.88	0.84	0.85	0.93
M3	0.9	0.88	1.7	0.85	0.93
M4	0.9	1.13	1.7	0.85	1.6
M5	0.9	3.0	1.7	2.5	2.56

Table 3.2: Comparison of Fabrication Processes.

no obvious relationship between these layers and higher layers. The other other extreme is IBM's essentially gridded approach for four higher layers. Only M1 is os smaller pitch, possibly to help cell density. Metal synergy may help in routing, as routes are on certain tracks and vias can be only on the routing grids. The importance of routing grid is related to methodology for wire sizing. In particular, the method, by which routers provide wide wire and tapering capability.

2. Local interconnect: While some manufacturers have provided local interconnect for better cell density (50%), other have opted to forego it since it may cost almost as much as a full layer but provides limited routing capability.

3. Use of higher metal layers: Table 3.2 clearly shows a divergence in the width of the higher metal layers. Note that AMD, TI, and Intel use very wide wires on higher metal layers. These layers provide high performance for long global interconnect. While IBM and DEC provide smaller width lines, possibly leaving the option for the designer to use a wider than minimum width wire, if necessary. However, this does allow process to provide optimal performance from wires in higher layers.

Table 3.3 shows details about the spacing between interconnect and aspect ratio of the interconnect for Intel's P856 process. First note that the thickness of metal lines varies widely. M1 is quite narrow at 0.48 um, while M5 is quite thick at 1.9 um. Aspect ratio of a wire is the ratio of its thickness to its width. Note that the interconnect aspect ratio is almost as high as 2.0 for some layers. That is, M2 (and M3) is twice as thick as it is wide. Higher aspect ratio provides better interconnect performance on smaller widths but also introduces the wall-to-wall capacitance.

Layer	Pitch(um)	Thickness(um)	Aspect Ratio
M1	0.64	0.48	1.5
M2	0.93	0.89	1.9
M3	0.93	0.89	1.9
M4	1.60	1.33	1.7
M5	2.56	1.90	1.5

Table 3.3: Intel's P856 Interconnect dimensions

3.3 Issues related to the Fabrication Process

Process scaling has allowed a high level of integration, better yields (for a constant die size), lower costs and allowed larger die sizes (for a constant yield). As a result, process has also introduced several problems and issues that need to be addressed. In this section, we will discuss these process related issues.

The first set of issues is related to parasitics effects, such as stray capacitances. The second set of issues is related to interconnect, which poses two type of problems, delay/noise and signal integrity which may limit maximum frequency. Other interconnect problems are associated with size and complexity of interconnect. The amount of interconnect needed to connect millions of transistors runs into hundreds of thousands of global signals and millions of local and mid-distance signals. The sheer size of interconnect needs to be addressed, otherwise the die size grows to accommodate the interconnect. This larger die size may make the design project more costly or even infeasible. This is due to that fact that larger die may have longer interconnect and may not allow the chip to reach its targeted frequency. Other issues include power dissipation and yield.

3.3.1 Parasitic Effects

The proximity of circuit elements in VLSI allows the inter-component capacitances to play a major role in the performance of these circuits. The stray capacitance and the capacitance between the signal paths and ground are two major parasitic capacitances. Another parasitic capacitance is the inherent capacitance of the MOS transistor. This capacitance has to be accounted for, as it has more effect than the parasitic capacitance to the ground. All MOS transistors have a parasitic capacitance between the drain edge of the gate and drain node. In an inverter, this capacitance will be charged in one direction for one polarity input and in the opposite direction for the opposite polarity input. Thus, on a gross scale its effect on the system is twice that of an equivalent parasitic capacitance to ground. Therefore, gate-to-drain capacitances should be doubled, and added to the gate capacitance and the stray capacitances, to account for the total capacitance of the node and thus for the effective delay of the inverter.

Interconnect capacitance is of two types; between wires across layers and between wires within layers. The former is more significant than the later. The interconnect capacitance within layers can be reduced by increasing the wire spacing and by using power lines shielding. Whereas, the interconnect capacitance across layers can be reduced by connecting wires in adjacent layers perpendicular to each other.

3.3.2 Interconnect Delay

The calculation of delay in a layout depends on a variety of parameters. The delays in a circuit can be classified as gate delay and interconnect delay. Both the gate and the interconnect delay depend on parameters such as the width and length of poly, thickness of oxide, width and length of metal lines, etc. Details on this topic can be found in [Bak90]. The process of extracting these parameters is called *extraction.* The tool that computes the delay using these parameters and a suitable delay model is often referred as an *RC-extractor.* We will restrict this discussion to the calculation of interconnect delays.

Historically, interconnect delay was considered to be electrically negligible. Currently interconnections are becoming a major concern in high performance ICs and the RC delay due to interconnect is the key factor in determining the performance of a chip. The resistance of wires increases rapidly as chip size grows larger and minimum feature size reduces. Resistance plays a vital role in determining RC delay of the interconnection.

The relative resistance values of metal, diffusion, poly, and drain-to-source paths of transistors are quite different. Diffused layers and polysilicon layers have more than one hundred times the resistance per square area of the metal layer. The resistance of a uniform slab of conducting material is given by:

$$R = \frac{\rho l_c}{h_c w_c}$$

where ρ is the resistivity, and w_c, h_c, and l_c are the width, thickness, and length of the conductor. The empirical formula for the interconnection capacitance is given by:

$$\begin{aligned} C = & \left[1.15\left(\tfrac{w_c}{t_o}\right) + 2.80\left(\tfrac{h_c}{t_o}\right)^{0.222}\right. \\ & \left. + \left[0.06\left(\tfrac{w_c}{t_o}\right) + 1.66\left(\tfrac{h_c}{t_o}\right) - 0.14\left(\tfrac{h_c}{t_o}\right)^{0.222}\right]\left(\tfrac{t_o}{w_{ic}}\right)^{1.34}\right] \epsilon_s\, \epsilon_o\, l_c \end{aligned}$$

where, C is the capacitance of the conductor, w_{ic} is the spacing of chip interconnections, t_o is the thickness of the oxide, $\epsilon_o = 3.9$ is the dielectric constant of the insulator, and $\epsilon_s = 8.854 \times 10^{-14}\ F/cm$ is the permittivity of free space. Various analytical models for two-dimensional interconnection capacitance calculations can be found in [Cha76, ST83]. Path capacitance could be computed by adding via capacitances to the net capacitances.

The expressions given above show that the interconnect delay is the more dominant delay in current technology. Consider a 2 cm long, 0.5 μm thick wire, having a 1.0 μm thick oxide beneath it in an chip fabricated using 2 μm technology. The resistance of such a wire is 600 Ω and its capacitance is approximately 4.0 pF. As a result, it has a distributed RC constant of 2.4 nsec. This delay is equivalent to a typical gate delay of 1 to 2 nsec in 2 μm technology. In this technology, the maximum die size was limited to 1 cm × 1 cm and therefore the gate delays dominated the interconnect delays. On the other hand, a similar calculation for 0.5-0.7 μm technology shows that if only sub-nanosecond delays are allowed, the maximum wire length can be at most 0.5 cm. Since, the gate delays are sub-nanosecond and the die size is 2.5 cm × 2.5 cm, it is clear that interconnect delays started dominating the gate delay in 0.5 μm process generation.

The delay problem is more significant for signals which are global in nature, such as clock signals. A detailed analysis of delay for clock lines is presented in Chapter 11.

3.3.3 Noise and Crosstalk

When feature sizes and signal magnitudes are reduced, circuits become more susceptible to outside disturbances, resulting in *noise.* Noise principally stems from resistive and capacitive coupling. Smaller feature sizes result in reduced node capacitances. This helps to improve circuit delays; however, these nodes also become more vulnerable to external noise, especially if they are dynamically charged. The coupling between neighboring circuits and interconnections and the inductive noise generated by simultaneous switching of circuits are most prevalent forms of internal noise. As chip dimensions and clock frequency increase, the wavelengths of the signals become comparable to interconnection lengths, and this makes interconnections better 'antennas.'

Noise generated by off-chip drivers and on-chip circuitry is a major problem in package and IC design for high-speed systems. The noise performance of a VLSI chip has three variables: noise margin, impedance levels, and characteristics of noise sources. Noise margin is a parameter closely related to the input-output voltage characteristics. This is a measure of the allowable noise voltage in the input of a gate such that the output will not be affected. Noise margin is defined in terms of two parameters: Low Noise Margin(LNM) and High Noise Margin(HNM). The LNM and HNM are given by:

$$LNM = \max(V_{IL}) - \max(V_{OL})$$

$$HNM = \min(V_{OH}) - \min(V_{IH})$$

Where V_{IL} and V_{IH} are low and high input voltages and V_{OL} and V_{OH} are low and high output voltages respectively.

One of the forms of noise is *crosstalk,* which is a result of mutual capacitance and inductance between neighboring lines. The amount of coupling between two lines depends on these factors. The closeness of lines, how far they are from the ground plane, and the distance they run close to each other. As a result of crosstalk, propagation delay increases and logic faults occur. The delay increases because the overall capacitance of the line increases which in turn augments the RC delay of the line.

3.3.4 Interconnect Size and Complexity

The number of nets on a chip increases as the number of transistors are increased. Rent's rule is typically used to estimated the number of pins in a block (unit, cluster, or chip) and number of transistors in that block (unit, cluster or chip). Rent's rule state that the number of I/Os needed are proportional to the number of transistors *N* and a constant *K,* which depends on the ability to share signals. Rent's rule is expressed as:

$$C = KN^n$$

where *C* is the average number of signal and control I/Os. *K* is typically 2.5 for high performance systems, and *n* is a constant in the range of 1.5 to 3.0. Originally, Rent's rule was observed by plotting I/Os versus transistors count of real systems. Since that time, several stochastic and geometric arguments have also been proposed that support Rent's rule.

It is quite clear from Rent's rule that the signal complexity at all levels of the chip stays ahead of the integration level (number of transistors, sub-circuits, etc).

3.3.5 Other Issues in Interconnect

Several other issues in interconnect may also cause some problems. Higher aspect ratio wires that are used to provide better performance for higher layers also cause more cross (wall to wall) capacitance. As a result, signals that may conflict need to be routed spaced from each other. Another issue is inductance modeling and design. As chip frequency reaches GHz and beyond, wires start acting like transmission lines and circuits behave like RLC circuits. In addition, use of different dielectrics, that are used on different layers complicates the delay, noise and inductance modeling.

3.3.6 Power Dissipation

Heat generation and its removal from a VLSI chip is a very serious concern. Heat generated is removed from the chip by heat transfer. The heat sources in a chip are the individual transistors. The heat removal system must be efficient and must keep the junction temperature below a certain value, which is determined by reliability constraints. With higher levels of integration, more and more transistors are being packed into smaller and smaller areas. As a result, for high levels of integration heat removal may become the dominant design factor. If all the generated heat is not removed, the chip temperature will rise and the chip may have a thermal breakdown. In addition, chips must be designed to avoid *hotspots,* that is, the temperature must be as uniform as possible over the entire chip surface.

CMOS technology is known for using low power at low frequency with high integration density. There are two main components that determine the power dissipation of a CMOS gate. The first component is the static power dissipation due to leakage current and the second component is dynamic power dissipation due to switching transient current and charging/discharging of load capacitances. A CMOS gate uses 0.003 mW/MHz/gate in 'off' state and 0.8 mW/MHz/gate during its operation. It is easy to see that with one million gates, the chip will produce less than a watt of heat. In order to accurately determine the heat produced in a chip, one must determine the power dissipated by the number of gates and the number of off chip drivers and receivers. For CMOS circuits, one must also determine the average percent of gates active at any given time, since heat generation in the 'off' state is different than that of 'on' state. In ECL systems, power consumption is typically 25 mW/gate irrespective of state and operating frequency. Current heat removal system can easily handle 25 to 100W in a high performance package. Recently, MCM systems have developed, which can dissipate as much as 600W per module.

Power dissipation has become a topic of intense research and development. A major reason is the development of lap-top computers. In lap-top computers, the battery life is limited and low power systems are required. Another reason is the development of massively parallel computers, where hundreds (or even thousands) of microprocessors are used. In such systems, power dissipation and corresponding heat removal can become a major concern if each chip dissipates a large amount of power.

In recent years, significant progress has been in made in development of low power circuits and several research projects have now demonstrated practical lower power chips operating at 0.5 V. In some microprocessors, 25-35% power is dissipated in the clock circuitry, so low power dissipation can be achieved by literally 'switching-off' blocks which are not needed for computation in any particular step.

3.3.7 Yield and Fabrication Costs

The cost of fabrication depends on the yield. Fabrication is a very complicated combination of chemical, mechanical and electrical processes. Fabrication process requires very strict tolerances and as a result, it is very rare that all the chips on a wafer are correctly fabricated. In fact, sometimes an entire wafer may turn out to be non-functional. If a fabrication process is new, its yield is typically low and design rules are relaxed to improve yield. As the fabrication process matures, design rules are also improved, which leads to higher density. In order to ensure that a certain number of chips per wafer will work, an upper limit is imposed on the chip dimension *X*. Technically speaking, an entire wafer can be a chip (wafer scale integration). The yield of such a process would however be very low, in fact it might be very close to zero.

Wafer yield accounts for wafers that are completely bad and need not be tested. The prediction of the number of good chips per wafer can be made on the basis of how many dies (chips) fit into a wafer (N_d) and the probability of a die being functional after processing (Y). The cost of an untested die C_{ud} is given by

$$C_{ud} = \frac{C_w}{N_d * Y}$$

where, C_w is the cost of wafer fabrication. The number of dies per wafer depends on wafer diameter and the maximum dimension of the chip. It should be noted that product $N_d * Y$ is equal to total number of "good" dies per wafer N_y. The number of dies of a wafer N_d is given by

$$N_d = \pi \frac{(D - \alpha)^2}{4X^2}$$

where D is the diameter of the wafer (usually 10 cm), and α is the useless scrap edge width of a wafer (mm). The yield is given by:

$$Y = (1 - A\delta/c)^c$$

where, A is the area of a single chip, δ is the defect density, that is, the *defects per square millimeter,* and c is a parameter that indicates defect clustering.

The cost of packaging depends on the material used, the number of pins, and the die area. The ability to dissipate the power generated by the die is the main factor which determines the cost of material used.

Die size depends on technology and gates required by the function and maximum number of dies on the chip, but it is also limited by the number of pins that can be placed on the border of a square die.

The number of gates N_g in a single IC is given by:

$$N_g = \frac{(X^2 - P * A_{io})}{A_g}$$

where, P is the total number of pads on the chip surface, A_{io} is the area of an I/O cell and A_g is the area of a logic gate. It should be noted that a gate is a group of transistors and depending on the architecture and technology, the number of transistors required to form a gate will vary. However, on the average there are 3 to 4 transistors per gate.

The number of pads required to connect the chip to the next level of interconnect, assuming that pads are only located at the periphery of the chip is

$$P = 4(X/S - 1)$$

where, S is the minimum pad to pad pitch.

An optimal design should have the necessary number of gates well distributed about the chip's surface. In addition, it must have minimum size, so as to improve yield.

The total fabrication cost for a VLSI chip includes costs for wafer preparation, lithography, die packaging and part testing. Again, two key factors determine the cost: the maturity of process and the size of the die. If the process is new, it is likely to have a low yield. As a result, price per chip will be higher. Similarly, if a chip has a large size, the yield is likely to be low due to uniform defect density, with corresponding increase in cost.

Fabrication facilities can be classified into three categories, prototyping facilities (fewer than 100 dies per lot), moderate size facilities (1,000 to 20,000 dies per lot) and large scale facilities, geared towards manufacturing 100,000+ parts per lot.

Prototyping facilities such as MOSIS, cost about $150 for tiny chips (2.22 mm × 2.26 mm) in a lot of 4, using 2.0 μm CMOS process. For bigger die sizes (7.9 mm × 9.2 mm) the cost is around $380 per die, including packaging. The process is significantly more expensive for 0.8 μm CMOS. The total cost for this process is around $600 per square mm. MOSIS accepts designs in CIF and GDS-II among other formats, and layouts may be submitted via electronic mail. For moderate lot sizes (1,000 to 20,000) the cost for tiny chips (2 mm × 2 mm) is about $27 in a lot size of 1,000. For large chips (7 mm × 7 mm), the cost averages $65 per chip in a lot of 1000. In addition, there is a lithography charge of $30,000 and a charge of $5,000 for second poly (if needed). These estimates are based on a 1.2 μm, two metal, single poly CMOS process and include the cost of packaging using a 132 Pin Grid Array (PGA) and the cost of testing. The large scale facilities are usually inexpensive and cost between $5 to $20 per part depending on yield and die size. The costs are included here to give the reader a real world perspective of VLSI fabrication. These cost estimates should not be used for actual budget analysis.

3.4 Future of Fabrication Process

In this section, we discuss the projected future for fabrication process. We also discuss several innovations in lithography, interconnect and devices.

3.4.1 SIA Roadmap

Fabrication process is very costly to develop and deploy. A production fab costs upwards of two billion dollars (circa 1998). In the early 1990's, it became clear that process innovations would require joint collaboration and innovations from all the semiconductor manufacturers. This was partly due to the cost of the process equipment and partly due to the long time it takes to innovate, complete research and development and use the developed equipment or methodologies in a real fab. Semiconductor Industry Association (SIA) and SRC started several projects to further research and development in fabrication process.

In 1994, SIA started publishing the National Technology Roadmap for Semiconductors. The roadmap provides a vision for process future. In 1997, it was revised and key features are listed in table 3.4. (note that 1000 nanometers = 1 micron).

Feature Size (nm)	250	180	130	100
Time Frame	1997	1999	2003	2006
Logic Transistors per area (Millions/sq. cm.)	3.7	6.2	18	39
Chip Frequency (MHz): Cost/Perf designs	> 300	> 400	> 500	> 650
Chip Frequency (MHz): High Perf designs	> 500	> 750	> 1100	> 1500
Chip Size (sq. mm)	300	360	430	520
Wiring Levels	6	6-7	7	7-8
Package Pins per Chip	512	512	768	768
Power Supply Voltage (desktop)	2.5	1.8	1.5	1.2
Power Supply Voltage (portable)	1.8-2.5	0.9-1.8	0.9	0.9
Interconnect	planar	planar	planar	planar
Min. Interconnect CD (nm)	250	180	130	100
Min. Contacted Interconnect Pitch (logic) nm	640	460	340	260
Contact/Via Critical Dimension(nm)	280/400	200/280	140/200	110/140
Via Aspect Ratio : Logic	2.0	2.0	2.3	2.7
Metal Aspect Ratio	1.8	1.8	2.1	2.4
Max. Interconnect Length (meters/chip)	820	1480	2840	5140

Table 3.4: Feature Size and Integrated Circuit Capability, 1997-2006. National Technology Roadmap for Semiconductors, 1997

3.4.2 Advances in Lithography

Currently, it is possible to integrate about 20 million transistors on a chip. Advances in X-ray lithography and electron-beam technology indicate that these technologies may soon replace photolithography. For features larger than 0.020 μm, X-ray beam lithography can be used. For smaller devices, a scanning tunneling electron microscope may be used. It has been shown that this technique has the possibility of constructing devices by moving individual atoms. However, such fabrication methods are very slow. Just based on X-ray lithography, there seems to be no fundamental obstacle to the fabrication of one billion transistor integrated circuits. For higher levels of integration there are some limits. These limits are set by practical and fundamental barriers. Consider a MOS transistor with a square gate 0.1 μm on a side and 0.005 μm oxide layer thickness. At 1V voltage, there are only 300 electrons under the gate, therefore, a fluctuation of only 30 electrons changes the voltage by 0.1 V. In addition, the wave nature of electrons poses problems. When device features reach $0.005\ \mu m$, electrons start behaving more like waves than particles. For such small devices, the wave nature as well as the particle nature of electrons has to be taken into account.

3.4.3 Innovations in Interconnect

As discussed earlier, interconnect poses a serious problem as we attempt to achieve higher levels of integration. As a result, several process innovations are targeted towards solution of the interconnect problems, such as delay, noise, and size/complexity.

3.4.3.1 More Layers of Metal

Due to planarization achieved by CMP, from a pure technology point of view, any number of metal layers is possible. Hence, it is purely a cost/benefit tradeoff which limits addition of layers. An increasing number of metal layers has a significant impact on the physical design tools. In particular, large numbers of layers stresses the need for signal planning tools.

3.4.3.2 Local Interconnect

The polysilicon layer used for the gates of the transistor is commonly used as an interconnect layer. However the resistance of doped polysilicon is quite high. If used as a long distance conductor, a polysilicon wire can represent significant delay. One method to improve this, that requires no extra mask levels, is to reduce the polysilicon resistance by combining it with a refractory metal. In this approach a Sillicide (silicon and tantalum) is used as the gate material. The sillicide itself can be used as a local interconnect layer for connections within the cells. Local interconnect allows a direct connection between polysilicon and diffusion, thus alleviating the need for area intensive contacts and metal. Also known as Metal 0, local interconnect is not a true metal layer. Cell layout can be improved by 25 to 50% by using local interconnect. However, CAD tools need to comprehend restrictions to use M0 effectively.

3.4.3.3 Copper Interconnect

Aluminum has long been the conductor of choice for interconnect, but as the chip size shrinks it is contributing to interconnect delay problem, due to its high resistance. Although Copper is a superior conductor of electricity, it was not being used earlier for interconnect because not only does copper rapidly diffuse into silicon, it also changes the electrical properties of silicon in such a way as to prevent the transistors from functioning. However, one by one, the hurdles standing in the way of this technology have been overcome. These ranged from a means of depositing copper on silicon, to the development of an ultra thin barrier to isolate copper wires from silicon. Several manufacturers have introduced a technology that allows chip makers to use copper wires, rather than the traditional aluminum interconnects, to link transistors in chips.

This technique has several advantages. Copper wires conduct electricity with about 40 percent less resistance than aluminum which translates into a speedup of as much as 15 percent in microprocessors that contain copper wires. Copper wires are also far less vulnerable than those made of aluminum to electro-migration, the movement of individual atoms through a wire, caused by high electrical currents, which creates voids and ultimately breaks the wires. Another advantage of copper becomes apparent when interconnect is scaled down. At small dimensions, the conventional aluminum alloys can't conduct electricity well enough, or withstand the higher current densities needed to make these circuits switch faster. Copper also has a significant problem of electro-migration. Without suitable barriers, copper atoms migrate into silicon and corrupt the whole chip.

Some manufacturers claim that the chips using Copper interconnects are less expensive than aluminum versions, partly because copper is slightly cheaper, but mainly because the process is simpler and the machinery needed to make the semiconductors is less expensive. However other chip manufacturing companies are continuing with aluminum for the time being. They believe that the newer dual Damascene process with copper requires newer, and more expensive equipment. The equipment is required to put down the barrier layer - typically tantalum or tantalum nitride - then a copper-seed layer, and the electroplating of the copper fill.

There a several possible impacts of copper on physical design. One impact could be to reduce the impact of interconnect on design. However, most designers feel that copper only postpones the interconnect problem by two years. Copper interconnect also may lead to reduction in total number of repeaters thereby reducing the impact on floorplan and overall convergence flow of the chip.

3.4.3.4 Unlanded Vias

The concept of unlanded via is quite simple. The via enclosures were required since wires were quite narrow and alignment methods were not accurate. For higher layers, where wire widths are wider, it is now possible to have via enclosure within the wire and, as a result, there is no need for an extended landing pad for vias. This simplifies the routing and related algorithms.

3.4.4 Innovations/Issues in Devices

In addition to performance gains in devices due to scaling, several new innovations are in store for devices as well. One notable innovation is Multi-Threshold devices. It is well known that the leakage current is inversely proportionate to the threshold voltage. As operating voltage drops, leakage current (as a ratio to operating voltage) increases. At the same time, lower V_t devices do have better performance, as a result these can be used to improve speed. In

Process Technology	0.1 u
Transistors	200 M
Logic Transistors	40 M
Size	520 mm^2
Clock frequency	2 - 3.5 GHz
Chip I/O's	4,000
Wiring levels (metals)	7 - 8
Voltage	0.9 - 1.2
Power	160 W
Supply current	160 Amps

Table 3.5: Aggressive projections for a chip in 2006.

critical paths, which might be limiting to design frequency, devices of lower V_t can be used. Process allows dual V_t devices. This is accomplished by multiple implant passes to create different implant densities. Most manufacturers plan to provide dual V_t rather than full adjustable multiple V_t, although techniques for such devices are now known.

3.4.5 Aggressive Projections for the Process

Several manufacturers have released roadmaps, which are far more aggressive than the SIA roadmap. In this section, we discuss two such projections.

Several manufacturers have indicated that the frequency projections of SIA are too conservative. Considering that several chips are already available at 600-650 MHz range and some experimental chips have been demonstrated in the one Gigahertz frequency range. It is quite possible that frequencies in the range of 2-3.5 Ghz may be possible by year the 2006. A more aggressive projection based on similar considerations is presented in Table 3.5.

Texas Instruments has recently announced a 0.10 micron process well ahead of the SIA roadmap. Drawn using 0.10-micron rules, the transistors feature an effective channel length of just 0.07-micron and will be able to pack more than 400 million transistors on a single chip, interconnected with up to seven l ayers of wiring. Operating frequencies exceeding 1 gigahertz (GHz), internal voltages as low as 1 V and below. The process uses copper and low *K* dielectrics. TI has also developed a series of ball grid array (BGA) packages that use fine pitch wire bond and flip chip interconnects and have pin counts ranging from 352 to 1300 pins. Packages are capable of high frequency operations in the range of 200 megahertz through more than one gigahertz. Power dissipation in these packages ranges from four watts to 150 watts. TI plans to initiate designs in the new 0.07-micron CMOS process starting in the year 2000, with volume production beginning in 2001.

3.4.6 Other Process Innovations

The slowing rate of progress in CMOS technology has made process technologists investigate its variants. The variants discussed below have little or no impact on physical design. These are being discussed here to provide a perspective on new process developments.

3.4.6.1 Silicon On Insulator

One important variant is Silicon On Insulator (SOI) technology. The key difference in SOI as compared to bulk CMOS process is the wafer preparation. In SOI process, oxygen is implanted on the wafer in very heavy doses, and then the wafer is annealed at a high temperature until a thin layer of SOI film is formed. It has been shown that there is no difference in yield between bulk and SOI wafers. Once the SOI film is made on the wafer, putting the transistor on the SOI film is straightforward. It basically goes through the same process as a similar bulk CMOS wafer. There are minor differences in the details of the process, but it uses the exact same lithography, tool set, and metalization.

There are two key benefits of SOI chips: Performance and power.

1. Performance: SOI-based chips have 20 to 25% cycle time and 25 to 35% improvement over equivalent bulk CMOS technology. This is equivalent to about two years of progress in bulk CMOS technology. The sources of increased SOI performance are elimination of area junction capacitance and elimination of "body effect" in bulk CMOS technology.

2. Low power: The ability of SOI as a low-power source originates from the fact that SOI circuits can operate at low voltage with the same performance as a bulk technology at high voltage.

Recently, fully-functional microprocessors and large static random access memory chips utilizing SOI have been developed. SOI allows designers to achieve a two-year performance gain over conventional bulk silicon technology.

3.4.6.2 Silicon Germaniun

Wireless consumer products have revolutionized the communications marketplace. In order to service this new high-volume market, faster, more powerful integrated circuit chips have been required. For many of these applications, silicon semiconductors have been pushed to the 1 to 2 GHz frequency domain. However, many new RF applications require circuit operation at frequencies up to 30 GHz, a regime well out of the realm of ordinary silicon materials. Compound III-V semiconductors (those made from elements from groups III and V in the periodic table) such as GaAs have been used successfully for such chips. These materials, however, are very expensive. Moreover, silicon technology has long proven to be very high-yielding by comparison to devices made with GaAs materials. So the chip manufacturers have chosen to enhance the performance of silicon to capitalize on the cost advantages and compete with

III-V compound semiconductor products. One such enhancement technique is to add small amounts of germanium to silicon in order to create a new material (SiGe) with very interesting semiconducting properties. Unfortunately, the germanium atom is 4% larger than the silicon atom; as a result, growing this material has been very tricky. A technique of using ultra-high vacuum chemical vapor deposition (UHV-CVD) to grow these films has proven to be key in overcoming these difficulties. The result has been the ability to produce high-performance SiGe heterojunction bipolar transistors. Conventional silicon devices have a fixed band gap of 1.12 eV (electron-Volt) which limits switching speed, compared to III-V compound materials such as GaAs. However, with the addition of germanium to the base region of a bipolar junction transistor (BJT) and grading the Ge concentration across the transistor base region, it is possible to modify the band gap to enhance performance of the silicon transistor. The potential benefits of SiGe technology are significant. One key benefit is that Silicon Germanium chips are made with the same tools as silicon chips. This means millions of dollars won't have to be invested in new semiconductor tools, as is typically the case when a shift is made from one generation of chip technology to the next. A number of circuit designs have been fabricated with SiGe technology in order to demonstrate its capability for RF chips. Among the circuits that have been measured are: voltage-controlled oscillators (VCOs), low-noise amplifiers (LNAs), power amplifiers, mixers, digital delay lines.

3.5 Solutions for Interconnect Issues

In this section, we briefly review possible options for the solution of interconnect problems.

1. **Solutions for Delay and Noise:** Several solutions exist for delay and noise problems. These include use of better interconnect material, such as copper, better planning to reduce long interconnect and use of wire size and repeaters.

   (a) Better interconnect materials: Table 3.6 gives the values of l_c at different w_c for two different materials when the value of R is 1000 Ω. The resistivity values for aluminum and polysilicon are 3 $\mu\Omega$-cm and 1000 $\mu\Omega$-cm respectively. From this table, it is clear that as compared to poly, aluminum is a far superior material for routing. Interconnect delays could also be reduced by using wider wires and inserting buffers. There is a wide range of possible values of polysilicon resistance (as shown above) for different commercial purposes. If a chip is to be run on a variety of fabrication lines, it is desirable for the circuit to be designed so that no appreciable current is drawn through a long thin line of polysilicon.

   As stated earlier, copper is the best option in terms of delay and incorporation into the leading processes.

w_c	Aluminum	Polysilicon
2.00 μm	44	0.13
1.00 μm	11	3.33×10^{-2}
0.75 μm	6.2	1.87×10^{-2}
0.50 μm	2.8	8.33×10^{-3}
0.25 μm	0.69	2.08×10^{-3}

Table 3.6: Comparison of wire length l_c (mm).

(b) Early Planning/Estimation: If interconnects needs are comprehended at an early stage, it is possible to avoid long interconnect and channels. Interconnect planning should plan all major busses and identify areas of congestion.

(c) Interconnect Sizing: Typically minimum size wire for a given layer (especially higher layer) are not capable of meeting the performance needs and they needs to be sized. Lower metal layers need to be sized typically for reliability (current carrying capacity) reasons. Several algorithms have now been proposed to optimally size (or taper) wires. Some router can size and taper on the fly.

(d) Repeater Insertion and design: If interconnect is too long, repeaters have to be inserted to meet delay or signal shape constraints. Typically, repeaters need to be planned, since they need to placed.

(e) Shielding and Cross Talk avoidance: Crosstalk, and hence noise, can be reduced by making sure that no two conflicting lines are laid out parallel for longer than a fixed length. The other method to reduce crosstalk is to place grounded lines between signals. The effect of crosstalk and noise in MCMs is discussed in Chapter 14.

2. **Solutions for size and complexity of Interconnect** There are fundamentally three ways to deal with the sheer amount of interconnect.

(a) More metal layers: Industry has steadily increased number of routing layers. Due to CMP, it is virtually possible to add as many layers as needed. Addition of a new layer is purely a cost/benefit trade-off question and not a technology question. Addition more layers have a negative effect on the lower layers, as vias cause obstacles in lower layers. In general, benefit of each additional layer decreases, as more and more layers are added.

(b) Local interconnect: As pointed out earlier, cell density can be improved by providing local interconnect layer. Local interconnect still remain questionable due to the routing limitations. Many manufacturers prefer to add lower layers for routing.

(c) Metal Synergy: To improve utilization of metal system, metal grids may be aligned and strictly enforced. This will allow maximum number of wires routable in a given die area. However, this scheme is applicable to lower frequency designs, where performance or reliability needs do not dictate wide ranges of wire sizes.

In addition to these ideas, one may consider more exotic ideas:

(a) Pre-Routed Grids: If all the wiring is pre-assigned on a six or eight layer substrate and routing problems is reduced to selection of appropriate tracks (and routes), then we can optimize placement to maximize usage of routing resources. This is akin to programmable MCM discussed in Chapter 14. However, unlike programmable MCM, such a routing grid will be used for planning and detailing the layout and only the used part of the grid will be actually fabricated.

(b) Encoding of signals: We can encode a signal in less number of bits for long busses. This will reduce the number of bits in long busses at the cost of time to transmit the message and overhead for encoding and decoding.

(c) Optical Interconnect: Although not feasible in the near future, optical interconnect may one day provide reliable means for distributing clock, thereby freeing up routing resources.

3.6 Tools for Process Development

In earlier sections of this chapter, we have discussed the impact of process innovations on CAD tools (in particular physical design tools). However, an emerging area of research and development is related to tools that help in design of the process. Given the complexity of choices and range of possibilities that technology offers, process decisions are very complex.

Currently, process decisions are made by a set of experts based on extrapolations from previous process generations. This extra-polation methods has worked up to this point mainly because the technology did not offer too many choices. It is significantly easier to decide if we want to migrate to a four layer process from a three layer process, if the bonding method, type of vias and dielectrics are not changing.

In general, two types of tools are needed.

1. Tools for Interconnect Design: The key decisions that need to be made for any process interconnect include: number of metal layers, line widths, spacing and thickness for each layer, type and thickness of ILD, type of vias, and bonding method. CAD tools are needed that can re-layout (or estimate re-layout of) a design, based on the options listed above. For example, should we have a seven layer process or a better designed six metal layer process can only be answered if a CAD tool can re-layout a target design with six and seven metal layers and compare the results.

2. Tools for Transistor design: Complex parameters are involved in transistor design. Lack of tools to help design the transistor may lead to non-optimal design. CAD tools are needed which can simulate the operation of a transistor by changing the settings of these parameters. Such tools may allow use of less efficient transistors for better area utilization.

3.7 Summary

Physical design is deeply impacted by the fabrication process. For example, chip size is limited by a consideration of yield: the larger the chip, the greater the probability of surface flaws and therefore the slimmer the chances of finding enough good chips per wafer to make batch production profitable. A good deal of physical design effort thus goes into the development of efficient structures that lead to high packing densities. As a result, physical designer must be very educated about the fabrication process. A designer must take into account many factors, including the total number of gates, the maximum size of die, number of I/O pads and power dissipation, in order to develop chips with good yield.

The fabrication process has made tremendous strides in the past few decades and it is continuing to reduce feature size and increase both the chip size and performance. While the SIA roadmap calls for 0.01 micron process in the year 2006, many feel that it may happen much sooner. Such aggressive approaches have kept the semi-conductor industry innovating for the last three decades and promise to continue to motivate them in the next decade.

In this chapter, we have reviewed scaling methods, innovations in fabrications process, parasitic effects, and the future of fabrication process. Interconnect is the most significant problem that needs to be solved. We need faster interconnect and more interconnect to complete all the wiring required by millions of transistors. In this direction, we have noted that innovations like copper interconnect, unlanded vias, and local interconnect have a significant effect on physical design.

3.8 Exercises

1. Given a die of size 25 mm × 25 mm and $\lambda = 0.7\ \mu\mathrm{m}$, estimate the total number of transistors that can be fabricated on the die to form a circuit.

2. Estimate the maximum number of transistors that can be fabricated on a die of size 25 mm × 25 mm when $\lambda = 0.1\ \mu\mathrm{m}$.

3. Estimate the total power required (and therefore heat that needs to be removed) by a maximally packed 19 mm × 23 mm chip in $0.75\ \mu m$ CMOS technology. (Allow 10% area for routing and assume 500 MHz clock frequency).

4. Assuming that the heat removal systems can only remove 80 watts from a 19 mm × 23 mm chip, compute the total number of CMOS transistors possible on such a chip, and compute the value of λ for such a level of integration. (Assume 500 MHz clock frequency.)

5. Assuming a 15 mm × 15 mm chip in 0.25 micron process where interconnect delay is 50% of the total delay, consider a net that traverses the full length of the chip diagonally. What is maximum frequency this chip can operate on ?

6. What will happen to the frequency of the chip in problem 3 if we migrate (process shift) it to 0.18 micron process? (assume 0.7 shrink factor and discuss assumptions made about the bonding pads and scribe lines which do not scale).

7. What will happen to the frequency of the chip in problem 3 if we migrate (process shift) it to 0.18 process?

Bibliographic Notes

The *International Solid-State Circuits Conference* (ISSCC) is the premier conference which deals with new developments in VLSI devices, microprocessors and memories. The *IEEE International Symposium on Circuits and Systems* also includes many papers on VLSI devices in its technical program. The *IEEE Journal of Solid-State Circuits* publishes papers of considerable interest on VLSI devices and fabrication. Microprocessor reports is a valuable source of information about process, comparisons between microprocessors and other related news of the semi-conductor industry. Several companies (such as IBM, DEC, Intel, AMD, TI) have internet sites that provide significant information about their process technology and processors.