Chapter 1

Introduction

1.1 Thesis aims

A study of three-dimensional integrated circuit layout is presented in this dissertation. Three-dimensional integrated circuits are those in which active devices such as transistors are fabricated in each of at least two vertically stacked semiconducting planes. An evaluation of the potential benefits of using three-dimensional integration is carried out. Such benefits include greater layout densities, shorter interconnection lengths and faster circuits. Emphasis is placed on considering the layout methods required to use three-dimensional fabrication techniques and to accrue the above benefits.

Three-dimensional circuits are one of a number of recent developments in integrated circuit construction techniques to have stimulated interest. Other developments include the use of a variety of semiconductor and wiring materials for faster transistors and connections, and enhanced processing techniques to improve yield and reduce transistor geometries. This enables both physically larger and logically more complex circuits to be integrated. These developments do not present any new layout problems. The search for better, more automated layout continues independently of such technological advances. The same is not true in the three-dimensional domain. The juxtaposition of devices in three dimensions presents an inherently different set of connection properties to the two-dimensional case, since a device may now have neighbours above and below in addition to those in the semiconducting plane. Layout in three dimensions requires new techniques to fully exploit this property.

Since three-dimensional fabrication techniques were first demonstrated in the early eighties [Gibbons 80], two areas of research have been explored. The primary research has concentrated on the necessary fabrication techniques. In the main, such work has not considered any of the broader issues of layout. The other research has been of a theoretical nature, using graph theory to put bounds on the expected reduction in interconnection length in three-dimensional layouts. A primary goal of the work described here is to present a quantitative measurement of the efficacy of three-dimensional circuit layout. This involves a comparison of two and three-dimensional layouts for several classes of circuit, and represents an approach somewhere between the two mentioned above. In order to perform such a comparison, three-dimensional layout algorithms have been developed to operate within a framework which allows the definition of a number of topologies. The topologies can be mapped against the constraints which different three-dimensional devices impose, such as the number of device layers and the availability and distribution of connections.

1.2 Thesis structure

The remainder of this chapter introduces two-dimensional chip technology. Beginning with a brief history of key developments in the manufacture of integrated circuits and continuing with the description of a typical two-dimensional fabrication sequence, the chapter concludes with a discussion of some of the recent trends in fabrication technology. This provides a background for the description of three-dimensional fabrication presented later. Chapter two is a discourse on two-dimensional chip design, highlighting the need for abstraction and automation in the design process. Gate array, standard cell and custom design are introduced and compared as possible integration routes. A framework for the classification of elements of design automation is introduced and used to describe a selection of design tools. Chapter two concludes with a case study detailing the design and implementation of a circuit in gate array and standard cell formats.

The remaining chapters are devoted to three-dimensional integration. Chapter three contains details of the techniques enabling three-dimensional devices to be constructed and of some of the devices which have been created. Included at this point is a discussion of the technological difficulties of constructing three-dimensional circuits. Chapter four concerns three-dimensional structures in a more abstract sense. A classification is introduced which maps combinations of the technological constraints of three-dimensional circuits into groups of layout topologies. This enables the effect of a technological constraint to be seen in the composition of the resulting structure. Also included at this point is a review of other work on three-dimensional layout.

Chapter five describes a model for experiments in three-dimensional layout. The model introduces cells of homogeneous size, shape and connection interface. The purpose of modelling layout with such specific cells is to explore the possibilities of connection by direct cell abutment, a popular technique in many two-dimensional layouts. Included at this point is the practical design of cells satisfying the required geometric properties of abutting cells. Chapter six contains details of the construction of a highly configurable experimental layout system based on the above model. Descriptions of the layout algorithms and merit functions which were developed are given, and the difficulties and limitations encountered are discussed.

Chapters seven and eight present the results from a large number of experiments carried out with a wide range of circuits and parameter values. The results are scaled with dimensions extracted from the cells designed earlier. In chapter seven, the configuration options of the system are examined. In chapter eight, the effect of varying the number and composition of layers on the quality of layout is presented. Also included for comparison are results from two-dimensional circuit layouts. Chapter nine contains concluding remarks and suggestions for further research.

1.3 Two-dimensional chip technology

1.3.1 A brief history of integration

The concept of the integrated circuit was first proposed by G.W.A. Dummer in 1952. He imagined a solid block containing layers of insulating, conducting and amplifying material with electrical functions being directly connected by cutting out areas on the various layers. At the time he had no idea how this could be realised. Transistor technology was still in its infancy, the first germanium bipolar transistor working in 1947. A number of advances in materials and processing technology prevailed before the integration imagined by Dummer was achieved. Four key developments were the controlled growth of high purity single crystal semiconductors, the understanding of semiconductor doping by the diffusion of impurities, the development of methods for selectively etching materials and the use of patterned insulating layers to mask the diffusion process.

The first functional integrated circuit was demonstrated and patented by J. Kilby in 1959. It consisted of a slice of single crystal germanium containing a bipolar transistor, a capacitor and three resistors. It demonstrated how several components could be integrated on the same semiconductor, but did not show any satisfactory method of connecting the components, this being achieved by hand using thin gold wires. The planar process, initially developed by J. Hoerni and R. Noyce in 1959 as an improved method of manufacturing discrete silicon transistors, combined the existing diffusion and masking techniques with a method of connection. A final layer of patterned oxide was used as a mask for connections between regions of the transistor and the outside world. Connections could then be formed by the deposition of aluminium in a batch production method. The techniques of the planar process when applied to the manufacture of integrated circuits caused a revolution in the electronics industry.

One of the first commercially available planar integrated circuits was a flip-flop containing four transistors and five resistors. It was one of a family of resistor-transistor logic chips offered by Fairchild in 1961. In 1964 the first linear integrated circuit was developed by R. Wildar, an operational amplifier called the μA702 which contained twelve transistors and five resistors. It was remarkable not only for being the first operational amplifier on a single chip, but also for the ingenuity of the design. Rather than attempting to translate a discrete circuit into silicon, Wildar thought in terms of the properties of the silicon components themselves. Where possible, he used DC biased transistors instead of resistors, and relied on matching component characteristics, only assuming approximate absolute values. Still in manufacture, the μA702 is the longest surviving integrated circuit to date. Transistor-transistor logic emerged as the permanent successor to resistor-transistor logic in 1964 with the advent of the Texas Instruments 5400 digital logic family.

For much of the decade or so between the invention of the transistor and the integrated circuit, research was directed at improving the characteristics and manufacturing techniques of bipolar devices. The perfection of the planar process solved many of the problems, and generated a new impetus in the experimentation with new devices. One such device was the field-effect transistor (FET). The first FETs had an aluminium gate insulated by a silicon dioxide layer from the underlying semiconductor channel, and the term metal-oxide-semiconductor (MOS) or MOS-FET was coined. In general, a MOS transistor is slower but smaller, cheaper and requires less power than a bipolar transistor. The first MOS integrated circuit, devised by S.R. Hofstein and F.P. Heiman in 1962, provided general purpose logic functions and contained sixteen transistors. MOS transistors can be constructed with either electrons as the majority carriers (NMOS) or holes (PMOS).

Integrated circuits combining both types of MOS transistor, complementary-metal-oxide-semiconductor (CMOS), can be arranged to consume even less power but require an increased number of processing steps. Demonstrated in 1963, they were available commercially in 1968.

By that time, the techniques of integrated circuit manufacture were maturing, and the pace of integration measured in terms of increased transistor count, greater manufacturing yield and reduced cost was marked. Finally, from a time roughly a decade after the development of the integrated circuit, three landmark designs deserve mention. First, the Micromosaic chip made by Fairchild in 1967 was a double innovation. It was the first gate array, that is a chip containing a number of unconnected logic gates in fixed positions which can be joined by specifying the pattern of the final layer of connections. It also involved the first use of computer-aided design (CAD) to translate a description of the customer's circuit into the necessary pattern for the connections. The second chip was the first sizeable integrated semiconductor memory, the Fairchild 4100 designed by H.T. Chua in 1970. Containing 256 bits of static random-access memory (SRAM) constructed from fast bipolar transistors, the chip was used extensively in the ILLIAC IV, one of the first mainframe computers to have a fully integrated primary store. The third chip was the first microprocessor, the Intel 4004 designed by M.E. Hoff in 1971. The four bit processor was part of the design for a series of calculators for a Japanese company. The processor contained an adder, an accumulator, sixteen four bit registers and a push down stack. The 4004, however, did not just appear in calculators, but opened whole new worldwide markets. Further details of the history of integrated circuits can be found in [Braun 82], [Atherton 83], [Augarten 83] and [Dummer 83].

1.3.2 A fabrication sequence

Some key elements of the planar process applied to the fabrication of a silicon MOS transistor are illustrated in Figure 1.1. Semiconductor wafers typically 1mm thick and 100mm in diameter are sliced from a cylindrical ingot of single crystal grown by the seeded annealing of molten polycrystalline silicon. One surface is polished to a mirror like finish with diamond powder (Figure 1.1(a)). A $1\mu m$ thick oxide is grown on the polished surface of the wafer, and this is then coated with photoresist material (Figure 1.1(b)). An ultraviolet light source exposes the photoresist through a wafer sized mask pattern (Figure 1.1(c)), which is then developed to leave areas of exposed oxide which are etched back to the silicon substrate. The undeveloped photoresist is removed (Figure 1.1(d)). Transistors will be created in the areas of

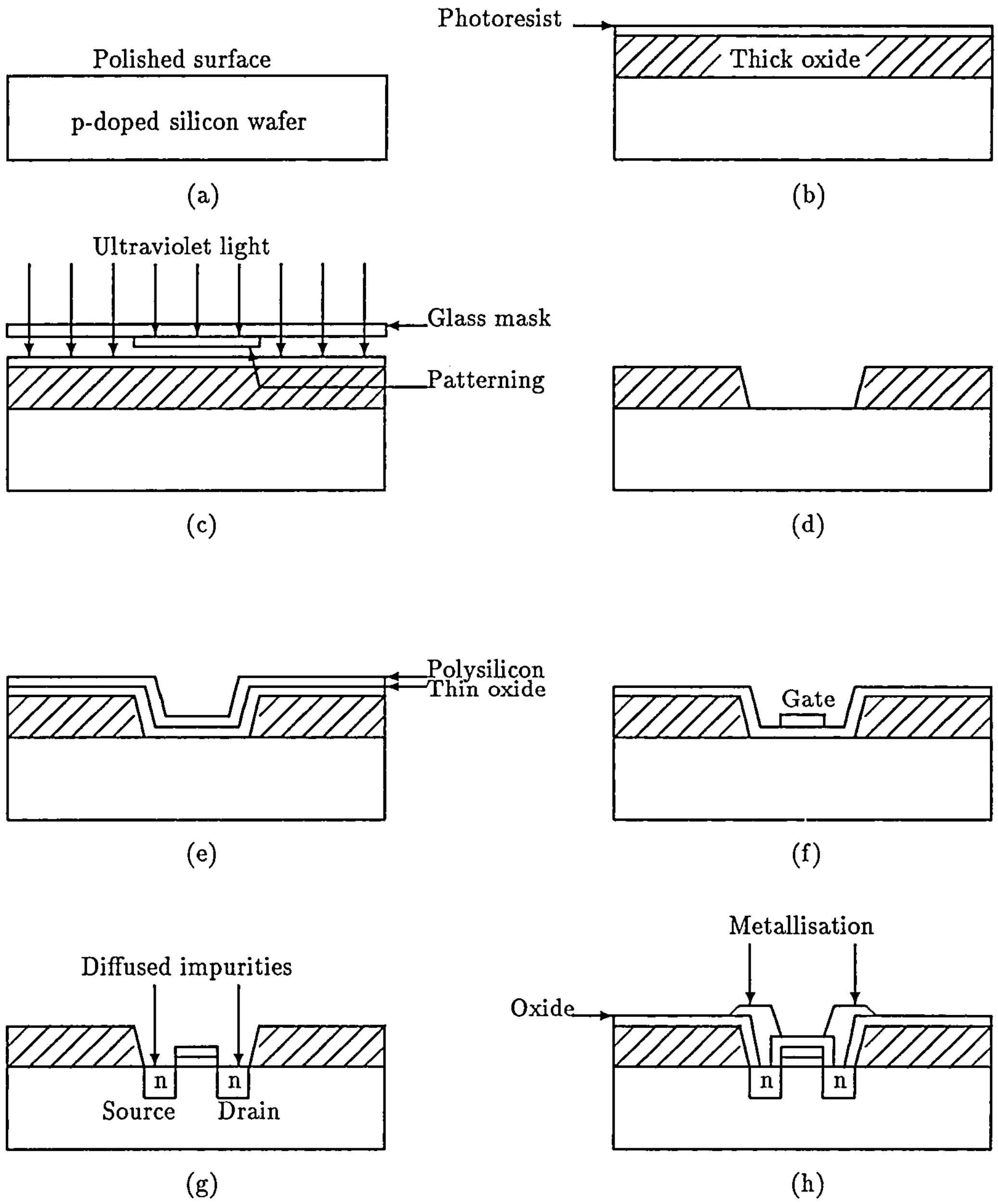

Figure 1.1: A simple fabrication sequence for a MOS transistor

exposed substrate formed by this initial patterning.

A 50nm thin oxide layer is then grown under highly controlled conditions, followed by the chemical vapour deposition of a 1μm polysilicon layer (Figure 1.1(e)). The polysilicon is patterned and etched, defining the transistor gates (Figure 1.1(f)). The thin oxide is etched away in areas not protected by the polysilicon and the whole structure is exposed to a gaseous source of dopant which diffuses into any exposed silicon. This defines 1μm deep transistor junctions in the substrate and increases the conductivity of the polysilicon (Figure 1.1(g)). The same mask is therefore used to define both the gate and junction regions, commonly known as a self-aligned process. A further patterned oxide layer opens contact holes to the substrate and polysilicon. These are filled with a layer of aluminium which is patterned to provide interconnect (Figure 1.1(h)).

1.4 Trends in integration

1.4.1 Faster and denser circuits

From the outset there has been demand from the electronics industry for faster and more highly integrated circuits. Chip manufacturers have responded with an approximate doubling in transistor count per device every year, often referred to as Moore's Law, and an increase in speed by a factor of one and a half every year. These trends arise from the desire to reduce the weight, bulk, power consumption and cost of integrated systems, while increasing the speed, manufacturability and ease of assembly. The communications and computer industries embody these trends. At each level of increased capacity, new applications present themselves and so the demand persists.

The limit on the number of transistors in an integrated circuit is largely an economic one, relating to the cost per working circuit. This cost rises when the number of working circuits per cm^2 of wafer falls. This is principally caused by a reduction in manufacturing yield, which is the number of working circuits as a percentage of the number of circuits manufactured. Circuits so large as to have a yield of less than 25% are unlikely to be economic to manufacture in quantity, and circuits with a yield of 1% are unlikely to be manufactured at all. In general, circuits fail to work when critical regions such as the gate and channel areas of a transistor, the contacts between connections and transistors or the connections themselves coincide with some structural fault. Such faults may be inherent in the materials involved, introduced by the processing techniques employed or caused by lithographic or mask alignment inaccuracies. Considering the maximum cost per working circuit to be invariant, there are three ways to increase the number of devices per circuit. First, device geometries can be reduced thereby increasing the device density. Second, fabrication technology can be improved to reduce defect density and thereby permit larger areas to be used. Third, techniques can be employed to prevent defects from causing the circuit to malfunction.

The limit on the switching speed of transistors in an integrated circuit is governed by the electron mobility of the semiconductor and the capacitances on the conducting path of the signal which is switching. Materials with a range of mobilities are used, and the choice of semiconductor material has a profound effect on the performance of the circuit. Capacitances are reduced by scaling down dimensions. The corresponding increase in circuit density also reduces interconnect length. Finally, at the system level, larger scale integration implies fewer chips, fewer connections between chips and less delay in input and output buffering and signal propagation.

1.4.2 Enhanced processing

The basic process described earlier suffers from a number of disadvantages which make it unsuitable for very small geometry devices (Figure 1.2). Transistor separation S is limited to around 1μm due to the lateral diffusion of the source and drain implanted regions under the thick oxide. Poor surface planarisation during gate oxide deposition leads to thinning of the photoresist at the channel/thick oxide step and possible gate disconnection, and a minimum channel width W of around 3μm is necessary. A number of enhanced processing sequences have been proposed to avoid such problems and to achieve smaller geometries and device spacing. Separations of 0.5μm and channel widths of 1.5μm are reported [Tsai 88].

Creation of smaller features also requires better lithography and alignment between mask steps. Conventional ultraviolet light step and repeat lithography is capable of defining channel lengths L down to about 1.5μm. This limit is imposed by the wavelength of the light, diffraction at the edge of mask patterns and alignment tolerances. Direct write electron beam lithography is capable of much finer geometries down to 0.15μm [Fichtner 82], and incorporates the flexibility to write different designs on a single wafer. However, the serial scanning nature of the beam makes throughput very low. X-ray or X-ray/photo lithography offers the best combination of high resolution and throughput with a line width of 0.2μm being reported [Mikaye 87].

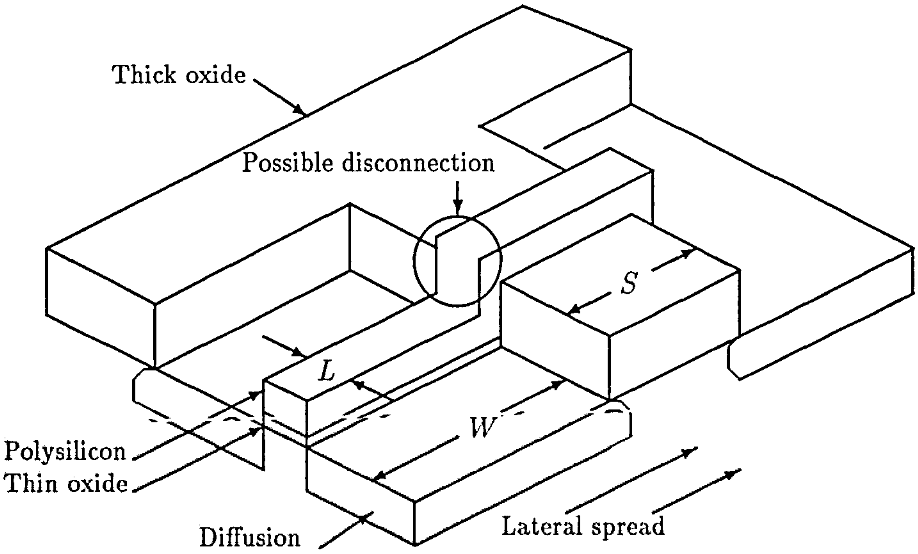

Figure 1.2: Problems with small geometry transistors

1.4.3 Alternative materials

There are alternative materials for the semiconducting, gate and wiring elements of integrated circuits. Of the various materials with semiconducting properties, gallium arsenide has the greatest potential for high performance. This is due to a high electron mobility, roughly an order of magnitude greater than that of silicon. Bulk gallium arsenide is also available in a semi-insulating form giving good device isolation, and reducing parasitic capacitances. The manufacture of gallium arsenide transistors is, however, subject to more complex and highly controlled processing due largely to the more stringent conditions required for correct transistor operation.

For the gate layer, polysilicon has a high resistance which limits its usefulness as a level of interconnect away from the gate region. In general, it is unsuitable for long distances, critical signals or power tracks. Refractory metals show promise as a gate and interconnect material either when compounded with silicon or in elemental form. Both polysilicon and diffusion can be capped with a titanium disilicide layer formed by depositing titanium and then performing a rapid thermal anneal. Requiring no extra mask steps, this process reduces the effective resistance of polysilicon and diffusion by a factor of fifty [Lai 86]. Alternatively, the gate material can be made entirely of molybdenum, which has a resistance five hundred times less than polysilicon [Kwasnick 88].

Aluminium wiring has some limitations. A course grain structure prohibits the creation of fine geometry lines, and planarisation problems make successive layers of metallisation increasingly error prone. Electromigration, which is the physical movement of the metal by the prolonged passage of current, becomes a problem for very fine lines and can lead to an open circuit after a time. Alternatives include tungsten and molybdenum/titanium tungsten schemes [Brown 87], [Kim 85]. In a novel scheme, grooves are etched in oxide and are filled by a blanket tungsten deposition which is then etched back to the oxide, so forming connections. This technique provides excellent planarity [Broadbent 88].

Optical interconnection is a radical approach to the interconnect problem. In various schemes, optical signals are received by integrated detectors or are injected directly into the transistor gate. The light source may be from a diode laser or light emitting diode transmitted through free space, optical fibres or integrated waveguides. Lenses and holographic routing elements can be employed [Goodman 84].

1.4.4 Mixed technologies

Bipolar and FET technologies have different speed, power consumption and current drive properties. They are frequently used together in systems containing separate bipolar and FET transistors, and it is natural to consider monolithic combination. The most mature and successful combination is of bipolar and CMOS, leading to the BiCMOS technology. This has an increased speed and drive capacity both for off-chip connections and between functions in the integrated circuit. Bipolar transistors are usually fabricated in an epitaxial silicon layer, that is a layer of silicon grown on a silicon wafer. The epitaxial layer is of controlled thickness and with precise levels of homogeneously distributed dopant. A number of selectively grown epitaxial layers can be used in BiCMOS circuits to optimise the doping and thickness of the active regions in each type of transistor [Washio 87], [O 88].

Of the many applications for BiCMOS technology, microprocessor design shows great potential since many of the standard functional elements of a processor benefit from performance in their output stages. A BiCMOS cell library containing adders, multipliers, read only and content addressable memory cells is reported which can be combined to form a processor roughly twice as fast as the CMOS equivalent [Hotta 88]. Such BiCMOS cells are larger than the CMOS counterparts