CAMBRIDGE

Cambridge University Press & Assessment
978-0-521-51613-6 — Electronic Thin-Film Reliability
King-Ning Tu
Excerpt
More Information

1 Thin-film applications to microelectronic technology

1.1 Introduction

Layered thin-film structures are used in microelectronic, opto-electronic, flat panel display, and electronic packaging technologies. A few examples are given below. Very large-scale integration (VLSI) of circuits on computer chips are made of multilayers of interconnects of thin metal films patterned into submicron-wide lines and vias. Semiconductor transistor devices rely on the growth of epitaxial thin layers on semiconductor substrates, such as the growth of a thin layer of p-type Si on a substrate of $n+$-type Si [1–3]. The gate of the transistor device is formed by the growth of a thin layer of oxide on the semiconductor. Solid-state lasers are made by sandwiching thin layers of light-emitting semiconductors between layers of a different semiconductor. In electronic and optical systems, the active device elements lie within the top few microns of the surface; this is the province of thin-film technology. Thin films bridge the gap between monolayer (or nanoscale structures) and bulk structures. They span thicknesses ranging from a few nanometers to a few microns. This book deals with the science of processing and reliability of thin films as they apply to electronic technology and devices [4]. To begin, this chapter describes the application of thin films to modern advanced technologies with examples.

1.2 Metal-oxide-semiconductor field-effect-transistor (MOSFET) devices

Advances in layered thin-film technology have been pivotal to the evolution of integrated circuits and opto-electronics. Today, we can fabricate hundreds of millions of transistors on a piece of Si chip the size of a fingernail. These transistors must be interconnected by thin-film lines to form circuits in order to function together. The basic circuit in a memory device is very simple. It consists of a transistor and a capacitor. A schematic cross-section of such a field-effect transistor is shown in Fig. 1.1, consisting of a transistor junction of $n+/p/n+$ type and a gate with a thin gate oxide over the p-type channel. The conductor on the gate is a bilayer structure consisting of a silicide and a heavily doped poly-Si, where the silicide is a metallic compound of metal and silicon. The $n+$ regions in the transistor junction are the source and drain regions and are connected by silicide contacts to the "word" line. Hence, silicide is used as a gate contact as well as source and drain contacts. There is a "bit" line connecting the source contact to the capacitor. The

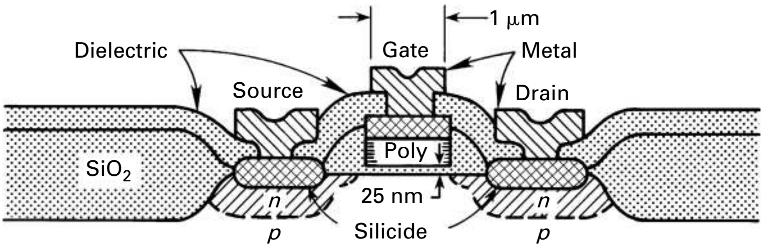

Fig. 1.1 Cross-section of a FET consisting of a transistor junction of $n+/p/n+$-type and a gate with a thin gate oxide over the p-type channel.

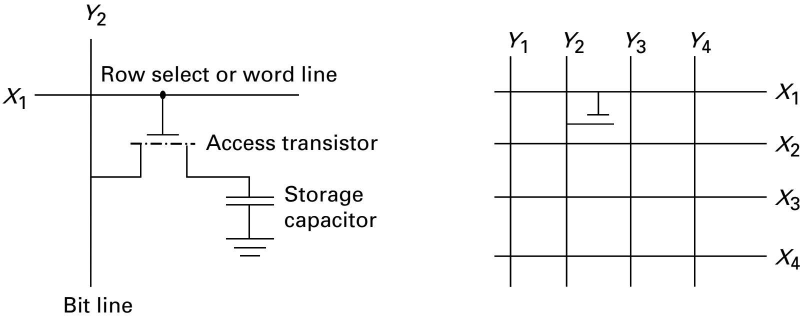

Fig. 1.2 Schematic diagram of an array of two-dimensional integrated circuits of MOSFET. Used with permission from *VLSI Technology*, S. M. Sze (1988), p. 494.

capacitor serves as a memory unit of either “1” (when the capacitor is full of charges) or “0” (when the capacitor is empty or stores no charge). The metal-oxide-semiconductor (MOS) field-effect transistor (FET) serves as a control (or gate) to allow the capacitor to discharge or not to discharge so that we can read or detect the two states of the capacitor, either full or empty [1–3].

Fig. 1.2 depicts an array of two-dimensional integrated circuits of MOSFET. In the x-coordinate, we have x_1, x_2, x_3, x_4, and so on, and in the y-coordinate, we have y_1, y_2, y_3, y_4, and so on. At each coordination point of (x, y), for example, take (x_1, y_2), we build a memory unit consisting of an FET and a capacitor. To operate the memory unit, a turn-on voltage is applied from the “word” line to open the gate. It attracts electrons to the p-type region and forms an inversion layer below the gate oxide. The inversion layer with electrons now electrically connects the two $n+$ regions. If the capacitor is full of stored charges, it will discharge so that a signal pulse can be detected at the end of the “bit” line. When this happens, we have identified a memory bit of “1” at the point (x_2, y_3). On the other hand, if the capacitor has no stored charges, there will be no discharge and no signal will be detected when we open the gate; then we have a memory bit of “0” at the point (x_2, y_3). The two-dimensional circuit integration as shown in Fig. 1.2 enables us to operate and detect every coordinate point on the two-dimensional integration of circuits, so it is called random access memory (RAM). Often, we use dynamic random

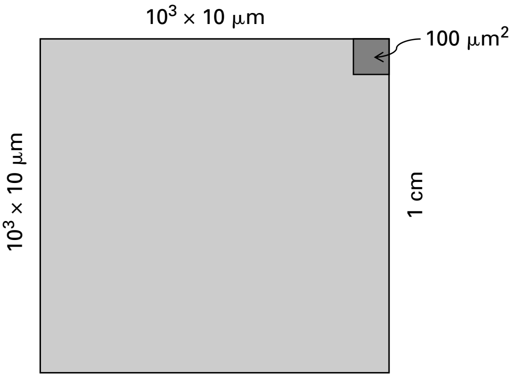

Fig. 1.3 Schematic diagram of a Si chip of size of 1 cm × 1 cm, divided into $10^3 \times 10^3 = 10^6$ small squares, so that each squares has an area 10 μm × 10 μm.

access memory (DRAM) to describe the device because the capacitors leak, since they are interconnected with lines, so we have to recharge them frequently and the recharge is a dynamic process. In certain devices, when the gate is isolated, we can use it as a floating gate.

In Fig. 1.3, we depict a Si chip of size 1 cm × 1 cm, and we divide it into $10^3 \times 10^3 = 10^6$ small squares, so that each of the squares has an area 10 μm × 10 μm. In each square, or cell area, if we can fabricate a FET and a capacitor, we have made a chip which has one million memory units. Needless to say, we should be able to interconnect them with their bit lines and word lines. In addition, we should also be able electrically to connect the chip to the outside circuit when we want to use it. The latter is a function of electronic packaging technology.

Next, we divide the 10 μm × 10 μm area into four smaller areas, i.e. cells about 5 μm × 5 μm. If we can shrink and build a FET and a capacitor in the smaller area, we will have a chip which has four million units of memory. This is the principle behind the miniaturization of the Si microelectronics industry in the last quarter century, or the essence of progress as suggested by Moore's law. The advance of one generation means the increase of a factor of four in circuit density on a chip. It goes from 1, 4, 16, 64, 256, to 1024 and so on. The industry started with about a one-thousand memory unit in the late 1960s and has advanced to about one billion memory units per chip today. Table 1.1 lists the dimensional changes of cells in several generations of devices. As the cell size becomes smaller, the feature size of the transistor, capacitor, and interconnects elements in the cell should also become smaller. There is a scaling law behind the shrinkage, which affects the electrical behavior of the transistor as well as the interconnect.

The VLSI of circuits is achieved by interconnecting all the transistors together using multilayers of Al or Cu thin-film interconnects. The process and reliability of multilayered thin-film interconnect structures are crucial to device applications. Today, there are eight or more layers of interconnects built on the transistors. Fig. 1.4 shows a scanning electron microscopic (SEM) image of a two-level Al interconnect on a Si surface after

Table 1.1. Dimensional changes of cells in several generations of devices

Cell density	Lithographic line width (μm)	Cell area (μm^2)
1 Mb	1	33
4 Mb	0.7	11
16 Mb	0.5	4.5
64 Mb	0.35 (deep UV)	1.5
256 Mb	0.25 (X-ray)	0.5
1 Gb	0.18	0.15

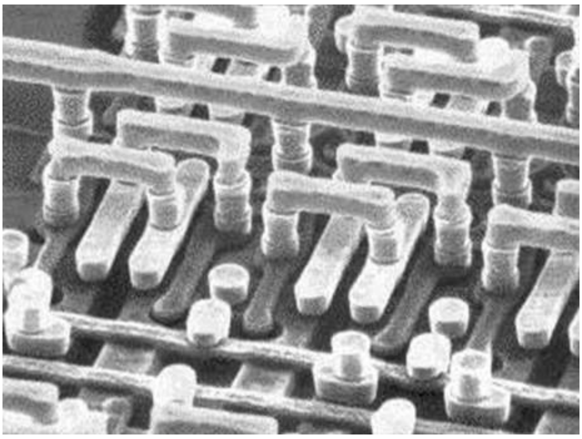

Fig. 1.4 SEM image of a two-level A1 interconnect on a Si surface after the insulating dielectric has been etched away. The width of the A1 lines is 0.5 μm and the spacing between them is 0.5 μm, so the pitch is 1 μm.

the insulating dielectric has been etched away. The width of the Al lines is 0.5 μm and the spacing between them is 0.5 μm, so that the pitch is 1 μm. On an area 1 cm × 1 cm, we can have 10^4 lines and each of them has a length of 1 cm, so the total length of interconnects in such a layer is 100 m. When we build eight such layers on a chip the size of a fingernail, the total length of interconnect is over 1 km, if we include those interlevel vias, i.e. the interconnects between layers. Fig. 1.5(a) shows a SEM image of an eight-level Cu interconnect structure taken after the interlevel dielectric was etched away. Fig. 1.5(b) shows a cross-sectional transmission electron microscopic image of a six-level Cu interconnect structure built on a Si surface. Here, the width of the narrowest interconnect via is 0.25 μm. The alignment of the vias between layers is a very challenging issue in device manufacturing.

The production cost of making the layered metallization structure is now more than half the cost of production of the whole Si wafer. In the interconnect metallization, silicide of C-54 $TiSi_2$, $CoSi_2$ or NiSi has been used as contacts to gate as well as contacts to source and drain of the FETs. Tungsten has been used as interlayer vias in the Al interconnect technology. It is also used as the first-level vias in Cu interconnect technology. To use

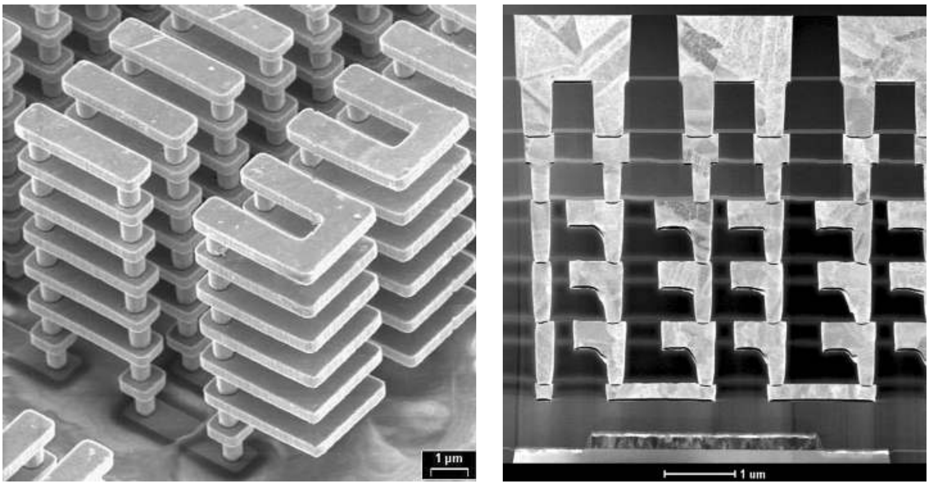

Fig. 1.5 (a) SEM image of an eight-level Cu interconnect structure taken after the interlevel dielectric was etched away. (b) A cross-sectional transmission electron microscopic image of a six-level Cu interconnect structure built on a Si surface. Here, the width of the narrowest interconnect via is 0.25 μm.

Cu in interconnect technology, the Cu must be coated with a very thin adhesion and diffusion barrier layer of Ta, or TiN. The processing, properties, and reliability of these thin films are relevant to the success of the technology.

Clearly, if the size of a cell is 1 μm × 1 μm, its circuit elements must be smaller than 1 μm. In the trend of miniaturization, not only the lateral dimension will become smaller; the vertical dimension such as the gate oxide thickness must be thinner too. No doubt we cannot keep shrinking the size in order to make smaller and smaller devices. The trend of miniaturization or the progress in scaling down device dimension has been modeled by Moore's law, which stated that the on-chip circuit density will double every 18 months. Fig. 1.6 shows a schematic curve of circuit density of memory and logic units on a chip in a central processing unit plotted against year. It is a log-linear plot to follow Moore's law.

1.2.1 Self-aligned silicide (salicide) contacts and gate

A critical dimension in the MOSFET device discussed above is the gate width or the distance between the source and drain contacts. The gate width is called the "feature size" of the device. Today, it is of nanoscale down to 45 nm and soon to be 33 nm and beyond. In fabrication, if the gate and contacts are made of different materials, it will require two different processing steps or two lithographic steps to manufacture them, so a high-precision alignment is required and it is difficult to control the feature size in nanoscale. The salicide process was invented to overcome this crucial step. Fig. 1.7 shows the schematic diagrams of the salicide process. Both gate and source/drain contacts are made of C-54 $TiSi_2$. Over the gate oxide is a layer of heavily doped poly-Si, and

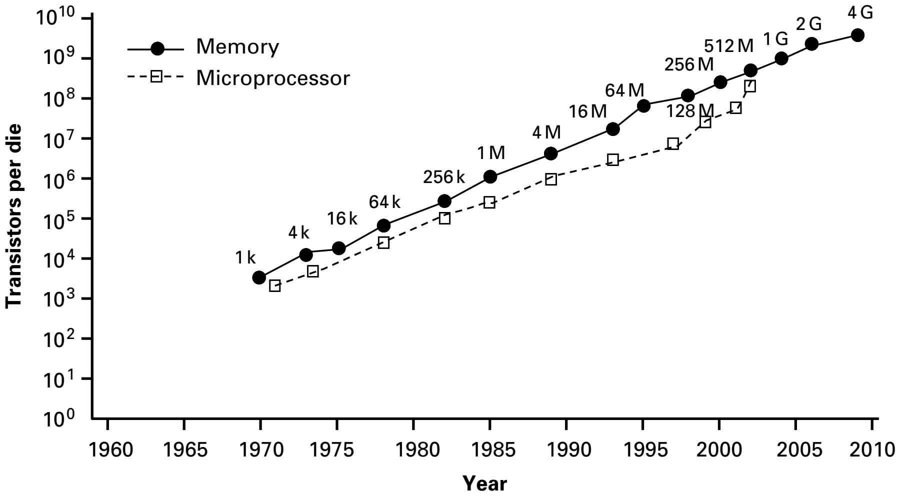

Fig. 1.6 Curve showing the circuit density of memory and logic units on a chip in a central processing unit plotted against the year. It is a log-linear plot and follows Moore's law.

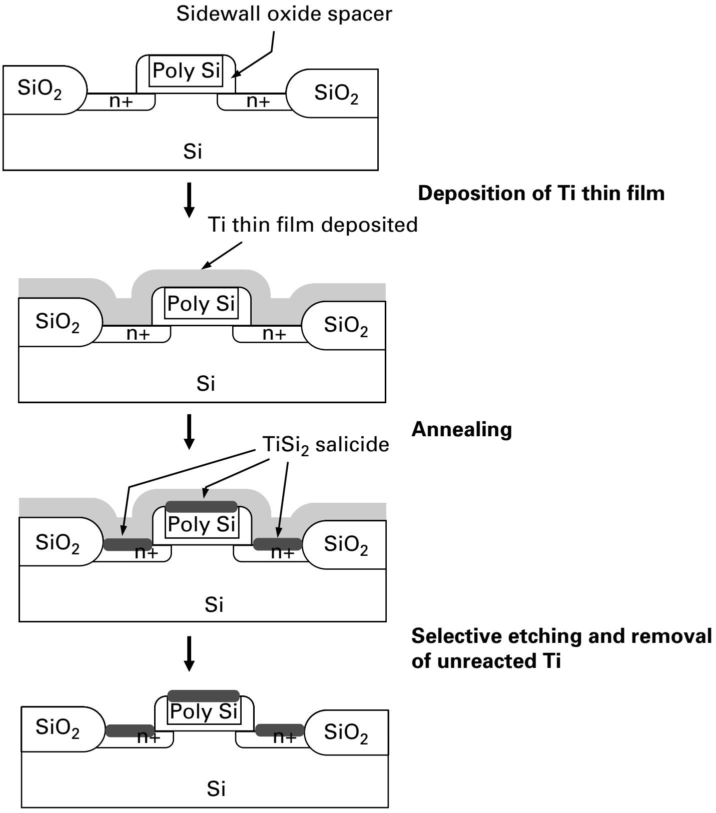

Fig. 1.7 Schematic diagrams showing the salicide process.

over the source/drain contact regions is a layer of heavily doped single crystal $n+$-Si. They are separated by two sidewalls of SiO_2. The spacing between the sidewalls and the thickness of the sidewall determine the lateral dimension of the feature size. When a Ti thin film is deposited and annealed, C-54 $TiSi_2$ forms on the gate contact and the source/drain contacts, simultaneously. But the Ti on the sidewall oxide does not react with the oxide to form silicide and the unreacted Ti can be selectively etched away, so that electrical insulation between the gate and the contacts can be achieved in one lithographic step. This "salicide" process avoids the high-precision alignment issue and enables the self-alignment of the gate and the contacts in production.

In the thin-film literature, there are many publications on silicide formation by the reaction between thin metal film and Si. Since there are hundreds of millions of silicide contacts and gates on a Si chip and they should have the same microstructure and electrical properties, a controlled salicide formation has been a very important processing step in VLSI device fabrication. The kinetics of silicide formation will be covered in Chapter 8.

1.3 Thin-film under-bump-metallization in flip-chip technology

How to connect the on-chip VLSI circuits to external circuits is the major function of electronic packaging technology [5–6]. To provide external electrical leads to all these on-chip interconnect wires, we may need several thousands of input/output (I/O) electrical contacts on the surface of the chip in a central processing unit. At present, the only practical and reliable way to provide such a high density of I/O contacts on the chip surface is to use an area array of tiny solder balls. We can have 50 μm diameter solder balls with a spacing of 50 μm between them, so the pitch is 100 μm. We place 100 of them along a length of 1 cm or 10 000 of them on an area of 1 cm^2. Typically the diameter of solder balls used today is about 100 μm, and the processing of solder balls in the electronic packaging industry is called "bumping technology." Because of the use of so small a size and large a number of solder balls, the International Technology Roadmap for Semiconductors (ITRS) has, since 1999, identified "solder joint in flip-chip technology" as an important subject of study concerning its yield in manufacturing and its reliability in application.

What is flip-chip technology? It is a technology to provide a large number of electrical connections between a Si chip and a packaging substrate using solder joints. The Si chip is flipped face down so that its circuits of very large-scale integration face the substrate. The electrical connections are achieved by an area array of solder bumps between the chip and its substrate. Fig. 1.8 shows an area array of solder balls on a chip surface. To join the chip to a substrate, the chip will be flipped over, so that the VLSI side of the chip is upside down, facing the substrate.

Flip-chip technology has been used for over 30 years in making mainframe computers. It originated from the "controlled collapse chip connection" or "C-4" technology in packaging chips on ceramic modules in the 1960s. Generally speaking, the advantages of flip-chip technology are: smaller packaging size, large I/O lead count, and higher

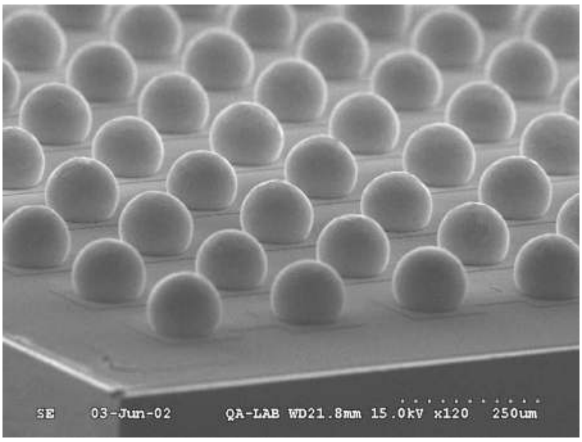

Fig. 1.8 SEM image of an area array of solder balls on a chip surface.

performance and reliability. Now it is used widely in consumer products such as in chip-size packaging, where the packaging substrate is of nearly the same size as the chip. The small packaging size is needed in handheld devices, where the form factor is important. In handheld terminals or computers, the demand for higher performance and greater functionality will require a large number of electrical I/O lead counts in the area array. The higher performance is because the solder bumps in the central part of the chip allow the device to operate at lower voltage and higher speed. Besides, flip-chip solder bumping is the only existing technology that can provide the reliability needed. We shall discuss reliability issues such as electromigration and stress-migration in later chapters.

At first, when VLSI chip technology was developed, a packaging technology of a high density of wiring and interconnection was required. This led to the development of multilevel metal–ceramic modules and multi-chip modules for mainframe computers. In a multilevel metal–ceramic module, many levels of Mo wire were buried in the ceramic substrate. Each of these modules could carry up to a hundred pieces of Si chip. Several of these ceramic modules were joined to a large printed circuit board and resulted in the two-level packaging scheme for mainframe computers shown in Fig. 1.9. It consisted of a first-level packaging of chip to ceramic module and a second-level packaging of ceramic module to polymer printed circuit board.

A schematic diagram of the cross-section of the first-level flip-chip C-4 solder joint is shown in Fig. 1.10. In the first-level packaging, the under-bump-metallization (UBM) on the chip side is a tri-layer thin film of Cr/Cu/Au. Actually, in the tri-layer the Cr/Cu has a phased-in microstructure for the purpose of improving the adhesion between the Cr and Cu and strengthening its resistance against solder reaction which may leach out the Cu so that the phase-in Cr/Cu, formed by the co-deposition of Cr and Cu having a composition gradient, can last several reflows in solder reaction. A “reflow” means that the solder is experiencing a temperature slightly above its melting point so that it melts. It is in the molten state that the solder will react with Cu to form the metallic bond or the formation of intermetallic compound (IMC) in the solder joint. On the substrate side

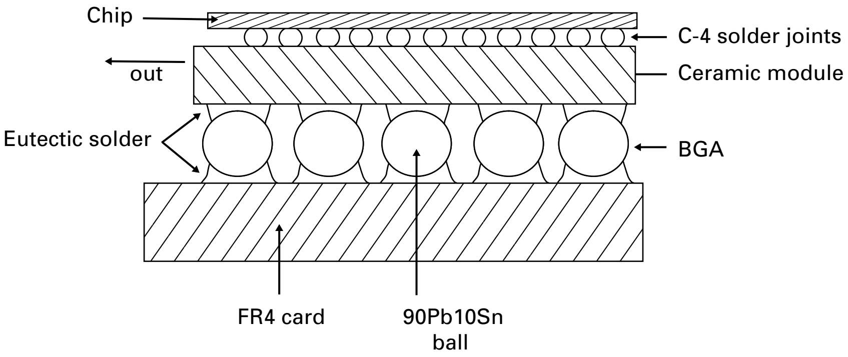

Fig. 1.9 Schematic diagram of the two-level packaging scheme for mainframe computers. It consists of a first level of chip to ceramic module packaging and a second level of ceramic module to polymer printed circuit board packaging.

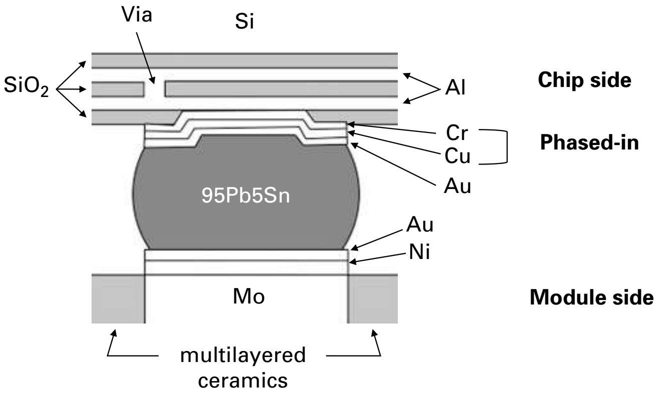

Fig. 1.10 Schematic diagram of the cross-section of a flip-chip solder joint.

of the joint, the metal bond-pad on the ceramic surface is typically Ni/Au. The solder which joins the UBM and the bond-pad is a high-Pb alloy such as 95Pb5Sn or 97Pb3Sn.

Initially, the on-chip solder bumps were deposited by evaporation and patterned by lift-off. Later, they were deposited by selective electroplating. Recently, in a new process of C-4, a template is used to form a two-dimensional array of solder balls, then a chip with a receiving array of UBMs is placed on the template so that all the balls can be transferred to the chip in a reflow, in which the balls react with the UBMs. The advantage of the new process is that no thick photo-resist is needed in selective electroplating, and the template can be used repeatedly to cut cost.

In the C-4 process using selective electroplating, after etching away the photo-resist and the electrical conducting line used for the plating, an array of cylindrical solder columns or bumps remains on the chip surface. The high-Pb bump has a melting point of over 300 °C. During the first reflow (around 350 °C), the column bumps change to ball-shape bumps on the UBM. Since the SiO_2 surface cannot be wetted by molten solder, the base of the molten solder bump is defined by the size of the UBM, thus the molten solder bump balls up or stands up on UBM contact. Therefore, the UBM contact controls the dimensions (height and diameter) of the solder ball when its volume is given. Often, the UBM contact is called “ball-limiting metallization” or BLM. As the BLM controls the height of the fixed volume of a solder ball when it melts, this is the meaning of the word “control” in “controlled collapse chip connection.” Without the control, the solder ball will spread on the UBM, and then the gap between the chip and the module is too small.

To join the chip which already has an array of solder balls to a ceramic module, a second reflow is used. During the second reflow, the surface energy of the molten solder balls provides a self-aligning force to position the chip on the module automatically. When the solder melts in order to join the chip to the module, the chip will drop slightly and rotate slightly. The drop and rotation are due to the reduction of surface tension of the molten solder balls, which achieve the alignment between the chip and its module, so it is a controlled collapse process. The word “collapse” means that the chip drops and rotates slightly when the area array of solder balls becomes molten and wets the pads on the module.

The high-Pb solder is a high melting-point solder, yet both the chip and the ceramic module can withstand the high temperature of reflow without problem. Additionally, the high-Pb solder reacts with Cu to form a layer-type Cu_3Sn, which can last several reflows without failure. It is worth noting that each of the metals in the tri-layer of Cr/Cu/Au has been chosen for a particular reason. First, solder does not wet the Al wire, so Cu is selected for its reaction with Sn to form IMCs to achieve a metallic joint. Second, Cu does not adhere well to the dielectric surface of SiO_2, so Cr is selected as a glue layer for the adhesion of Cu to SiO_2. The phased-in Cu/Cr UBM was developed to improve the adhesion between Cu and Cr. Since Cr and Cu are immiscible, their grains form an interlocking microstructure when they are co-deposited. In such a phased-in microstructure, the Cu adheres better to the Cr, and also it will be harder for the Cu to be leached out to form IMC with Sn during reflow. Furthermore, the phase-in microstructure provides a mechanical locking of the IMC. Finally, Au is used as a surface passivation coating to prevent the oxidation or corrosion of Cu. It also serves as a surface finish to enhance solder wetting. In the literature, much study on interdiffusion and reactions in the bilayer thin films of Cr/Cu and Cu/Au has been performed.

In the second-level packaging of ceramic module to polymer board, i.e. to join the ceramic substrate to a polymer printed circuit board, another area array of solder balls is placed on the back side of the ceramic substrate. They are called ball-grid-array (BGA) solder balls, which have a much large diameter than the C-4 solder balls. Typically the BGA solder ball diameter is about 760 μm. They are eutectic SnPb solder with a lower melting point (183 °C), which is reflowed around 220 °C. Sometimes, composite solder balls of high-Pb and eutectic SnPb are used, with the high-Pb as the core of the ball. It is obvious that during this reflow (the third reflow) of the eutectic solder, the high-Pb solder joints in the first-level packaging or the high-Pb core in the composite solder balls will not melt. In certain applications, the high-Pb core in the composite can be replaced by a Cu ball.