CHAPTER 3: Epitaxy

Epitaxy (*epi* means "upon" and *taxis* means "ordered") is a term applied to processes used to grow a thin crystalline layer on a crystalline substrate. The seed crystal in epitaxial processes is the substrate. Unlike the Czochralski process, crystalline thin films can be grown below the melting point using techniques such as chemical vapor deposition (CVD), molecular beam epitaxy (MBE), etc. When a material is grown epitaxially on a substrate of the same material, the process is called homoepitaxy, an example of which is depicted in ***Figure 3.1***. On the contrary, if the layer and substrate are of different materials, such as $Al_xGa_{1-x}As$ on GaAs, the process is termed heteroepitaxy. Naturally, in heteroepitaxy, the crystal structures of the layer and the substrate must be similar in order to achieve good crystalline integrity.

Performance of many devices, such as CMOS (Complementary MOS) and DRAM (Dynamic Random Access Memory), can be improved by employing epitaxial wafers. Succinctly speaking, epitaxial wafers have two fundamental advantages over bulk wafers. First, they offer device engineers a means to control the doping profiles not attainable through other conventional means such as diffusion or ion implantation. Secondly, the physical and chemical properties of the epitaxial layers can be made different from the bulk materials. A DRAM (dynamic random access memory) circuit employing MOS devices (***Figure 3.2-1***) is vulnerable to soft errors created by alpha particles originating from packaging materials and the environment. These alpha particles can cause electron-hole pairs in the bulk of the wafer. If these charges migrate to the storage cell of a DRAM structure, the data stored can be wiped out. A DRAM fabricated in the structure shown in ***Figure 3.1(b)*** has higher immunity against alpha-particle soft error. A heavily doped substrate increases the rate of electron-hole pair recombination and the DRAM is less prone to alpha-particle soft errors.

Another important type of MOS device is the non-volatile memory. In the structure shown in ***Figure 3.2-2***, the device consists of two gate electrodes. The top electrode that can be directly accessed is the control gate whereas the one underneath is the floating gate. To program the cell, a high longitudinal electric field is applied to the control gate and on account of the hot carrier effects, some electrons will tunnel into the sandwiched dielectric C_{ONO} and are permanently stored there due to the high energy barrier that exists between the conduction bands of Si and SiO_2. To erase the cell, a high positive voltage is applied to the source with the control gate grounded. Fowler-Nordheim tunnel occurs draining the electrons through the floating gate to the source.

A CMOS circuit built in a structure such as the one displayed in ***Figure 3.1(b)*** minimizes the possibility of latch-up. As shown in ***Figure 3.3*** and ***Figure 3.4***, latch-up is due to a regenerative bipolar-transistor action used by a clamped, low-resistance path between the power supply and ground. Avoiding latch-up is especially challenging in small-dimension CMOS for dense VLSI applications.

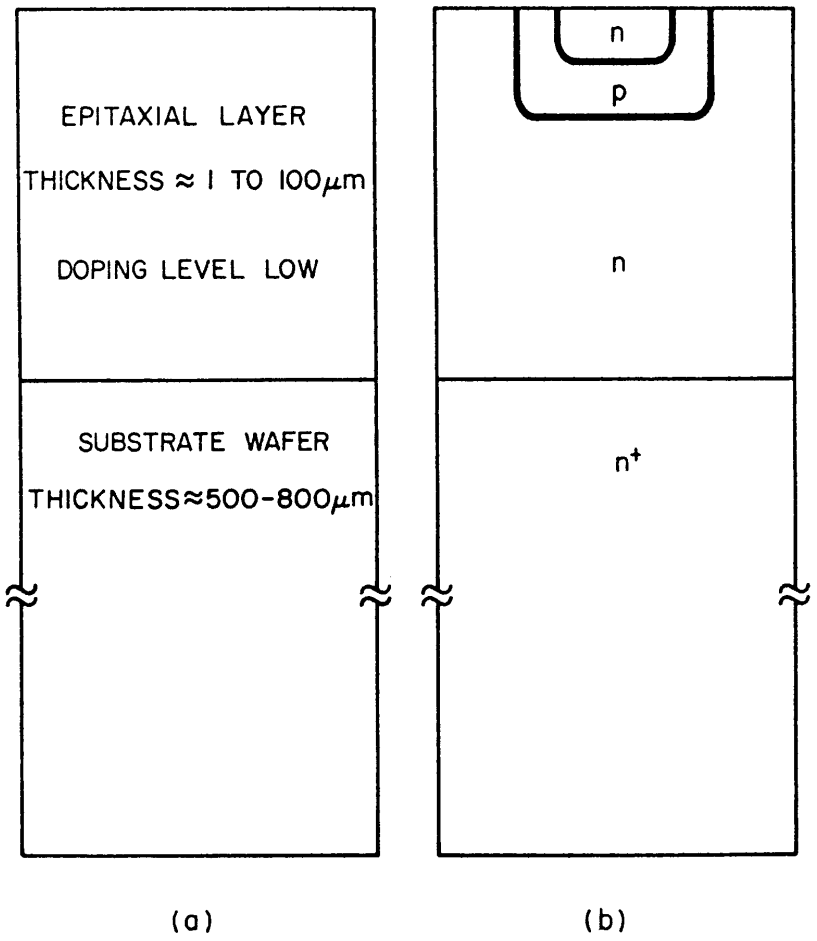

Figure 3.1: (**a**) Cross-sectional schematic of a typical epitaxial layer and substrate. (**b**) Wafer with an npn discrete transistor fabricated in a lightly doped n-type epitaxial layer grown on a heavily doped n-type substrate.

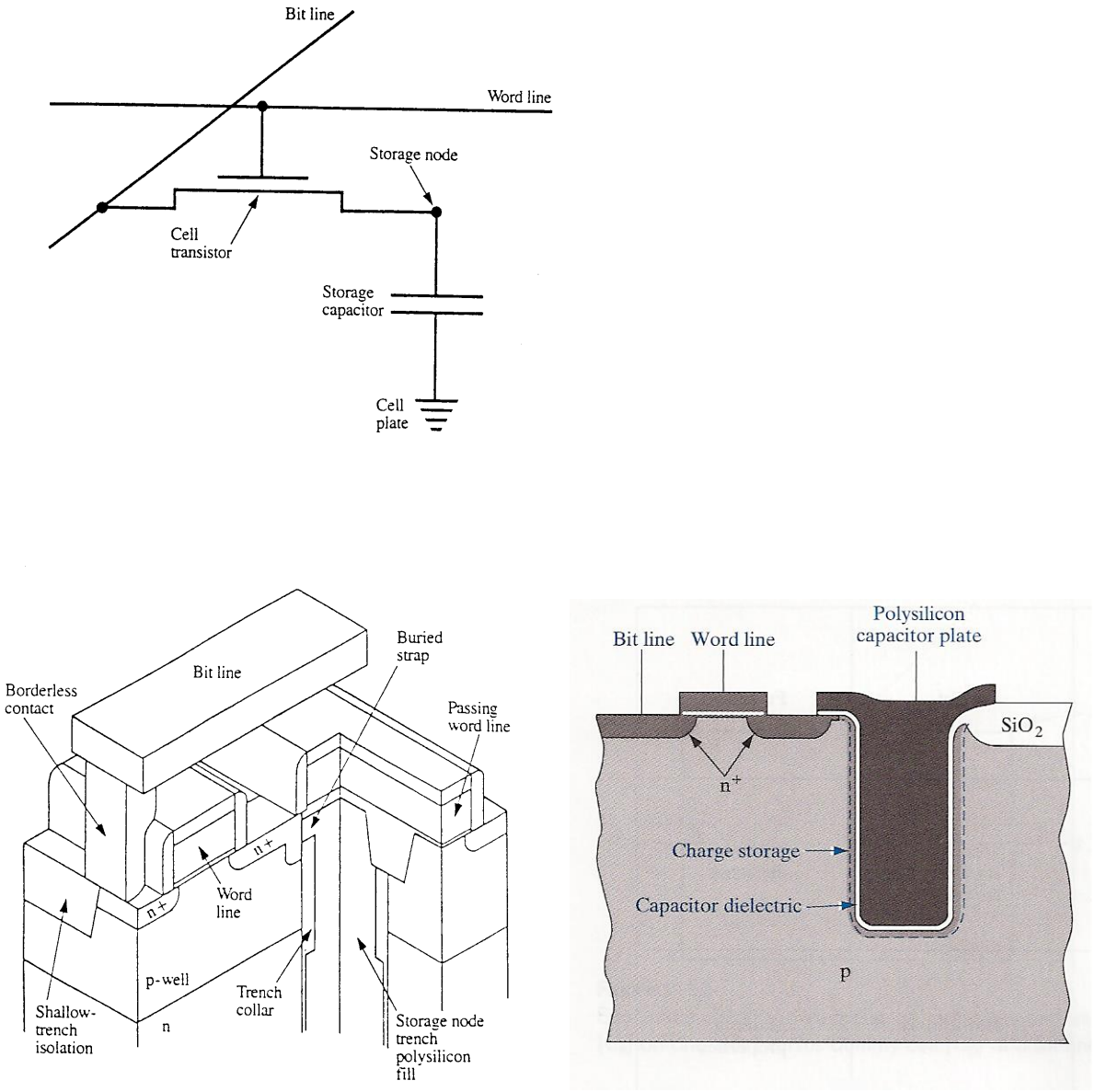

Figure 3.2-1: (**Top**) Schematic of a simple DRAM structure consisting of one MOS device for switching and one capacitor for charge storage (memory). (**Bottom**) Three-dimensional cross section of a 256Mb buried strap trench (BEST) DRAM cell on the left and schematic of the trench DRAM on the right.

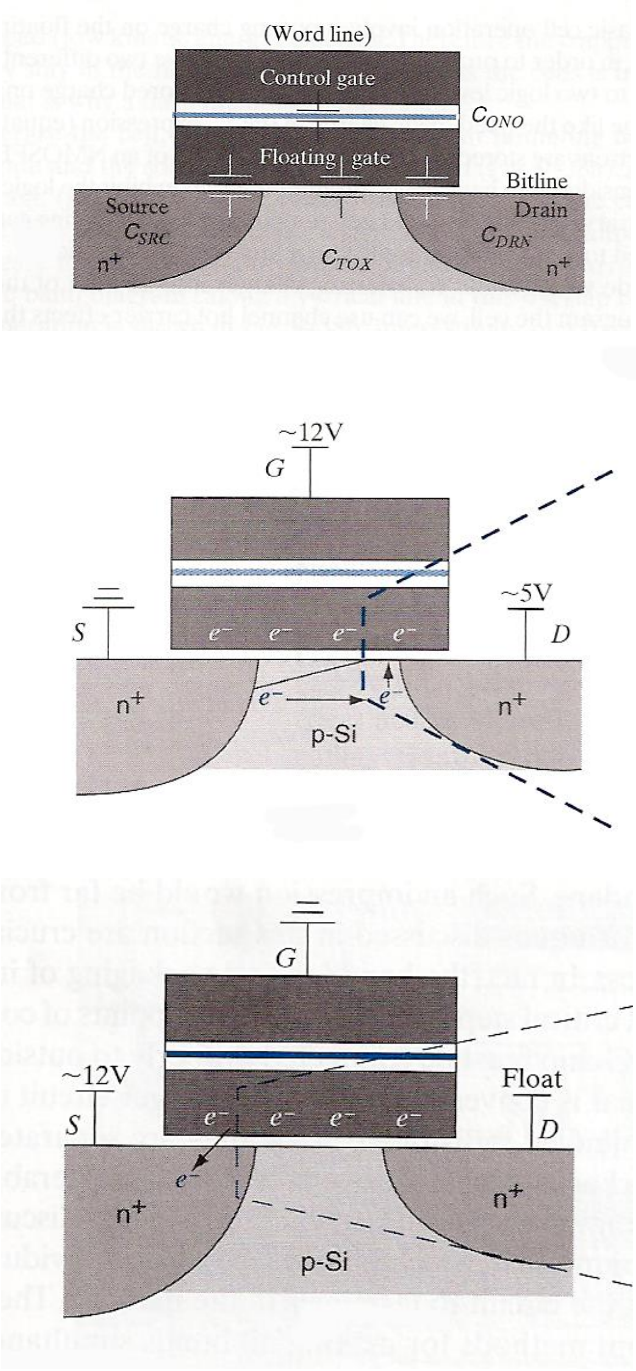

Figure 3.2-2: (**Top**) Schematic of a simple non-volatile flash memory cell employing a double gate structure. (**Middle**) Hot carrier programming of the flash memory cell. (**Bottom**) Fowler-Nordheim tunneling erasure.

Latch-up

To understand the basic latch-up phenomena, consider the p-well CMOS structure shown in ***Figure 3.3***. Superimposed on the MOS cross sections are unwanted or parasitic npn and pnp bipolar transistors. The transistors are cross-connected so that the base-collector junctions are common. From the equivalent circuit illustrated in ***Figure 3.4***, it can be seen that under active bias, the pnp collector delivers current to the npn base, and the npn collector delivers current to the pnp base. If these bipolar transitors have even moderate current gains (β), this interconnection can easily lead both devices to saturate so that the supply voltages become connected across a low resistance in series with two voltage drops: (i) the voltage across a saturated base-collector junction V_{CEsat} and (ii) the voltage across a saturated base-emitter junction V_{BEsat}.

Under normal CMOS operating conditions, the base-emitter junctions for both bipolar transistors are reversed-biased, which makes latch-up impossible. A successful circuit design must, however, preclude latch-up under any conditions that might be experienced by the circuit. To understand the ways in which latch-up can be initiated, we refer to ***Figure 3.5***, in which the cross-connected bipolar pair is drawn and two elements – a capacitor C_{PS} and a current source I_o – are added in parallel across the base-collector junctions. The capacitance C_{PS} is much larger than that of a typical base-collector junction because this capacitor represents the large junction between the p-well and the substrate. The current source I_o normally models only junction leakage and is very small in magnitude. Several mechanisms, however, can cause I_o to increase markedly.

Among the possible sources for current through I_o are (1) minority carries injected into the substrate by transient forward bias on pn junctions (typically in input and output circuits), (2) photogeneration by ionizing radiation, and (3) impact generation by hot carriers. The large capacitor C_{PS} can also deliver currents when voltage transients occur, especially during the power-up phase of the circuit. Any of these sources of current can turn on one or both of the bipolar devices. Therefore, latch-up will take place if the gain of the cross-coupled bipolar pair is sufficient and if the V_{DD} power supply can deliver enough current.

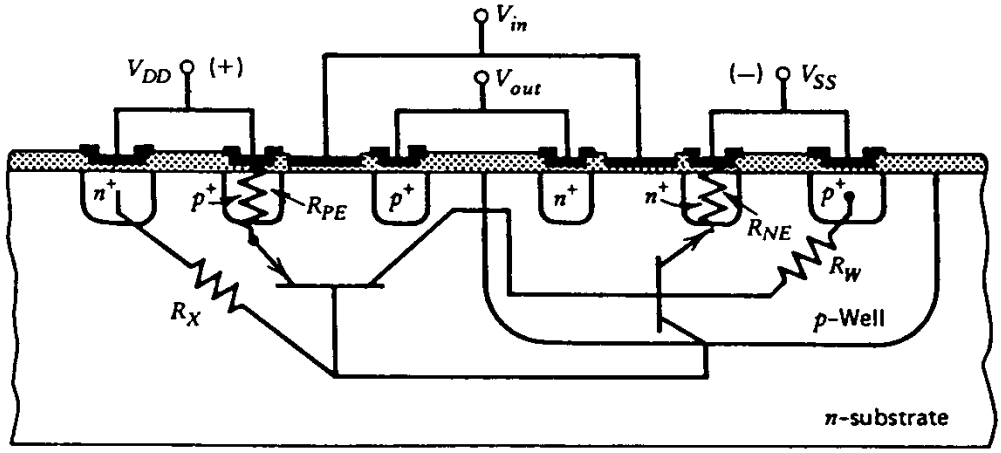

Figure 3.3: Cross section of a p-well CMOS inverter. The parasitic pnp andnpn bipolar transistors are indicated along with associated substrate resistor R_x and well resistor R_W. The two resistors R_{PE} and R_{NE} represent contact and diffused-region resistance in the emitters.

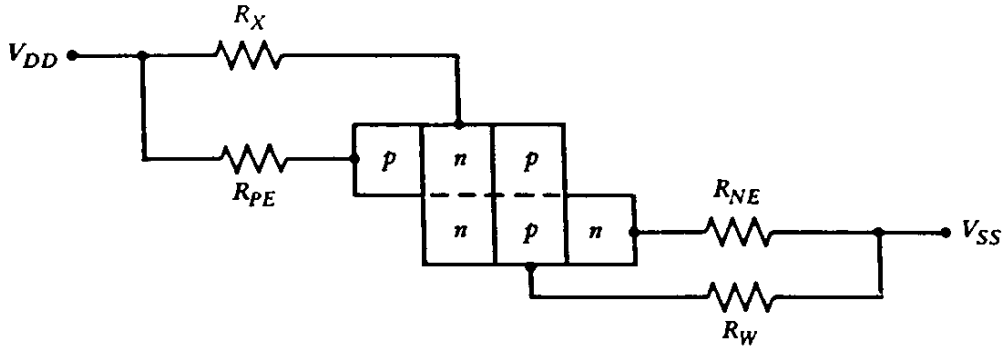

Figure 3.4: Circuit and schematic representation of the cross-coupled parasitic npn and pnp transistors.

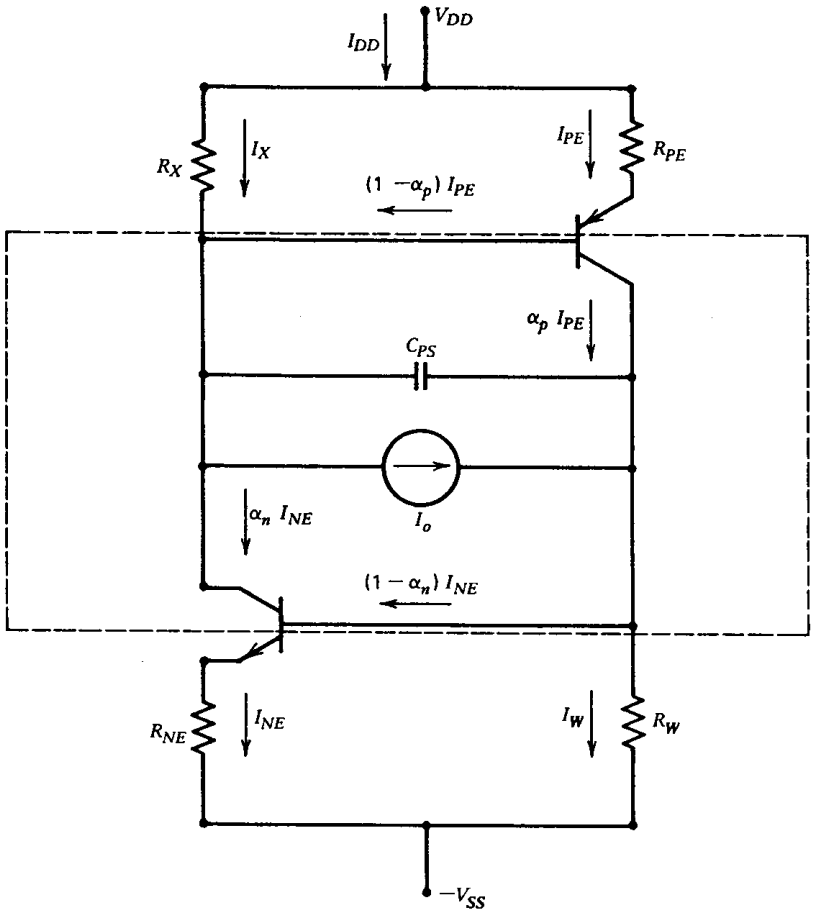

Figure 3.5: Latch-up equivalent circuit including well-to-substrate capacitor C_{PS} and parasitic current I_o. The dashed lines surround all elements connected between the well and substrate nodes.

3.1 Chemical Vapor Deposition (CVD)

Epitaxial growth can be achieved from solid-phase, liquid-phase, vapor-phase, and molecular-beam deposition. For a Si epilayer, vapor-phase epitaxy (VPE), which is a form of chemical vapor deposition (CVD), is the most common. Chemical vapor deposition (CVD) of single-crystal silicon is usually performed in a quartz reactor into which a susceptor is placed. The susceptor provides physical support for the substrate wafers and provides a more uniform thermal environment. Deposition occurs at a high temperature at which several chemical reactions take place when process gases flow into the chamber.

A typical VPE process is illustrated schematically in ***Figure 3.6***. Prior to the layer deposition, the growth system is purged by nitrogen or hydrogen for a short period, and followed by a vapor HCl etching. The deposition process is then initiated by directing the reactant gases into the reactor, where the substrate is located and heated to a temperature T_g. As shown in ***Figure 3.7***, the growth process can be broken down into the following steps:

(1) Introduction of the reactant species to the substrate region

(2) Transfer of the reactant species to the substrate surface

(3) Adsorption of the reactant species on the substrate surface

(4) Surface diffusion, site accommodation, chemical reaction, and layer deposition

(5) Desorption of residual reactants and by-products

(6) Transfer of residual reactants and by-products from the substrate surface

(7) Removal of residual reactants and by-products from the substrate region

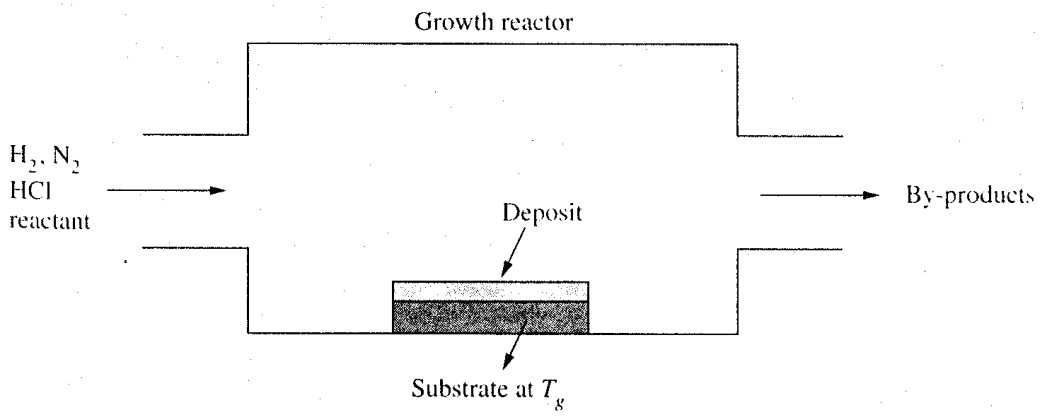

Figure 3.6: Schematic illustration of a typical VPE process.

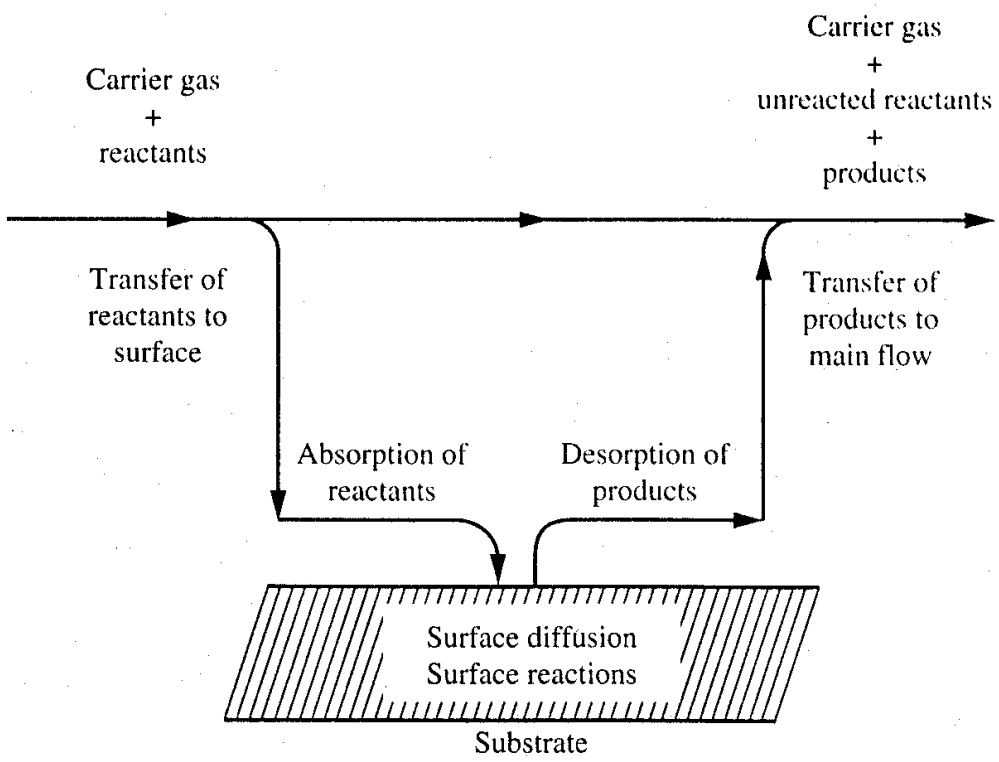

Figure 3.7: Schematic description of the sequence of steps in a VPE process.

3.1.1 Growth Model and Simple Theoretical Treatment

The Reynolds number, R_e, characterizes the type of fluid flow in a reactor:

$$R_e = D_r v\rho/\mu \qquad \textit{(Equation 3.1)}$$

where D_r denotes the diameter of the reaction tube, v is the gas velocity, ρ represents the gas density, and μ stands for the gas viscosity. Values of D_r and v are generally several centimeters and tens of cm/s, respectively. The carrier gas is usually H_2, and using typical values for ρ and μ, the value of R_e is about 100. These parameters result in gas flow in the laminar regime. That is, the gases flow in a regular, continuous, and non-turbulent mode and in a specific direction. Accordingly, a boundary layer of reduced gas velocity will form above the susceptor and at the walls of the reaction chamber. The thickness of the boundary layer, y, is defined as:

$$y = \left[\frac{D_r x}{R_e}\right]^{1/2} \qquad \textit{(Equation 3.2)}$$

where x is the distance along the reactor. ***Figure 3.8*** shows the development of this boundary layer. It is across this boundary layer that reactants are transported to the substrate surface and reaction by-products diffuse back into the main gas stream. The fluxes of species going to and coming from the wafer surface are complex functions of the temperature, pressure, reactant, concentration, layer thickness, etc. By convention, the flux, J, is defined to be the product of D and dn/dy, and is approximated as:

$$J = \frac{D(n_g - n_s)}{y} \qquad \textit{(Equation 3.3)}$$

where n_g and n_s are the gas stream and surface reactant concentrations, respectively, D is the gas-phase diffusivity, which is function of pressure and temperature, y is the boundary layer thickness, and J is the reactant flux of molecules per unit area per unit time.

In a steady state, the reactant flux across the boundary layer is equal to the chemical reaction rate, k_s, at the specimen surface. Therefore,

$$J = k_s n_s \qquad \textit{(Equation 3.4)}$$

and

$$n_s = \frac{n_g}{1 + \frac{k_s y}{D}} \qquad \textit{(Equation 3.5)}$$

The quantity D/y is often called the gas phase mass-transfer coefficient, h_g. In the limiting case when $k_s >> h_g$, n_s approaches zero, thereby implying that the overall reaction is limited by transport of reactant across the boundary layer. Conversely, if $k_s << h_g$, n_s is roughly equal to n_g, and the growth process will be dominated by the surface chemical reaction rate.

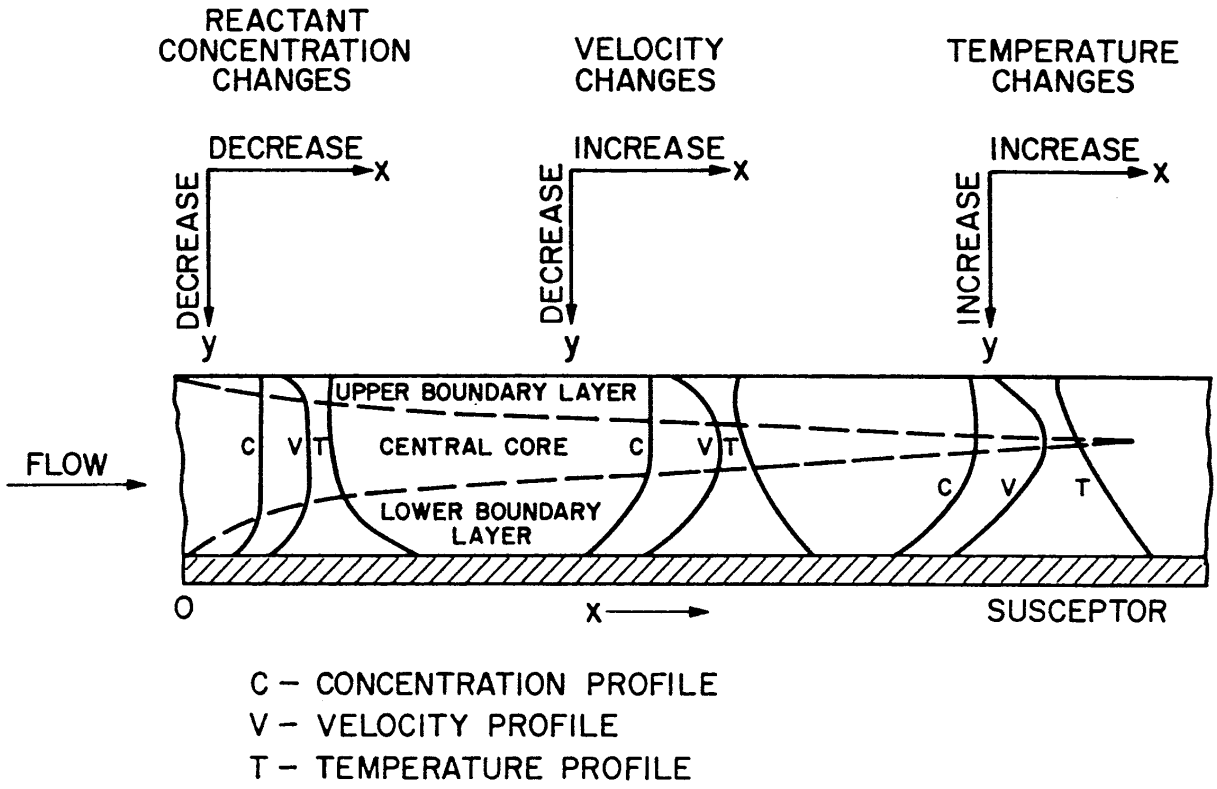

Figure 3.8: Boundary layer formation in a horizontal reactor.

3.1.2 Growth Chemistry

The most common starting chemical is silicon tetrachloride ($SiCl_4$) as it has a lower reactivity with respect to oxidizers in the carrier gas than the other silicon hydrogen chloride compounds, such as SiH_4, $SiHCl_3$, etc. The overall reaction is:

$$SiCl_4\,(gas) + 2H_2\,(gas) \Leftrightarrow Si\ (solid) + 4HCl\ (gas)$$

Experimental results indicate the presence of many intermediate chemical species. In particular, at a reaction temperature of 1200°C, four species have been observed using FTIR. ***Figure 3.9*** illustrates the concentrations of these species at different positions along the horizontal reactor. The detailed reaction mechanism is postulated to be:

$$SiCl_4 + H_2 \Leftrightarrow SiHCl_3 + HCl$$

$$SiHCl_3 + H_2 \Leftrightarrow SiH_2Cl_2 + HCl$$

$$SiH_2Cl_2 \Leftrightarrow SiCl_2 + H_2$$

$$SiHCl_3 \Leftrightarrow SiCl_2 + HCl$$

$$SiCl_2 + H_2 \Leftrightarrow Si + 2HCl$$

All the above reactions are reversible and at certain conditions, the overall reaction rate can become negative. That is, etching occurs in lieu of deposition.

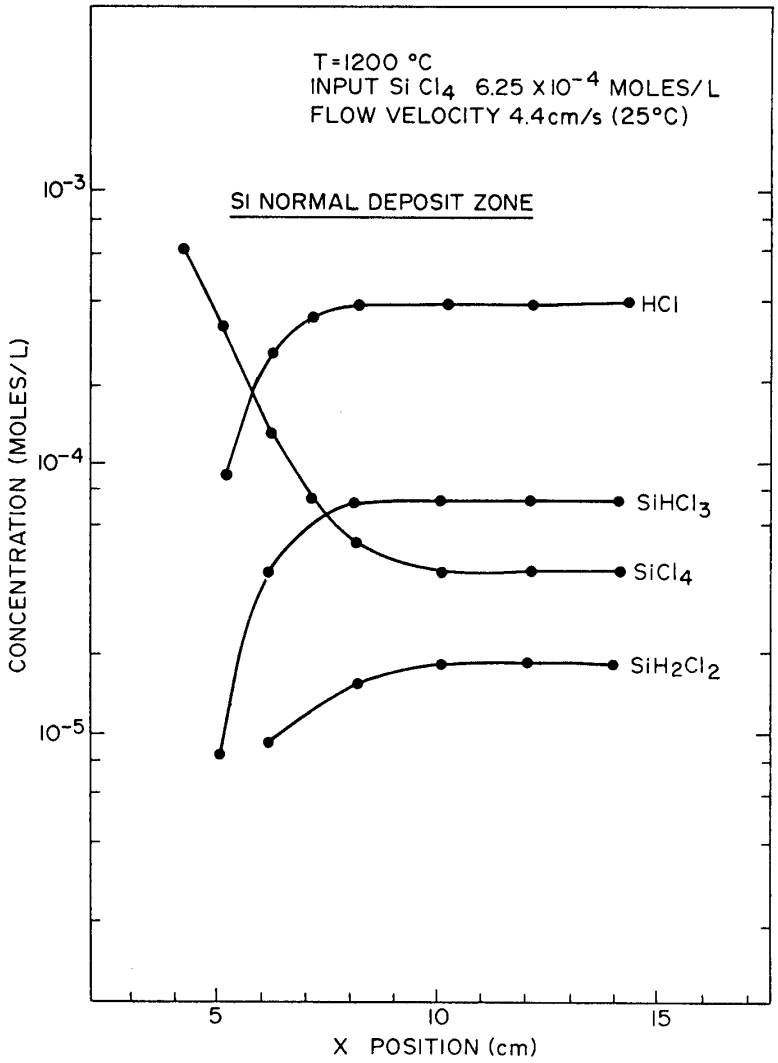

Figure 3.9: Species detected by IR spectroscopy in a horizontal reactor using $SiCl_4$ and H_2.

3.1.3 Doping and Autodoping

Hydrides of the impurity atoms are usually used as the source of dopants during epitaxial growth. For instance,

$$2AsH_3\,(gas) \rightarrow 2As\,(solid) + 3H_2\,(gas)$$

$$2As\,(solid) \rightarrow 2As^+\,(solid) + 2e^-$$

The dopant incorporation process is depicted schematically in ***Figure 3.10***.

In addition to intentional dopants, unintentional dopants are introduced from the substrate via a process called autodoping, in which the dopant is released from the substrate through solid-state diffusion or evaporation, and is reincorporated into the growing layer either by diffusion through the interface or through the gas. Autodoping is manifested as an enhanced region between the layer and the substrate. Zone A in ***Figure 3.11*** is due to solid-state out-diffusion from the substrate, and can be approximated by the complementary error function if the growth velocity is less than $2(D/t)^{1/2}$, where D is the dopant diffusion constant and t denotes the deposition time. Zone B in ***Figure 3.11*** originates from gas-phase autodoping. Because the dopant evaporating from the wafer surface is supplied from the wafer interior by solid-state diffusion, the flux of dopant from an exposed surface decreases with time. Once autodoping diminishes, the intentional doping predominates and the profile becomes flat. Autodoping thus limits the minimum layer thickness that can be grown with controlled doping as well as the minimum dopant level.

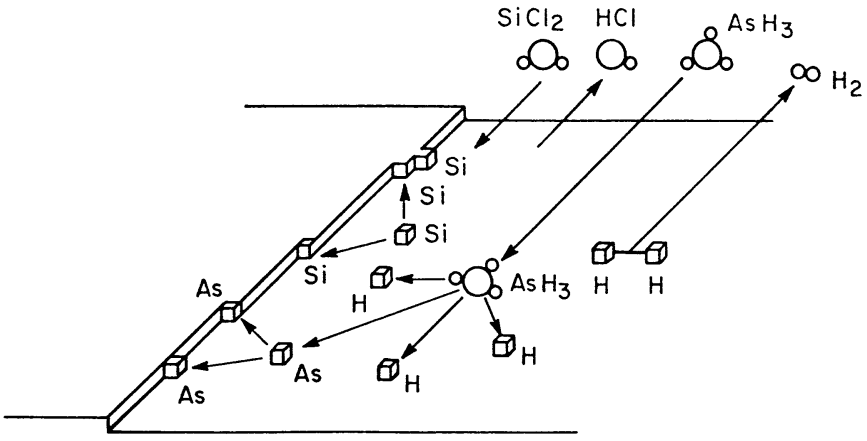

Figure 3.10: Schematic representation of arsine doping and growth process.

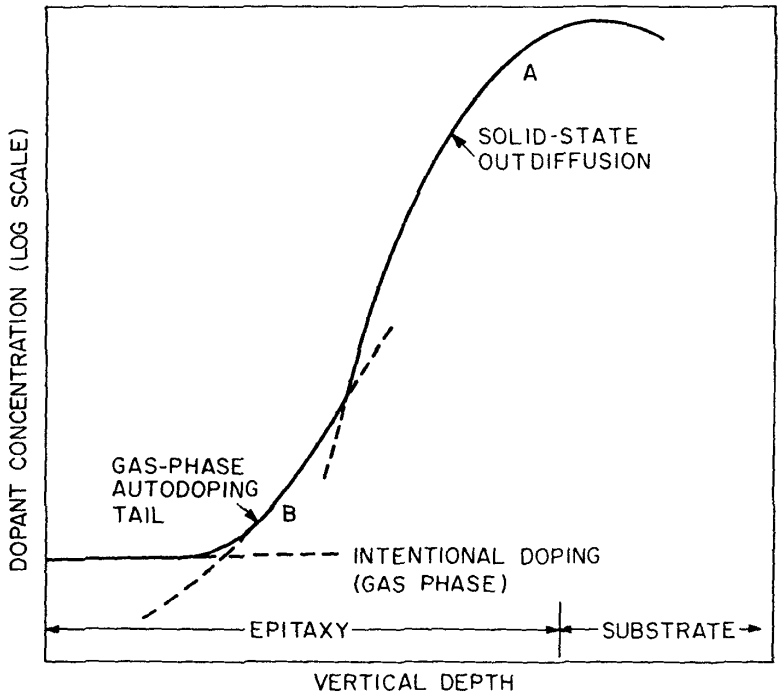

Figure 3.11: Generalized doping profile of an epitaxial layer detailing the various regions of autodoping.

Example 3.1

If the intrinsic diffusivity of boron in silicon is expressed by $D = 0.76e^{-\frac{3.46(eV)}{kT}}$, calculate the minimum growth rate that is required when a silicon epilayer is grown on a heavily boron-doped silicon substrate for 20 minutes at 1200°C in order that autodoping becomes insignificant. Discuss why a low growth temperature is essential to achieve a sub-micrometer thick silicon epitaxial layer.

Solution

$$D = 0.76e^{-\frac{3.46(eV)}{kT}}$$

At 1200°C,

$$D = 0.76e^{-\frac{3.46}{\frac{1.38x10^{-23}}{1.6x10^{-19}}x1473}} = 0.76e^{-\frac{3.46}{0.127}} = 0.76x1.47^{-12} = 1.12x10^{-12}(cm/s)$$

The minimum growth rate is: $v > 2(\frac{D}{t})^{1/2} = 2\left[\frac{1.12x10^{-12}}{20x60}\right]^{1/2} = 0.6nm/s$

A low temperature is necessary to minimize the autodoping effect for submicron Si-epi layer deposition at reasonably low growth rates for tight thickness and doping control.

3.1.4 Reactors

There are three common epitaxial reactor designs: barrel, vertical, and horizontal, as demonstrated in ***Figure 3.12***. The horizontal reactor is the most commonly used system. It offers high capacity and throughput. However, it suffers from the inability to achieve a uniform deposition over the entire susceptor. Tilting the susceptor by 1.5 to 3 degrees mitigate the non-uniformity substantially. In contrast, the vertical pancake reactor is capable of very uniform growth with minimal autodoping problems. Disadvantages of this system include mechanical complexity, low throughput, and susceptibility to particulate incorporation. The barrel reactor is an expanded version of the horizontal reactor in a different configuration. When used with a tilted susceptor, radiant-heated barrel reactors allow high-volume production and uniform growth.

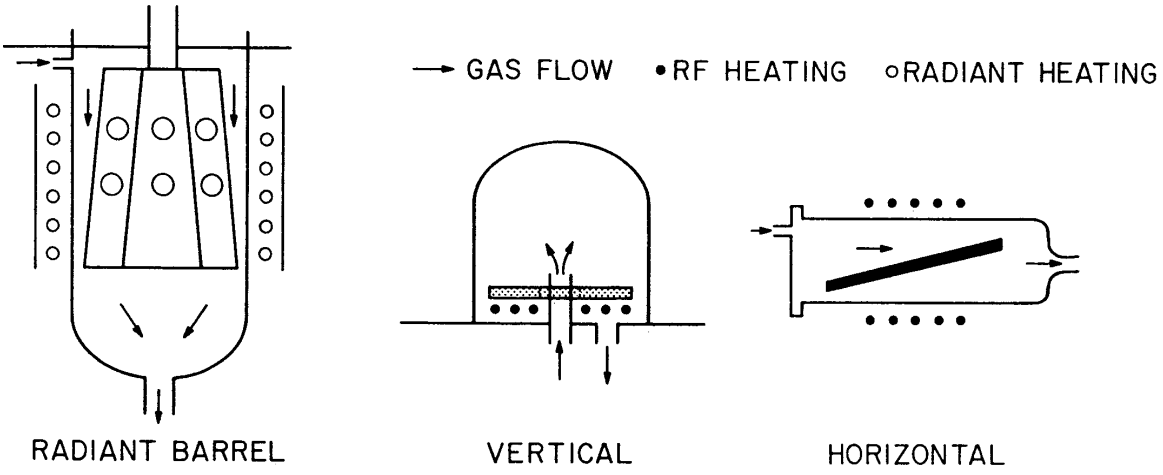

Figure 3.12: Schematics of three common reactors.

3.1.5 Defects

The crystal perfection of an epitaxial layer never exceeds that of the substrate and is frequently inferior. ***Figure 3.13*** depicts some of the common structural defects in an epitaxial layer. Generally, defects can be reduced by a higher growth temperature, reduced gas pressure, lower growth rate, and cleaner substrate surface. A typical pre-epitaxy substrate cleaning process consists of a wet clean followed by a dilute HF dip and an in-situ HCl, HF, or SF_6 vapor etch.

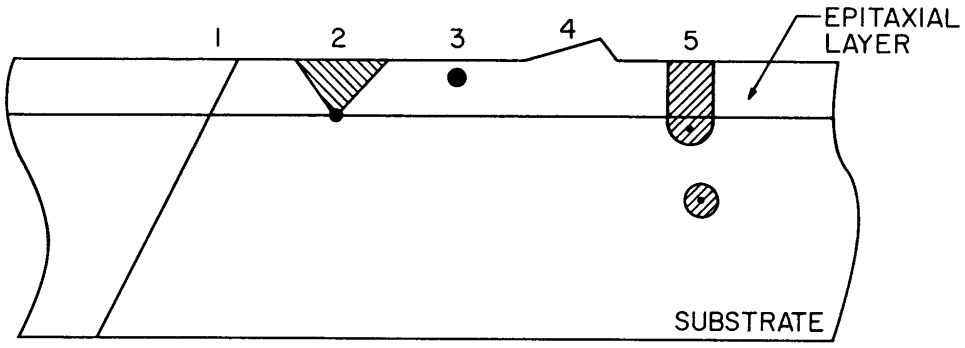

Figure 3.13: Schematic representation of common defects occurring in epitaxial layers: (1) line (or edge) dislocation initially present in the substrate and extending into the epitaxial layer, (2) epitaxial stacking fault nucleated by an impurity precipitate on the substrate surface, (3) impurity precipitate caused by epitaxial process contamination, (4) growth hillock, and (5) bulk stacking faults, one of which intersects the substrate surface, thereby being extended into the layer.

3.1.6 Selective Epitaxy Growth (SEG)

Selective epitaxy is a technique by which single-crystal silicon is fabricated in a small designated area as exemplified in ***Figure 3.14***. The process allows for the deposition of a Si epitaxial layer on a bare Si-substrate surface without the simultaneous growth of amorphous Si thin film on the silicon dioxide or silicon nitride surface. This is usually accomplished at reduced partial pressure of the reactant in order to suppress the nucleation of silicon on the dielectric film, thereby resulting in nucleation only on the exposed silicon surface.

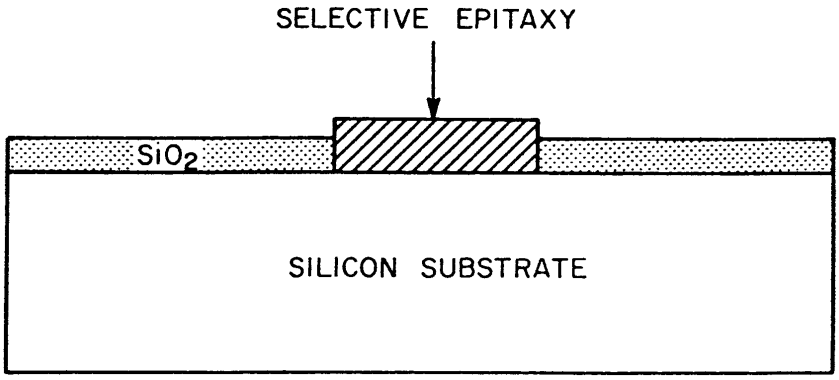

Figure 3.14: Cross-sectional schematic of a selective epitaxy process using an oxide mask.

3.2 Low-Temperature Epitaxy (LTE) and Molecular Beam Epitaxy (MBE)

Low-temperature epitaxy (LTE) of Si produces epitaxial growth at temperature of 550ºC or less, much lower than that in conventional epitaxial processes. A low temperature is required to minimize thermal diffusion and mass-transport-controlled processes. CVD and molecular beam epitaxy (MBE) are the most popular methods. The success of these techniques relies on both an ultra-clean growth environment and a unique Si surface-cleaning process.

Molecular beam epitaxy, which utilizes evaporation, is a non-CVD epitaxial growth process. MBE is therefore not complicated by boundary-layer transport effects, nor are there chemical reactions to consider. The essence of the process is evaporation of silicon and one or more dopants, as depicted in ***Figure 3.15***. Silicon MBE is performed under ultra-high vacuum (UHV) conditions of 10^{-8} to 10^{-10} Torr, where the mean free path of the atom is given by $5x10^{-3}/P$ where P is the system pressure in Torr. At a typical pressure of 10^{-9} Torr, L is $5x10^{6}$ cm, transport velocity is dominated by thermal energy effects. The lack of intermediate reactions and diffusion effects, coupled with relatively high thermal velocities, results in film properties changing rapidly with any change of the source. The typical growth temperature is between 400ºC and 800ºC in order to reduce out-diffusion and autodoping. Growth rates are in the range of 0.01 to 0.3 μm/minute.

Despite the slow growth rate and relatively expensive instrumentation, MBE offers several advantages over conventional CVD for VLSI. For example, MBE is a low-temperature process that minimizes dopant diffusion and autodoping. Moreover, MBE allows more precise control of doping and layer thickness, because CVD is limited by reactant introduction and pumping time constants. Presently, these advantages are not exploited extensively in silicon IC technology, but MBE has found tremendous usages in microwave and photonic devices made of III-V semiconductors.

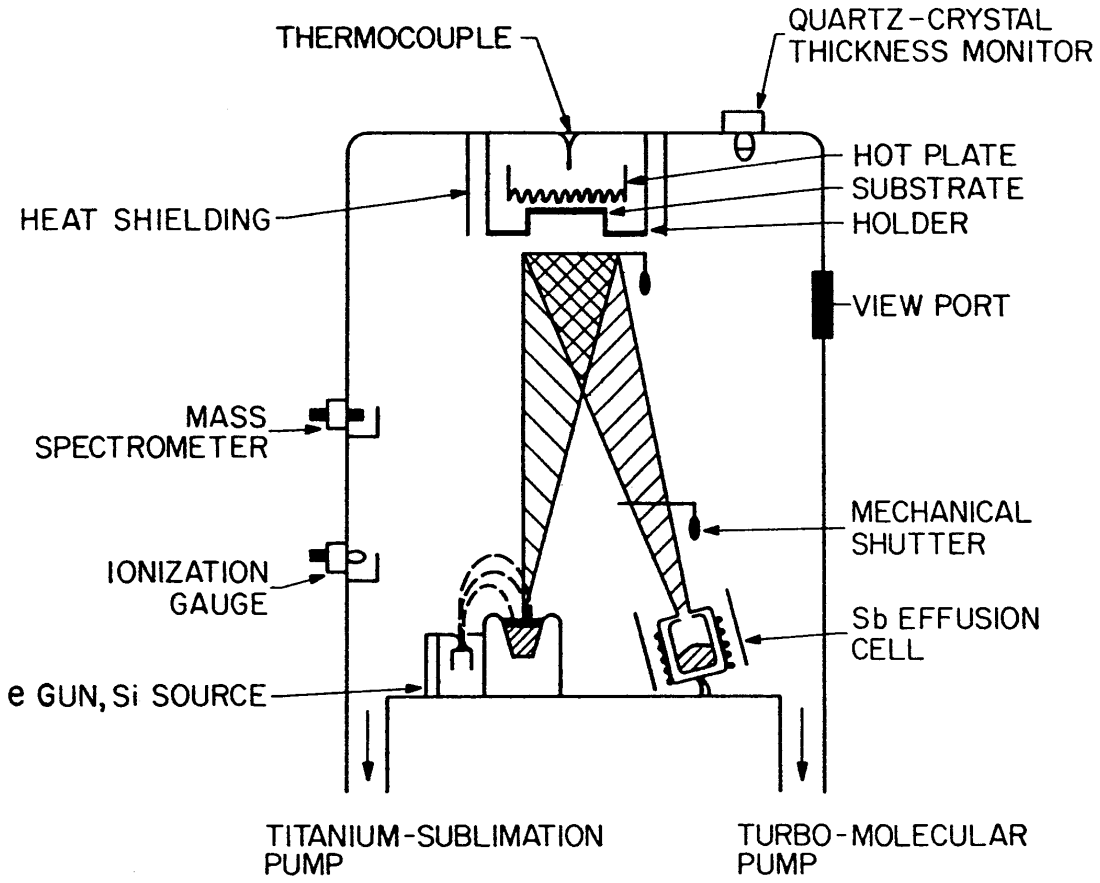

Figure 3.15: Schematic of MBE growth system.

3.3 Rapid Thermal Processing (RTP)

Chemical and physical processes applied to silicon wafers are generally thermally activated. Typical silicon-based processes use batch furnaces for thermal fabrication steps, where a batch consists of 20 to 100 wafers that are simultaneously processed in a single system. Processing of wafers requires tight control of contamination, process parameters, and reduced manufacturing costs, and some producers are now using single-wafer processing in some steps. In rapid thermal processing (RTP) using typically transient lamp heating (***Figure 3.16***) or a continuous heat source, vertical furnace (***Figure 3.17***), a single wafer can be heated very quickly to reduce the thermal cycle and mitigate undesirable effects such as dopant diffusion.

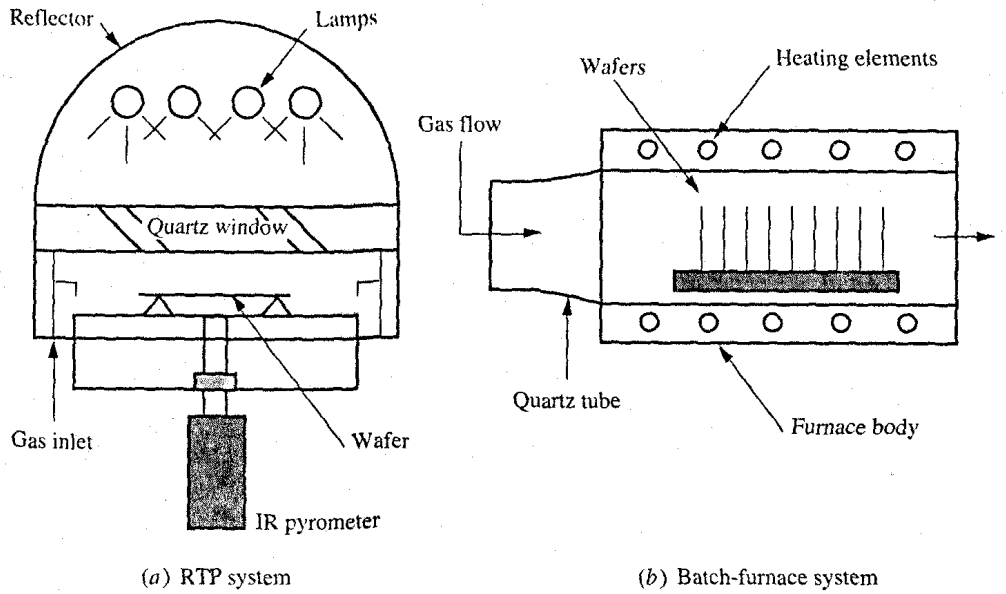

Figure 3.16: (a) Rapid thermal processing (RTP) system that is optically heated, and (b) batch-furnace that is resistively heated.

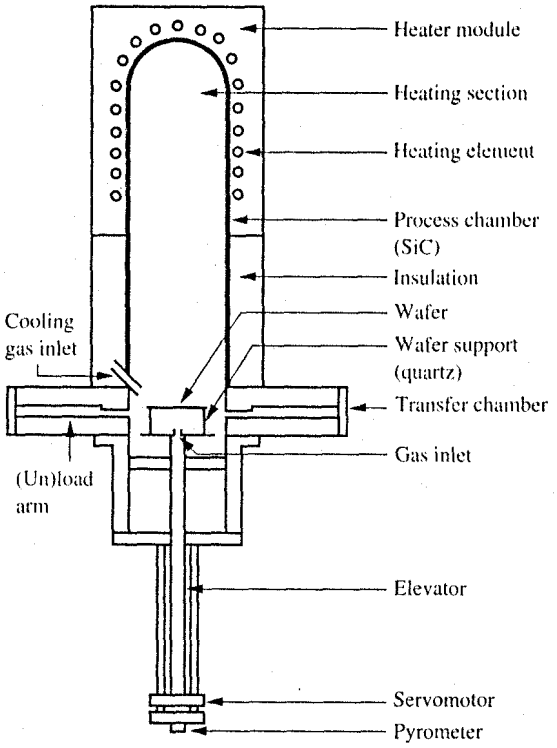

Figure 3.17: Schematic of a continuous heat source, vertical furnace RTP system.

The most important feature of a rapid thermal annealing processing system consisting of tungsten-halogen lamps is its generation and quick delivery of radiant energy to the wafer (large dT/dt) in a wavelength band of 0.3 to 4.0 μm. Because of the optical character and wavelength of the energy transfer, the quartz walls do not absorb light efficiently, whereas the silicon wafer does. Thus, the wafer is not in thermal equilibrium with the cold walls of the system, thereby allowing for short processing times (seconds to minutes) compared to minutes to hours for conventional furnaces. The reduction in temperature-time exposure afforded by RTP is dramatic. However, rapid heating with large temperature gradients can cause wafer damage in the form of slip dislocations induced by thermal stress and heating can be laterally non-uniform across the wafer. On the other hand, conventional furnace processes bring with them significant problems such as particle generation from the hot walls, limited ambient control in an open system, and a large thermal mass that restricts controlled heating times to tens of minutes. Requirements on contamination, process control, cost, and space are driving a paradigm shift to RTP.

RTP processing demands on the growth of high-purity epitaxial Si include ambient purity (oxygen and water concentrations in the parts per billion range), optimization of gas flow patterns, minimum wall deposition, and vacuum compatibility. An example of a low-pressure, epitaxial Si reactor based on RTP technology is exhibited in ***Figure 3.18***. The deposition process comprises a mass-transport process with a weak temperature dependence and a sequential surface-reaction process that is exponentially dependent on wafer temperature.

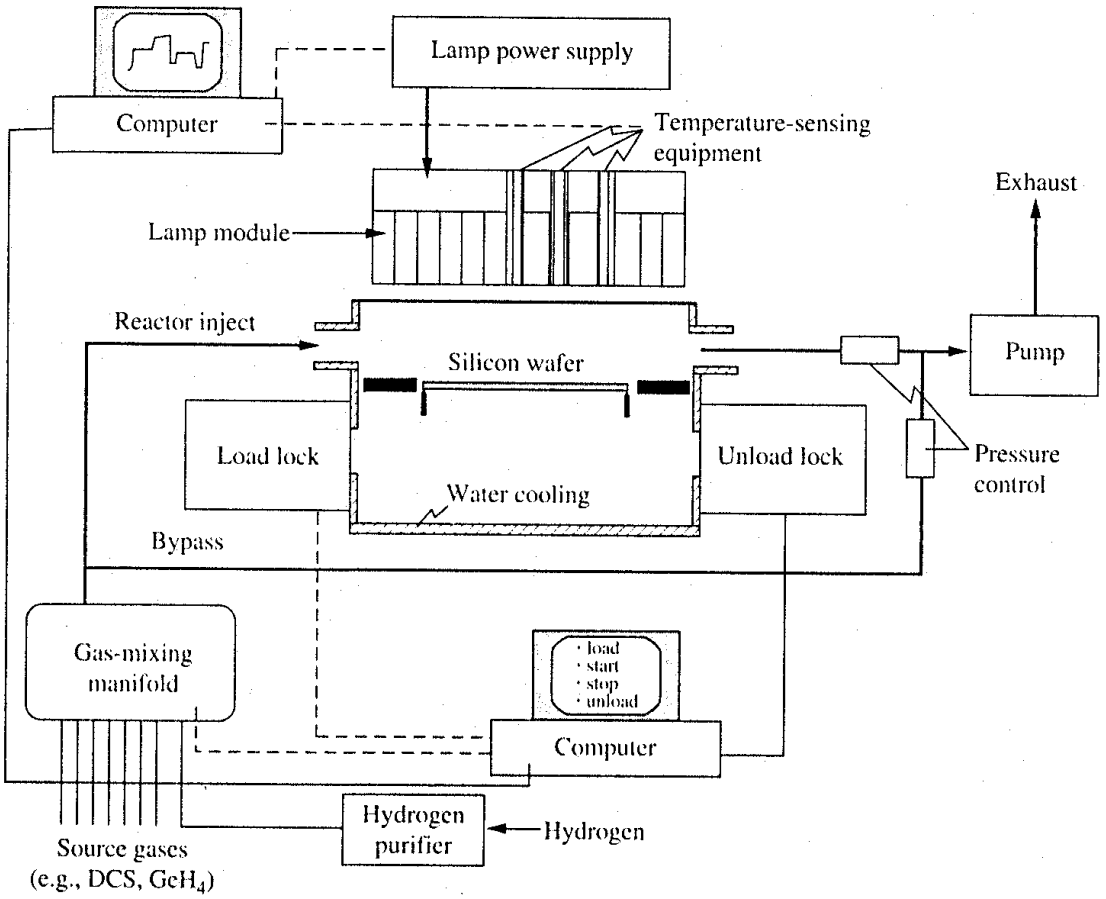

Figure 3.18: Diagram of a low-pressure RTP epitaxial Si reactor.

3.4 Silicon-on-insulator (SOI)

Silicon device structures have inherent problems that are associated with parasitic circuit elements arising from junction capacitance. These effects become more severe as device dimensions shrink. A viable means to circumvent the problem is to fabricate devices in small islands of silicon on an insulating substrate as shown in ***Figure 3.19***. The traditional approach is to fabricate such a structure in a silicon epitaxial thin film grown on sapphire (Al_2O_3). Since the lattice parameters of silicon and sapphire are quite similar, high quality SOS (silicon-on-sapphire) epitaxial layers can be fabricated. However, the high cost of sapphire substrates, low yield, and lack of commercially viable applications limit the use of SOS to primarily military applications.

There are several alternative silicon-on-insulator (SOI) approaches. SIMOX (separation by implantation of oxygen) utilizes high dose blanket oxygen ion implantation to form a sandwiched buried oxide layer to isolate devices from the wafer substrate. Another interesting approach is wafer bonding. This process utilizes Van der Waals forces to bond two polished silicon wafers, at least one of which is covered with thermal oxide, in a very clean environment at about 1000°C. Mechanical or electrochemical thinning has achieved 1 μm thickness with 0.1 μm deviations. More recent approaches include the combination of wafer bonding and layer cleavage using hydrogen or helium ion implantation (ion-cut) as well as epitaxial growth on porous silicon and wafer bonding.

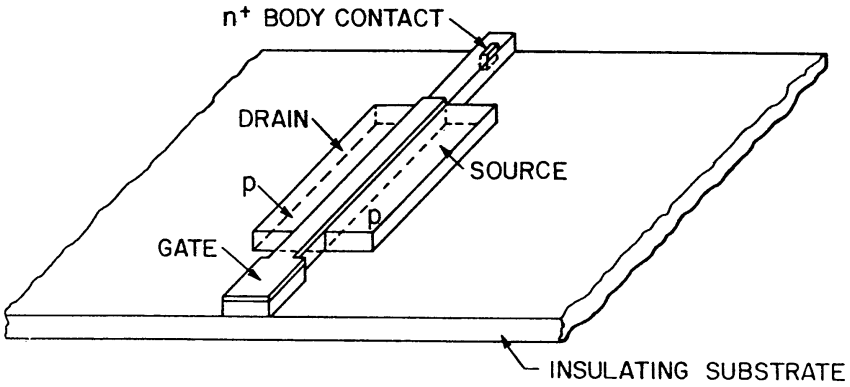

Figure 3.19: MOSFET device fabricated in a silicon island on sapphire substrate.