Dr D M Holburn and Prof M J Kelly

1 (a) and (b): diagrams and notes straight from lectures.

(c) The annealing process after deep implantation required very high time temperature profiles for the wafer, so that its must be performed first, and not destroy later processes.
Processes from the bottom up.
Drain engineering before metal and other overlayers are put on.
Local interconnects before some passivation and global interconnects.

(d) Previous technology ends in transport have been when cost and technology difficulties have coincided. This likely to happen here.
Here the physics of the operation of transistors with few electrons presents real difficulties with fan out.
The control of the variability of transistor parameters gets more difficult below about 22nm gate-length.
The cost of manufacturing equipment grows exponentially as the size shrinks.

The present recession is giving rise to good enough computing, and low cost is the driver rather than greater computing power, much of which is not used in daily applications. This may interfere with the investment/return cycles of the manufacturers.

Once the limit of physical smallness is reached, the concern will shift to a better use of the silicon real estate, with architecture playing a greater role.
Even more important, some of the long-lost disciplines of memory-conscious software writing will be rediscovered.

2 (a) Critical success factors in manufacturing

(i) In house quality to be able to meet customer specifications
(ii) Cost: in a competitive environment the cost-performance trade-off is optimised.
(iii) Delivery: rapid response is key to a competitive edge.
(iv) Service: flexibility, troubleshooting, product selection etc.

(b) Layout:

(i) Relationship to connection to utilities (DI water, chemicals, gases, pumps, ...)
(ii) Wafer movement path – minimise movement between process equipments and inspection stations to minimise time and defects
(iii) Routing of supplies to each step – dummy wafers, quartzware
(iv) Inventory storage – particularly the case when different batches are going through a line
(v) Cross-contamination phosphorous, and metal/non-metal processes well segregated.

(c) Metrology and analysis in-line.
Thicknesses of layers: ellipsometers, alpha step, ...
Dimensions in x-y plane: SEM
Electrical: resistance, carrier lifetime, C-V
Particle./defects: particle counters, microscopic inspection
Others: stress, FTIR for boron and phosphorous content, stepper alignment

In-line process control and improvement always essential.
Used for problem solving
Need to check incoming quality.

Off-line: there is a bewildering array of techniques that can deployed off-line as a means of verifying improvements and continuity
Spectrometers to check the presence of trace elements in liquids, and in gases
Moisture measurements,
X-rays for structure and materials
Device testers at AC/DC
High bias and elevated temperature geat for rapid testing of potential failures.
Laser spot size for feature size checking.

(d) Calibration at start-up and afterwards

All processes must be tightly specified with clear tolerances.
Equipment capable of delivering to these specifications must be used.
The materials must be qualified on input
A method of learning from past mistakes must be instituted.
All utilities must be tested and confirmed to be within specification, and capable of being run as an ultra-stable system
Each item of equipment must be checked, its cleanliness established, its integrity confirmed,
Each process has to be fully characterised and established, with the variations being fully understood and to be within tolerances
The final product must be characterised in terms of parameters uniform, reproducible and in spec.

The product will have to be proved reliable – i.e. do what it does repeatedly through accelerated life tests.
Finally, the product will need to be qualified internally before release. It will be separately qualified by the customer on receipt and a common understanding of these processes is essential to prevent claims and disputes.

3. Based on lecture notes. A good response should include the following points.

Design rules allow a ready translation from schematic circuit to actual geometry in silicon. They are the effective interface between the circuit/system designer and the fab/process engineer. They provide a workable and reliable compromise which is friendly to both sides.

The designer is concerned to achieve:

- best possible electrical performance – speed, noise margins, linearity, gain
- minimum area of Si per circuit – lower costs, better yield, reliability

The process engineer on the other hand seeks:

- to maximise tolerances on all 'parts' - easier fabrication, better yield

There are 3 basic tolerances that set limits to the shapes that the designer can specify:

- dimensional resolution governed by λ of light used in lithography, photoresist characteristics, ...
- alignment errors – registration, temperature variation, bowing/distortion
- reproducibility of processing – wet etching, plasma, layer thickness control

For practical purposes all three effects can be reduced to linear dimensions on a plan view of the mask layout. The permissible dimensions are often highly specific to a manufacturer's process.

The simplest rules originate from the need for continuity of layers and for avoidance of unintended short-circuits. Layers such as polySi, metal and diffusion are associated with

- minimum dimensions and
- minimum separations

They may also be associated with ohmic resistance (electrical origin). Violation of these rules may lead – as in PWB technology – to open-circuits in conducting traces, or short-circuits, where tracks are too close.

With metal Al interconnect it is necessary to ensure that the current density does not exceed about 10^9 Am-2 otherwise there is risk of **electromigration**, which is iduced by transfer of momentum from the electronics carriers to metal atoms, and causes progressive thinning of interconnect at current bottlenecks – e.g. as metal crosses a step. Interconnect width is hence governed by the anticipated peak (rather than mean) current and not simply by lithographic considerations.

Plasma processing involves application of high energy RF fields that can induce high voltages in previously fabricated circuit interconnections, which act as 'antennas'. These so-called **process-induced gate-oxide damage** instances gives rise to the so-called 'antenna rules'. The fields can be enough to cause breakdown in gate oxide and other fragile structures. The risk is greatest when extensive metal interconnect is coupled to a transistor gate (e.g. long clock line), but is reduced when p-n junctions (source/drain diffusions are also connected, since these assist conducting excess energy to ground. Careful design strategies to ensure the area exposed is not too great taking into account the fragility of the coupled structures.

Since fab involves several sequentially masked steps there is a need to accommodate the possibility of misregistration between successive masks. For this reason

- implant masks overlap the active areas/diffusions to which they correspond, by a significant specified amount
- polysi gates extend beyond the edge of the underlying diffusion
- metal, diffusion and polysi are required to surround contact cuts and vias by a significant margin

Not that the inception of self-aligned processes has greatly relaxed the requirements for registration of implant masks. In one example the polySi gate acts itself as a mask for implantation, guaranteeing correct positioning of those implants relative to the gate.

It is possible to define an 'alignment tree' which summarises the statistical probability of misregistration between related mask layers.

The use of metal rather than polySi is dictated for

- power distribution
- signal transmission over long distances – e.g. clock lines, to avoid skew

owing to its lower resistivity/sheet resistance, and (to a lesser degree) lower C.

Where significant currents are transmitted from one metal layer to another, or to transistor diffusion, the **contact structure** must be capable of carrying the current. For lithographical reasons, a minimum contact cut lateral dimension is mandated; but since contact conductance is proportional to the cut perimeter (not area) this is achieved through use of many minimum-geometry cuts, filling the available space. A maximum current is often associated with a min-geom cut or via.

For **power pads**, a substantial rectangle of superposed aluminium interconnect layers is required, robust enough and dimensions sufficient to allow a fine gold wire to be ultrasonically bonded to it during packaging. Bond pads have to be around 80-100 microns square, irrespective of the process geometry. Interconnect linked to them has to be dimensioned according to contact/via considerations, and bearing in mind electromigration, and to mimimise series resistance and inductance.

Minimum geometry transistors may be undesirable where high voltage immunity is required – e.g. output or input pads.

A number of precautions are required to alleviate the risk of latch-up. As far as design rules are concerned, this calls for use of **well and substrate taps** at a specified maximum pitch and frequency, and in close proximity to power devices (for example, I/O pads).

4 (a). Notes straight from lectures.

(i) EM (Electromigration)
(ii) Corrosion
(iii) HCI (Hot Carrier Degradation)
(iv) TDDB (Time dependent Dielectric Breakdown)
(v) ESD (Electrostatic discharge)

(i) EM (electro-migration): EM is caused by migration of metal atoms of a conductor that have acquired momentum from the electrons passing through the conductor.
The scenario is: high current density (J > ~10e6 A/cm2), momentum transfer between the electrons and metal atoms, migration of metal atoms, void & hillock formation, open & short circuit failure
EM wear-out mechanism: (i) thermal: high temperature (major factor), temp gradients, (ii) electrical: high current density(major factor), current gradients, (iii) metal characteristic: grain size, metal length, metal width, metal additives.

(ii) Corrosion: Thin and small structures, operating and condition of high electrical and thermal stress are susceptible to any forms of attack by the atmosphere or liquids, especially through any cracks in encapsulation induced by wear and tear.

(iii) HCI (Hot Carrier Degradation) HCI (Hot Carrier Degradation)
It is generated in MOSFET by the large channel electric fields near the Drain region.
It HVICs it can also appear in the field oxides or buried oxides (in the case of SOI).
Mechanism: Carriers (electrons or holes) that flow into the high electric field area are accelerated by the strong field and gain substantial energy. Some of the carriers have enough energy (that is to say they are hot) to overcome the electric potential barrier existing between the Si substrate and gate oxide film. These hot carriers are injected and subsequently trapped into the gate oxide film. They form a space charge or inversion layer and over a period of time they can affect the threshold voltage (Vth) , transconductance (gm) or specifically in high voltage VLSI circuits they can alter the field distribution resulting in breakdown failures.

(iv) However, oxide film failure over time even in lower electric-field intensity (conditions of practical use) is a major cause of failure. Destruction occurring over time is called TDDB (time dependent dielectric breakdown). The time-dependent destruction of the oxide film (dielectric film) is one of major causes of failure.
Gate oxides are generally affected by this but in HVICs we can also have the BOX or FOX or ILDs affected by this.
An electric field applied to an oxide film causes the injection of holes into the oxide film to occur and it consequently causes traps to be made in the oxide film. As the number of traps increases, an electric current via the traps is observed as an SILC (Stress Induced Leakage Current) due to hopping or tunneling. It has been reported that if the number of traps continues to increase and the traps connect between the high voltage and low voltage terminal, the connection carries a high current that causes the field oxide/ILD or gate oxide film to break down
Because the level of the traps (i.e. defect) in an oxide film strongly influences TDDB, it is necessary to characterize the oxide film quality with accelerated tests such as HTRB and feed the results into design rules.

(v) ESD (Electrostatic discharge) Devices can be damaged or destroyed by electrostatic discharge (ESD) due to local heating caused by the discharge current flowing in the device and/or by device breakdown caused by the electric field.
The Si and SiO2 used as the primary materials in semiconductor devices are able to withstand heat and voltage stress very well, but under ESD the current densities are so high that melting and dielectric breakdown occur.
Dielectric breakdown is caused by the voltage drop due to currents flowing through resistances and by direct voltage application to the dielectric film.

4 (b) Slew rate is determined from the available output current:

$$i = C_L \frac{dv_{out}}{dt}$$

The current needed to charge or discharge C_L at the given rate 3×10^6 Vs^{-1} is:

$$+/- 200 \times 10^{-12} \times 3 \times 10^6 \text{ A} \quad = \quad 0.6 \text{ mA}$$

We shall assume that M2, which has fixed bias, delivers this current continuously and all its current is available to charge C_L – i.e. M1 draws negligible current when this happens. We'll also assume that vout = 0 (the middle of the required range), ± 2 volts. Hence $V_{DS2} = -5$ V.

Considering M2, we see that provided Vout < +2 V, it remains in saturation at all times. Rearranging the SH equations given,

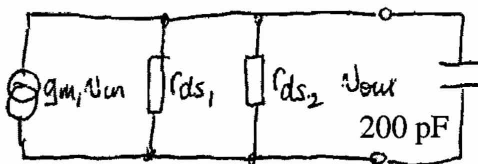
$$W_2 = L_2 \times \frac{i_{D2}}{\frac{1}{2}(V_{GS} - V_T)^2 (1 + \lambda V_{DS}) \mu\varepsilon / t_{ox}}$$

$$\frac{W_2}{L_2} = \frac{0.6 \times 10^{-3}}{\frac{1}{2}(2-1)^2 (1 + 0.02 \times 3) \times 6 \times 10^{-6}} = 47.2 \text{ and } W_2 = 472\ \mu\text{m}$$

Under quiescent conditions, M1 draws the same current; however, since M2 delivers 0.6 mA over the whole of the range while in saturation, M1 must be sized for twice this to allow C_L to be discharged at the required rate. When Vout = 0V, $V_{DS1} = 5$ V. M1 will stay in saturation provided $V_{GS1} < V_{DS1} + V_T$, or +6V. Under these conditions we can determine a suitable size for M1:

$$\frac{W_1}{L_1} = \frac{1.2 \times 10^{-3}}{\frac{1}{2}(5-1)^2 (1 + 0.01 \times 5) \times 15 \times 10^{-6}} = 9.5 \text{ and } W_1 = 95\ \mu\text{m}$$

To calculate the a.c. or small-signal gain, note the SSEC for the output:



Using the approximations:

$g_{ds} \sim I_D \lambda$ and

$$g_{m1} \sim \sqrt{2 \frac{\mu\varepsilon}{t_{ox}} \frac{W}{L} I_D}$$

We shall assume that I_{D1} is 0.6 mA quiescent

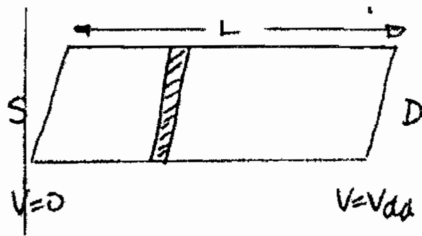
Hence $g_{m1} = \sqrt{2\times15\times10^{-6}\times9.5\times0.6\times10^{-3}} = 1.3\times10^{4}$

and $\dfrac{1}{g_{ds1}+g_{ds2}} = \dfrac{1}{0.6\times10^{-3}}\dfrac{1}{0.01+0.02} = \dfrac{1}{0.6\times0.03\times10^{-3}} = 55.5\ \text{k}\Omega$

By inspection of the SSEC, gain = $-\dfrac{g_{m1}}{g_{ds1}+g_{ds2}} = -1.3\times10^{-4}\times55\times10^{3} = -7.15$

and the small-signal output impedance = 55 kΩ.

5.



Consider the shaded channel element near the source.

Let Vgate = Vdd to make the device conduct

Let the charge density in the shaded element be Q per unit length

Q	=	C V W	V is strictly the excess of voltage above VT
	~	Cox Vdd W	where Cox is the oxide capacitance pua
Current I	=	Q μn E	where the field E is assumed invariant along
	~	Q μn Vdd/L	

Hence conductance

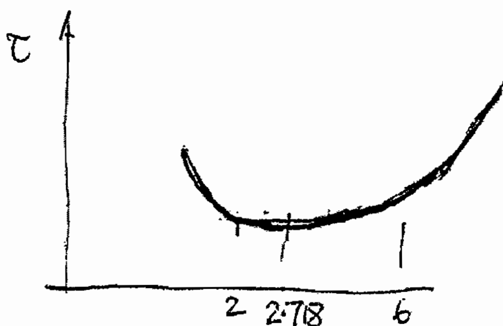
G = I/V = Cox Vdd μn W/L

(b) To drive a high C load it is necessary in order to minimise delay, to use stages of progressively increasing W/L. Later devices are able to conduct higher current to charge/discharge nodal capacitances, which are themselves bigger because of the use of larger devices.

It can be shown that the optimum number of stages to mimimise delay is ln (Cpad/Cgate), where Cpad is the pad capacitance (including external driven capacitance, and Cgate is the capacitance at the input to the pad driver.

Hence for minimum delay, for the values given, the number of stages would be:

ln (75/0.2) = ln 375 = 5.9 or 6, rounding up. This suggests 6 stages each a magnification factor U = e = 2.718 larger than the previous one. Other values of U may be advantageous and could result in reduced area, but strictly, such alternatives do not give minimum overall gate delay.



A graph of gate delay versus U has its minimum at U=2.718, but is fairly flat between about 2 and 6.

Hence larger values of U, say 4-6, give only only slightly longer delays but may reduce the number of stages and significantly reduce the total pad driver area.

This optimisation is not required here.

Note that the total of 6 stages includes the minimum geometry gate generating the signal. Hence the driver itself consists of 5 further inverter stages.

If channel length is maintained at the mimimum dimension, 0.5 μm, successive stages are designed with channel widths inflated by 2.718. Hence those stages drive capacitances inflated by that same factor (assuming that gate and pad capacitances dominate e.g. interconnect), so the delay remains constant in each stage.

We shall also assume that the stages are designed for symmetrical rising & falling delays, meaning that the p-channel devices are a factor $\mu n/\mu p = 2$ larger than the n-channel devices.

The delay in the first min-geom inverter, if connected directly to the 75 pF pad is:

$$\tau\text{direct} = \frac{3\times75\times10^{-12}}{10^{-4}\times2\times3} = 375\text{ nS}$$

Using the graded drivers, the delay observed in the min geom. stage – and in each subsequent stage – is

$$\tau 1 = \frac{3\times2.718\times0.2\times10^{-12}}{10^{-4}\times2\times3} = 2.7\text{ ns}$$

Hence the delay in the remaining 5 stages is $5 \times 3.7 =$ 13.5 ns

This is the characteristic minimum delay for the driver