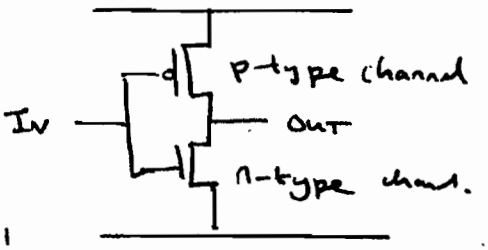
1.

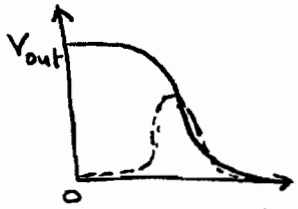
4B7 Q1

The key to CMOS technology is the low power consumption. Since only the n-channel or the p-channel device is on except during the switching transient.



Inverter device

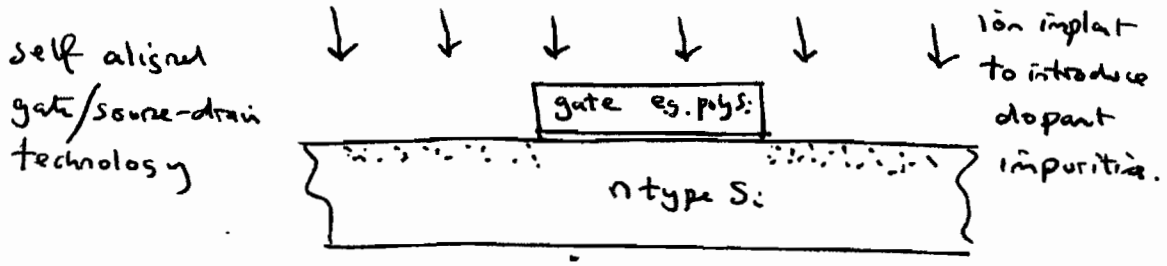
IDm 1/2006 *



input-output characteristic

In silicon technology the thermal oxidation gate fabrication process produces uniform high quality gate dielectric and 10^7 devices can be reliably produced on one chip.

The clear input-output switching characteristic gives wide margins of operation compared with competing technologies allowing for some variation in device switching threshold across a VLSI chip.



The finest available lithographic technology and reactive ion etching is used to define the device gate structure and good alignment to the doped source and drain regions is achieved by using the polycrystalline silicon gate as a mask.

After annealing to thermally activate the implanted impurity atoms the small gate/source overlap ensure good device switching performance

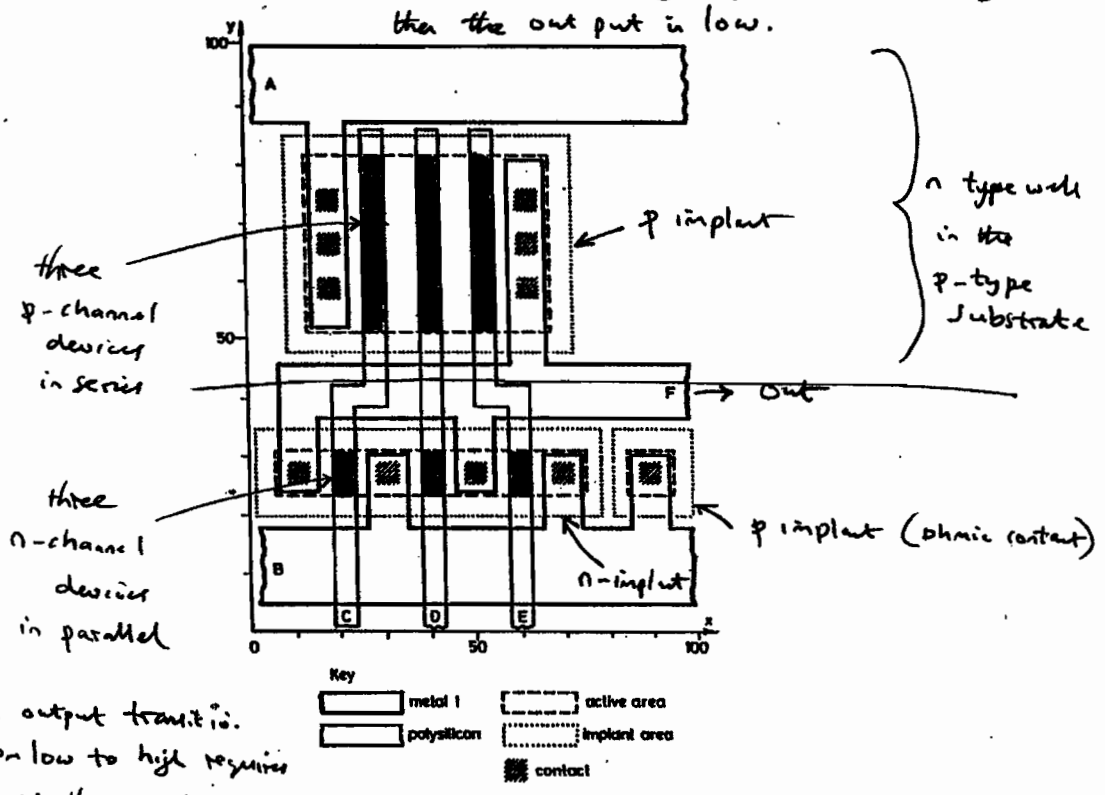
Device miniaturisation down to 100 nm source-drain separation is being achieved because of advances in the lithographic processes. Possible future limits in the 30 nm range may be due to the cost of lithography equipment or due to small channel and statistical variations

1 (cont)

4B7 Q1 (cont.)
(b)

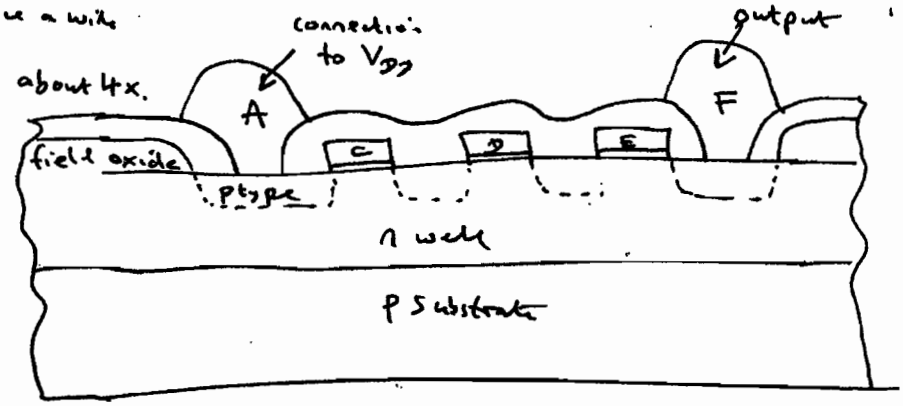
Three input NOR function. Dfm 1/2006 *

If at least one of C, D or E is high then the output is low.



The output transition from low to high requires current through the three p-devices which have lower mobility than the n-devices – hence a wide p-device is used about 4x.

CROSS SECTION.

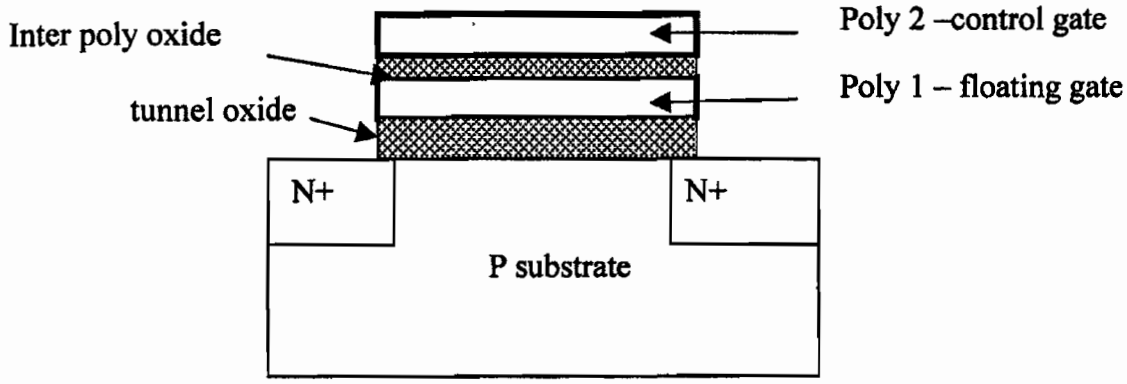


The three gates in series have self-aligned source drain implants

2.

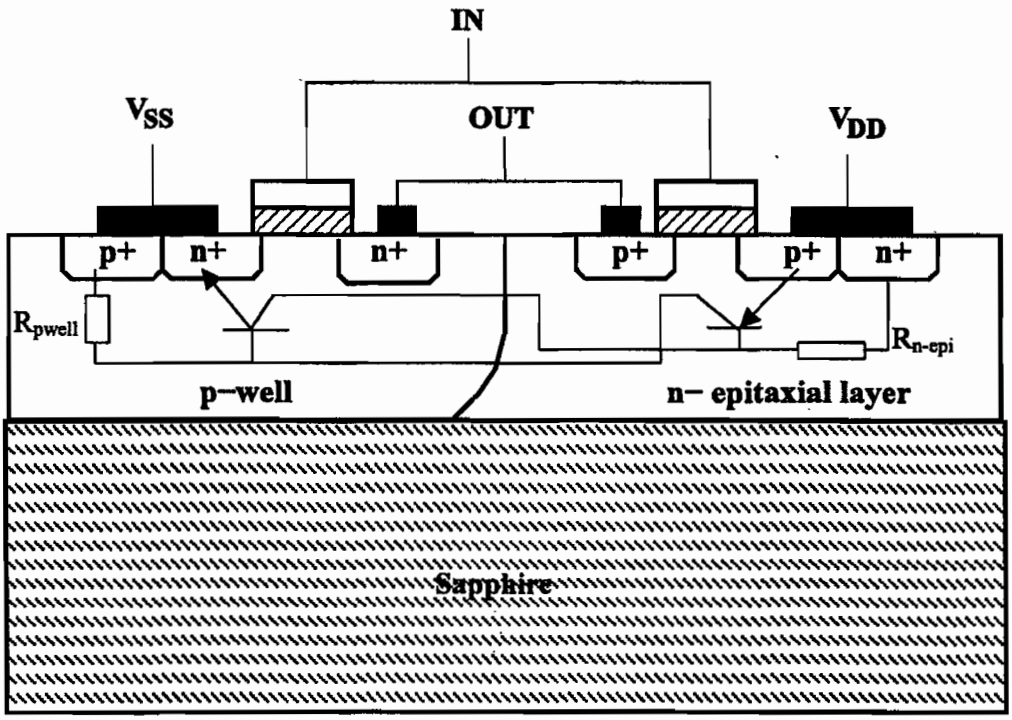
2. (a) The bulk resistivity of copper is about 1.67ohm-cm. That is almost half of that of Aluminium. Thus, the sheet resistance of a layer of Copper would be half of that of Al with the same thickness. In addition Copper has superior resistance to electromigration and thus can allow higher current densities than Al. This means that layers of metal interconnect. which in practical designs occupy a considerable area, can be made narrower, while still being more conductive. Copper however cannot be patterned and etched using a standard process. *Damascene* process is the solution for advanced interconnection. The process eliminates the need for metal etch and dielectric gap fill that becomes very challenging as dimensions continue to shrink. The *Damascene process* is based on defining a trench pattern; etching through the dielectric material, depositing copper on the surface through electroplating, and planarising the surface through CMP (Chemical Mechanical Polishing).

Good quality capacitors can be obtained by using an additional layer of polysilicon. The polysilicon layers need not be separated by a thick layer of dielectric (like subsequent layers of metals) and therefore a very thin layer of oxide can be used between the two poly layers to make large capacitors occupying a very small area. These are useful both in analogue circuits where RC filters or switched capacitors are used (like in RF) or in dynamic memory cells.
Reprogrammable memories (EEPROM) can also be made using two polysilicon layers separated by a thin inter-poly oxide.



(b) The twin-tub technology is based on a highly conductive substrate onto which an epitaxial layer is grown. This is followed by the n-well and p-well formation. The values of the parasitic p-well and n-well resistors which are placed between the base and emitter of the parasitic NPN and PNP transistor respectively are considerably reduced, thus providing a more effective short to the base-emitter junction. The technology also provides an effective sink (conductive substarte) for collecting the hole current of the parasitic pnp transistor.

3. (cont)



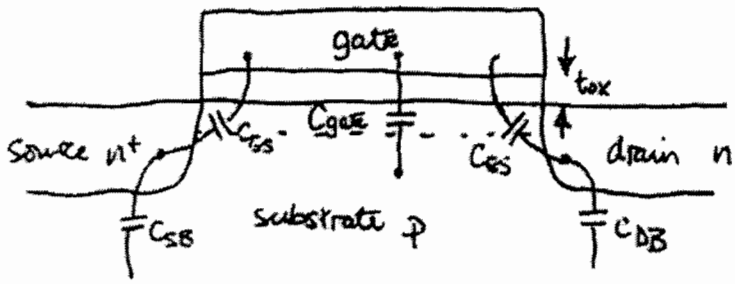
- NPN transistor
  - Emitter: n+ source of the n-channel MOSFET
  - Base: p-well
  - Collector: n-epi layer
- PNP transistor
  - Emitter: p+ source of the p-channel MOSFET
  - Base: n-epi
  - Collector: p-well

The 'latch-up' parasitic structure is comprised of two bipolar transistors (NPN & PNP) in a thyristor configuration. The PNP transistor is more likely to be turned-on (leading eventually to latch-up) because the $R_{n\text{-}epi} > R_{pwell}$ and because the current gain of the NPB transistor α_{NPN} is greater than that of the PNP transistor, α_{PNP}.

The latch-up condition can be written as :

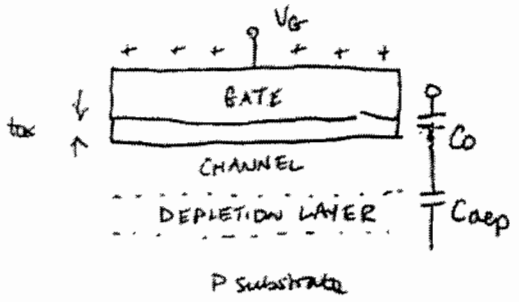
$I_{trigger} = V_{pnp\text{-}on} / (\alpha_{NPN} \cdot R_{n\text{-}epi}) = 0.7\ V / (\alpha_{NPN} \cdot R_{n\text{-}epi})$

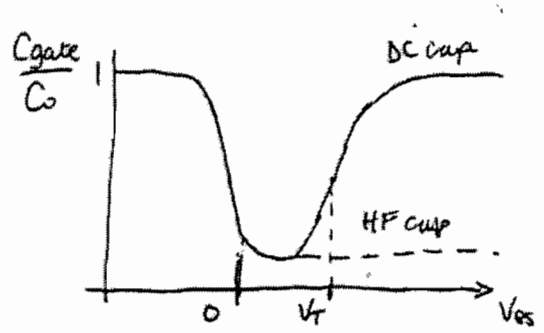
3



n-channel enhancement mode MOSFET

(i) C_{SB}, C_{DB} are source and drain « diffusion capacitances to substrate caused by formation of p-n junctions at drain-substrate and source-substrate interfaces
For each of these two components can be distinguished:
(a) an area-dependent component proportional to the plan-view area of the source/drain
(b) a peripheral component, due to the side-walls of the source/drain, proportional to the perimeter of the 'diffusion'
As device areas shrink, being $\propto$ to L^2, the peripheral component is of increasing significance ($\propto$ L). Both depend on junction bias, doping profile etc.
(ii) C_{GS}, C_{GD} are gate-source & gate-drain capacitances due to proximity of these electrodes and to process-dependent overlaps
(iii) Cgate is a parallel-plate capacitance between gate and substrate. This depends on floor area, but is strongly dependent on gate potential and whether or not a channel has been formed.





When $V_G \prec 0$, an accumulation layer is formed as the gate attracts holes. The structure behaves like a capacitor of dielectric thickness t_{ox} and of capacitance Co where Co $\propto$ (ε/t_{ox}) × gate area.

As V_G is raised above 0, holes are repelled leaving a region depleted of carriers. The depth of the depletion region, d, increases as V_G increases. This gives rise to a capacitance C_{dep} in series with Co, causing a reduction in measured gate capacitance.

As V_G approaches V_T (threshold voltage) surface inversion gives a relatively high conductivity layer (the channel) restoring the capacitance progressively back towards Co. Only at very high frequencies, or in the absence of a nearby source/drain (which provide carriers), will the channel be unable to form sufficiently rapidly, and a lower capacitance (dashed line) will be observed.

For C_{GD}: in CMOS gates, which are intrinsically inverting structures, as the input swings, the output swings in the opposite direction and the large signal gain is effectively about –1. The opposing swing of V_G and V_D causes an increase in the apparent capacitance being driven at both gate and drain owing to Miller effect. To account for this the static value for C_{GD} is typically doubled.

Total gate capacitance is thus

$Cg = Cgate + C_{GS} + 2\,C_{GD}$

The polysilicon gate electrode may also serve as short-range interconnect, where it is not superimposed on the channel, the specific capacitance is much lower, and it is not much affected by potential.

Total drain capacitance or source capacitance is the sum of the area and peripheral components for each. Metal interconnect also contributes capacitance, and other inter-layer capacitances (e.g. between adjacent signal interconnects) may also be identified.

Numerical part. We consider only those capacitances that are driven with signals. Hence the V_{DD}/V_{SS} lines are not evaluated. The two transistor channels have the same dimensions. Hence:

Cinput = Cpoly-substr + 2 × Cgate-substr + (C_{GS} + 2 × C_{GD}) × 2
Coutput = Cmetal-substr + 2 × C_D-substr + (2 × C_{DG}) × 2

The factors of 2 in the brackets arise from Miller effect. Another factor of 2 comes from the two identical MOSFETs which have gates (input) and drains (output) connected together. For Cmetal-substr and Cpoly-substr there is an *area* and a *peripheral* component.

Input: consider first the poly not over active, then the gates themselves::

Apoly = (50 – 8) × 2	=	84×10^{-12} m^2
Ppoly = (50 – 8) × 2 + 4 × 2	=	92×10^{-6} m
Agate = (4 × 2)	=	8×10^{-12} m^2
Pgate = 4 + 4	=	8×10^{-6} m^2

For Pgate we consider only the part over the channel, since this is where the gate-drain and gate-source overlaps occur. Exactly half the length is associated with the drain, half with the source. The part of the gate perimeter at the edge of the channel is accounted for in Cpoly-substr.

Cpoly-substr	$= 84 \times 10^{-12} \times 4 \times 10^{-5} + 92 \times 10^{-6} \times 5 \times 10^{-11} =$		7.96 fF
Cgate-substr	$= 8 \times 10^{-12} \times 7 \times 10^{-4}$	=	5.6 fF
C_{GS}	$= 4 \times 10^{-6} \times 3 \times 10^{-10}$	=	1.2 fF
C_{GD}	$= 4 \times 10^{-6} \times 3 \times 10^{-10}$	=	1.2 fF
Hence Cinput	$= 7.96 + 2 \times 5.6 + (1.2 + 2 \times 1.2) \times 2 =$		26.4 fF

Output: consider first the metal interconnect, then the drain diffusion (single device), and we assume the gate is centred on the active region. C_{DG} was dealt with. above.

Amet	$= (80 - 8) \times 4$	=	288×10^{-12} m^2
Pmet	$= (80 - 8) \times 2 + 2 \times 4$	=	152×10^{-6} m
A_D	$= 7 \times 4$	=	28×10^{-12} m^2
P_D	$= (7 + 4) \times 2$	=	22×10^{-6} m

Hence Cmetal-sub	$= 288 \times 10^{-12} \times 3 \times 10^{-5} + 152 \times 10^{-6} \times 4 \times 10^{-11}$	=	14.7 fF
Cdrain-sub	$= 28 \times 10^{-12} \times 10^{-4} + 22 \times 10^{-6} \times 4 \times 10^{-10}$	=	11.6 fF
Hence Coutput	$= 14.7 + 2 \times 11.6 + 2 \times (2 \times 1.2)$	=	42.7 fF

C_D-subst is expected to fall as V_D rises and the degree of reverse bias increases. Metal and poly-substr capacitances are substantially constant.
Cgate varies as per the discussion above.

4. **Clock Skew**. In many VLSI systems operations are synchronised to a Master Clock. This might be generated by on-chip circuitry or introduced from outside via an input pad. The clock is distributed to all parts of the circuit by means of interconnect, which may be as long as 1-2 chip diameters
The interconnect introduces R-C delay (and the L element also introduces distortion). Hence different destinations on the chip receive clock signals delayed by different amounts. These different delays relative to the master clock are called CLOCK SKEW. Clock skew can therefore arise from the following:

- different lengths and types of interconnect between the master generator and locations where the clock is used
- passage through different numbers / configurations of control gates
- the need for extra inverters to form $\overline{\varphi}$, e.g. for transmission gates

In design of sequential circuits, designers need to specify a minimum HOLD time to guarantee proper latching of data to alleviate the effects of clock skew

Clock skew is reduced by:

- keeping interconnect paths short and direct
- avoid the use of high resistivity conductors (e.g. polySi) for all but the shortest interconnect runs
- split clock lines into short segments separated by buffers
- user of pipe-lining – an enabled T-gate may be placed in series with a signal to compensate for delays in other paths
- use of silicide, copper, to minimise inserted resistance
- use of organoSi glass or SoI to minimise capacitance to substrate and hence the delay

Numerical Part

The given formula, $T = \frac{rcl^2}{2}$, is in terms of R/unit length, C/unit length

Alternatively, $T = \frac{1}{2}RC$, where R = total resistance, C = total capacitance

Total resistance 10 mm $= \frac{10\times10^3}{2}\times 40 = 2\times10^5\,\Omega$

Total capacitance 10 mm $= 2\times10^{-10}\times10^{-2} = 2\,\text{pF}$

Hence for 10 mm trace, $T = \frac{1}{2}\times 2\times10^5\times2\times10^{-12} = 200$ ns

(i) Using a double width interconnect is expected to double the conductance, and to increase the capacitance. The capacitance will not double, since there are both area and peripheral components, of which the peripheral component scarcely changes with the doubled width. This is more noticeable with smaller geometries. Hence the delay falls slightly, since T $\propto$ RC.

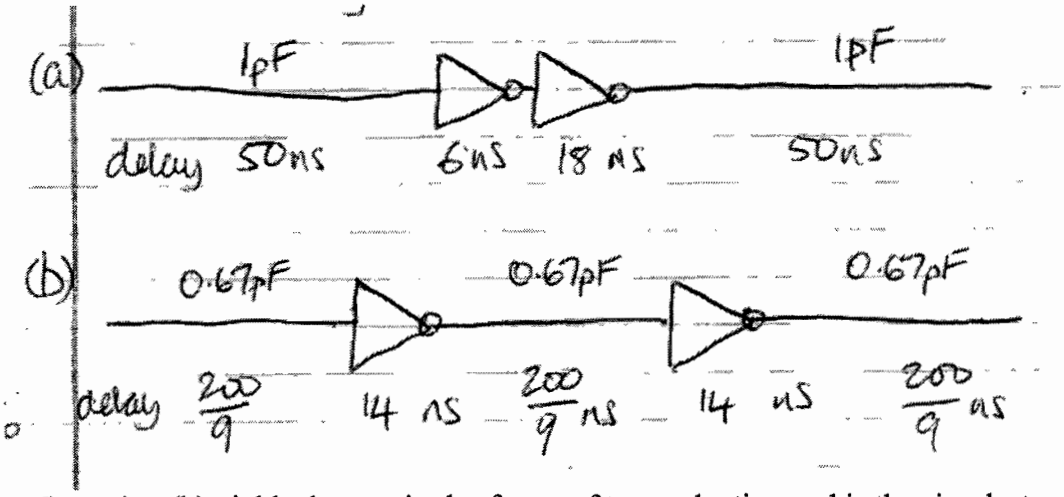
(ii) Using the silicide reduces the resistance by a factor 40/4 = 10, but leaves the capacitance effectively unchanged since there is no change in geometry. Hence T is reduced by a factor 10, to 20 ns.

The use of buffers to separate the clock line into shorter segments *may* reduce the overall delay, but this depends on the relative delays arising from the increased number of shorter segments and from the buffers driving the capacitance of the line.

A number of configurations are worth considering. They must have an even number of gates, since a non-inverted clock is mandated.

In (a) two inverters are inserted at the centre of the line. The first drives effectively zero line capacitance so its delay is 6 ns. The second drives 1pF so its delay is 18 ns. Each 5 mm segment has 50 ns delay. The total is 50 + 6 + 18 + 50 = 124 ns – too much.

In (b) the two inverters are placed so they break the line into 3 equal segments.

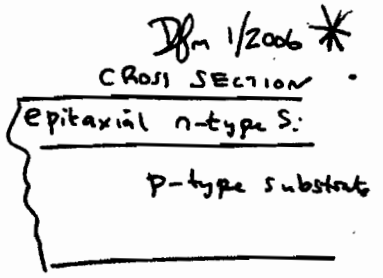


Configuration (b) yields the required > factor of two reduction and is the simplest arrangement to do so.

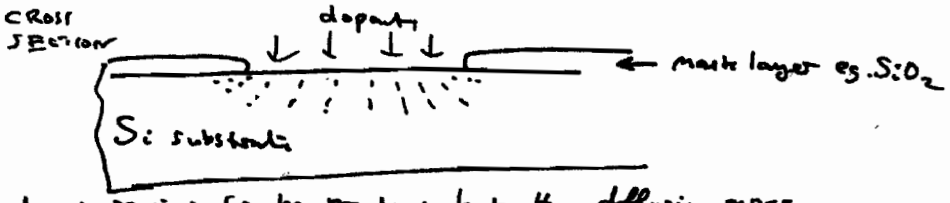
5(a)

Q 5(a)(i) In growth of silicon by chemical vapour deposition (CVD) impurity dopant atoms are introduced into the feed gas going into the reactor, for example to produce a 2 μm thick epitaxial n-type device layer on a substrate.

Dfm 1/2006 *

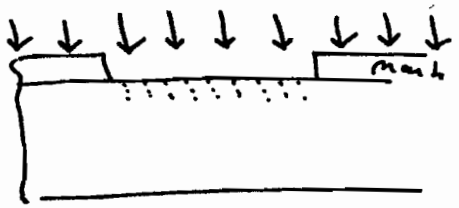


(ii) In diffusion impurities are introduced by spinning on dopant precursor to a masked silicon wafer to produce for example n-wells for CMOS



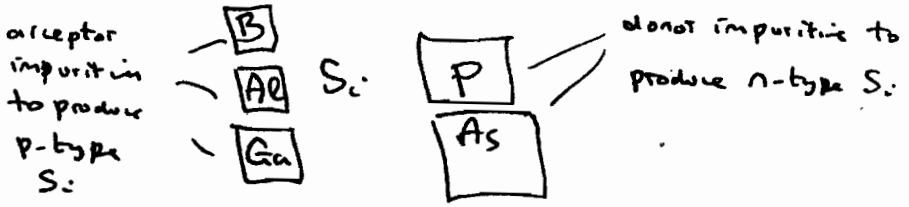
Uniformly doped regions can be produced but the diffusion goes sideways under the mask as well as down into the substrate.

(iii) In implantation high energy dopant ions are accelerated into the silicon substrate using mask layers of photoresist, polysilicon etc.



the process is precise, directional and the total dose is easily controlled.

Shallow source and drain implants are achieved by implanting at relatively low energy. Nevertheless a short post anneal is required to activate the implant.

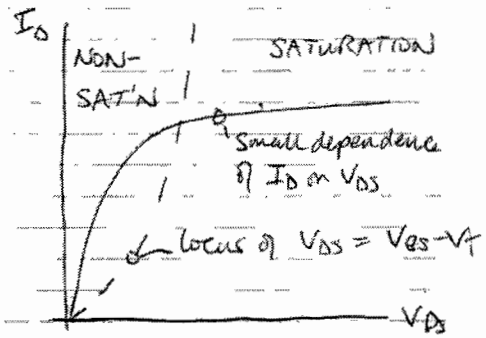


For a p-type well boron would be used because the low atomic number results in a large diffusion constant. Arsenic would be used for the n-type source drain implant. Because of the high atomic number diffusion is small during the activation anneal.

5(b) **Current Sink** – a two-terminal component whose current at any instant is independent of the potential difference across its terminals. Its Thevenin equivalent should therefore have very high (strictly infinite) Thevenin resistance. The MOS form relies on the fact that in the SATURATION region, I_D is substantially independent of V_{DS} and is controlled by V_{GS}. NB there is a small dependence of $I_D$on V_{DS} in this region owing to channel-length modulation, which accounts for the λ coefficient in the given equation.

Hence $V_{DS} > V_{GS} - V_T$
The given equation shows the dependence on λV_{DS} in the saturation region.
For $V_{GG} = 3$ V, $I_D = 100$ mA, and with M1 in saturation:

$$I_D = \frac{1}{2}\frac{\mu\varepsilon}{t_{ox}}\frac{W}{L}(3-1)^2(1+\lambda V_{DS})$$



At the onset of saturation, $V_{DS} = V_{GS} - V_T = 3 - 1 = 2$ V
The circuit will operate satisfactorily from this V_{DS} up to the onset of avalanche breakdown! At the lower limit of V_{DS},

$$10^{-4} = \frac{1}{2}\times 15\times 10^{-6}\frac{W}{L}\times 2^2\times(1+0.02), \text{ and}$$

$$\frac{W}{L} = \frac{2\times 10^{-4}}{15\times 10^{-6}\times 2^2\times 1.02} = 3.27$$

The current will rise by about 1% per volt increase in V_{DS}.

Small signal resistance is $\frac{1}{g_d}$. $g_d = 1\Big/\frac{\partial V_{DS}}{\partial I_D}$

From above, $I_D = A(1+\lambda V_{DS})$ - A is a constant. Hence $\frac{\partial I_D}{\partial V_{DS}} = \lambda A = \frac{\lambda I_D}{1+\lambda V_{DS}} \approx \lambda I_D$

Since $\lambda V_{DS} \ll 1$, hence $r_d = \frac{1}{\lambda I_D} = \frac{1}{0.01\times 10^{-4}} = 1\,\text{M}\Omega$