DEPARTMENT OF ELECTRONIC AND ELECTRICAL ENGINEERING

Spring Semester 2013-14

Answers to EEE348 Electronics and Devices 3, Questions 1…8

1. *The pull-down network for a CMOS, digital circuit is shown in* ***Figure 1****.*

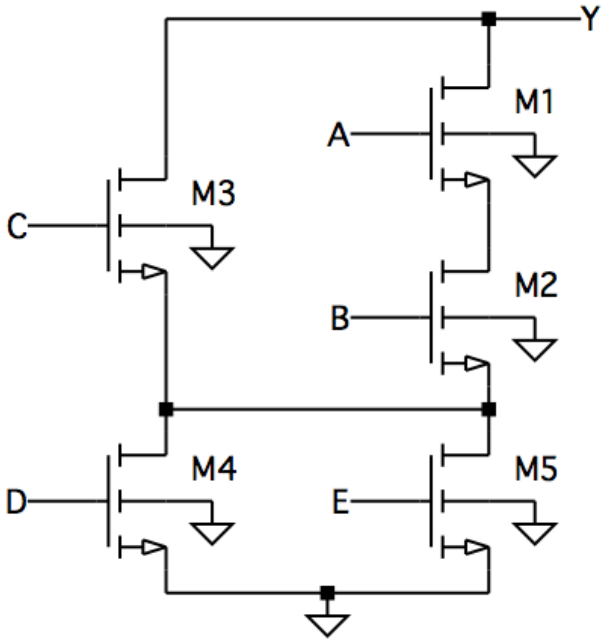
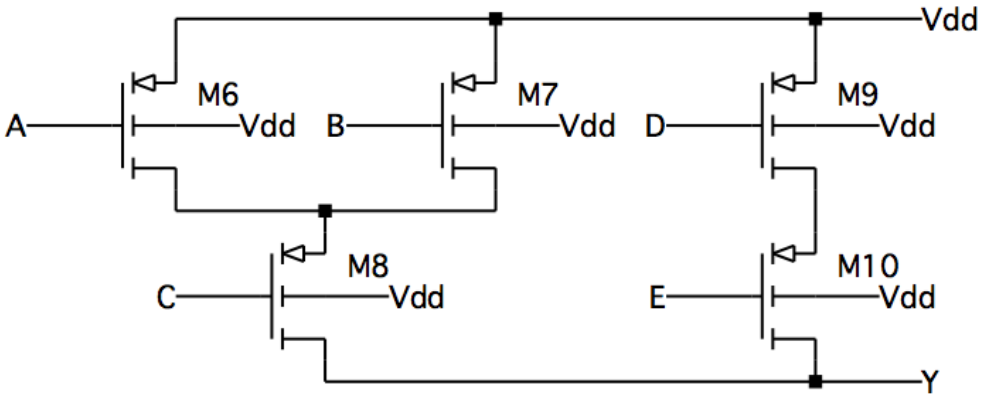


Figure 1: CMOS Pull-Down Network

a. *Draw the corresponding pull-up network for this circuit (remembering to show how the substrates are connected).*



(5)

b. *Determine the function,* ***Y****, in terms of* ***A****,* ***B****,* ***C****,* ***D****, and* ***E****.*

$Y = \overline{(AB + C)(D + E)}$ **(3)**

c. *Size the transistors* ***M1****…****M5*** *(as multiples of the minimum width of an n-type FET), assuming that the gate is 'minimum-sized'.*

Two approaches here to yield a worst case on-state resistance between the supply and the output equivalent to a minimum sized *n*-type FET – either is acceptable:

M1 = 3, M2 = 3, M3 = 3/2, M4 = 3, M5 = 3 or

M1 = 4, M2 = 4, M3 = 2, M4 = 2, M5 = 2

There is little to choose between these two solutions (one mark for each Tran.) **(5)**

d. *Why are all the substrates of these n-type FETs be connected together and to the most negative point in the circuit?*

The substrate of an *n*-type FET should be connected to a voltage equal to or more negative that the source or the drain to prevent the diodes associated with these junctions from forward-biassing. It would be helpful, therefore to connect each transistor's substrate to its source. However, this would require each transistor to be placed into a separate region of *p*-type material – isolated from the regions associated with any other transistor. This can be done but it is costly in terms of area. Alternatively, transistors can be placed in common regions of *p*-type material but must share a common substrate voltage. To ensure that this voltage is such that it does not forward-bias diodes in any transistor, it is normally connected to the most negative point in the circuit: V_{SS} or GND. **(1) (1) (1) (1)**

e. *The wire connecting the drains of **M4** and **M5** is removed. What is the new function of **Y**?*

$Y = \overline{ABE + CD}$ **(3)**

2. a. *An IC designer has to decide how to design and fabricate an ASIC:*

i) *What types of fabrication technologies are available and what are their attributes?*

Answer should include a range of approaches to implementing ASICs that are available (with some elaboration):

CPLD – few 1000 gates, deterministic behaviour/limited complexity, short design time, commodity part (10s of $), no lead time, glue logic

FPGA – up to 5M gates, non-deterministic (need full simulation), highly complex (could include large standard blocks: processors, memory, signal processing); front-end design processes follow standard 'flow', commodity parts (10-100s of $), no lead time, prototyping, low volume runs

SA – migration path from FPGA design to foundry-produced, specific IC with minimal risk for same performance. NRE in £10k+ range, lower device cost, mini, modest volume runs

MGA – pre-diffused wafers, design decomposed to NAND gates (for example) supplies metallisation, up to M gates, non-deterministic (need full simulation). Front-end and back end design process required. Foundry made but fast because wafers are standard (few weeks). NRE could be £100k+, but parts are low cost ($s) minimum order quantities e.g. 50k parts.

CBIC – from blank wafers using library of components. Full mask set required for foundry fabrication so very high NRE (up to $M). Can take up to 12 weeks to fabricate. Highest performance parts, most efficient, lowest area. Best for very big volumes but highest risk if there's a problem. **(4)**

ii) *What factors will influence the choice of technologies and the approach to design?*

The issues that affect choice will include (with some elaboration, possibly):

Size of design
Performance required
Power dissipation
Market volume

Target cost
Time-to-market
Skills/experience of design team
Any pre-existing IP (either external or internal) **(4)**

b. *Currently, the design approach often adopted by designers (at an intermediate level) is termed register-transfer level (RTL).*

i) *What does this mean and why is it important?*

RTL is a methodology that helps to ensure that designs can be implemented reliability so that they will work, as intended, when implemented in devices across the whole gamut of variation (process, temperature, voltage). Looking at a snippet of code (in VHDL):

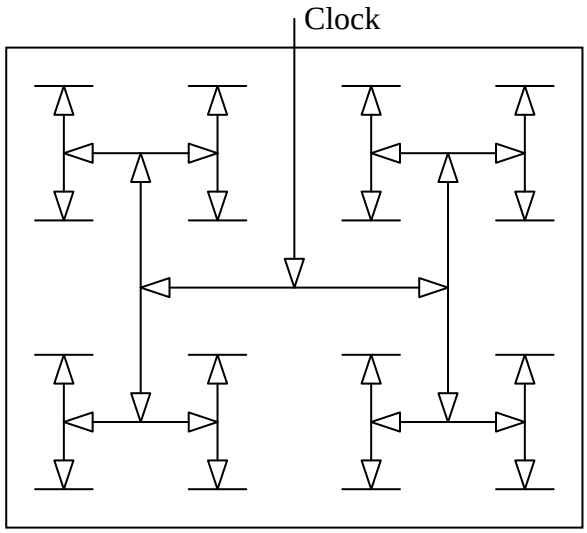
```
Process (clk)
begin
   if clk'event and (clk = '1') then
      reg3 <= reg1 OP reg2;
   end if;
end process;
```

`reg3` is a register, updated on the rising edge of the clock (`clk`) signal. Similarly, `reg1` and `reg2` are other registers written to – at the same point in time – somewhere else in the design.

With RTL, the assumption when designing, is that the value written to `reg3` on the N+1th clock cycle is the value computed from `reg1 OP reg2` using the values written to these registers at the Nth clock cycle. That is, the values written to `reg1` and `reg2` at the Nth clock cycle have a whole clock cycle to propagate through intervening logic and arrive at the register holding `reg3` in time for the correct value to by stored at this register at the N+1th clock cycle. This is predicated on the assumption that the clock signals arrive co-incidentally at all flip-flops across the IC at the same time (i.e. a signal propagating from `reg1` after the Nth clock cycle updates this flip-flop cannot arrive at the flip flop associated with `reg3` before the Nth clock cycle has updated this register). **(4)**

ii) *How might a clock tree be constructed to support the implementation of a reliable design?*

A single clock signal should be propagated across the surface of an IC by a clock network (tree) so that the delay experienced between the input of the network and any output (where flip-flops are clocked) is equal. e.g.



The clock is routed to the centre of the design, buffered, and distributed in a recursive H pattern so that each route from the input to any of the outputs is electrically the same length. The intermediate buffers allow the network to be fast by limiting the capacitance seen by any individual buffer. **(4)**

c. *Why is verification important and why is the time spent on verification beginning to dominate design activity?*

The costs of modern ASIC design and implementation can be extremely high, along with the lost opportunity cost of a device, intended for a product heading to market fails to function and results in a delay in the product entering a market. Consequently, it makes sense to verify (simulate) the behaviour and performance of the ASIC to ensure that it does what it is supposed to do, functionally, meets all the design constraints (e.g. clock frequency) and will work reliably over all process and environmental conditions. Moreover, it must also incorporate sufficient testability to ensure that malfunctioning devices can be weeded-out before they are incorporated in a product. This burden already consumes upwards of 50% of the design effort because there is a lot of verification to undertake on a large ASIC (getting correspondingly larger with larger designs) and it can be very time-consuming in its own right. However, the burden is growing, particularly as designs shrink because the an increasing number of potential problems need to be simulated to ensure that there are no problems. So, traditionally, whilst verification covered logic errors and timing, this might be extended to cover the variability of device behaviour and the variability of timing arising from signal interactions (e.g. coupled capacitance) up to partial/full electromagnetic simulations. Clearly, the time spent on verification must, at some point, be traded-off: the cost/time of verifying comprehensively .v. the risk/costs of missing errors by not simulating comprehensively. **(4)**

3. a. i)

$$V_{GS} = 10 \times \left(\frac{0.5}{4.01}\right) = 1.25V$$ **(1)**

$$V_{OV} = V_{GS} - V_{TO} = 1.25 - 0.6 = 0.65V$$

$$I_D = \frac{K_n}{2}\frac{W}{L}V_{OV}^2 = 40 \times 10^{-6} \times 20 \times 0.43 = 338\mu A$$ **(1)**

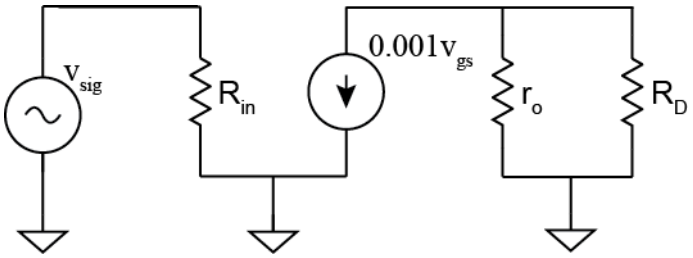
$$g_m = K_n\frac{W}{L}V_{OV} = 80 \times 10^{-6} \times 20 \times 0.65 = 1mS$$ **(1)**

The Q-point voltage at Vout is

$$V_{out} = 5 - I_D R_D = 5 - 338\mu A \times 15k\Omega \approx 0V$$

So that V_{DS} for M1 is 5V>>V_{OV}, and the FET is in saturation. **(1)**

ii)



$$R_{in} = (3.51M\Omega \parallel 500k\Omega) = 438k\Omega$$

$$r_o = \frac{1}{\lambda I_D} = \frac{1}{0.025 \times 338 \times 10^{-6}} = 118.3k\Omega$$ **(3)**

Ignoring channel modulation, the gain is:

$$A_v = -g_m R_D = -0.001 \times 15k\Omega = -15$$ **(1)**

iii)

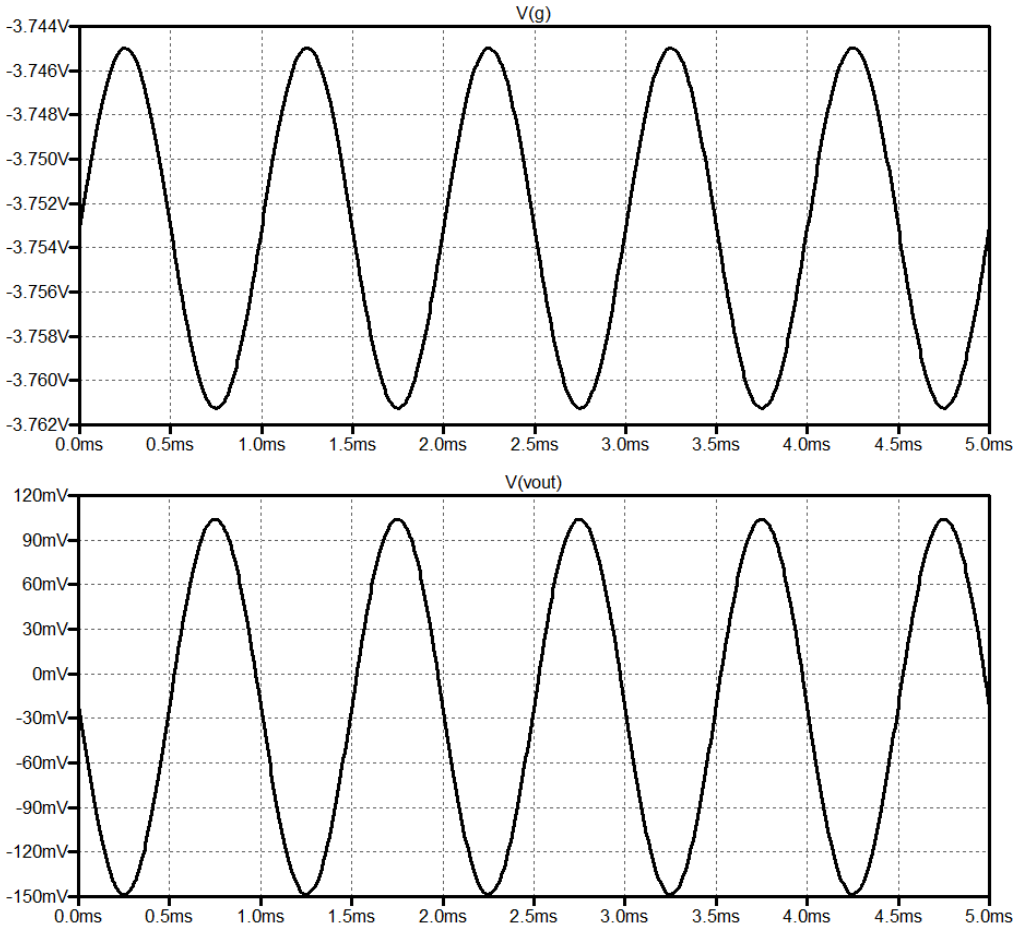
The gate-source voltage is $v_{gs} = \frac{R_{in}}{100k + R_{in}} v_{sig} = 0.814 v_{sig}$

So the gain becomes:

$$A_v = -g_m R_{out} \times 0.81 = -0.001 \times 15k\Omega \times 0.814 = -12.2$$

(2)

b.



Marks for correct offset voltages, correct sinewave magnitudes, correct x-axis labeling and inversion of output.

(4)

c. **i)** The performance of the amplifier is highly dependent upon the parameters of the FET itself. These can vary between devices with the same part number, between transistors on the same chip, or as a function of temperature. The voltage divider biasing holds the gate voltage constant regardless of the actual current that flows through the FET, and so the amplifier's Q-point and gain are both unpredictable. Better biasing requires some form of feedback, so that variations in the drain current of the FET are controlled by appropriate variations in V_{GS}. **(3)**

Voltage divider biasing is unsuitable for implementation on an IC because large value resistors take up huge amounts of room on a semiconductor substrate. The coupling capacitor also takes up valuable chip area. A better solution is to use a self-biased differential amplifier that doesn't require any capacitive signal coupling.

ii.) Since the curve of I_D vs V_{DS} for M2 in the diagram above is proportional to the reciprocal of the output resistance of the mirror, maximizing this resistance means that V_{DS}(M2) can vary over a large range of voltages (as may be the case when the mirror is attached to a circuit), without a significant change in the current it pulls. When used for amplifier biasing, this means that Q-points within the amplifier can be made extremely stable, since the current mirror will automatically set a value of Vout that delivers the designed bias current over a large range of operating conditions. **(3)**

4. a. i)

$$V_{OV} = \frac{1}{\lambda\sqrt{2500}} = \frac{1}{0.05 \times 50} = 0.4V$$ **(1)**

For FETS M3,4,5,6:

$$r_{o1} = \frac{2}{\lambda I_{ref1}} = \frac{2}{0.05 \times 150 \times 10^{-6}} = 266.7k\Omega$$ **(1)**

For FETs M7,8:

$$r_{o2} = \frac{1}{\lambda I_{ref2}} = \frac{1}{0.05 \times 300 \times 10^{-6}} = 66.7k\Omega$$ **(1)**

ii)

The transconductance of the differential amplifier is:

$$g_{m1} = \frac{150 \times 10^{-6}}{0.4} = 0.375mS$$

Whilst that of the CS amplifier stage is:

$$g_{m2} = \frac{600 \times 10^{-6}}{0.4} = 1.5mS$$ **(1)**

The gain of the differential pair is then:

$$A_{diff} = g_{m1}(r_{o1} \parallel r_{o1}) = \frac{0.000375 \times 266700}{2} = 50$$ **(1)**

The gain of the CS amp is:

$$A_{CS} = g_{m2}(r_{o2} \parallel r_{o2}) = \frac{0.0015 \times 66700}{2} = 50$$ **(1)**

b. To find the channel aspect ratios:

For M1:

$$300\mu A = \frac{1}{2} K_p \frac{W_1}{L_1} V_{OV}^2$$

$$300\mu A = \frac{1}{2} 150 \times 10^{-6} \times \frac{W_1}{L_1} 0.4^2$$

$$\frac{W_1}{L_1} = 25, \quad W_1 = 18.75\mu m$$ **(1)**

M2 must be chosen to provide I_{ref1}=150µA, i.e half the master current, I1, of the current mirror, so:

$$\frac{W_2}{L_2} = 12.5, \quad W_2 = 9.38\mu m$$ **(1)**

For M3,4:

$$\frac{I_{ref1}}{2} = \frac{1}{2} K_p \frac{W_{3,4}}{L_{3,4}} 0.4^2$$ **(2)**

$$\Rightarrow \frac{W_{3,4}}{L_{3,4}} = 6.25, \quad W_{3,4} = 4.69\mu m$$

For M5,6:

$$\frac{I_{ref1}}{2} = \frac{1}{2} K_n \frac{W_{5,6}}{L_{5,6}} 0.4^2$$

$$75\mu A = \frac{1}{2} 200 \times 10^{-6} \times \frac{W_{5,6}}{L_{5,6}} 0.4^2$$ **(1)**

$$\Rightarrow \frac{W_{5,6}}{L_{5,6}} = 4.69, \quad W_{5,6} = 3.52\mu m$$

For M7:

$$I_{ref2} = \frac{1}{2} K_n \frac{W_7}{L_7} 0.4^2$$

$$\Rightarrow \frac{W_7}{L_7} = 18.75, \; W_7 = 14.06\mu m$$ **(1)**

For M8:

$$I_{ref2} = I1 \frac{W_8}{L_8} \Big/ \frac{W_1}{L_1}$$

$$\Rightarrow \frac{W_8}{L_8} = \frac{I_{ref2}}{I1} \frac{W_1}{L_1} = \frac{300}{300} 25 = 25$$

$$\Rightarrow W_8 = W_1 = 18.75\mu m$$ **(1)**

c. A virtual ground exists at the drain of M2, or at the combined source of M3 & M4. **(1)**

Since the op-amp is driven by a balanced load with no common-mode component, the Q-point voltage at this point is $0V+V_{TO}+V_{OV}=0.9V$. **(2)**

d. The Miller multiplier is equal to 1 + the gain of the CS amp = 51. The capacitor **(4)**

C1 appears in parallel with C_{GD}=10fF, so that the total Miller-multiplied capacitance is:

$$C_M = 51 \times (0.02 \times 10^{-12} + C1)$$

The total capacitance at the gate of M7 is then:

$$C_{tot} = C_M + C_{GS} = 51 \times (0.02 \times 10^{-12} + C1) + 0.05 \times 10^{-12} \qquad (*)$$

The resistance seen by C_{tot} is $r_{o1} \parallel r_{o1}$=133.3kΩ, so the upper cutoff frequency is approximately:

$$f_{cutoff} \approx \frac{1}{2\pi \times 133333 \times C_{tot}} = 100kHz,$$

$$C_{tot} \approx \frac{1}{2\pi \times 133333 \times 100000} = 11.9pF$$

Rearranging * to find C1:

$$C1 = \frac{C_{tot} - 0.05 \times 10^{-12}}{51} - 0.02 \times 10^{-12} = 0.21pF$$

NLS AMM

PART B.

Q1

a)

i) Light intensity in a semiconductor is described by

$I = I_o \exp(-\alpha x)$

If 90% of light is absorbed within a length L, we have

$0.1I_o = I_o \exp(-\alpha L)$

$\exp(-\alpha L) = 0.1$

$L = -\frac{1}{\alpha}\ln(0.1)$

For GaAs, $L = -\frac{1}{5\times10^4}\ln(0.1) = 4.6\times10^{-5} cm = 0.46\,\mu m$

For Si, $L = -\frac{1}{2\times10^3}\ln(0.1) = 1.15\times10^{-3} cm = 11.5\,\mu m$

GaAs and Si layers with these thicknesses can be grown with high crystal quality and are widely available.

ii) Any two of the followings

-High purity bulk Si, with extremely low impurity concentration and long minority electron diffusion length (due to low defect density) is readily available. The former allows a large depletion region to be obtained while the latter increases the carrier collection efficiency, leading to very efficient Si solar cell.

-Large, cheap and high quality Si substrates significantly reduce the cost of Si solar cells. Si substrate size of up to 300 mm diameter (11.8 inch) is routinely used while 450 mm diameter substrate is being developed. These are much larger than the largest 150 mm (6 inch) GaAs substrate.

-The abundance of Si coupled with the state-of-the art fabrication technology and very large scale manufacturing facilities are important for manufacturing of solar cell for mass market applications. For this reason GaAs based solar cells find more application where efficiency is more important than cost.

-In Si, the dangling bonds at the surface can be successfully passivated with hydrogen. This reduces the surface recombination.

-SiO_2 which is a good electrical insulator is a native oxide that can be easily grown. When deposited with appropriate thickness it also reduces the reflection of the incident light and improves the conversion efficiency of the solar cell.

b) Using a simple case of a solar well with a depletion with of *W* and assuming that the carrier generation rate, *G*, is constant, the photocurrent in a solar cell is given by

$I_{ph} = qAG(L_e + W + L_h)$, where *L* is the minority carrier diffusion length and *A* is the area.

The depletion width is defined by the thickness of the undoped i-region. To ensure that the i-region is fully depleted, its unintentional doping needs to be very low. In the case of Si, unintentional background doping below 10^{14} cm^{-3} will ensure that a large depletion width can be achieved.

The minority carrier diffusion lengths L_e and L_h can be maximised by reducing the doping concentrations in the p and n regions. However this is not ideal since high doping

concentration in the p and n-regions are required to minimise the series resistance. Therefore in practice, a relatively thin highly doped p^+ region is used to minimise the series resistance.

c) Inverted pyramid : Sunlight incident on a side slope is partially transmitted into the cell and partially reflected to the other slope, increasing the probability of light being absorbed.

Oxide layer: Designed to minimise reflection of incident sunlight. It also passivates the surface to minimise surface recombination.

p-region: Should have long minority carrier diffusion length to ensure high output current. Minority carrier lifetime in ms is routinely achieved.

P+ regions: Minority carriers are collected and is highly doped to minimise series resistance. Metal-semiconductor area (defined by the small Boron diffused p+ region) is small to minimise surface recombination loss.

N+ regions: Highly doped to minimise series resistance

Si-Oxide-Aluminium: Act as mirror to reflect light (particularly at long wavelengths) back into the cell.

d) As the name suggests, three junctions are stacked and connected in series to provide a closer match of the solar cell spectral response to that solar emission spectral. The bandgap is reduced from the top to the bottom. For example the junctions, from top to bottom, are InGaP, InGaAs and Ge with bandgaps of 1.8, 1.4 and 0.65 eV respectively. Very highly doped pn junctions are inserted between each junction. These pn junctions are designed such that carriers can tunnel through the different bandgap junctions. The triple junction can be designed to achieve high absorption efficiency from UV to 1.9 μm and achieves much higher efficiency than Si.

Q2 a) The empirical expression for avalanche multiplication factor is given by

$$M = \frac{1}{1-\left(\frac{V-IR}{V_b}\right)^{n_m}}.$$

Using the parameter values provided the multiplication factor in the dark is

$$M = \frac{1}{1-\left(\frac{29-10\times10^{-9}\times10}{30}\right)^{2}} = 30.$$

Note that since the dark current is low, the voltage drop across R is negligible compare to the bias voltage of 29 V.

The photocurrent produced is

$$I_{ph} = \frac{\eta\lambda P_{opt}}{1.24} = \frac{0.9\times0.85\times5\times10^{-9}}{1.24} = 3.1\times10^{-9}\,A$$

Total current flowing = 10 + 3.1 = 13.1 nA. IR = 0.13 μV is much smaller than 29 V. The multiplication factor is not affected. The total photocurrent is therefore 3.1 x 30 = 93 nA.

b) The noise signal at the output of the amplifier is given by

$v_{noise} = 10\times10^{-12}\times\sqrt{0.1\times10^{9}}\times G_{amp}$ where G_{amp} is the amplifier gain factor.

The photocurrent signal is $v_{photo} = 3.08\times10^{-9}\times M\times G_{amp}$. Therefore to overcome the amplifier noise $M > \dfrac{10\times10^{-12}\times\sqrt{0.1\times10^{9}}\times G_{amp}}{3.08\times10^{-9}\times G_{amp}} = 32.5$.

The gain of 32.5 is moderate and is below the breakdown voltage. Therefore the gain stability can easily be achieved by using a thermistor and feedback circuit to maintain the bias required to achieve the required gain.

c) Current commercial APDs for high speed optical communication employs an InGaAs absorption region and an InP multiplication region. The electric field profile is maintained by a field control layer with very small doping concentration tolerance. To achieve 100 Gb/s and above the total thickness of the APD has to be very thin. For instance to achieve a transit time limited bandwidth of 50 GHz, the thickness has to be

$$W = \frac{0.4v_s}{f_{3dB\_tr}} \sim \frac{0.4\times10^{5}}{50\times10^{9}} = 0.8\mu m.$$

Due to multiple transit times required to achieve the required multiplication factor, the APD has to be thinner than 0.8 μm and the field control layer and the avalanche layer will be much thinner. Growth of such APD is not mature, hence commercial APD operating at 100 Gb/s is not available.

d) To achieve 100 Gb/s operation, the diode should have a small RC time constant and a small transit time. Assuming that a minimum bandwidth of 50 GHs is required the depletion region thickness of 0.8 μm is required. Since the quantum efficiency also depends on the depletion width, a vertically illuminated photodiode is not suitable. A laterally illuminated waveguide photodiode is therefore the preferred technology. In practice InGaAs is usually used as the absorption active region. The waveguide photodiode has wide bandgap cladding layers, InP, to achieve high optical confinement. The area is typically small with width of less than 10 μm to maintain low capacitance while the contact layer is highly doped to minimise series resistance.

Q3

a)

i) In a homojunction laser the optical confinement is provided by the change of refractive index in the highly doped p and n regions. This change is relatively small, up to 1 %, and therefore light is poorly confined within the active region, leading to high opical loss. The potential barriers are also small such that carriers leak out from the active region and recombine non-radiatively in the highly doped p and n neutral regions. These factors lead to a large threshold current in the homojunction laser.

ii) In a double heterojunction laser the wide bandgap p and n regions can provide higher refractive index change of typically 5% (assuming GaAs and $Al_{0.3}Ga_{0.7}As$ combination). When combined with appropriate thickness of the active region, close to unity confinement factor can be achieved. Since light is fully confined within the active region, the photon intensity increases much more easily. The large conduction and valence band offsets also ensure that carriers are confined in the active region to increase the electrons and holes densities available for radiative recombination. Because of these the double heterojunction laser can achieve much lower threshold current than the homojunction laser.

b)

i) In a quantum well laser energy states are closely packed into sub-bands such that they can be approximated by a series of fixed energy states. In bulk, the density of states is zero at the band edge and it increases gradually with energy. On the other hand the density of states is constant at a fixed energy level. Because of the constant density of states, a group of electrons of nearly the same energy exist. There is also a corresponding group of holes in the valence band. These higher electron and hole concentration allows the population inversion to be much more easily achieved, leading to much lower threshold current. The threshold current is also much lower since the active region, d_{las}, is much thinner as described by

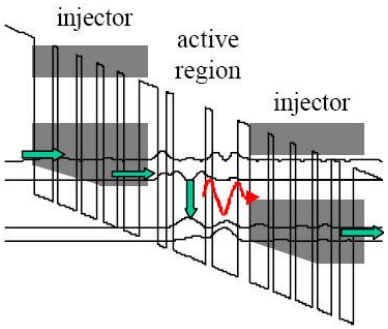
$$J_{th} = \frac{qd_{las}n_{th}}{\tau_r(J_{th})},$$

where nth is the carrier density when the cavity gain equals to the cavity loss and $\tau_r(J_{th})$ is the radiative recombination time.

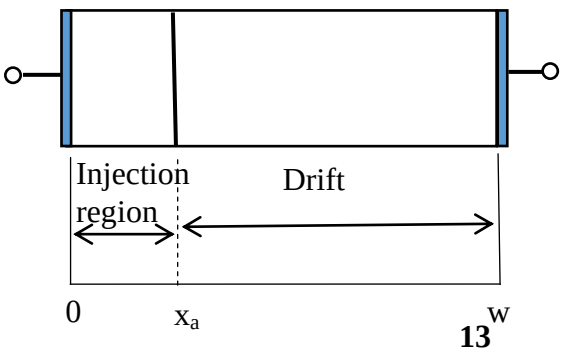
ii) The energy states can also easily be modified by changing the width of the well. Although less desirable using high current injection conditions, excited states will also be occupied and can contribute to radiative recombination. Therefore more than one wavelength can be emitted. When combined with cavity of different lengths or with Bragg reflectors, a wide lasing wavelengths can be easily obtained.

c) Auger recombination increases exponentially as the bandgap decreases. Hence it is a very dominant non-radiative recombination mechanism in narrow bandgap materials such as InSb. In addition there is no wide bandgap material that is lattice matched to InSb. Therefore it is not possible to fabricate an InSb based laser.

d) The quantum cascade laser (QCL) is an important laser for the 3-5 μm. The band diagram is shown below. In the injector region, the quantum wells are designed such that under appropriate biasing condition, the energy states are aligned to allow electron to tunnel through the barriers. This enables the electron to move from a low energy state to a higher energy state. The electron is subsequently injected into the active region containing quantum well with energy state separation that will produce the desired wavelength when the electron make the intra-band transition from the high to low energy states. After the photon emission the electron from the low energy state is transported to the injector region and the process is repeated. Typically the injector and active regions are repeated for 20-100 times to produce a QCL.



Q4.



A schematic of an idealised IMPATT diode is shown above. Charge carriers are generated by impact ionisation in the injection region. The IMPATT diode is biased such that the field in the injection region is very close to the breakdown field. These charges are then injected into the drift region which has low electric field to ensure that the charges do not experience impact ionisation while they drift. The time delay created by the impact ionisation and drift process causes a phase lag between the output current and the input voltage, leading to a negative resistance.

b)

i) The oscillation frequency is given by

$$f_{IMP} = \frac{v_{sat}}{2(w - x_a)} = \frac{1\times10^5}{2\times500\times10^{-9}} = 100GHz\,.$$

ii) The breakdown voltage is given by

$$V_B = E_m x_a + \left(E_m - \frac{qQ_c}{\varepsilon_s}\right)(w - x_a) = 6\times10^5\times100\times10^{-7} + \left(6\times10^5 - \frac{q\times4\times10^{12}}{12.4\times8.85\times10^{-14}}\right)(500\times10^{-7})$$

$$V_B = 6.0 + 0.84 = 6.84V$$

iii) The voltage drop across the injection region is 6V. hence the voltage across the drift region is 29.6 V. The electric field is therefore $0.84/500\text{x}10^{-7}$ = 16.8 kV/cm.

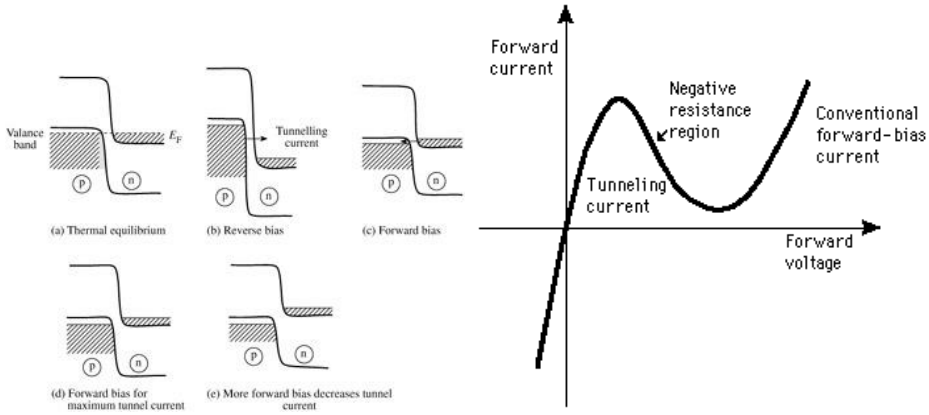
c) To operate at 1 THz, the drift region has to be very thin.

$$w - x_a = \frac{v_{sat}}{2(1\times10^{12})} = \frac{1\times10^5}{2\times10^{12}} = 50nm\,.$$

To ensure that the impact ionisation occurs over a short duration than the drift time the injection region and the field control region will have to much shorter than 50 nm. Accurate control of doping across a few nm of field control layer is challenging. In addition if the injection region is below 50 nm, very high band to band tunnelling current will dominate making the IMPATT diode very noisy.

d) A tunnel diode consists of a p-n junction in which both p and n regions are highly doped such that the Fermi levels are in the valence band in the p region and in the conduction band in the n region. Because of the very high doping the depletion region is very thin and hence the tunnelling distance is only 5-10 nm thick.

Under forward bias there is a band of energy states that are fully occupied in the n side and a corresponding empty states in the p side. The electrons can tunnel from n to p side. A peak current is produced when the tunnelling is maximum. At higher forward bias, the occupied and empty states are not fully "aligned" and hence the tunnelling current drops. At even higher forward bias the occupied and empty states are no longer aligned. Electrons will have to overcome the potential barrier via thermionic effect and normal thermal current flows.


Valance band
E_F
Tunnelling current
(a) Thermal equilibrium
(b) Reverse bias
(c) Forward bias
(d) Forward bias for maximum tunnel current
(e) More forward bias decreases tunnel current
Forward current
Negative resistance region
Conventional forward-bias current
Tunneling current
Forward voltage

CHT