DEPARTMENT OF ELECTRONIC AND ELECTRICAL ENGINEERING

Spring Semester 2004-2005 (2 hours)

Answers to Introduction to VLSI Design

1. a. i) *Explain why the interconnect (or wiring) between gates on a typical ASIC is becoming increasingly important and problematic and how the problems worsen as technology shrinks.*

Answer should include the following points:

Below 0.25μm delay in wire is more than in gates – achieving timing closure in design is difficult because layout has significant influence and optimisation at synthesis is increasingly unreliable.

With high clock frequencies and long wires the delay along a wire can be a significant percentage of the clock period.

Power consumption driving wiring capacitance represents a significant part of overall power consumption.

As technology shrinks, spanning wires become relatively longer and the difference between wiring delay and gate delay increases. For a spanning wire (IC size remains constant), if λ scaling were to be employed then you might expect $r \rightarrow r/\lambda^2$, $c \rightarrow c$, and gate delay, $t_{gate} \rightarrow t_{gate}.\lambda$ (maybe). Consequently, $rc/t_{gate} \rightarrow rc/(\lambda^3 t_{gate})$. **(6)**

ii) *what can be done to address the problems*

Firstly, the problem is not quite as bad as presented because local lines scale in length too. Consequently, $r \rightarrow r/\lambda$, $c \rightarrow c\lambda$, and $rc/t_{gate} \rightarrow rc/(\lambda t_{gate})$. Additionally, λ scaling would not be employed, tiered interconnect would be used with minimum sized wires for local connections, larger wires (lower r) for longer wires spanning blocks and large wires (lowest r) for long wires spanning the IC. Moreover, Al is being replaced by Cu (despite increased difficulty in processing) ro reduce r and lower k dielectric separating wires is being investigated to reduce c. **(2)**

b. i) *Show how using repeaters can reduce the delay experienced by a signal passing along a long wire.*

Let us assume that the intrinsic delay of a gate can be modelled as:

$$t_{gate} = R_o\left(C_{out} + C_{wire} + C_{in}\right)$$

where R_o represents the output resistance of the gate, C_{out} represents the lumped capacitance associated with the output node of the gate, C_{wire} represents the wiring capacitance, and C_{in} represents the lumped capacitance associated with the input of the gate. Consequently, a 1st order approximation to the overall delay of the gate driving a wire of length L will be the sum of the gate delay (the time taken for the driving voltage to change state) and the delay along the wire, which is rcL^2. Furthermore, $C_{wire}=cL$ (that is the total effective capacitance on the wire).

From this, the overall delay is, simply:

$$t_{delay} = t_{gate} + t_{wire} = R_o\left(C_{out} + cL + C_{in}\right) + rcL^2$$

Assume, now, that instead of there being only one driver at the source end, there are a total of *N* buffers at equal intervals (excluding the final buffer loading the wire) – each driving a length of wire equal to *L/N*. In this case:

$$t_{delay|N} = Nt_{gate} + Nt_{wire} = NR_o\left(C_{out} + C_{in} + c\frac{L}{N}\right) + rcN\left(\frac{L}{N}\right)^2$$

$$= NR_o\left(C_{out} + C_{in}\right) + cR_oL + rc\frac{L^2}{N}$$

By differentiating this *w.r.t. N* we can determine if there is an optimum point for operation:

$$\frac{d}{dN}t_{delay} = R_o\left(C_{out} + C_{in}\right) - rc\frac{L^2}{N^2} = 0$$

$$R_o\left(C_{out} + C_{in}\right) = rc\frac{L^2}{N^2}$$ **(4)**

ii) *What other benefits arise from inserting repeaters*

Most ICs are wire bound – that is, the area of the IC is determined by the area occupied by the wiring. Consequently, adding buffers will not increase the area of the IC. Moreover, at some point, by reducing the length of individual wires, the need for larger spanning wires will decrease and the area of the IC will probably fall (this will at some point be countered by the increase in area as the IC becomes logic bound as buffers are inserted). Finally, the size of individual buffers will fall and, paradoxically, the power consumption will fall. **(2)**

c. *In a particular technology, for the only available buffer, the output capacitance, C_{out}, is 10fF whilst its input capacitance, C_{in}, is 1.5fF and its effective output resistance is 400Ω. The capacitance per unit length of the wiring is 0.3fF/μm and its resistance per unit length is 2.9Ω/μm. A circuit is set up as shown in* ***Figure 1****.*

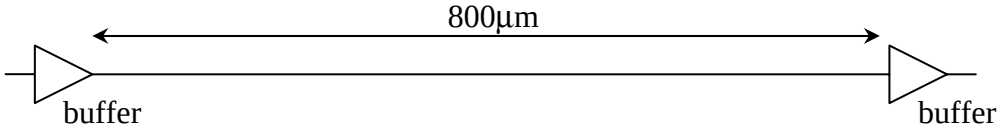


Figure 1

i) *Estimate the original delay*

$$t_{delay} = t_{gate} + t_{wire}$$

$$= 400\left(10\times10^{-15} + 800\cdot0.3\times10^{-15} + 1.5\times10^{-15}\right) + 2.9\cdot0.3\times10^{-15}800^2$$

$$= 100.6\times10^{-12} + 556.8\times10^{-12} = 657.4\times10^{-12}\,s$$ **(2)**

ii) *Identify how the circuit might be improved to minimise its delay.*

$$R_o\left(C_{in}+C_{out}\right)=400\cdot\left(1.5\times10^{-15}+10\times10^{-15}\right)=4.6\times10^{-12}$$

$$=0.3\times10^{-15}\cdot2.9\cdot800^2/N^2=556.8\times10^{-12}/N^2$$

N^2=556.8/4.6=121

N = 11 **(2)**

iii) *Estimate the minimised delay*

$$t_{delay|N}=NR_o\left(C_{out}+C_{in}\right)+cR_oL+rc\frac{L^2}{N}$$

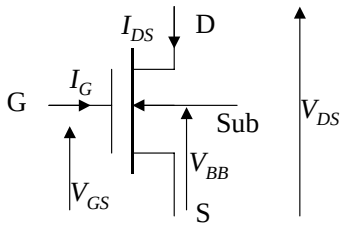
$$=11\cdot400\cdot11.5\times10^{-15}+400\cdot0.3\times10^{-15}\cdot800+2.9\cdot0.3\times10^{-15}\cdot800^2/11$$

$$=50.7\times10^{-12}+96\times10^{-12}+50.6\times10^{-12}=197.3\times10^{-12}s$$ **(2)**

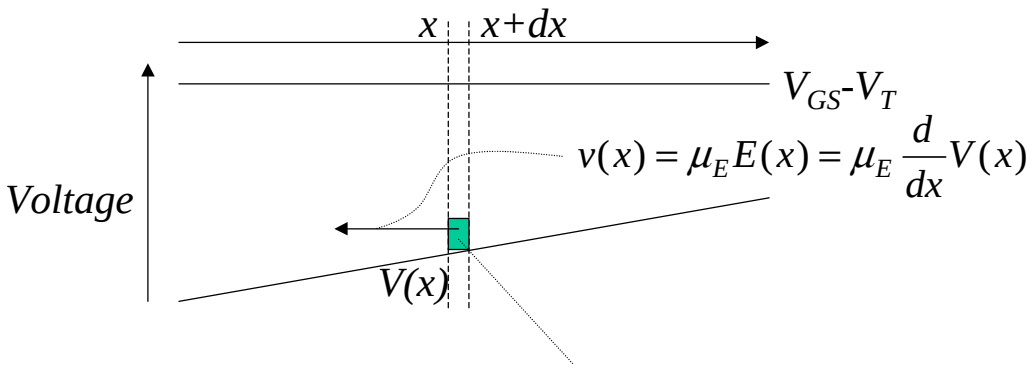
2. a. *Show that the relationship between the applied voltages and drain current for a n-channel FET with a long channel can be expressed as:*

$$I_{DS} = \frac{\mu_E \cdot C_{OX} \cdot W}{L} \cdot \left(V_{GS} - V_T - \frac{V_{DS}}{2} \right) \cdot V_{DS}$$

where the terms have their normal meanings – ensure that you answer includes a schematic of the FET with the senses of the voltages and current shown.



Definitions



$$Q / \text{unit length} = \frac{\varepsilon_0 \varepsilon_r W}{t_{OX}} \left(V_{GS} - V_T - V(x) \right)$$

Looking at the voltage along the channel of the n-channel FET and the charge/unit length in an elemental point along the channel. Where the charge is related to the voltage falling across the oxide using a parallel plate capacitance model and the velocity of the charge is related to the electric field via the carrier mobility.

These can be equated:

$$I_{DS} = Q / \text{unit length} \cdot v(x) = \frac{\varepsilon_0 \varepsilon_r W}{t_{OX}} \left(V_{GS} - V_T - V(x) \right) \mu_E \frac{d}{dx} V(x)$$

and integrating across the channel distance and voltage limits (remembering that the current is deemed to be continuous along the channel.

$$\int_0^L I_{DS} dx = \frac{\varepsilon_0 \varepsilon_r \mu_E W}{t_{OX}} \int_0^{V_{DS}} \left(V_{GS} - V_T - V(x) \right) dV(x)$$

Collecting terms and identifying that C_{OX} , the capacitance per unit area of the gate, is $\varepsilon_0\varepsilon_r/t_{OX}$ yields the expression being sought.

$$I_{DS}L = \frac{\varepsilon_0\varepsilon_r\mu_E W}{t_{OX}}\left[(V_{GS}-V_T)V(x) - \frac{V(x)^2}{2}\right]_0^{V_{DS}}$$

$$I_{DS} = \frac{\varepsilon_0\varepsilon_r\mu_E W}{t_{OX}L}\left((V_{GS}-V_T)V_{DS} - \frac{{V_{DS}}^2}{2}\right)$$ **(6)**

b. *A n-channel transistor in a digital CMOS circuit might be simply modelled as a resistance whose effective value, R_o, is:*

$$R_o = \frac{2}{\beta(V_{DD}-V_T)}$$

where, again, the terms have their usual meaning. What is the significance of this equation for circuit performance (especially when sub-threshold conduction is considered).

Essentially, the drive of the FET – the ability for it to deliver current (modelled by a resistance above) is related to the gain of the transistor, β, and, critically, the difference between the supply voltage and the threshold voltage (V_{DD}-V_T). As V_T approaches V_{DD}, the effective resistance increases towards infinity. Consequently, as V_{DD} is reduced as technology is scaled down, there is a requirement to scale V_T correspondingly to ensure that current drive remain at a *commensurate* level. However, when the gate-source voltage, V_{GS} falls below V_T, the transistor still conducts (sub-threshold conduction) and as V_{GS} is reduced then this current falls by a decade for every 80-90mV of reduction in V_{GS} for Si. Thus, when V_{GS} reaches 0V there will still be a leakage current flowing. For every 80-90mV reduction in the value of V_T, therefore, there will be a 10x increase in leakage current and this must be considered in the light of performance and static power consumption. **(4)**

c. *A CMOS circuit has the pull-up network shown in **Figure 2**:*

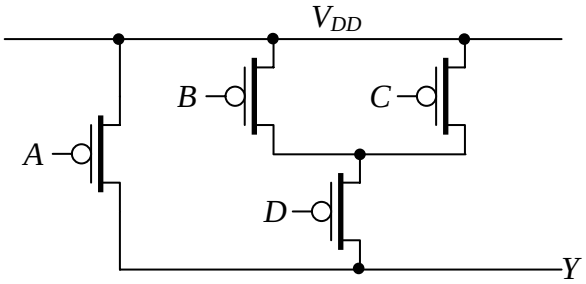
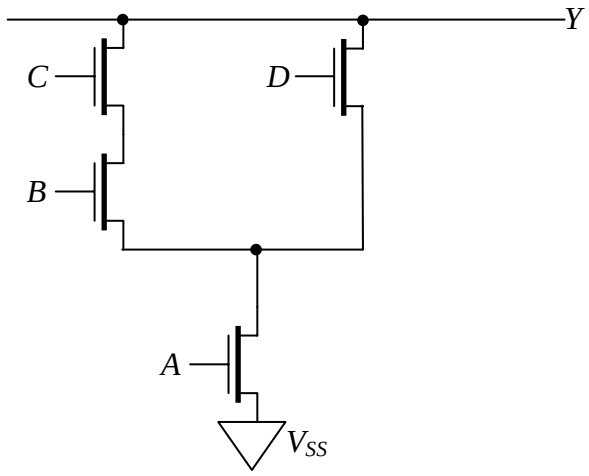


Figure 2: CMOS Pull-Up Network

For this pull-up network:

i) *draw the corresponding pull-down network;*



(3)

ii) *determine the function, Y, of the logic gate;*

$Y = \overline{A \cdot (B \cdot C + D)}$ **(3)**

iii) *size all of the transistors for a minimum-sized gate (assuming that the mobility of holes is half that of electrons) in terms of the minimum transistor width, W.*

For the pull down network:

A=2W, B=4W, C=4W, D=2W

For the pull-up network:

A=2, B=4, C=4, D=4 **(4)**

3. *An IC fabricated in 0.13μm CMOS technology contains 20M two-input, minimum sized NAND gate equivalents and approximately 500m of interconnect. For the fabrication technology, you know that a minimum-sized n-channel FET is 0.13μm long and 0.3μm wide, the gate oxide ($\varepsilon = 3.45\times10^{-11}$F/m) is 5nm thick, and that the hole mobility is half that of electrons. Furthermore, it is known that the average width of the wiring is 0.75μm and the ratio of height to width of a wire is 1.6 and it is estimated that the average thickness of dialectric ($\varepsilon = 3.1\times10^{-11}$F/m) around each wire is approximately 0.75μm also.*

a. *For a NAND gate equivalent (without any wiring attached), estimate:*

i) *the input capacitance of each input;*

For a minimum sized NAND gate, the width of the n-channel FETs are 2x minumum width as are the p-channel FETs. Consequently, the input capacitance is 4x that of a minimum sized n-channel FET. The capacitance is:

$$C_{in} = 4 \cdot \frac{\varepsilon_0 \varepsilon_0 W_{\min} L_{\min}}{t_{OX}} = 4 \cdot \frac{3.45\times10^{-11} \cdot 0.3\times10^{-6} \cdot 0.13\times10^{-6}}{5\times10^{-9}} = 1.076\, fF$$

(3)

ii) *the output capacitance*

Assuming that the output node is loaded by ½ the gate capacitances connected to it – this is one n-channel FET and both p-channel FETs. ½

the gate capacitance of each FET is 1.076fF/4 and three of these contribute ¾ * 1.076fF = 0.807fF. **(3)**

b. *Estimate the total capacitance due to the wiring on the IC.*

Each wire is 0.75μm wide and 1.6*0.75μm high. Consequently, the perimeter – the 'width' of the capacitor is 2*(0.75+0.75*1.6)μm = 3.9μm. The 'length' is 500m and the thickness of the dielectric is 0.75μm. Neglecting fringing fields, the capacitance is $3.9\text{x}10^{-6}\text{x}500*3.1\text{x}10^{-11}/0.75\text{x}10^{-6}$= 80.9nF. **(3)**

The power supply voltage to the core logic is 1.2V. It is estimated that I/O pads consume 30% of the total power consumed by the IC and that leakage currents in the core logic consume another 10% of the total power. The core logic in the IC is clocked at 1GHz and 5% of the interconnect and 5% of the logic circuits are dedicated to clock distribution. It is estimated that in the remainder of the core logic, the signals change state on the rising edge of the 1GHz clock with a probability of 0.2.

c. *Show that the dynamic power consumption as a consequence of switched capacitance in a CMOS circuit is:*

$$P_{sw} = \alpha . f_{clk} . V_{DD}^2 \sum_{i=1}^{n} C_i$$

ensure that you define the terms and any approximations that are made in reaching this result.

Substantially, the inputs to a gate and the interconnect between gates appears to be capacitive. As a wire connected to an input cycles from 0V→V_{DD}→0V, the capacitance is charged and discharged. The charge comes from V_{DD} and is discharged to 0V. This is a current.

Consider charging a capacitor, C to V_{DD}. The charge on the capacitor will be CV_{DD}. As the capacitor is discharged to 0V, this charge will flow down to earth. The net charge moved through V_{DD} is, therefore, CV_{DD}. If this operation is being done f times a second then the charge moved per second is fCV_{DD}. and this is current, axiomatically. This current flows across V_{DD} and so the power dissipated by this switching activity, P_{sw}, is fCV_{DD}^2. To put this in terms of a circuit, $C=C_{in}+C_{wire}$ (the sum of the gate's input capacitance and the capacitance of the wire driving the gate input), and f is the frequency at which the input is being driven.

If we are to extend this expression from a single gate to an entire circuit we must perform a summation across all of the gates and interconnect in the circuit. So if we assume that there are n wires in the design and the total capacitance associated with wire$_i$ and the load that it is driving is C_i then the total switching power dissipated by the circuit should be:

$$P_{sw} = \sum_{i=1}^{n} f_i C_i V_{DD}^2$$

This expression assumes that each wire is switching at its own frequency, f_i. However, in practice, the switching of all the wires will be controlled by a single frequency f_{clk} and each wire will change state of the rising clock edge with a defined probability α_i. In this case, the expression becomes:

$$P_{sw} = f_{clk} V_{DD}^2 \sum_{i=1}^{n} \alpha_i C_i$$

In many cases, it is possible to simplify the expression further by assuming a value for α that is representative for the whole circuit rather than an individual wire.

$$P_{sw} = \alpha f_{clk} V_{DD}^2 \sum_{i=1}^{n} C_i$$

(4)

d. *Estimate the total amount of power consumed by the IC and comment on the figure.*

If 95% of the logic is not driven by the clock then this is equivalent to 19M NAND gates. Assume that all the wires switch unrelatedly and if the probability of a transition is 0.2 then α=0.1. The total capacitance being switched is $20x10^6x(2x1.076x10^{-15}+8.07x10^{-16})+80.6x10^{-9}=139.78x10^{-9}$F. Therefore, $P_{sw|core}$ = $0.95x0.1x1x10^9x1.2^2x139.78x10^{-9}$ = 19.12W.

The logic/interconnect driven by the clock is 5% but α = 1. Therefore, $P_{sw|clock}$ = $0.05x1x1x10^9x1.2^2x139.78x10^{-9}$ = 10.06W.

Therefore, the total power consumed in the core is 29.18W but this is only 60% of the power consumed by the IC and so the total power is 29.18/0.6 = 48.63W.

This is not an unreasonable figure. The complexity of the device is akin to a large microprocessor (such as a Pentium) with approximately 80M transistors and a 1GHz internal clock frequency. Such a power consumption would require relatively sophisticated heat removal techniques. **(7)**

4. a. *Show how the cost associated with a processed Si wafer might be estimated.*

Looking at capital investment and depreciation, a new 300mm, 0.13μm Fab may cost \$2.7B (of which 85% will be equipment). Lets assume that this cost is amortised over 10 years and that 5% of the equipment needs to be replaced on a yearly basis. Furthermore, assume that the capital is borrowed. Consequently, the cost of the investment per annum might be \$300M. The cost of equipment in the fab is 0.85x\$2.7B = \$2.3B and 5% of this needs to be re-invested annually. Consquently, the yearly equipment bill is \$110M.

These costs will be a significant proportion of the cost of running the Fab. However, there will be additional costs such as labour. The industry calculates that the productivity of each worker is 55 wafer layers/day/person and a high-volume Fab will start 20000 wafers per month with 25 layers in a set of masks (involving in excess of 900 individual operations in total). Consequently, the number of workers needed is:

$$Wkrs = \frac{WafPerMonth \cdot Masks}{WafLayDayPer \cdot 28} = \frac{20000 \cdot 25}{55 \cdot 28} = 325$$

Assume that the annual labour cost is \$100K per person then the labour cost will be \$33M. The figure for capital, depreciation, and labour is \$475M and there might be an additional 13% on top of this for cost of operating the facility. Thus the overall operating cost of the Fab will be \$475Mx1.13 = \$540M per annum.

This cost can be amortised over the number of wafers produced to give an approximate cost to produce each individual wafer. However, 16% of wafer starts will not result in finished wafers (these may be test, dummy or monitor wafers). Consequently, the total number of wafers per annum is 20000*12*0.84 = 202000 and so the cost to produce an individual finished wafer is \$540M / 202000 = \$2700. **(4)**

b. **i)** *What is yield and what factors influence it?*

Yield is the proportion of working die that emerge from the fabrication process. An IC may be deemed to be defective because it fails to perform the function correctly or because some parameter of its behaviour (e.g. speed) does not lie within the specifications set. Every stage of the manufacturing process can suffer problems. For example:

problems with the process itself may result in a systematic yield that affects all of the ICs on a wafer.

Point defects on the wafer, dust, etc. result in failures on individual ICs. A single defect in a single transistor or part of the interconnect can result in a IC to fail.

Dicing ICs can result in failure and the attach and bonding process

ii) *Give one expression that has been used to model random yield.* **(2)**

One of the following, for example:

Poisson Model $Y_r = e^{-DieArea \cdot DefectDensity}$

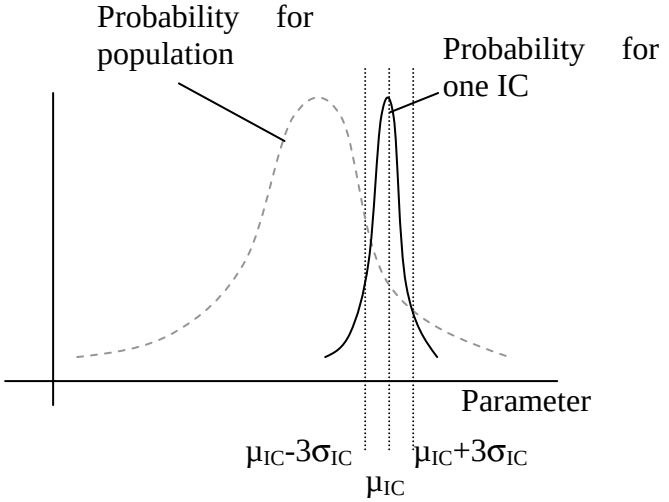
Seeds' Model $Y_r = e^{-\sqrt{DieArea \cdot DefectDensity}}$

Murphy's Model $Y_r = \left(\frac{1 - e^{-DieArea \cdot DefectDensity}}{DieArea \cdot DefectDensity} \right)^2$

ITRS Model $Y_r = \left(\frac{1}{1 + \frac{DieArea \cdot DefectDensity}{5}} \right)^5$ **(2)**

c. *How does process variation affect the functioning of an IC? Ensure that you distinguish between variation across a population of ICs and within an IC, and give an example.*

Any aspect of an IC's behaviour might be affected by variation in process conditions – e.g. leakage current, threshold voltage. All of these variations can manifest themselves as changes in behaviour. However, if a number of ICs are examined – for any parameter, then there will be variation. The parameter may well vary within an IC but the extent of the variation may be smaller.



For example, consider an 8 bit flash Analogue to Digital Converter (ADC) that is fabricated in a process where the accuracy of a resistor fabricated on the IC is governed by $3\sigma_{IC}$=0.5% of the nominal resistor value. The ADC contains 256 matched resistors and the accuracy of the ADC depends upon the matching of the resistors. Let us assume that the overall accuracy of the ADC (defined in whatever way) can be met to an acceptable level if the resistors are matched to better than 0.5%: this information can be used to determine the yield of ADCs meeting the specification?

Assuming a Gaussian distribution, 99.73% of all resistors will meet the matching constraint and so:

$$yield = 0.9973^{256} = 0.50$$

i.e. 50% of ADCs meet the specification. **(4)**

d. Distinguish between the following and identify the factors that are making each of them increasingly difficult and problematic:

i) *Verification*

Verification is the process of determining – prior to implementation – that the design will function correctly when implemented. Both in terms of functionality and in terms of all other specifications being met. For example, power consumption, clock frequency. This must be done into all corners of process, temperature, and supply voltage variation. Verification is becoming the major challenge as technology shrinks and ICs grow in size. This is because the computational difficulty of functionally verifying large designs is outstripping the computing resource. As technology shrinks the variation in behaviour increases and the uncertainty attached to the resultant behaviour of the design once fabricated makes it more difficult to determine whether the IC once fabricated will function correctly. In the future, this may require that designs become fault tolerant. **(4)**

ii) *Manufacturing Test*

Because manufacturing is prone to failure and prone to variation (and this, as was said in the previous section is increasingly the case) it is necessary to test ICs, one manufactured, to ensure that they function as required. This is becoming increasingly difficult because size and speed of designs is going up – there is more to test and if it must be tested at speed then this is an additional burden on test equipment. The length of tests (in terms of states) goes up *at least* linearly with the size of a design. Pin counts are rising and parametric tests require increasing bandwidths from test equipment adding to cost. Fault coverage, the proportion of possible faults uncovered by tests, also needs to rise towards 100% to ensure that acceptable levels of yield can be achieved. Furthermore, the effect of essentially analogue faults – delay variation, I_{CCQ}, etc. will require increasingly sophisticated, and probably slower tests. Finally, the problem of signal integrity may result in problems that manifest themselves under only certain conditions making exhaustive test impractical – requiring other solutions. **(4)**

END OF ANSWERS