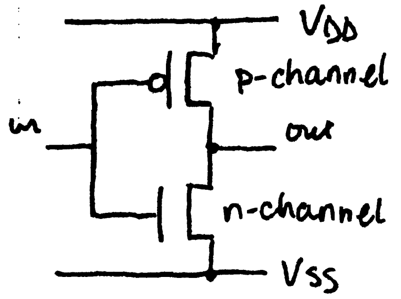
4B7 2005 Qn 1

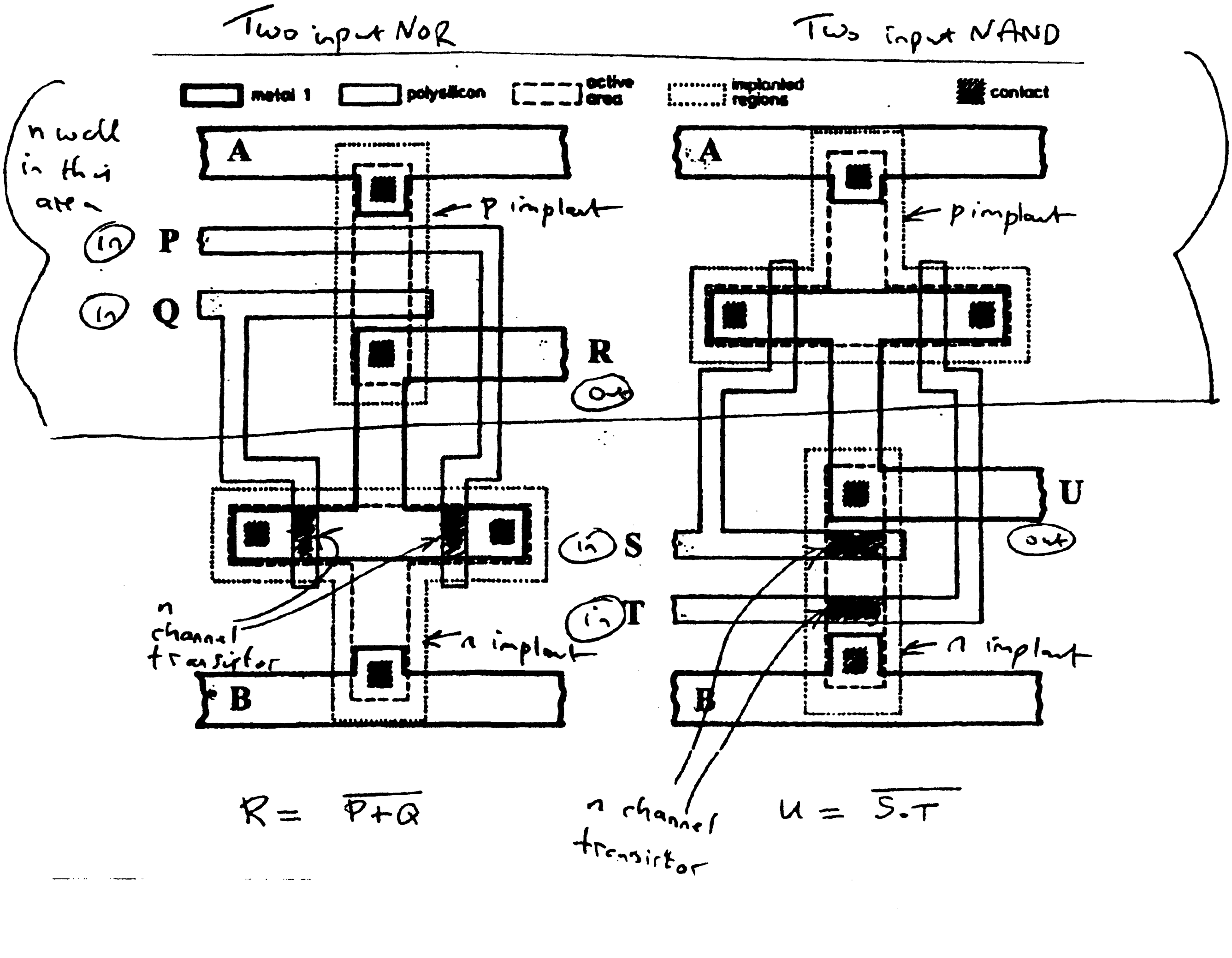
a) CMOS inverter



An important advantage of CMOS technology is the lower power consumption since very little current is drawn except during a switching transient. This allows dense circuits to be fabricated without exceeding thermal limitations during operation

One disadvantage is the relatively low hole mobility requiring wider p-channel devices to achieve high performance and symmetrical operation. This leads to increased capacitance. GaAs devices are faster but the materials technology is more complex.

Silicon dominates because of low cost and a well-developed technology



R = $\overline{P+Q}$ U = $\overline{S.T}$

4B7 2005 Qu 1 (cont.)

From inspection of Fig. 1,

2NOR	$W_n = 9mm$	$W_p = 9mm$	$r_{np} = 1$
2NAND	$W_n = 9mm$	$W_p = 9mm$	$r_{np} = 1$

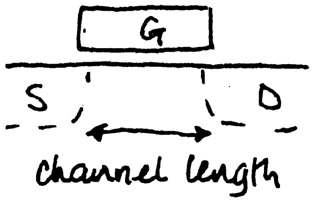
c) Since $\mu_n/\mu_p = 2$, the worst-case fall time for the 2NOR device is via one n-channel device. The worst-case rise time is determined by 2 series p-channel devices. Hence the rise time is much slower.
Similar rise/fall delays can be achieved by ensuring
$r_{np} = 1/4$ for the 2NOR

For the 2NAND device, the worst-case for rising outputs is with conduction via a single p-channel device: the worst case for falling output involves 2 series n devices. Hence for similar delays we require:
$r_{np} = 1$ for the 2NAND - as drawn

In fact for a general purpose circuit we are interested in minimising the overall delay, so one approach might be to consider how to minimise the sum of rising & falling delays (i.e. for a pair of gates in cascade); in which case it is found that (for the 2NOR) r_{np} is somewhat greater than 1/4, to allow for capacitive effects.

d) The most important dimension determining switching speed is the channel length (source-drain separation). Reducing this gives faster carrier transit time and improved device performance, limited ultimately by short-channel & punch-through effects in sub-100nm devices

X-section

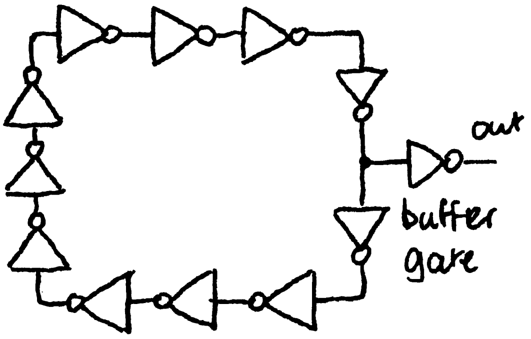


Presently manufactured devices have an effective electrical channel length of ~130nm. Research devices have been demonstrated with 20nm gate lengths and good electrical characteristics.

If the lithographic manufacturing technology continues to improve over the next decade 20nm channel length devices will be manufactured.

4B7 2005 Qn 2

a) Ring oscillator

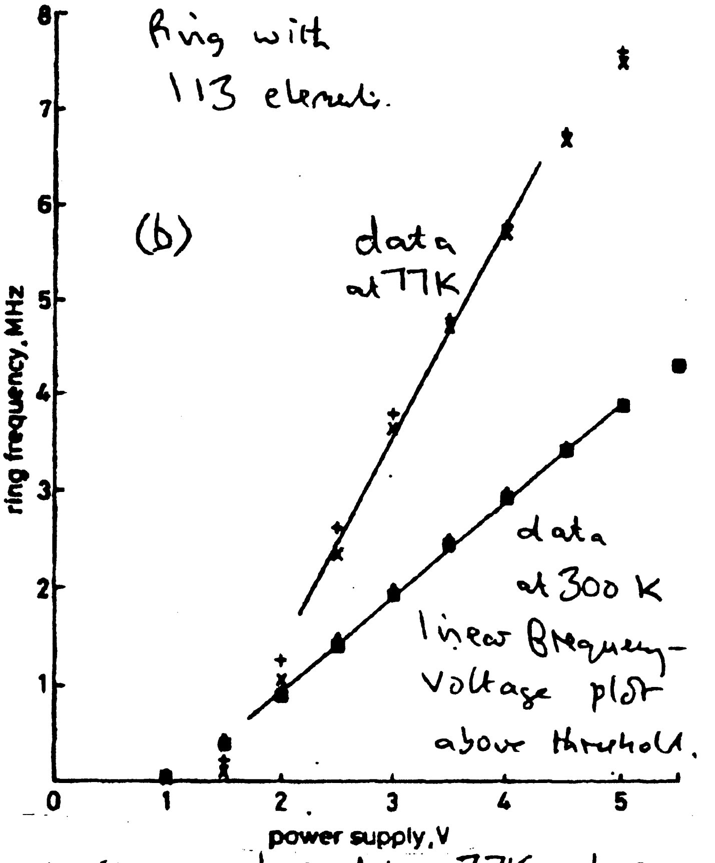


The switching speed of individual minimum-geometry CMOS transistors is much faster than that of a pad driver which may have to drive a high-capacitance load.

A ring comprising an ODD number of inverting gates connected in cascade is unstable and will oscillate at a frequency determined by the number of gates and their individual.
An output buffer is then used to buffer the output and drive a pad.

The switching speed is largely determined by the transistor channel length (S-D distance) and capacitance of the next stage to be driven.

b) From the graphs, the two different chips have similar characteristics, and the data neatly overlay



d) At 3V and 300K, 2MHz corresponds to a gate delay

$$\frac{1}{2\times10^6}\times\frac{1}{2\times113}=2.2nS$$

At 3V and 77K, 3MHz corresponds to a gate delay of 1.5nS

At 3V, the switching speed increases by one third as the temperature is reduced to 77K, because

4B7 2005 Qn 2 (cont.)

d) of reduced carrier scattering and hence increased mobility at low temperature
This voltage is well above the threshold at 77K, about 1.5 V, hence changes in device threshold do not play a dominant role

At 1.5V, however, the ring slows down on cooling since the power supply is now comparable with the device threshold voltage.

c) Modelling the input to the device as a lumped capacitor Cinput implies that the charge required to switch a device between one power supply rail and the other will increase linearly with Vsupply. However, the frequency of operation (i.e. of charge/discharge) is actually seen to increase linearly above the threshold voltage.

Hence we conclude total current $\propto (V_{supply})^2$ above threshold

e) An 'unloaded' inverter (with fan out of unity) switches unrealistically fast because of the small capacitive load being driven. With more typical fan out of 3, actual switching times are 2-4 times slower than this.

A more realistic ring could have 'dummy' devices, similar to the ring elements themselves, acting as loads in addition to the ring stages themselves.

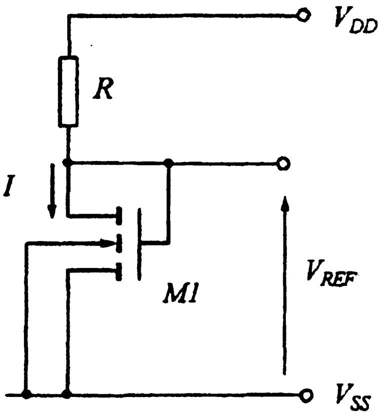
f) Switching speed is expected to improve as the devices are miniaturised, roughly in proportion to the square of the channel length (shorter channel -> faster switching). Switching speeds of order 10ps have been observed in research chips with channel lengths of 20nm

Present commercial technology is about three times slower.

4B7 2005 Qn3

a) It can be seen that since G is shorted to D,

$V_{GS} \equiv V_{DS}$ and M1 is always in saturation



Hence the second given equation, for

$0 < V_{GS} - V_T < V_{DS}$

applies

From Kirchhoff $\quad V_{ref} = V_{DD} - IR \quad (1)$

Sat. MOS Eqn $\quad I = \frac{1}{2}\frac{\mu\epsilon}{t_{ox}}\frac{W_1}{L_1}(V_{ref} - V_{tn})^2 \quad (2)$

Rearrange (2) $\quad \underline{V_{ref} = V_{tn} + \sqrt{\frac{I}{\frac{W_1}{L_1}\frac{\mu\epsilon}{t_{ox}}}}} = V_{DD} - IR$

Hence $\quad \underline{V_{ref} = V_{tn} + \sqrt{\frac{V_{DD} - V_{ref}}{\frac{RW_1}{2L_1}\cdot\frac{\mu\epsilon}{t_{ox}}}}}$

b) Substitute the given values:

$$2 = 1 + \sqrt{\frac{12-2}{\frac{R}{2}\frac{80}{4}\times 2.0\times10^{-5}}}, \quad \text{and} \quad \frac{12-2}{10R + 2.0\times10^{-5}} = (2-1)^2 = 1$$

Thus $R = 50\,k\Omega$ and $I = (12-2)/50\times10^3 = 200\mu A$

Using a standard poly Si of approx. 50Ω square, this resistor would require L/W = 1000, and would occupy a great deal of space

4B7 2005 Qn 3 (cont)

c) An ideal voltage reference is independent of the power supply that services. Practical references fall short of this ideal however, and 'P.S. sensitivity' expresses this in a quantitative way - it is defined:

$$S_X^{V_{REF}} = \frac{\partial V_{REF}}{V_{REF}} \Big/ \frac{\partial X}{X} = \frac{X}{V_{REF}} \times \frac{\partial V_{REF}}{\partial X}$$

Where V_{REF} is the voltage reference, X is the parameter under consideration, say, V_{DD} for power supply. If S is unity, a 10% change in V_{DD} will result in a 10% change in V_{REF}. The objective is to produce a circuit design in which $S_X^{V_{REF}}$ is as small as possible for relevant parameters X, such as V_{DD}.

An ideal voltage reference will also be independent of ambient temperature. Since most components & materials used in electronic circuits have temperature-dependent properties, and in different ways, this is a non-trivial problem. To quantify the effect of components due to temperature change we define 'fractional temperature coefficient TC_F, defined as follows for a component of value X:

$$TC_F = \frac{1}{X}\frac{\partial X}{\partial T} = \frac{1}{T} S_T^X \quad \text{using the above notation}$$

TC_F is typically expressed as parts per million per deg. C, or ppm/degC. In circuits comprising several components that determine the output, all may contribute to the TC_F of the output itself, and the various contributory values of TC_F must be taken into consideration with the governing equations that determine the output. TC_F may be +ve or -ve depending on materials of which the component is made, & on mode of operation. This opens up the possibility of using devices in combination in circuits such that the temperature dependent effects cancel to a significant extent.

Achieving minimum $S_{V_{DD}}^{V_{REF}}$ and TC_F simultaneously for a voltage reference is a considerable challenge

4B7 2005 Qn 3 (cont...)

d) Some of the techniques available are:

(i) use of a zener diode where the reverse breakdown voltage characteristic is almost independent of current. Most zener diodes have a significant temperature coefficient, minimised by use of devices in which the different physical effects responsible (zener and avalanche effect) wholly or partially cancel. Only achieved at certain voltages. Zener diodes may not be compatible with commodity CMOS processes & may have to be provided off-chip

(ii) use a band-gap reference - an arrangement in which the V_F of a pair of forward-biased diodes are operating at different current densities are subtracted in an amplifier to give a result almost independent of temperature

(iii) use an XFET, where the difference in pinch-off voltage of two similar FETs is used to provide a reference

- both (ii) and (iii) call for special processing that may not be compatible with commodity CMOS

(iv) A reference proposed by Kwon et al (covered in lectures) uses only components available in a regular CMOS process.

A thermal voltage is generated by use of a pair of MOSFETs operating sub-threshold. This has a positive temperature coefficient.

A voltage is derived from a self-biased β-multiplier circuit, again based on MOS transistors which has output proportional to V_T: this has a -ve. temp. co.

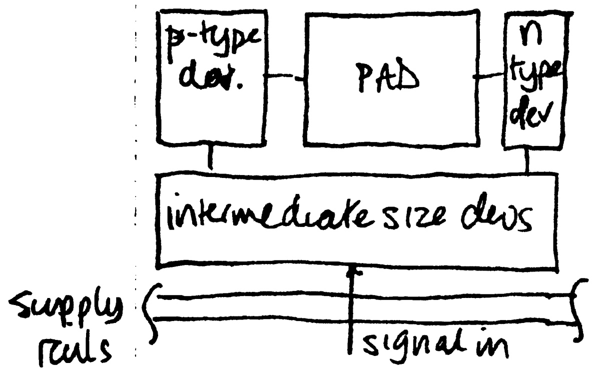
When these are scaled appropriately and summed in a suitable amplifier, the result has a near-zero tempco.

4B7 2005 Qn 4

a) CMOS VLSI designs for digital applications are normally implemented with minimum geometry devices wherever possible
(i) to economise on space
(ii) for faster operation
(iii) for lower power consumption
Devices like these are unsuitable for direct connection to external circuits since these may impose large capacitive, resistive or inductive loads, and may give rise to transient voltages or currents outside the safe operating range, which can induce latch-up, and in extreme cases can cause permanent damage thru' static discharge
Also, to facilitate wired connections it has become the custom to use bonding pads - squares/rectangles of metallisation typically O(100 μm) in dimension to which fine wires may be bonded (ultrasonic cold-welding). These wires contribut R, L, C (as do external elements)

Output drivers must therefore be provided to interface between the min-geom. devices and the pads - pad drivers They consist of high current-capacity (large W/L) devices which can supply the transient current surges needed to charge/discharge the pad & attached external circuit. They may also have to supply static or changing currents to a resistive/inductive load. However, because of their large channel area such devices have high input capacitance and themselves need to be driven by transistors of greater size

b)



Typical pad driver layout

To reduce the area occupied:
(i) use interdigitated / folded structures to economise on space
(ii) successive stages are progressively increased in size & drive ability
The number of stages is minimised subject to constraints of acceptable delay.

To reduce risk of latchup all devs must have multiple well/substrate taps to V_{DD} / V_{SS}. Guard rings may be incorporated to minimise risk of injecting minority carriers into other latchup-sensitive circuits

c) To drive the high-C output load and minimise delay, it is necessary to use stages of progressively greater W/L. Later stages have higher conductance to charge/discharge nodal capacitances, which are themselves larger because of the use of larger devices

It can be shown the optimum number of stages to minimise delay depends purely on the ratio of the output driven capacitance to the input capacitance : hence this is

$\ln \left(C_{load} / C_{input} \right)$ C_{load} is the driven capacitance
C_{input} is the capacitance at the input to the driver

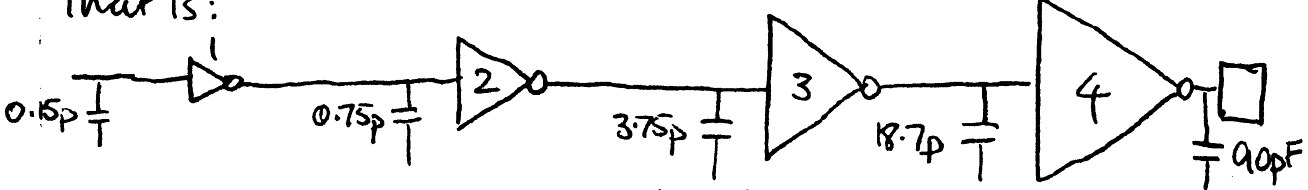
Each successive stage should have its W/L increased in the same ratio.

Here 4 stages are to be used. If each stage has W/L U times that of the previous stage:

$$U^4 = C_{load} / C_{input} = 90/0.15 = 600$$

Take logs: $4 \log_{10} U = \log_{10} 600 = 2.778$, so $U = 4.95$
(Say 5)

Since the capacitance at the input of the first stage of the driver is 0.15pF, the successive driven capacitances will be inflated by $U \sim 5$. Hence each device will have W/L 5 times greater than that in the previous stage. That is:



And W, L values must therefore be

		1	2	3	4	
n	L	0.5	0.5	0.5	0.5	µm
	W	1	5	25	125	
p	L	0.5	0.5	0.5	0.5	
	W	2	10	50	250	

The p devices are scaled in proportion to μ_n / μ_p to achieve equal delays for rising/falling signals

4B7 2005 Qn 4 (cont...)

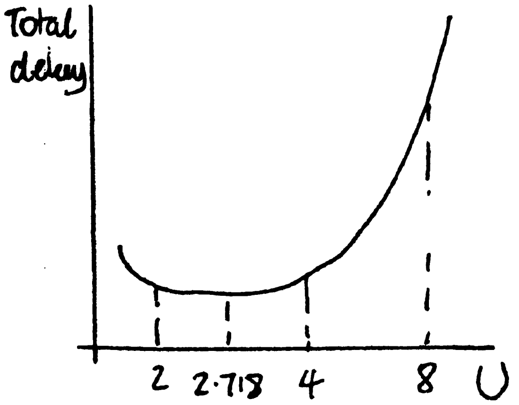
(C)(i) The delay through each stage is the same, and is equal to:

$$\frac{3 \times 0.75 \times 10^{-12}}{2.5 \times 10^{-4} \times (1/0.5) \times 3} \simeq 1.5nS$$

Each subsequent stage has to drive 5× the capacitance but its W/L is scaled by ×5 so the delay is unchanged. This assumes that the capacitance of the transistors dominates all other sources within the driver

Hence total delay is $4 \times 1.5nS \sim 6nS$

(C)(ii) If we plot total delay vs U, it has the form below:



There is a clear minimum at U = 2.718. This is achieved with the number of stages N:

$$N = \log_e {}^{90}/_{0.15} = 6.39$$

Clearly N must be an integer

N=6 is closest & gives a delay very close to the minimum and has the required non-invert characteristic. However, note that the area occupied is expected to be considerably greater than for the case of U=5. The total delay curve is fairly flat near the minimum, so there is a trade off to be made of area versus delay, but a big economy of area can be obtained with acceptable increase in delay.

For N=6, the total delay T_6 can be determined:

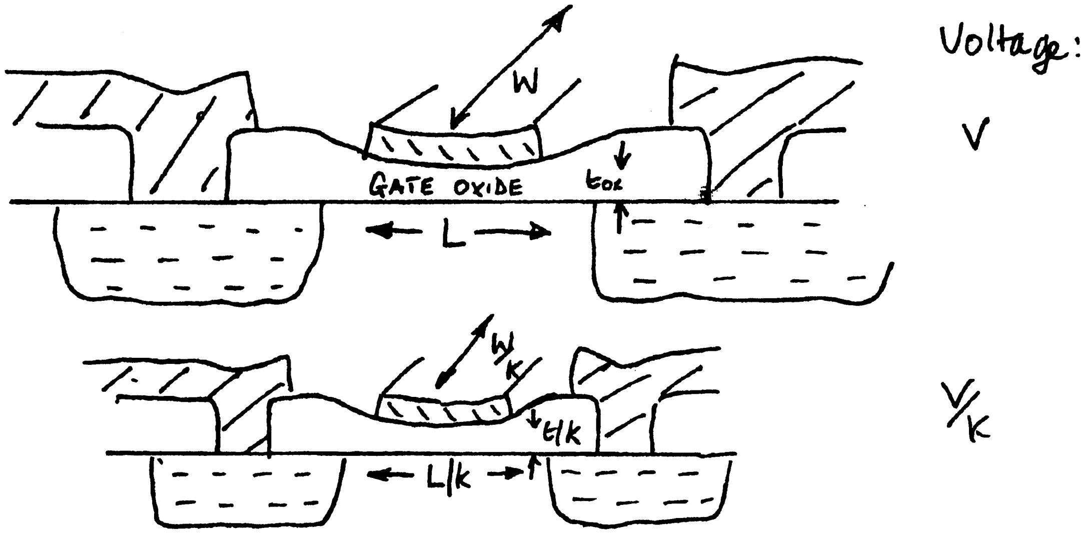
$$6 \log_{10} U = \log_{10} (90/0.15) = 2.778 \rightarrow U = 2.9$$

Hence $T_6 = \frac{2.9}{5.0} \times 1.5 \times 6 = 5.2nS$ only about 12% faster than the four stage design, but with an area penalty

4B7 2005 Qn 5

a) In constant-field scaling, geometric dimensions are progressively reduced with the evolution of foundry technology. Electrode voltages are also scaled to maintain electrostatic fields. Unless this is also done MOSFET operation is impaired, requiring major process alterations, in doping density, etc, etc.

Assume all dimensions and voltages are reduced as below:



For the MOS devices: the effects are

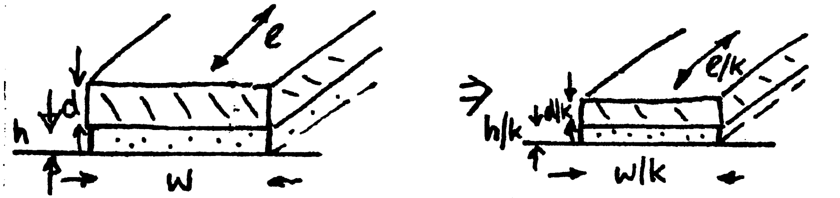
(i)	Gate area	A	$\propto LW$	decreased by	k^2
	Field at channel	E	$\propto V/t_{ox}$	unchanged	
	Transit time	τ	$\propto L^2/V$	decreased by	k

since carrier velocity $= \mu E$ and distance $= L$. This affects speed

(iii)	Capacitance	C	$\propto LW/t_{ox}$	decreased by	k
(iv)	Static current $I \propto CV/\tau$		$\propto V^2LW/L^2t_{ox}$	" by	k
(ii)	Gate delay	$\propto \tau$	$\propto L^2/V$	" by	k
(v)	Static pwr $\propto IV$		$\propto V^3LW/L^2t_{ox}$	" by	k^2
(vi)	Powerdensity	IV/A	$\propto V^3LW/L^3Wt_{ox}$	unchanged	

We consider interconnect separately:

4B7 2005 Qn 5 (cont)



$l \rightarrow l/k$
$w \rightarrow w/k$
$d \rightarrow d/k$
$h \rightarrow h/k$

interconnect length, width l, w, thickness d;
dielectric thickness h

Wire resistance R $\propto l/dw$ increased by k
Capacitance C $\propto lw/h$ decreased by k

(vii) Current density J $\propto I/dw$ dncreased by k

(note that I is decreased by k in normal scaling)

(vi) Signal delay $t = RC$ $\propto l^2 w/dwh$ unchanged
or time constant
This is so for wire interconnects whose lengths scale. Note that in terms of clock cycles, this is effectively an increase, since a faster clock is liable to be used with the scaled process

b) Main points: scaled devices offer

- higher packing density, but NB trend towards larger, more complex chips mean that some elements do not scale in the same way as MOS devices

- higher speed of operation, with faster clock - but note that this nullifies certain other performance enhancements brought about by scaling

- lower current (and much lower power consumption - down by k^2) provided the clock is unchanged; but if advantage is taken of improved speed and the clock is scaled, current remains unchanged. Hence bigger voltage drops across scaled interconnect.

- There are many other subtle consequences that need careful evaluation when a scaled process is to be adopted.

4B7 2005 Qn 5 (cont...)

c) Problem areas - lead to failure of simple model for larger K

- charge stored in gate reduced by k^2 so dynamic devices more liable to soft errors
- faster clock speeds coupled with trend to larger Ks leads to longer interconnect delays in comparison with the faster clock. Clock skews (differential delays) may rise as k^3
- Roff/Ron declines as dimensions decrease so role of sub-threshold conduction becomes more critical
- Contact resistances rise as contact structures are scaled
- Yields reduce at smaller geometries (economic issue)
- Onset of short-channel effects requires different MOSFET models that account for quantum tunneling effects.
- Velocity saturation in channel leads to reduction in μ and increase in channel resistance
- Oxide breakdown likely as local E-fields approach ~700 MV/m
- Drain-induced barrier lowering
- Substrate currents which may affect V_T (body effect)