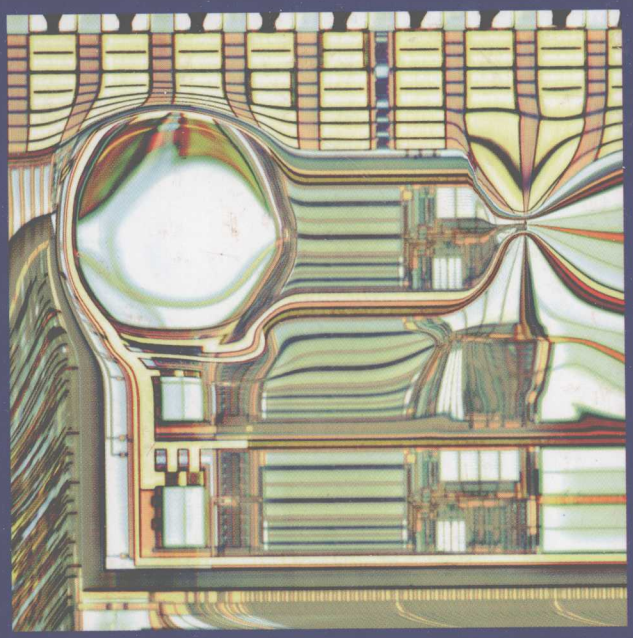
ANALOG DESIGN ESSENTIALS

Willy M.C. Sansen



Springer

CD-ROM
LUDED

ANALOG DESIGN ESSENTIALS

by

Willy M. C. Sansen

Catholic University, Leuven, Belgium

Springer

A C.I.P. Catalogue record is available from the Library of Congress.

ISBN-10 0-387-25746-2 (HB)
ISBN-13 978-0-387-25746-4 (HB)
ISBN-10 0-387-25747-0 (e-book)
ISBN-13 978-0-387-25747-1 (e-book)

Published by Springer,
PO Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Printed on acid-free paper

Printed in the Netherlands.

Dedication

This book is dedicated to my wife
Hadewych Hammenecker

Contents

Comparison of MOST and Bipolar transistor models

Willy Sansen
KULeuven, ESAT-MICAS
Leuven, Belgium

willy.sansen@esat.kuleuven.be

Willy Sansen 10-05 011

011

Analog design is art and science at the same time.

It is art because it requires creativity to strike the right compromises between the specifications imposed and the ones forgotten.

It is also science because it requires a certain level of methodology to carry out a design, inevitably leading to more insight in the compromises taken.

This book is a guide through this wonderful world of art and science. It claims to provide the novice designers with all aspects of analog design, which are essential to this understanding.

As teaching is the best way to learn, all slides are added on a CD-ROM, with and without the comments added as notes in the pdf files. The reader is suggested to try to explain parts of this course to his fellow designers. This is the way to experience and to cultivate the circles of art and science embedded in this book.

All design is about circuits. All circuits contain transistors. Hand-models are required of these devices in order to be able to predict circuit performance. CAD tools such as SPICE, ELDO, SPECTRE, etc. are then used to verify the predicted performance. This feedback loop is essential to converge to a real design. This loop will be used continuously in this book.

Table of contents

Ref.: W. Sansen : Analog Design Essentials, Springer 2006

Willy Sansen 10-05 012

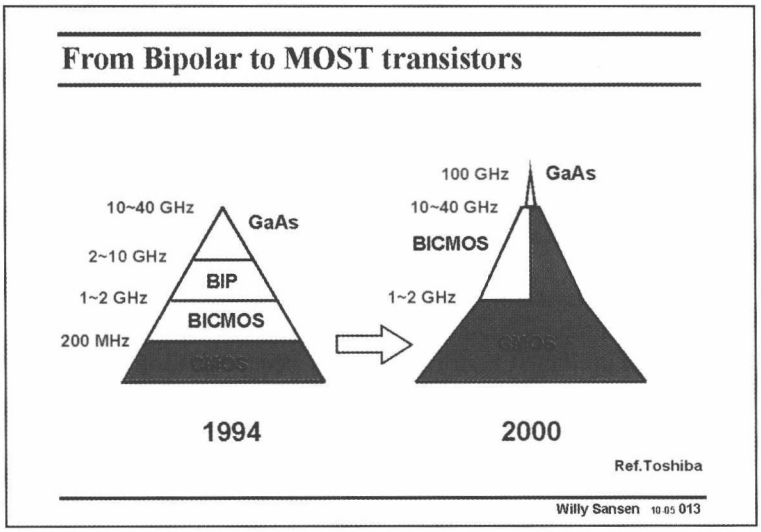
012

For the design of analog integrated circuits, we need to be able to predict the performance by means of simple expressions. As a result, simple models are required. This means that the small-signal operation of each transistor must be described by means of as few equations as possible. Clearly the performance of the circuit can then only be described in an approximate way. The main advantage however, is that transistor

sizing and current levels can easily be derived from such simple expressions. They can then be used to simulate the circuit performance by means of a conventional circuit simulator such as SPICE or ELDO.

In these simulators, models are used which are much more accurate but also much more complicated. These simulations are required afterwards to verify the circuit performance. The initial design with simple models is the first step in the design procedure. They are aimed indeed at the determination of all transistor currents and sizes, according to the specifications imposed.

We start with MOST devices, although the bipolar transistor are historically first. Nowadays the number of MOS transistors integrated on chips, vastly outnumber the bipolar ones.

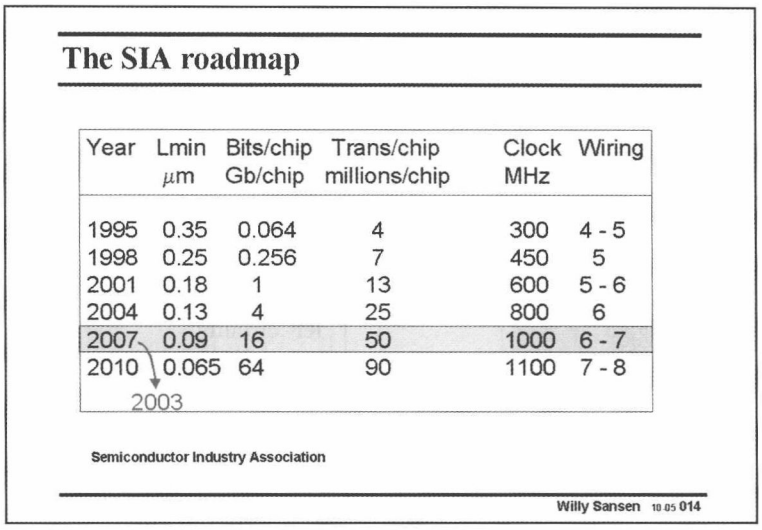


013

Indeed, previously CMOS devices were reserved for logic as they offer the highest density (in gates/mm^2). Most high-frequency circuitry was carried out in bipolar technology. As a result, a lot of analog functions were realized in bipolar technology. The highest-frequency circuits have been realized in exotic technologies such as GaAs and now InP technologies. They are quite expensive however and really reserved for the high frequency end.

The channel length of CMOS transistors shrinks continuously however. In 2004, a channel length of 0.13 micrometer is standard but several circuits using 90 nm have already been published (see ISSCC). This ever decreasing channel length gives rise to ever increasing speeds. As a result, CMOS devices are capable of gain at ever higher frequencies.

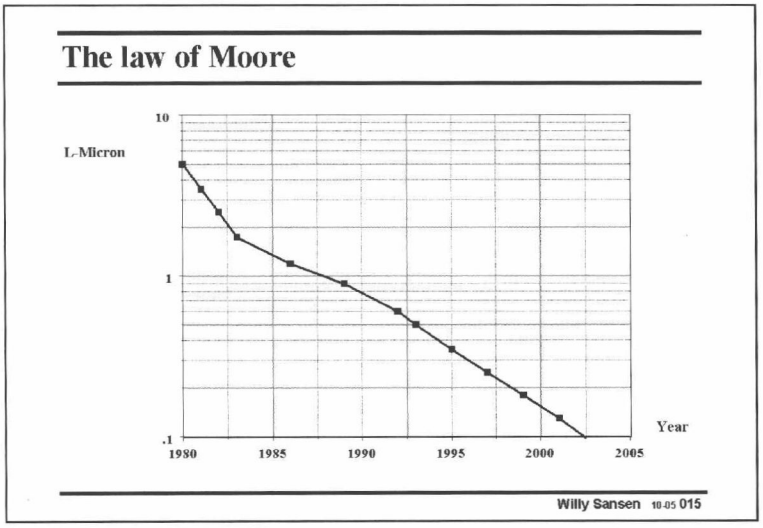
Today CMOS and bipolar technologies are in competition over a wide frequency region, extending all the way to 10 and even 40 GHz, as predicted in this slide. For these frequencies the question is indeed, which technology fulfills best the system and circuit requirements at a reasonable cost. BICMOS is always more expensive than standard CMOS technology. The question is, whether the increase in cost compensates the increase in performance?



014

This ever decreasing channel length has been predicted by the SIA roadmap. It tries to predict what the channel length will be in a few years, by extrapolating the past evolution.

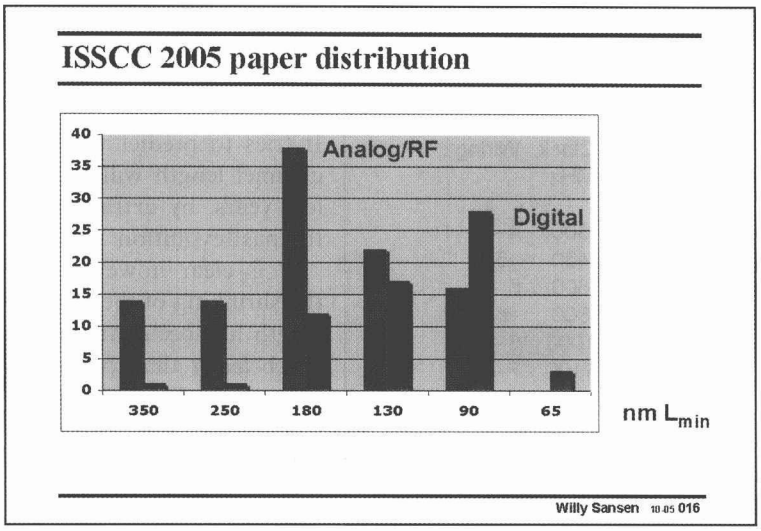
It is clear however, that the shrinking of the channel length has been carried out much faster than predicted. For example, the 90 nm technology was originally expected only in 2007, but was already offered in 2003. This technology was expected to allow 50 million transistors to be integrated on one single chip. Present day processors and memories offer double that amount. Moreover, this technology was expected to give rise to clock speeds around 1 GHz. High end PC's already clock speeds beyond 3 GHz!



015

This ever decreasing channel length has also been predicted by the curve of Moore. This is simply a sketch of channel length versus time. It is a graphic representation of the numbers of the SIA road map. Indeed 90 nm is reached in 2003!

The slope of that curve has not always been the same. Indeed, the slope was higher in the early eighties, but has declined a bit as a result of economic recessions. Also, the cost of the production equipment and the mask making grows exponentially, delaying the introduction of ever newer technologies somewhat.

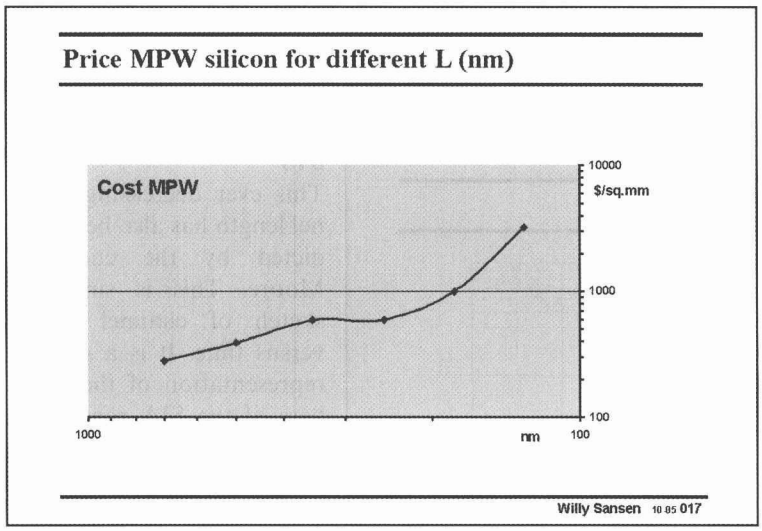


016

Which are the most used channel lengths today?

To explore this, the number of papers is shown of the last IEEE ISSCC conference (held at San Francisco in February) for two categories, the digital circuits and the analog or RF circuits.

It is clear that the digital circuits peak at 90 nm channel length, whereas the analog ones lag behind by two generations; they peak at about 180 nm.



017

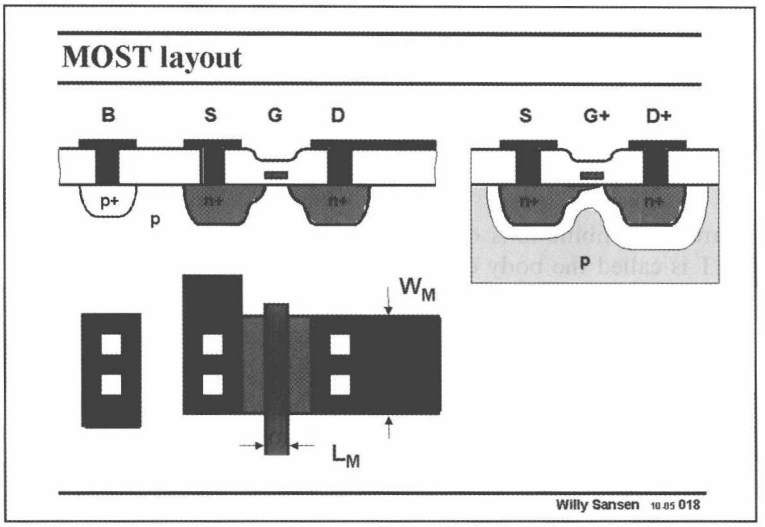
Indeed for small quantities, silicon foundries offer silicon at higher cost if the channel lengths are smaller. This is clearly illustrated by this cost of a Multi Project Wafer chip versus channel length. In such a MPW run, many designs are assembled and put together in one single mask and run. As a result the total cost is divided over all participants of this run. This has been the source of cheap silicon for many universities and fabless design centers.

The cost in $\$/mm^2$ is reasonable up to about 0.18 μm. From 0.13 μm on the cost increases dramatically, depriving many universities from cheap silicon.

What the cost will be of 90 nm and 65 nm is easily found by extrapolation! This shows very clearly that a crisis is at hand!

018

Let us have a closer look now at a MOST device. What are the main parameters involved, and what are the simplest possible model equations that still describe the transistor models in an adequate way for hand analysis.

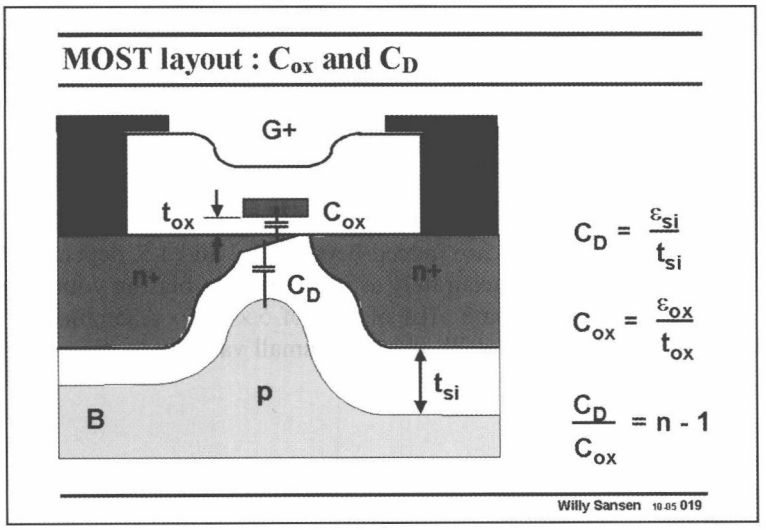


The cross-section of a MOS transistor is shown with its layout. On the left, the MOST is shown without biasing. On the right, voltages are applied to Gate and Drain.

The main dimensions of a MOST are the Length and Width. Both are drawn dimensions on the mask. In practice they are usually a bit smaller. This is a result of underdiffusion and some more technological steps. In this layout the W/L on the mask is about 5.

Application of a positive voltage at the Gate V_{GS} causes a negatively charged inversion layer, which connects the Source and Drain n+ islands. It is a conducting channel between Source and Drain and thus acts as a resistor between Source and Drain.

Application of a positive voltage at the Drain V_{DS}, with respect to the Source, allows some current to flow from Drain to Source (or electrons from Source to Drain). This current is I_{DS}. As a result, the channel becomes non-homogeneous. It conducts better on the Source side than on the Drain side. The channel may even disappear on the Drain side. Nevertheless, the electrons always manage to make it to the other side, because they have acquired sufficient speed along the channel.



019

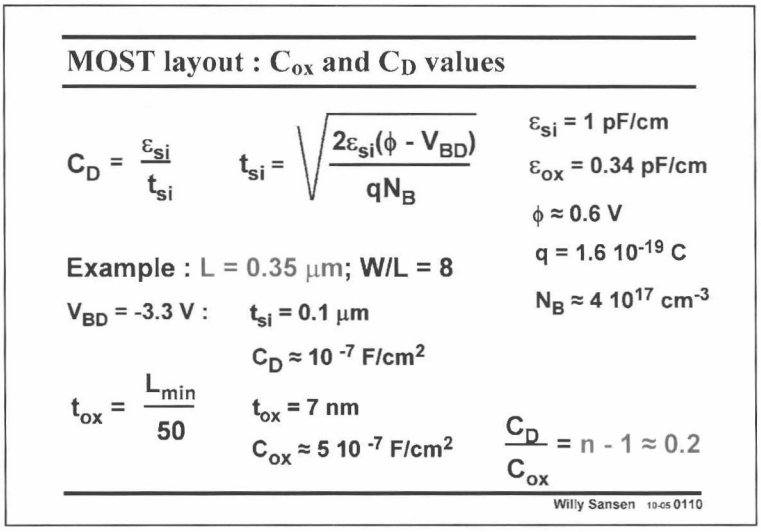
Zooming in on the channel region, disappears once V_{DS} is too high.

The channel region, together with the two n+ islands of Source and Drain, are enveloped by an isolation layer. Indeed, in a pn junction the p and n regions are always separated by an isolation region, which is called depletion region. The silicon is depleted of electrons or holes; it is non-conducting, it is an isolator, very much as oxide an isolator is.

The oxide has a thickness t_{ox}, whereas the depletion layer has a thickness of t_{si}. Both give rise to capacitances C_{ox} and C_D, respectively. Both have dimensions F/cm^{-2}. Normally C_D is about one third of C_{ox} as we will calculate in detail on the next slide. Their ratio is n-1 [Tsividis].

It is mportant, however, to note that the channel inversion layer is coupled to the Gate by means of C_{ox}, but as much coupled to the Bulk by C_D. Changing the Gate voltage will thus change the conductivity of the channel and hence the current I_{DS}. In a similar way, changing the Bulk voltage will thus also change the conductivity of the channel and will thus change the current I_{DS} as well. The top gate gives the MOST operation, whereas the bulk gives JFET operation. Indeed, a Junction FET is by definition a FET in which the current is controlled by a junction capacitance.

All MOST devices are thus parallel combinations of MOSTs and JFETs. We normally use only the MOST whereas the JFET is called the body effect, and is treated as a parasitic effect.



0110

The width of the depletion region depends to a large extent on doping levels and the voltage across it. The larger the doping levels on both sides of the junction, the narrower the depletion region is. On the other hand, the larger the voltage is across the depletion region, the wider this region becomes, as shown by the equation.

It includes the silicon dielectric constant ε_{si}, the junction built-in voltage φ, the charge of an electron q and the bulk doping level N_B. Values are given in this slide.

For example for a 0.35 mm technology, a drain-bulk voltage V_{BD} yields a depletion layer thickness of about 0.1 nm. It is about 14 times thicker than the gate oxide. This is offset somewhat by the fact that the silicon dielectric constant is three times higher than the oxide one. Silicon is three times more efficient to make capacitors with than oxide. Silicon capacitances are very nonlinear because they depend on the voltage, whereas oxides capacitances do not.

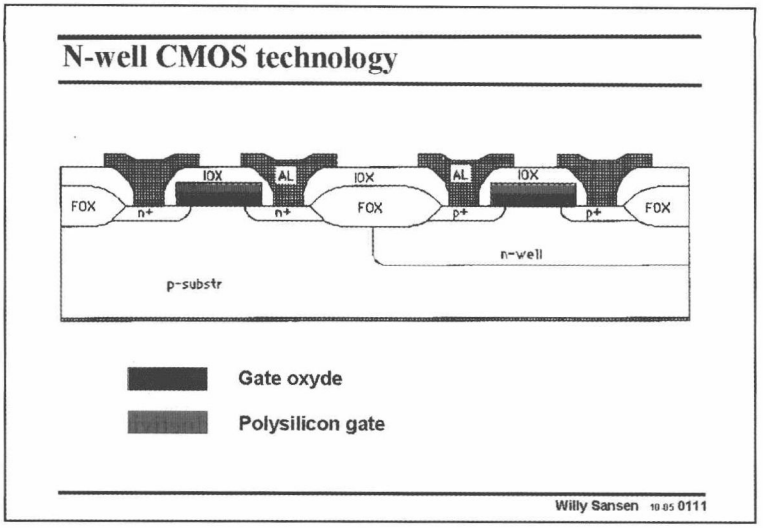
The ratio $n-1$ is then about 0.2. Most values of n are indeed between 1.2 and 1.5, depending on the value of t_{si}. Parameter n is thus never known accurately as it depends on biasing voltages.

Note that all capacitances are in F/cm^2. For a Gate area of WL of $5 \times 0.35 \times 0.35\ mm^2$ the total Gate oxide capacitance is thus $C_{ox}WL \approx 5$ fF, which is quite a small value indeed!

0111

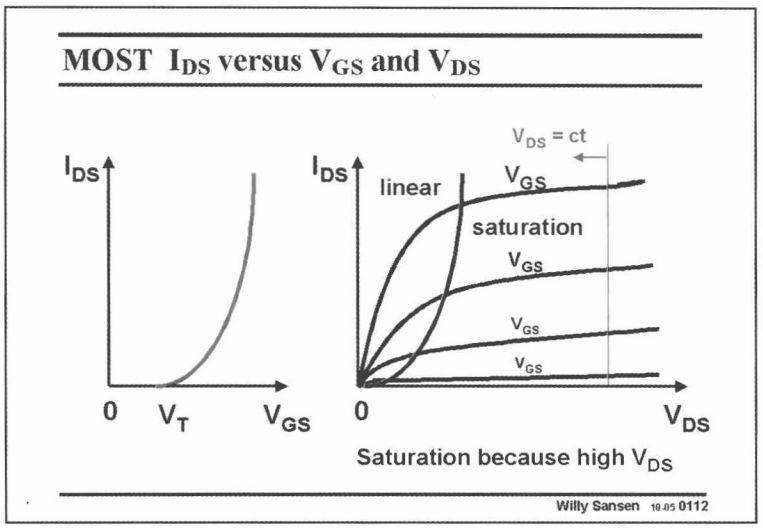
The bulk doping level N_B is not the same for a nMOST and a pMOST device. Indeed normally nMOST devices are implemented directly on the p-substrate. This substrate is thus common to all nMOS transistors on that chip.

The pMOS transistor has a p-channel however and has to be implemented in a n-tub or n-well, which is always higher doped than the common p-substrate. The disadvantage is that the bulk doping for a pMOST is always higher than the bulk doping of a nMOST. The pMOST



C_D will be higher and so is its n factor. The advantage of a pMOST however is that its bulk is isolated from the substrate and can be used to control the transistor current I_{DS}. Such pMOST devices have two gates, i.e. a top gate and a bottom gate. Both can be driven independently.

Most technologies are n-well CMOS technologies although some p-well ones are still around.



0112

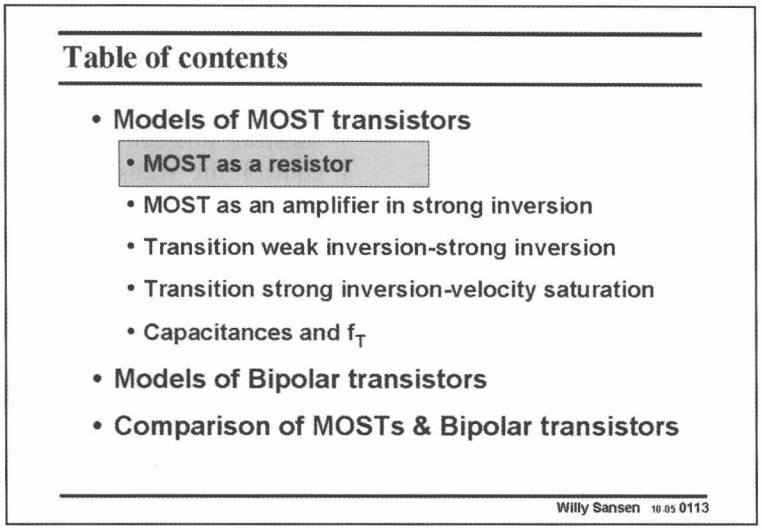
Application of a positive Gate voltage V_{GS} causes an inversion layer (or channel) which connects Source to Drain. Application of a positive voltage V_{DS} causes some current I_{DS} to flow from Drain to Source. Now we want to find simple expressions for this current, so that we can use them for design purposes.

The curve of I_{DS} versus V_{GS} is sketched on the left. The current starts flowing as soon as V_{GS} exceeds V_T, called the threshold voltage. For larger values of V_{GS}, the current increases in a nonlinear way. How much we actually exceed V_T is $V_{GS}-V_T$; this will be the most important design parameter later on!

The curve of I_{DS} versus V_{DS} is sketched on the left. For small values of V_{DS}, the current increases linearly. Indeed, the transistor behaves as a resistor. This is called the linear region.

For larger values of V_{DS} the current stops increasing but levels off towards nearly constant values: the current is said to saturate. This is called the saturation region. Curves are given for four different values of V_{GS}.

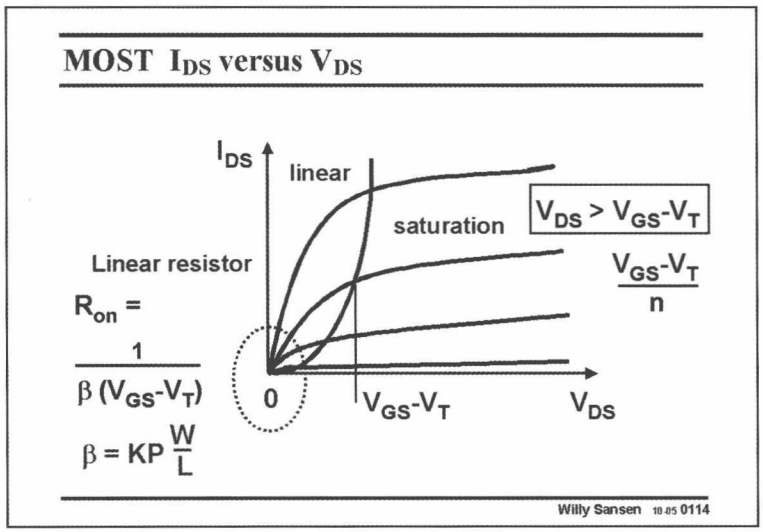
The linear and saturation regions are separated by a parabola, which is described by $V_{DS}=V_{GS}-V_T$. We will concentrate on the linear region first.



0113

In many applications a MOST is simply used as a switch. Its voltage V_{DS} is then very small. The MOST is then operating in the linear region (sometimes called the ohmic region). In this region the MOST transistor is really a small resistor. It provides a linear voltage-current characteristic. The channel has the same conductivity at both sides – the Source side and the Drain side.

Let us investigate what the actual resistance then is.



0114

Zooming in on the corner, for very small values of V_{DS}, we find that indeed the $I_{DS}-V_{DS}$ curves are very linear. The MOST behaves as a pure resistor.

The resistance value R_{on} is given in this slide. In addition to the dimensions W and L, a technological parameter appears, called KP.

This parameter characterizes a certain CMOS technology as will be explained on the next slide. Its dimension is A/V^2.

It is clear that the transistor turns very nonlinear when we apply larger V_{DS} voltages. The crossover value towards the saturation is reached for $V_{DS}=V_{GS}-V_T$, or more accurately for $V_{DS}=(V_{GS}-V_T)/n$. We will drop this factor n however, as a kind of safety factor. We will, from now on, assume that a transistor is operating in the saturation region provided $V_{DS}>V_{GS}-V_T$.

0115

For sake of illustration let us a have a closer look at this resistor "in the corner". For this purpose we have to find an easy approximation for KP. It is given in this slide. Factor β (Greek beta) contains both the parameter KP and the dimensions of the resistor W and L.

Actually, KP contains the oxide capacitance C_{ox}, and the mobility μ (Greek mu). This factor

MOST parameters β , KP , C_{ox} , ...

$\beta = KP \frac{W}{L}$

$KP = \mu C_{ox}$

$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}$

$t_{ox} = \frac{L_{min}}{50}$

$KP_n \approx 300\ \mu A/V^2$

$C_{ox} \approx 5\ 10^{-7}\ F/cm^2$

$\varepsilon_{ox} = 0.34\ pF/cm$

$\varepsilon_{si} = 1\ pF/cm$

$t_{ox} = 7\ nm$

$L_{min} = 0.35\ \mu m$

$\mu_p \approx 250\ cm^2/Vs$

$\mu_n \approx 600\ cm^2/Vs$

Willy Sansen 10-05 0115

shows what speed (cm/s) an electron can develop, subject to an electric field (V/cm). It is given in cm^2/Vs. Electrons travel about twice as fast as holes.

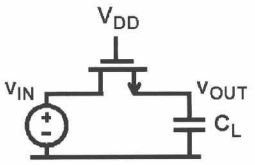
Values for a standard 0.35 μm CMOS technology are given in this slide.

Note that the oxide thickness is about L/50, as has been checked on most standard CMOS processes of the last 20 years.

As a rule of thumb, the resistor of a square transistor (W/L = 1) for a drive voltage $V_{GS} - V_T = 1$ V is about 3.4 kΩ in 0.35 μm CMOS.

For deeper submicron CMOS technologies, KP is higher because of C_{ox}. This square resistor now decreases!

Example : Analog switch on CL



We want to switch 0.6 V to a load capacitance C_L of 4 pF.
We want to do this fast, with time constant 0.5 ns.
Supply voltage V_{DD} = 2.5 V
V_T = 0.5 V
Use standard 0.35 μm CMOS.

Choose minimum channel length and find an average V_{GS} !

Willy Sansen 10-05 0116

0116

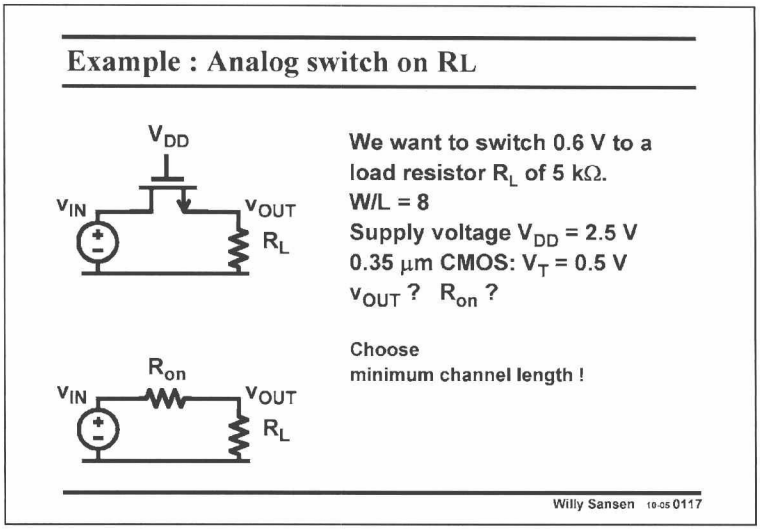
To have a time constant of 0.5 ns with 4 pF we need a switch resistance of 125 Ω.

This will, to a large extent, depend on the value of $V_{GS} - V_T$ used. Indeed, as soon as the switch turns on, the output voltage is still at zero Volt and $V_{GS} - V_T = 2$ V. At the end of the switching in, the output voltage has risen to 0.6 V: it has become the same as the input voltage. The $V_{GS} - V_T$ has decreased by 0.6 V towards $V_{GS} - V_T = 1.4$ V. The average value is now $V_{GS} - V_T = 1.7$ V.

For a transistor size W/L = 1, the on-resistance is thus 2 kΩ (using $KP = 300\ \mu A/V^2$). This is 8× larger than what we can allow. We thus have to take a W/L of 8.

Note that we will have great difficulties in switching large input voltages. Indeed, for $v_{OUT} = v_{IN} = 2$ V, the V_{GS} has become zero. As a result, the resistor has become infinity: the switch cannot be switched on any more!!

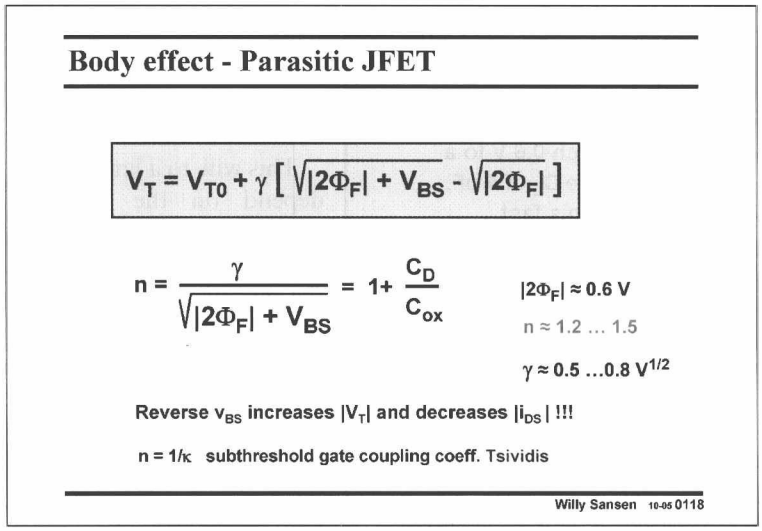
Note also that we have forgotten to take into account the bulk effect. Indeed, V_{BS} is not zero, it is 0.6 V. The parasitic JFET will play as well as we will see later.



0117

For a transistor size W/L = 8, the KP × W/L product is 2.4×10^{-3} S (using KP = 300 µA/V^2). Taking the switch as a resistor with value R_{on}, as shown in this slide and substitution of R_{on} by its expression, requires an iteration, which yields a value of R_{on} of 216 Ω and an output voltage of 0.575 V.

Note that we have forgotten to take into account the bulk effect. Indeed, V_{BS} is not zero, it is about 0.575 V. The parasitic JFET will become active as well, as we will see next.



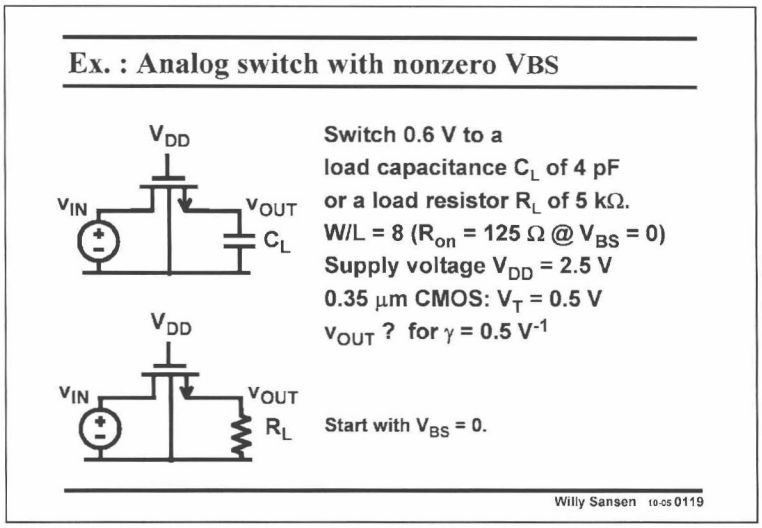
0118

The drain-source current I_{DS} and the channel resistance R_{on} show the influence of V_{GS} in an explicit way, but not that of the bulk-source voltage V_{BS}. Indeed, the effect of V_{BS} is embedded in the threshold voltage V_T.

Increasing the V_{BS} will increase the depletion layer width under the channel and will increase V_T. More reverse biasing that junction will increase V_T in absolute value and decrease the current. For zero V_{BS}, V_T evidently equals V_{T0}.

Parameter γ (Greek gamma) has to do with the junction depletion region and is linked to parameter n. Actually, factor γ depends on the technology used (such as the bulk doping N_B) but is not voltage dependent. The denominator of n now shows explicitly the voltage dependence of n.

Some approximate parameter values are given as well for a 0.7 µm CMOS.

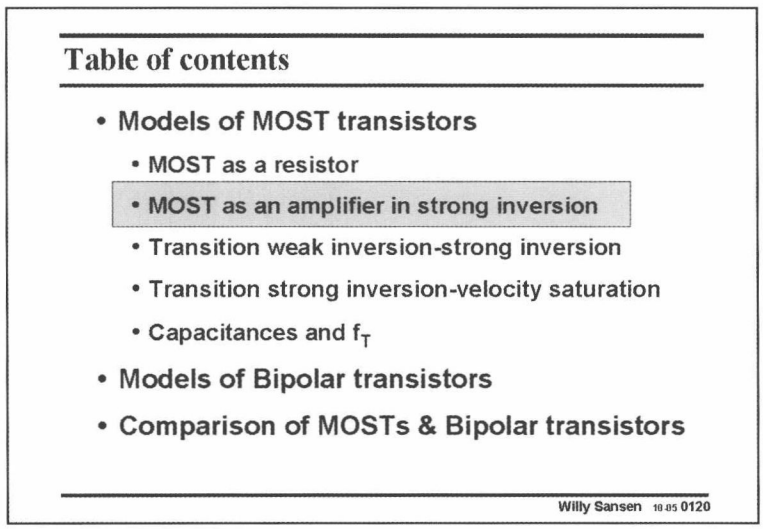


0119

For a transistor size W/L = 8, the KP × W/L product is again 2.4×10^{-3} S (using KP = 300 μA/V²).

Taking the switch as a resistor with value R_{on}, as shown in this slide, and substitution of R_{on} by its expression requires another iteration, as now V_T depends on the output voltage. This yields a value of R_{on} which is now larger. It is 291 Ω, instead of 216 Ω.

Also, the output voltage is a little bit lower. It is 0.567 V instead of 0.575 V. The time constant is now simply the product of 216 Ω and 4 pF.



0120

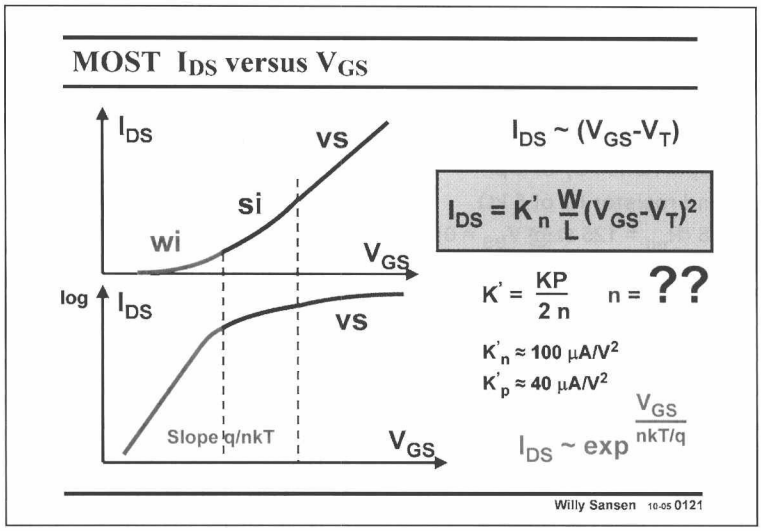
In most applications the MOST is used as an amplifier. This means that its V_{DS} is larger than $V_{GS} - V_T$. Its transconductance is then higher than at low V_{DS} values. It is used to generate gain.

The value of $V_{GS} - V_T$ itself, however, determines in which region the MOST is operating. For medium currents the MOST is operating in the strong inversion region. This is used most of the time.

At lower currents the MOST ends up in weak inversion. This is especially important for portable and low-power applications in general.

If we bias the MOST at the highest possible transconductance (for example in RF applications and very-low-noise applications), then the current densities are higher. The transconductance of the NMOST is then limited by velocity saturation. Again, another model is required.

All three regions are now discussed.



0121

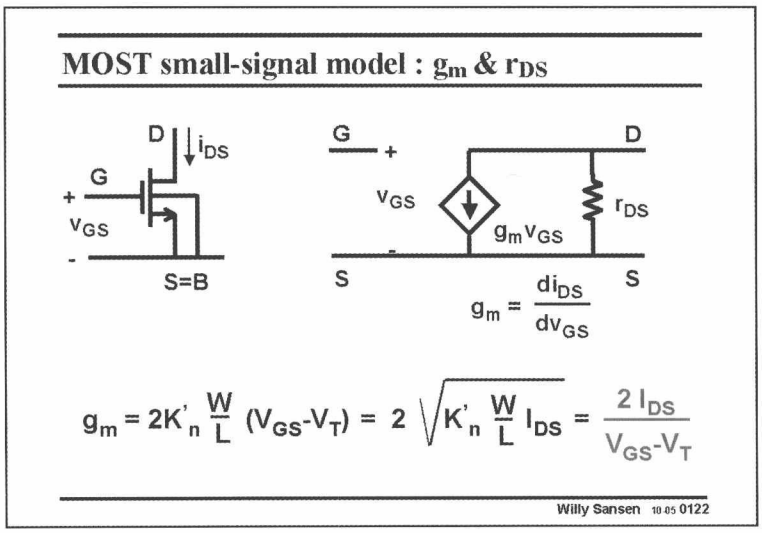
In most amplifiers the MOST operates in the saturation region, i.e. we maintain $V_{DS} > V_{GS} - V_T$ at all times. We obtain the $I_{DS} - V_{GS}$ curve shown before. A closer look however, reveals that this curve has three distinctive regions. The one in the middle is called the strong-inversion region or square-law region as the current expression contains the factor $(V_{GS} - V_T)^2$.

At lower currents we find the weak-inversion region, or exponential region because the current expression now contains an exponential in V_{GS}. Indeed a $\log(I_{DS})$ curve is linear in that region.

At higher currents the $I_{DS} - V_{GS}$ curve becomes linear, because of several physical phenomena. The most important one is velocity saturation: all electrons reach their maximum speed v_{sat}.

Most transistors are biased in the strong-inversion region because this is a good compromise between current efficiency and speed, as explained later. In this region, the current expression is simply proportional to $(V_{GS} - V_T)^2$, but includes a technological parameter as well K′.

This parameter K′ is linked to the one in the linear region KP by the ratio 2n. It is thus always smaller than KP. It is not very accurately known however, because of the mobility (in KP) and especially n. Remember that n depends on biasing voltages and so does K′.



0122

Let us now make an amplifier using one single nMOST.

We assume that transistor is biased at some DC current I_{DS}. We want to know now what is the small-signal or AC current i_{DS} superimposed on it by application of a small-signal input voltage v_{GS}.

For this purpose we have to find the transistor transconductance g_m. This is nothing else than the derivative of the drain current to the Gate-Source voltage, as shown by the expression on the left.