NSS 2006 Short course #3
Integrated Circuit Front Ends for Nuclear Pulse Processing
Lecture "Semiconductor technology for Integrated Circuit Front Ends"
Slide notes

Giovanni Anelli, CERN

Section: Operation and characteristics of MOS and Bipolar transistors

Slide 3
In this section we will see (or maybe review for many of you!) the basic operating principles and most important formulas for MOS and bipolar devices.

Slide 5
This slide shows a simplified NMOS transistor. It is a 4 terminal device. The gate is on top of a thin layer of silicon dioxide. Underneath the silicon dioxide we have the channel, which puts in communication source and drain. The bulk (or body or substrate) is the fourth terminal (not to be forgotten).
The direction y is also called "vertical", direction x "horizontal". The distance between the source and drain diffusions in the x direction is called gate length L. The channel width W is the width of the source and drain diffusions in the z direction.
The gate is normally made with highly doped polysilicon.
The example in the slide is for a NMOS. For a PMOS we have something very similar, but with all the dopings changed. Source and drain are p-doped diffusions for a PMOS. The gate is normally polysilicon with a high n-doping for NMOS and p-doping for PMOS. The "substrate" of a PMOS will be n-doped. In a p-substrate CMOS technology the PMOS transistors will therefore be built in an n-well. In advanced CMOS processes we might have a well for the NMOS transistors too.
The symbol for a PMOS is similar but the arrow for the source terminal goes towards the gate. In textbooks and papers you can also find other symbols.
We will spend quite some time explaining how a MOS transistor works, but essentially we can say that for analog circuits it is normally used as a voltage to current amplifier (more precisely as a voltage to current converter), and for digital circuits as a switch.

Slide 6
Rising Vg first depletes the p-type silicon under the gate from holes. This means that the acceptor ions of the p-doped substrate become negatively charged. Rising Vg again will call more electrons form the source and drain, where they are zillions! A conductive channel is formed.
When we apply a Vds we will have less electrons close to the drain than to the source. Why? Because the space charge region close to the drain is larger, and therefore it has already more negative charges (ions) to compensate for the positive charges on the gate. Therefore we will have less free electrons close to the drain than to the source. Rising Vds at a certain point the channel close to the drain is pinched off, and the device enters the saturation region.
For Vds small enough compared to Vgs (I will be more quantitative later) the current flowing between source and drain depends linearly on Vds. We have a resistor. The resistance depends on the "size" of the channel. The more voltage on the gate, the more electrons in the channel, the less resistance we will have.
For Vds big enough compared to Vgs the channel is "pinched-off", Ids does not depend as a first approximation on Vds anymore and we have a current source. The amount of current is again controlled by Vgs.
Another way to look at saturation is the following. The current is constant along the channel: charge carriers are slower closer to the source and faster closer to the drain. Thus more electrons are needed on the source side than on the drain side, and the channel is pinched-off.

Slide 7
Here we better illustrate with a real measurement what explained in the previous slide.
Let's first focus on one of the 3 curves, for example the topmost. Here we have a fixed V_{GS} and we increase the V_{DS} from 0 to V_{DD}. When $V_{DS} < V_{DS_SAT}$ (the drain-to-source saturation voltage), the transistor is in linear region and it behaves as a Voltage Controlled Resistor (VCR). The value of the resistor is the inverse of the slope of the curve. When V_{DS} is high enough ($> V_{DS_SAT}$) then the channel is pinched off close to the drain. From now on, the current does not depend from V_{DS} anymore (but only as a first approximation), and the device is a Voltage Controlled Current Source (VCCS). As it is seen in the curves, the current still increases with V_{DS}, due to an effect called Channel Length Modulation (see later). The slope of the I_{DS} vs V_{DS} characteristics in the saturation region is called Output conductance.
In the linear region (VCR), the value of the resistor is controlled with the gate voltage. In the plots, the higher the slope the higher the gate-to-source voltage. This makes sense, as a bigger V_{GS} means that the channel is richer in electrons, so less resistive, so the conductance (slope of the plot in linear regions) is higher.

In the saturation region (VCCS), the device behaves as an almost ideal VCCS, and the current is controlled again by V_{GS}. The higher V_{GS}, the higher the current.

Slide 8

This plot shows how the drain-to-source current varies with V_{GS}, for two different V_{DS} voltages: the lower plot is with $V_{DS} < V_{DS_SAT}$ (device in the linear region), the higher is for $V_{DS} > V_{DS_SAT}$ (device in saturation).

In both cases, we can see that the current is practically 0 for V_{GS} voltages below a value called threshold voltage (V_T). For V_{GS} voltages above V_T, the device current increases, more or less rapidly depending on the value of V_{DS}.

The region in which V_{GS} is below V_T is called Subthreshold region. For $V_{GS} > V_T$ we will be in the linear region or in saturation, depending on the value of V_{DS}.

The slope of the I_{DS} vs V_{GS} plot is called transconductance, which tells us how much the drain-to-source current varies when varying the gate-to-source voltage. It expresses the "gain" of our voltage to current "amplifier", and it is one of the most important parameters in analog design.

When V_{GS} and/or V_{DS} are sufficiently high, the vertical and longitudinal electric fields in the device become high enough to cause some problems, such as, for example, a reduction of the transconductance and the saturation of the carrier speed. We will discuss these issues a bit more in detail later.

Slide 9

This plot is the same as before (Ids current vs Vgs voltage), but the y axis is in log scale. We can see from this plot that there is a current flowing in the device even for voltages Vgs smaller than Vt.

This current (called subthreshold current) is orders of magnitude smaller than the current flowing above threshold, but it has the important property of varying exponentially with the Vgs voltage (linear part of the characteristic). The MOS transistor behaves here almost as a bipolar (as we will see later).

The subthreshold region (Vgs < Vt) is also called weak inversion region, opposed to the strong inversion region (for Vgs > Vt).

Another important parameter to be defined is the leakage current, which is the current flowing in the device when Vgs = 0 V.

Slide 10

Here we have the equations for the MOS transistor in Strong Inversion (Vgs > Vt).

The voltage (Vgs – Vt) is also called gate overdrive, and tells us how much we have biased the gate above threshold. The drain-to-source saturation voltage is linked to the gate overdrive.

Linear region: for small Vds voltages, the current varies (almost) linearly with Vds (we have a resistor). How much it deviates from linearity depends on the value of Vds compared to the gate overdrive (and therefore Vds-sat). The closer we are to the transition between linear region and saturation, the more it deviates from linearity. The slope of the Ids vs Vds plot (inverse of the resistance of the VCR) is given by Beta*(Vgs – Vt – n*Vds). For a fixed gate overdrive, we have an almost constant value resistor only when Vds <<< Vgs – Vt. The transconductance in this region is not very interesting, as it is very small.

Saturation region: here the current does not depend on Vds anymore (we have already mentioned that this is true only as a first approximation), and it depends on the square of the gate overdrive voltage. The transconductance can be expressed as a function of the gate overdrive voltage or as a function of the drain-to-source current (this is a useful equation when designing a circuit).

In all the equations we have the parameter Beta. This is given by the product of the aspect ratio W/L of the transistor, of the mobility and of the gate oxide capacitance per unit area Cox. W and L are respectively the transistor gate width and length. The mobility of the carriers depends on the kind of carrier: electrons are faster than holes. Cox depends on the gate oxide dielectric constant and is inversely proportional to the gate oxide thickness.

n is a number which is close to one. It is expressed as a function of the bulk transconductance and of the gate transconductance. What is the bulk transconductance? We can vary the current in the device also varying the bulk-to-source voltage, not only varying the gate-to-source voltage. This is a somewhat less effective way of modulating the drain-to-source current. Therefore, the bulk transconductance is normally only a fraction of the gate transconductance.

Another thing to notice: the transconductance (we always refer to the gate transconductance if not differently specified) can be expressed as a function of Beta and of the drain-to-source current. Therefore it depends on:

uCox, a technology parameter: this is fixed once the technology is chosen

W/L, a geometrical parameter: this is specified by the circuit designer, but is fixed once the circuit is fabricated

I_{DS}, a design parameter: this is also specified by the designer, and can be made variable in the real circuit.

One last important aspect to stress is that all these equations are valid only as a first approximation. They are all good for back-of-an-envelope calculations and to understand better how the device works and how the parameters, voltages and currents are linked between each other. But in a modern MOS device there are many effects which make the real device behavior departing from these ideal formulae.

Slide 11
In weak inversion, the drain-to-source current depends exponentially on Vgs. The (normally unwanted) dependency of the current on Vds can be eliminated at relatively low Vds voltages, just a few thermal voltages (this becomes very interesting at low temperatures!). When Vds is high enough, the device becomes an excellent VCCS.
What is very important here is that the transconductance depends only on the drain-to-source current, and not on the device dimensions anymore (as long as the device is in weak inversion).
Also, in this region the transconductance is proportional to the current (and not to the square root of it, as it is the case in strong inversion).

Slide 12
As we have already mentioned, in strong inversion and in saturation our MOS transistor is a VCCS, but it is not ideal. The current still changes with Vds. The slope of the Ids vs Vds characteristics in saturation is called output conductance of the transistor, and gives an indication of the amount of non ideality of the current source. As we will see later, together with the transconductance the output conductance is another very important parameter for analog design.
The fact that Ids increases with Vds is due to a phenomenon called channel length modulation. Increasing the drain voltage increases the space-charge region around the drain. This moves the pinch-off point in the channel towards the source; it is like shortening the channel, and therefore the current increases.
It can be shown that the output conductance decreases (i.e. becomes better) for longer channels: this can be easily understood, as in a longer channel a variation in the position of the pinch-off point will have a smaller effect.
For a given device dimension, the output conductance increases with the current.

Slide 13
The dependency of the drain-to-source current in saturation on the drain-to-source voltage has to be modeled in the equations. This is done with a parameter called “lambda”, which in the ideal case is = 0.
The output conductance can then be expressed as the product of lambda and of Ids at the onset of saturation (i.e. @ Vds = Vds_sat).
The inverse of the output conductance is called output resistance (ideally infinite).
As we have already mentioned, lambda depends on the channel length L. lambda can be expressed as a function of L and of a voltage (per unit channel length !!!), in a way similar to what is done for the bipolar transistor (Early voltage),
The output conductance is a very difficult parameter to model, especially in advanced deep submicron processes, as it is influenced by many effects.

Slide 14
There are a few more effects / equations which we need to add:
- Bulk effect: the threshold voltage depends on the source-to-bulk voltage.
- There are several parasitic resistors in an MOS transistor. One of the most disturbing is the source parasitic resistance, which results in a decrease in the transconductance.
- The voltage applied to the gate results in a vertical electric field, which reduces the mobility of the carriers in the channel (and therefore the transconductance). The exact value of the carrier mobility depends on the gate overdrive and, of course, on the doping in the channel region
- Maximum operation frequency: it depends on the transconductance and on the gate-to-source capacitance. It can also be expressed as a function of gate overdrive and gate length L. As it can be seen, the dependency on length is quite strong (square): as long as the device stays in strong inversion and does not enter velocity saturation (see later), there is a gain in speed going to more advanced processes (shorter L).

Slide 15
The saturation of the carrier velocity is a phenomenon which became more and more important in more advanced CMOS processes.
The equations valid in strong inversion do not hold anymore in velocity saturation.
Ids, for example, does not depend anymore on the square of the overdrive voltage but only on the overdrive voltage. Also, Ids does not depend on L anymore.
The transconductance loses the dependence on L and on Ids. This means that once in velocity saturation it is useless to increase Ids again: we would waste power and not gain anything.
The maximum frequency of operation is not depending on the gate overdrive anymore and has a weaker dependence on L.

Slide 16
An interesting plot can be obtained showing how the g_m / Ids ratio varies with the current density (plot in a log scale). The g_m / Ids ratio is an indication of the “power efficiency” of the device: it tells us how much transconductance I get for a given current (power) expense.

We can see that this ratio varies depending on the inversion region of the device: it is maximum in weak inversion, and then decreases going to moderate inversion, strong inversion and velocity saturation.
This is a quite useful information, especially for low power design: for a given device dimension, in weak inversion we get the maximum power efficiency. The absolute value of the transconductance is much smaller than what we can get for the same device in strong inversion, but the power efficiency is much higher. But, as always in analog design, there is a trade off here: for a given current, making the device larger pushes it towards weak inversion, increasing the g_m but decreasing the speed at the same time (due to the increase in the size).
The g_m / Ids ratio does not change with Ids in weak inversion, is proportional to the inverse of the square root of Ids in strong inversion and is proportional to the inverse of Ids in velocity saturation. These trends can be better seen plotting the curve with log scales for both axis, as it is done in the next slide.

Slide 18
Let's do some justice to the poor PMOS transistor!
There is not much to say here, a part from the fact that almost everything is the same but opposite!
One important difference: the mobility of holes is smaller than the one of electrons. In pure silicon (which is not the case of a MOS transistor channel!!!) there is a factor 3 difference: 1417 cm^2/V*s for the electrons, 471 cm^2/V*s for the holes. These values are much smaller for the electrons and holes in a MOS transistor. The ratio between them depends on the technology, and can vary from 2 to 4 (roughly).
There are a few more subtle differences which we will not discuss here.
NOTE: usually schematics are designed in a way that node potentials rise going from the bottom to the top of the sheet. This is why I have drawn the source above the drain.

Slide 19
This slide illustrates the simplest small-signal equivalent circuit of a MOS transistor. Small signals are small variations of the bias currents and voltages.
In the equations in the slides, capital letters are used for bias quantities, and small letters for small signals.
This circuit shows that the transistor can be seen as a Voltage Controlled Current Source (VCCS) between drain and source. The control voltage is the gate-to-source small signal voltage. r_0 is the output resistance of the transistor. It is in parallel to the VCCS and represents the non ideality of the current source.
At low frequencies the gate is completely isolated from all the other terminals of the device.
In the slide we have taken the equation for the bias current in strong inversion. In other inversion regions the correct equations have to be taken.
The equation which defines the small signal behavior is valid no matter which inversion region we are in: it is just the value of the transconductance which will have to be calculated accordingly with the equations we have already seen. From the small signal equation we can see that the transistor can be seen as a voltage-to-current amplifier: modulating the gate-to-source voltage with a small signal, we obtain a modulation of the drain-to-source current. The transconductance is the gain of the amplifier.

Slide 20
Here we see a simple voltage-to-voltage amplifier. A small signal on the gate is converted through the transconductance into a signal current. The resistor R_D converts this signal current into a signal voltage at the output.
The plot illustrates well that the transistor is biased in the bias point Q, and that the small signal is a variation of this bias situation. The voltage at the output can be found on these plots as the intercept between the transistor and the resistor characteristics.
Here we did not consider the bulk effect, as we assumed the bulk and the source shorted.
Also, as it can be seen in the Ids vs Vds characteristics, we assumed the output conductance of the transistor to be zero (output resistance infinite). In fact this is never the case: the gain of the amplifier is therefore not g_m*R_D but $g_m*(r_0//R_D)$, where r_0 is the output resistance of the transistor and // means the parallel of the two resistors: this means that the maximum (theoretical) gain of a single transistor MOS amplifier is given by g_m*r_0 (and this quantity is called intrinsic gain). This tells us how important the output resistance is!

Slide 21
This is a much better small signal equivalent circuit of a MOS transistor in saturation.
First of all we note that when source and bulk are not shorted, there is another Voltage Controlled Current Source (VCCS) between source and drain, controlled by the source-to-bulk small signal voltage, in parallel to the VCCS controlled by the gate-to-source voltage.
Then we see that we have capacitances connecting all the nodes between each other. The two most important are Cgs and Cgd, as they make a path for the signal between the gate and the source and drain. This has a very important impact on the frequency behavior of CMOS circuits.

Slide 22
The first point contact transistor was created by Bardeen and Brattain in December 1947.
The junction transistor was invented by Shockley in July 1951.

William Shockley, John Bardeen and Walter Brattain won the Nobel price in Physics in 1956, “for their researches on semiconductors and their discovery of the transistor effect”.
Bardeen won it again in 1972, together with Leon Cooper and John Schrieffer “for their jointly developed theory of superconductivity, usually called the BCS-theory” .
Shockley joined Beckman Instruments in 1955, where he was appointed as the Director of the Shockley Semiconductor Laboratory division.
In late 1957 eight of his researchers left and started Fairchild Semiconductor. This moment is often considered as the birth of Silicon Valley. Among these researchers there were Robert Noyce and Gordon E. Moore, who will then leave Fairchild to create Intel. Other offspring companies of Fairchild were National Semiconductor and Advanced Micro Devices.

Slide 23
A bipolar transistor consists basically of two diodes connected back to back. The central terminal is in common between the two diodes (it is the same piece of silicon!) and must be very narrow for the transistor to work (this will be clearer later).
A basic explanation of the working principle of the transistor is given in the slide.
Note that a bipolar transistor is a minority carrier device, as the working mechanisms is determined by electrons in the p doped base.
In the slide we discuss the case of an npn transistor. Pnp transistors also exist, they are similar to npn transistors but with the dopings and the applied voltages changed.
The symbol of a npn transistor is shown in the slide. The arrow on the emitter indicates the direction in which the current flows.

Slide 24
Here we see all the current components flowing in a bipolar transistor. I_{nE} is the electron current injected from the Emitter into the Base. Some of the electrons recombine with holes in the base, but the majority of them (I_{nC}) gets to the collector. For this to happen efficiently the base must be narrow enough. I_{nE} - I_{nC} represents one component of the hole base current. The other components of the (unwanted) base current are the holes injected from the base into the emitter and the recombination current in the emitter space charge layer.
The equations of the TOTAL base, emitter and collector currents are shown on the slide.

Slide 25
The emitter injection efficiency gamma specifies the portion of minority carriers injected into the base. Only these carriers may be collected by the collector junction to become a useful output current. The other components (I_{pE} and I_{rg}) should be reduced as much as possible. The space-charge recombination current I_{rg} is determined by the quality of the starting wafer, and is normally quite small. I_{pE} is determined mainly by the dopings.
The carriers injected into the base will traverse it by diffusion, and some of them will recombine. The base transport factor tells us how many will survive. This is determined by the carrier lifetime and by the base width. The carrier lifetime (and so the diffusion length) must be much larger than the base width.
The common base current gain is given by the product of the emitter injection efficiency and of the base transport factor, and it tells us how much of the input current gets to the output.
The common emitter current gain tells us how much current is flowing in the base for a given collector current. This is a very important difference from the MOS, in which we had no gate current.

Slide 26
The electron current injected into the base depends on the emitter-base area, on the inverse of the base doping concentration and exponentially on the base-to-emitter voltage.
This exponential behavior is one of the MOST important and useful and well known characteristic of a bipolar transistor.
The emitter injection efficiency gamma is determined mainly by the ratio between emitter and base doping (as the emitter and the base widths and diffusion constants are similar). To have gamma close to one the emitter must be doped much more than the base.
As already mentioned, the base transport factor is optimized through the carrier lifetime and the base width.
The total base current (which includes the emitter-base space charge layer recombination current) depends on the exponential of V_{BE}/(n*phi), where n is a number between 1 and 2 and phi is the thermal voltage. This change in the slope can be noticed plotting the log of the current as a function of the base-emitter voltage (see next slide).

Slide 27
The left plot shows the collector current as a function of the collector – emitter voltage for a bipolar transistor with the emitter grounded (common emitter).
In the cutoff region both emitter and collector junctions are reverse-biased. The normal active region is when the emitter junction is forward-biased and the collector junction is reverse-biased.

In the saturation region (which has nothing to do with the saturation region of a MOS transistor!!!) both junctions are forward-biased.
A bipolar transistor is used as an amplifier in the normal active region.
The maximum power dissipation allowed to the transistor is also shown. The transistor should not be operated beyond this curve due to overheating.
As we can see in this plot, a bipolar transistor is also, like a MOS transistor, a non ideal Voltage Controlled Current Source (VCCS). The non ideality comes from the modulation of the base-collector space charge region width with the collector-emitter voltage.
In the right plot we find the base and collector currents (in log scale) as a function of the base-emitter voltage. We can see the importance of the space-charge recombination current at low base currents (change in the slope). Note also how the collector current depends EXPONENTIALLY on the V_{BE} for many decades of current.

Slide 28
Here is the list of the important equations for bipolar transistors.
The key issues to be retained are:

- The collector current varies exponentially with the base-emitter voltage.
- The transconductance is linearly proportional to the collector current, and the transconductance to current ratio is equal to almost 40 at room temperature
- The input resistance is given by the inverse of the transconductance multiplied by the common emitter current gain.
- The output resistance is given by a voltage called Early Voltage divided by the collector current. The active region part of all the output characteristics (I_C vs V_{CE}) projected towards the left would cross the V_{CE} axis on the same point. The absolute value of this voltage is the Early voltage.
- The speed of the device is proportional to the inverse of the square of the base width
- A variation in the base-emitter voltage causes a variation in the injected charge Q_B. This represents a capacitance, called diffusion capacitance.

Slide 29
In this slide we present the small signal equivalent circuit of a bipolar transistor in the active region.
As already mentioned, we have a non ideal voltage controlled current source. The output resistance r_0 represents the non ideality.
Another important resistor is the base-emitter resistor r_π, which is the input resistance of the transistor for a signal.
r_B is the base resistance. It is a physical resistor which depends on the actual transistor layout.
C_{jC} and C_{jE} are two capacitances associated with the junctions, and C_D is the already discussed diffusion capacitance.
The diffusion capacitance is proportional to the transistor current, and in normal operation is bigger than the junction capacitance.

Slide 30
Here we have an interesting comparison between MOS and bipolars. The most noticeable differences are:

- The input resistance of a bipolar is much smaller than the one of a MOS.
- The saturation voltage of a bipolar does not depend on the transistor size, as it is the case for a MOS. Also, the saturation voltage is normally smaller for a bipolar.
- A bipolar transistor is always more power efficient than a MOS.
- A MOS transistor has more design parameters to play with (not necessarily an advantage!). In a bipolar transconductance and current are linked only by the thermal voltage (which is fixed). In a MOS we have the transistor sizes, which have to be determined by the designer.
- A bipolar transistor has a “reliable” exponential characteristic over many decades of current. A MOS transistor is in strong inversion for only a few decades, and this range becomes smaller and smaller in newer technologies.
- At low currents a MOS can be faster, as C_{gs} can be made smaller than C_D (making the transistor small). At high currents a bipolar is normally faster (but MOS are catching up!)
- Noise: the thermal noise is comparable in the two transistors. The 1/f noise is higher for MOS.

Section: Sub-micron CMOS and BiCMOS technologies

Slide 31
In this section we will see a few details on the technologies in which MOS and Bipolar transistors are fabricated.

Slide 32
This slide shows a (very simplified) representation of a NMOS and a PMOS transistors in a (p-substrate) CMOS technology.
CMOS stands for Complementary MOS
This is a simplified representation because many things are missing: Epitaxial layers, double wells, retrograde wells, …..
The simplified layout of an NMOS is also shown: with one mask we define the active area (S/D diffusions). The polysilicon gate shields the silicon from the S/D implant (self-aligned processes).
To keep the source, drain and gate resistances low, these regions are covered with metals (for example, Ti, Ta, or Co), which react with silicon and form a "disilicide". This process is referred to as a self aligned silicide (or salicide) process.

Slide 33
To better understand the beauty of CMOS, let's see how a CMOS inverter (the simplest logic gate) works. I hope I do not offend anybody…
It is made by an NMOS and a PMOS transistors connected as in the picture: common gates (input), common drains (output), source of the NMOS to GND (logic 0) and source of the PMOS to VDD (logic 1).
When the input is at 0, the NMOS is off. The PMOS has the channel formed, and therefore the output is at 1 (as the channel connects it to V_{DD}).
Vice versa for the input at 1.
The two transistors can therefore be seen as two switches.
Since one of the two switches is always open (off, open circuit) with the input at 0 or 1, there is no current flowing in these two conditions. So, when CMOS logic is not doing something, it is not dissipating any power! This is one of the most important advantages of CMOS!!
There will be a current flowing only when the gate is switching, as at a certain moment the two transistors will be on at the same time. This can be seen in the following slide. Most current is used to charge and discharge the capacitive load represented by the following stages and interconnects.

Slide 34
Here we simulate a chain of two inverters: the second inverter is there just to load the first one. We can see here that the dynamic current which we have while the transistors are switching can be quite large, but it is normally very fast. For example, in a 0.13 um technology the propagation delay of an inverter with fanout=1 (loaded with another inverter) and no interconnection parasitics is ~25ps.

Slide 35
This is a fairly reasonable representation of a modern integrated bipolar transistor in a p-substrate CMOS technology. A better one will be shown later. It is called vertical NPN because the current is flowing in a direction perpendicular the the wafer surface. Practically all high-speed bipolar transistors are made this way.
The n+ region underneath the collector is called subcollector, and is used to make an electrical contact to the collector at the surface.
Starting from a p-type silicon wafer, the subcollector is formed by ion implantation and diffusion. The n-type collector is then formed by epitaxial growth. The p-type base is then ion implanted in the epitaxial layer. The n+ emitter is finally made by ion implantation and diffusion or by depositing a heavily doped n-type polysilicon layer on top of the base region.
Transistors are isolated from one another by p-type pockets or by oxide-filled trenches.

Slide 36
Vertical PNP are also called substrate PNP, as the substrate acts as collector. All the vertical PNPs have therefore the same collector: this is an important limitation when we use them in circuit design. The epitaxial n region which is normally used as collector for a vertical npn is here the base, which is contacted by a n+ diffusion (the emitter in a npn transistor). The emitter is realized with the same diffusion which is used for the base of a vertical npn.
The lateral pnp uses again the epitaxial layer n as base. Emitter and collector are p diffusion (the base of a vertical npn).
The dopings of all these regions are optimized for npn transistors. Therefore, the performance of pnp transistors fabricated as just explained is quite poor.
Technologies with good performance npn and pnp transistors exists, but they are normally quite expensive and/or might not be easily accessible. In these technologies vertical PNPs are normally not substrate PNPs, as the collector is a separate terminal (they are built differently).

Slide 37
This is an example of a modern npn bipolar transistor. The important features to be noticed are:

- Deep-trench isolation between transistors: this results in a much higher density. The area of a diffusion isolated bipolar transistor is completely dominated by the isolation. Trenches can be made much smaller than diffusions (which are normally as wide as deep).
- Polysilicon emitter
- Self-aligned polysilicon base contact

Slide 41
Bipolar logic circuits have a too large power consumption per logic gate. This puts an upper limit on the numbers of gates that can be reliably integrated on a single die.
Despite the fact that the MOSFET original principle came in 1925 (and the first patent in 1930), manufacturing MOSFETs reliably came only after a few decades.
MOS digital integrated circuits took off in 1970. In the late 1970s NMOS-only logic circuits were limited again (as with the bipolar before) by power consumption.
The enormous progresses in technology manufacturing made CMOS a reality.
Before about 1977 practically all analog circuits were done with bipolars, and MOS where used for memories and digital circuits. After that date many people started to do everything in CMOS.

Slide 42
These prices have been taken from the web, and are intended for prototyping. Prices taken without discounts.
UMC: dimensions should be multiple of 5 mm. For example, a 6 mm by 6 mm chip would cost, in 0.13 um, 55.4K * 4!!!

Section: Feature size scaling

Slide 45
Transistors became and are becoming smaller and smaller and smaller…
In this section we will understand why and we will explain how scaling works and what is its impact on integrated circuits performance (mainly CMOS).

Slide 46
The first IC had 5 components (transistors, resistor and cap)

Slide 47
In April 1965 Moore was director of research at Fairchild Semiconductor (created in 1957).
Intel was created in 1968. Moore was one of the co-founders.
Examples of the broad definition of Moore's law (1996) are, for example, Intel's net revenue and Intel's investment in R&D.

Slide 48
For each microprocessor picture, we have indicated the name (blue), the date of introduction (red), the number of transistors (light blue), the minimum feature size of the technology used (green) and the clock speed (pink).

Slide 50
Note here that the right scale is log.
Physical gate length and on chip clock refer to the left scale, transistors per chip and minimum supply voltage to the right scale.
The density doubles every 3 years.
Transistors per chip: data taken for high volume microprocessors at production.

Slide 51
The static power dissipation is given by the drain-source subthreshold leakage, the junction leakage (negligible) and by the leakage through the gate.
The dynamic power dissipation is given by the charging and discharging of the load capacitance, and it is frequency dependent.
With scaling we reduce the gate oxide thickness t_{ox}. This means that the power supply voltage has to be reduced as well (otherwise we risk breaking the oxide). Reducing t_{ox} increases the gate capacitance per unit area C_{ox}, but the capacitance of a gate is nevertheless reduced as both W and L of each transistor are reduced.

Slide 52
To scale the depletion region width x_d the substrate doping N_A has to be increased.
The constant field scaling is called like this because the electric fields in the device are kept constant, to avoid undesirable effects.

Slide 54
The problem of constant field scaling is that the width of the moderate inversion region and the subthreshold slope do not change.
With scaling, the power supply voltage is reduced, and so the threshold voltage of the transistors. This means that with constant field scaling the leakage current of the transistors increases.
For digital circuits with millions of gates this can be a big problem.
This issue has caused the departure from the classical constant field scaling (see next slide).
Another important remark is that in modern CMOS technologies very often we find different kind of devices with different threshold voltage values. Lower threshold voltages (high performance devices) give higher speed but also higher power consumption. Higher threshold voltages (Low Power devices) are used when power is an issue. This has been done to give more flexibility to designers and to partially solve the problem given by the fact that with ANY kind of scaling the threshold voltages tend to become smaller and the leakage currents tend to increase.

Slide 55
Let's review a few of the already mentioned issues: the aim of scaling is to reduce the device dimensions (to improve the circuits performance) without introducing effects which can disturb the good operation of the device.
Constant field scaling does this keeping the electric field in the device constant. The problem is that the subthreshold slope and the width of the moderate inversion region do not scale. Due to the problem mentioned before, several other scaling approaches have been proposed.
One is the constant voltage scaling. This approach has been proposed to make the compatibility between different logic families easier. The power supply voltage is not scaled and the oxide thickness is scaled less drastically to avoid too large electric fields in the oxide.
The constant voltage scaling still has several inconveniences, the major being that the electric field in the thin gate oxide increases too much. To overcome them, the generalized scaling has been proposed.
With the generalized scaling the voltages are scaled but not at the same rate as the oxide thickness. This allows having acceptable electric fields in the device and at the same time a useful voltage swing for the signals. Also, the threshold voltage scales less abruptly than t_{ox} and also than V_{dd}, to limit the subthreshold currents.

Slide 56
The delay caused by interconnections was negligible in the early days of CMOS logic circuits. Now that circuits are bigger and logic gates are much faster, this is not true anymore, and interconnections can be the limiting factor of the speed of a digital circuit. The plot shows the delay of a wire with square section as a function of its length, for three different sections of the wire. From this plot we can see that:
If the wire is not made big enough, a 1 mm long wire can easily cause delays above several tens of picoseconds
For a given wire length, increasing the wire size can help up to a certain limit, as we hit the limit determined by the speed of an electro-magnetic wave
For wires of the order of a few tens of microns the delay of the wire is not a problem as it can easily be kept below 1 ps.

Slide 57
To scale the wires, we could shrink everything by alpha (>1): the wire dimensions W and t_m, the wire length L and the distance between different metal planes (oxide thickness) t_{ox} (not to be confused here with the gate oxide thickness of a MOS transistor). With scaling the currents are reduced by a factor alpha, but if we scale the wire section by alpha squared we increase the current density by alpha. An increase of the current density in the wire is normally very dangerous, as it can cause electromigration, which has a deadly impact on reliability (see also later). Also, if we shrink everything by alpha we reduce the capacitance of the wire by alpha but we increase its resistance by the same factor. This means that the delay of the wire (the RC delay) does not scale.
So, shrinking everything does not work.
In practice, different scaling rules apply to wires at different metallization levels.
For the metal levels closer to the transistors, where we need relatively short wires for local interconnection, we can shrink everything by alpha EXCEPT the wire thickness. This allows having narrow wires, which are needed to connect the smaller and smaller transistors, but at the same time we have that the current density does not change (good for reliability) and that the resistance does not change while the capacitance is reduced by alpha: we have then a faster wire (faster by a factor alpha). This of course as long as we do not have another wire close to it (the lateral C would dominate!)
At higher metal levels, we need even faster wires, as chips tend to become bigger. So for "global" interconnection not only we do not scale the thickness of the wires and of the oxide, but we might even want to increase it: this is why this is called "reverse" scaling.

Slide 58
I have only this slide on bipolar transistor scaling and its impact on the transistor performance. We will see the scaling of CMOS much more in detail.

I use here base width / depth because it depends on how we look at the device (vertical – horizontal). What I mean is that it is beneficial to reduce the path of the minority carriers in the device.
The breakdown voltage is the avalanche breakdown voltage of the reverse biased collector-base junction.

Slide 59
Here we have a summary of some of the important analog performance parameters which IMPROVE with scaling of CMOS technologies.
We assume here that we are moving to a more advanced process (the gate oxide thickness t_{ox} scales) BUT we keep the same transistor dimensions. This is a choice that we always have in analog design.

- Threshold voltage matching: identically designed transistors will newer be exactly identical. There are a number of things which can be slightly different in the two devices. As an example, the number of dopant atoms in the channel will never be identical, as ion implantation is a statistical process. These differences in the devices give, for example, a difference in the threshold voltages. If we fabricate 100 transistor pairs and we measure the threshold voltage differences between the 2 transistors of each pair, we should normally find a gaussian distribution of these differences. The mean, if the designer has done his job properly, should be zero. A non zero mean means that there is a systematic effect, affecting all the time the same device in the pair. The sigma of the gaussian is what we call mismatch, and we can not eliminate it. The sigma of the distribution of the threshold voltage differences is given in the slide. We can see that bigger devices match better. We can also see that matching improves going to more advanced technologies (for the same device dimensions), as it is proportional to t_{ox}.
- The input referred 1/f noise formula is given in the slide. K_a is a technology dependent parameter. To make a comparison between technologies is fundamental to know its value for each technology. Assuming that it does not change with scaling, we see that 1/f noise improves with scaling (for the same device dimensions), as C_{ox} increases. Note that C_{ox} is not always found squared in the formula (it depends on the textbook).
- Transconductance: for the same device dimensions and the same bias current, the transconductance increases as uC_{ox} increases.
- For the same reason, the input referred white noise (proportional to $1/g_m$) decreases with scaling. We assume here that we do not have phenomena which can cause noise in excess.

Slide 60
Here we have a summary of some of the important analog performance parameters which DO NOT IMPROVE with scaling.

- With scaling we can have new noise mechanisms which can increase the 1/f noise parameter K_a or increase the white noise excess factor.
- Deep submicron transistors are difficult to model, especially for some analog parameters
- Especially in their early stages, deep submicron CMOS processes did not have devices which are used in analog design (high precision linear capacitors, resistors, …). CMOS processes were optimized, at the beginning, for memories and for digital circuits. Analog components were normally added later. This is practically not a problem anymore, as most CMOS processes have today a very large offer of devices.
- The power supply reduction with scaling imposes new circuit architectures (we can not stack too many transistors between V_{DD} and GND)
- CMOS processes were used, at the beginning, mostly for digital circuits, as the analog performance of the transistors was much worse than the one of bipolar transistors. Scaling of CMOS technologies and the improvement in the transistor performance pushed towards system integration on chip using only CMOS processes. In modern CMOS technologies it is possible to integrate analog and digital blocks on the same substrate. This imposes to deal with the problem of noise propagating from the digital blocks to the sensitive analog blocks. This point will be addressed in more detail at the end of the presentation.
- The problem of velocity saturation (that we have already seen) becomes more and more important in scaled down devices.

Slide 61
This slide shows a resume of the I_{ds} and g_m formulae for week inversion, strong inversion and velocity saturation.
We have also calculated here the V_{gs} voltages which delimit the weak inversion to strong inversion transition and the strong inversion to velocity saturation transition.
We can see that the first does not change with scaling, whereas the second decreases. This means that the width of the strong inversion region becomes smaller, and that, as already mentioned, it is easier to end up working in velocity saturation (and we do not want to!).

Slide 62
In this slide I made a little “game”: using the basic equations that we have seen and applying constant field scaling (all this is obviously a rough approximation) I have calculated how several quantities vary with scaling, for different choices of the W and L dimensions of the transistor. We refer here to analog circuits.

Each row in the table represents a different choice of W and L. We consider here only thermal noise, and the Signal-to-Noise ratio (SNR) is defined as V_{DD} divided by the thermal noise.
A few interesting things to note:

- We always gain in power (except last line)
- We always loose in SNR (except again last line)
- Speed is linked to L only (not W): we gain in speed only decreasing L
- The last line in the table represents the case of constant power consumption, constant dynamic range in a thermal noise limited system. We see that to keep the SNR constant with scaling we have to increase W! We also see that we do not gain in power. In the literature we can find that, in the reality, things are even worse! To keep the same SNR with scaling we even have to burn more power at the certain point. There is a kind of optimum, which seems to be around 0.25 um technologies. All this is of course pretty much dependent on the circuit type.

Section: Radiation effects and reliability

Slide 63
This section is about:

- radiation effects on integrated circuits
- how to make radiation tolerant circuits in non rad hard CMOS technologies
- reliability

I will most likely NOT have the time to explain all this during the course, but I have put the slides anyway because I believe it is an interesting topic when talking about electronics for nuclear physics. I hope you will find the slides + the comments useful.
This is a quite complex topic, so some of the slides (like the first ones) are simplifying the matter a bit.

Slide 64
Note: It is obviously quite important to know the radiation environment in which our circuits will have to work (and survive): kind of particles, energies, fluences…

Slide 65
TID: total ionizing dose
Cumulative effects are gradual effects taking place during the whole lifetime of the electronics exposed in a radiation environment. A device sensitive to TID or displacement damage will exhibit failure in a radiation environment when the accumulated TID (or particle fluence) has reached its tolerance limits. It is therefore in principle possible to foresee when the failure will happen for a given, well known and characterized component.
On the contrary, Single Event Effects are due to the energy deposited by one single particle in the electronic device. Therefore, they can happen in any moment, and their probability is expressed in terms of cross-section. A device sensitive to SEE can exhibit failure at any moment since the beginning of its operation in a radiation environment.
From this slide we can see that MOS transistors are sensitive to TID and SEE but not to nuclear displacement.

Slide 66
Let's now start to see the problems given by TID. We will see the SEE later.
We start with TID effects in a MOS. The slide shows the band diagram of a MOS structure under bias (positive bias applied to the gate with respect to the bulk).
When ionizing radiation goes through the MOS layers, it creates electron-hole pairs everywhere. In polysilicon (M) and in silicon (S) which are conductive (with lots of free carriers) the e-h pairs recombine immediately.
In the oxide (O) only a fraction of them recombines. Most of them are separated by the electric field. Electrons have a mobility in the oxide which is many orders of magnitude higher than the holes, and they leave immediately the oxide. Holes move more slowly (the transport phenomenon is called "polaron hopping") through the oxide towards the Silicon Dioxide – Silicon interface. Some of the holes are trapped in the oxide (normally close to the interface where the trap density is higher). Some other holes gets to the interface where they create some new interface traps.
What is the electrical result of all this mess?
Holes trapped in the oxide move the threshold voltage: one way to understand this is to think that these holes will attract some electrons at the interface. This means that, for example in an NMOS, we need a smaller V_{GS} to get the same channel, which means that the threshold voltage is smaller. As we always trap holes, the threshold voltage shift is negative in both NMOS and PMOS transistors.
The new traps created at the interface can also trap charges. Under normal bias conditions, they trap electrons in NMOS transistors and holes in PMOS. This means that the threshold voltage shift is positive in NMOS and negative in PMOS.
It is important to mention that the two processes:

- Hole trapping in the oxide

- Interface states creation

Have a very different temporal dynamic. The hole trapping is much faster, interface states come normally later. This means that, after irradiation, we can measure different threshold voltage shifts in time.
Since interface traps degrade the Silicon Dioxide – Silicon interface, they also degrade the mobility of the carries (and then the transconductance), the slope of the subthreshold region, the noise (especially 1/f) and the matching.

Slide 67
The problems we have explained in the previous slide refer to TID damage in the thin gate oxide of a MOS transistor. But in CMOS ICs there is plenty of oxide everywhere.
Oxide is also used to isolate devices between each other. At the edges of the channel, the thin gate oxide has to become thick isolation oxide. This used to be done with a process called LOCOS (LOCal Oxidation of Silicon), which gave to this region the characteristic shape of a bird's beak. Below roughly 0.35 um, this process was substituted by another one called STI (Shallow Trench Isolation). In both cases, we have at the channel edges a region with a quite thick oxide. This oxide can trap quite a lot of holes during irradiation, which can then attract electrons just underneath. These electrons can form a parasitic path between source and drain. It is just like having two parasitic transistors in parallel to the main one, one on each side of the main transistor. This phenomenon is important only for NMOS transistor (as we always trap holes in the oxide).

Slide 68
To understand a bit better the effect of irradiation at the bird's beaks, we can think at this region as if it was made up by many transistor is parallel, each one with a different gate oxide thickness. Being this thickness much larger than the one of the main transistor, the I_D vs V_{GS} characteristic of these transistors is much different: the threshold voltage is much higher and the saturation current is much smaller.
But with the irradiation the threshold voltage of these parasitic transistors can shift so much that their characteristics get overlapped to the one of the main transistor. Two possible examples are shown in the slide. This is a qualitative explanation of the shape of the characteristics that we can measure after irradiation.

Slide 69
This slide shows the measurement of a NMOS fabricated in a 0.7 um CMOS technology. We can see the I_{DS} vs V_{GS} characteristic before irradiation and after 1 Mrd. We can easily notice the negative threshold voltage shift and the devastating impact of the leakage at the bird's beaks. The leakage problem is an issue especially for digital circuits. For analog circuits with transistors working in strong inversion it has a smaller impact.

Slide 70
As already mentioned, in a CMOS IC we have thick oxide everywhere (wherever we do not have transistors). This oxide is called field oxide, and it can be charged by trapped holes. This induces leakage paths a bit everywhere!
The effect is lethal. Here we show an example in which we have leakage between an N-well (connected to V_{DD}) and the source of a NMOS, connected to GND (V_{SS}). The radiation induced leakage path creates a short between V_{DD} and GND!
This can be easily verified experimentally: if we take a rad soft CMOS IC and we irradiate it monitoring at the same time the power supply current, we will see it rising sharply after a given dose (which is technology dependent).
The effect of radiation can be made even worse if there is an interconnection on top of the field oxide. This acts as a gate, making the effect of radiation even more pronounced (see next slide).

Slide 71
To better study the problems created in the field oxide, we have fabricated and measured some "field oxide transistors". These transistors have as a gate oxide the thick field oxide used for isolation. As a gate one can use polysilicon or metal.
In the left plot wee see that these transistors behave exactly as standard transistors, but their threshold voltage is in the order of a few tens V.
After irradiation we immediately see that the leakage current (the current at $V_{GS} = 0$) increases by many orders of magnitude.
We can also see that the case of a polysilicon gate is much worse than the one of a metal gate: this is because, for the same applied bias, the electric field in the oxide is higher when the gate is in polysilicon. A higher electric field is more efficient in avoiding recombination between the electrons and the holes created by irradiation.

Slide 72
The text for this slide is from Federico Faccio, CERN.
The TID effects in bipolar devices are also due to charge trapping in the oxide and creation of interface states. The effects can be divided in two categories:

1) inversion of the silicon under a thick oxide, opening a conductive channel
2) effects decreasing the gain of the transistor
The inversion channel can be formed in several places, depending on the technological characteristics:
a) Substrate: opening of the channel between two buried layers
b) Sidewall: inversion near the sidewall oxide, shorting collector and emitter of npn transistors
c) Surface: inversion of the surface. Even though the surface oxide is generally thinner than the isolation recess oxide, the effect might be important when the transistors are working at low current levels.

Slide 73
The text for this slide is from Federico Faccio, CERN
TID acts on the gain by increasing the surface component of the base current (the bulk component being mainly sensitive to displacement damage).
The increase in this surface current component is mainly due to an increase of interface states at the surface of the base and a positive charge buildup near the emitter-base junction (both increasing the minority carrier recombination rate).
Excess base current is in general the dominant effect, with the collector current being constant.
The sensitivity is higher at lower injection levels, as in this case there is more sensitivity to surface phenomena.

Slide 74
The text for this slide is from Federico Faccio, CERN
Oxide thickness: the thicker the oxide above emitter-base junction area, the grater the TID effects
Oxide trap efficiency: the more degraded the oxide (during manufacturing), the grater the TID effects
Electric field: difficult to have a clear picture of the electric fields, and to generalize to several technologies
Surface doping concentration: the more heavily doped the base or emitter surface, the lower the TID effects
Emitter perimeter-to-area ratio: the grater the ratio, the grater the TID effects
Transistor geometry: vertical structures have lower sensitivity to TID effects than surface lateral structures (or substrate PNP, where 20% of the current is lateral)
Injection levels: in almost all cases, degradation is higher at low injection

Slide 75
Up to now we have seen the total ionizing dose effects in MOS and Bipolars.
In this slide we briefly mention the displacement damage effects.

Slide 76
We now see 3 Single Event Effects, which are very important for CMOS processes.
The first is the Single Event Latch-up (SEL).
In every CMOS process there is a parasitic structure called thyristor (shown in the slide). Electrical latch up in a CMOS process happens when this parasitic thyristor turns on, shorting V_{DD} and GND. This can destroy the circuit if the power supply is not cut on time. Electrical latch up is something which is normally addressed by the manufacturer, which tries to reduce resistances R1 and R6 (which are responsible for turning on the device) and tries to reduce as much as possible the current gain (beta) of the two bipolar transistors making up the thyristor.
SEL is a Latch-up which can be initiated by ionizing particles, which can create electron hole pairs in strategic points of the IC and turn on the thyristor. As already mentioned, this can be a destructive effect.

Slide 77
We explain now the Single Event Upset (SEU), illustrating it in the case of a Static RAM cell.
A highly ionizing particle can deposit enough energy to produce a “soft error”. A SEU occurs if the charge deposited on the critical node where the information (represented by a charge) is stored is higher than a minimum charge quantity, called critical charge. To produce a SEU particles with a sufficiently high Linear Energy Transfer (LET) are therefore needed. The Linear Energy Transfer expresses the energy per unit length which the particle deposits in the material.
The SEU is NOT a destructive effect.
What is very important with SEU is to know well the radiation environment in which the IC has to work, to be able to estimate the SEU rate. We have then to decide if we can afford to have this rate of SEU in every block of the IC.
The third SEE (no slide for this one) is called Single Event Gate Rupture. It is a destructive effect, which is caused by an ionizing particle which crossing the gate of a transistor causes the rupture of the thin gate oxide (we do not enter here in the physics of this phenomenon).

Slide 78
We now enter a new “chapter”: what is the effect of scaling on the radiation tolerance of CMOS processes?
Let’s start with the Total Ionizing Dose (TID) damage of the thin gate oxide.

Already more than 20 years ago it was understood that reducing the gate oxide thickness t_{ox} was improving the radiation tolerance of the gate oxide.
The two plots show the damage in the oxide (left) and at the interface (right) normalized to 1 Mrd dose as a function of the gate oxide thickness.
The y axis of the left plot represents the flat-band voltage shift with radiation (normalized to 1 Mrd dose), which is an indication of the amount of holes trapped in the oxide. The y axis of the right plot represents the increase of the interface state densities. The two series of points in this plot correspond to two different electric fields in the oxide during irradiation.
Let me remind you the typical gate oxide thicknesses of a few CMOS technologies:

- 0.7 um technology: t_{ox} = 17 nm
- 0.5 um technology: t_{ox} = 10 nm
- 0.35 um technology: t_{ox} = 7 nm
- 0.25 um technology: t_{ox} = 5.5 nm
- 0.13 um technology: t_{ox} = 2.2 nm

As we can see from the two plots, we have a huge increase of the radiation tolerance with scaling. For oxide thicknesses below 10 nm this improvement becomes even larger.
We can therefore expect, from this slide, that the gate oxide of a deep submicron CMOS technology is inherently radiation hard! To be verified…

Slide 79
So, decreasing the gate oxide thickness t_{ox} we reduce the following radiation induced degradations:

- Threshold voltage shift.
- The threshold voltage shift normalized to the absolute threshold voltage value or to V_{DD} also decreases. This is important because with scaling V_T and V_{DD} decrease.
- The radiation induced mobility degradation, which is a function of the increase in the interface traps density, is also reduced with scaling.
- In general, all the radiation induced degradations linked to the thin gate oxide or to its interface with the silicon become smaller with scaling.

Slide 80
This is an interesting plot which shows several measurements done in the Microelectronics Group of CERN (only 4 points are taken from the literature: the 1.6, 1.2, 0.8 and one of the 0.5). What was shown already 20 years ago is really true!
All the points except the two for the 0.13 um technology were taken immediately after irradiation and after doses of the order of 100 krd. This was done to try to have only the component from the oxide traps (we remind here that the threshold voltage shift is given by two components: the one from the oxide and the one from the interface states. These components are both negative for a PMOS and have different signs for a NMOS. Also, the dynamic behavior of the creation of these two components is very different: the interface state component comes normally more slowly).
The two points for the 0.13 um technology need some comments: they were measured after 100 Mrd, which means that here we have the components from the oxide traps and from the interface states summed.
The reason why we have taken the values after 100 Mrd was to have a better measurement (at least for the PMOS), as the values are very small!
The point for the PMOS is probably higher than expected because of the interface states (whose contribution goes in the same direction as the contribution from the oxide traps).
The point for the NMOS is not so much representative, as after 100 Mrd the shift was only 1 mV and as we have the sum of the two components!
Anyway, the key message here is that even after doses of the order of several tens of Mrd we can expect, using a deep submicron CMOS technology, a threshold voltage shift of the order of a few mV or tens of mV.

Slide 81
Modern CMOS technologies became more and more tolerant to electrical latch-up (and therefore to Single Event Latch-up). This is linked to the introduction of several novelties in the process: retrograde wells (which are wells doped more heavily at the bottom than at the top), thinner epi layers, shallow trench isolation (STI). All this reduces resistances R1 and R6, making it more difficult for the parasitic thyristor to turn on. Also the fact that the power supply voltage V_{DD} is reduced with scaling helps, as to turn the thyristor on and keep it on we need at least 1 V.
All this is very much technology dependent. What we can also say is that in general it is good to have as many substrate and well contacts as possible, because this in any technology helps preventing the SEL.

Slide 82
The plot shown in this slide illustrates that the SEU was becoming more important with scaling. On the y axis we have the threshold LET, which is the LET at which we start to have SEUs.

An explanation for this plot is that with scaling we reduce both the power supply voltages and the node capacitances. The critical charge is therefore reduced, which means that an ionizing particle needs to deposit less charge to create an SEU.
But there is also another effect which is competing with the above one: due to the reduction in the device dimensions the charge deposited by the ionizing particle which is collected in the critical nodes becomes smaller.
These two effects are compensating each other. In the literature there is not a very clear trend, but SEU could become worse with scaling.

Slide 83
The necessity imposed by scaling to make thinner and thinner oxides results in an improvement in the oxide quality.
This has a beneficial impact on Single Event Gate Rupture (SEGR).
This plot shows what is the electric field that we have to apply to the gate to have a SEGR (called critical field E_{CR}) as a function of the LET of the ionizing particle. This is done for 4 different oxide thicknesses.
For each plot we see that the critical field decreases increasing the LET.
And we also see that thinner oxides seem to be better, as they have a higher critical field.
On the plot we have added the maximum electric field in the gate of a 0.25 um CMOS technology (5.5 nm gate oxide thickness, 2.5 V power supply voltage). We can see that even for quite high LETs the critical field is much higher than the one we can have in an even thinner oxide.
Which value of LET can we expect in different environments? To give a couple of examples, in LHC we can have up to 15, in space up to 100.

Slide 84
Newer technologies have even thinner oxides than the ones considered in the previous slide. Also, simple silicon dioxide is not anymore the oxide being used in modern CMOS processes. Nitrided oxides are already in use since a few years, and several other high-k dielectric materials are under study for future generation CMOS processes. High-k materials allow increasing the gate capacitance per unit area without decreasing too much the gate oxide thickness (and therefore without increasing too much the gate leakage).
These two plots show results on oxides as thin as 2.2 nm and on different kind of oxides.
On the left plot we see the critical field that led to SEGR under exposure of particles with a LET = 80. We can indeed see that the critical field is increasing for thinner oxides. On the same plot we have the field needed to break the oxide before irradiation, were we see a similar trend (but an even bigger slope): this is an indication of the increasing quality of these this oxides.
On the right plot, the y axis represents the voltage which needs to be applied to the gate to break it (always under irradiation). On the same plot the maximum power supply voltage V_{DD} for technologies with those gate oxide thicknesses is shown.
From these two plots one can see that SEGR does not seem to represent a problem.

Slide 85
Let's now make a summary of radiation induced problems in CMOS and of the effects of scaling on them.
As we can see from this slide, using a deep submicron technology solves most of the problems.
The two problems left are:
- Leakage currents (Total Dose in the field oxide and at the edges of the NMOS transistors)
- SEU

Slide 86
Here we explain a way to solve the leakage problem in an NMOS.
On the left we have a standard transistor layout, with the indication of the leakage paths. On the right we see a different layout: the gate is annular, and it is completely surrounding the source (or it could be the drain).
We call these transistors Enclosed Layout Transistors.
This layout works very efficiently in cutting any possible leakage path between source and drain: the source is completely surrounded by thin oxide!
To cut leakage paths between different transistors, we will have to use guard rings (see next slide).

Slide 87
P+ guard rings have to be used around each n-well or n+ diffusion (at different potentials). This method is very effective in cutting the radiation induced leakage path (which is always made by electrons as we always trap holes in the oxide).

Slide 88
This plot shows the effectiveness of Enclosed Layout Transistors. We have designed two transistors with exactly the same W/L ratio. One has a standard layout, the other is an ELT.
Before irradiation the two characteristics are coincident.

After irradiation we can see that the ELT completely eliminates the parasitic leakage paths at the edges of the channel.
The threshold voltage shift is nevertheless still present in this technology, as the gate oxide is still relatively thick. This still gives a quite high leakage current, even for the ELT.

Slide 89
Going to a smaller feature size, we completely eliminate the problem!
Here we see two measurements done on an ELT before irradiation and after 13 Mrd. The two plots are coincident!

Slide 90
This slide shows the layout of a radiation tolerant inverter. We can notice:
- The NMOS is an ELT
- The PMOS is not, as it is not necessary
- The p+ guard ring is essential to cut parasitic paths between the external n+ diffusion of the ELT and any other n region around (at different potential)
- The gates of the NMOS and of the PMOS have to be connected by metal. A connection made by polysilicon would open the guard rings
- The n+ guard ring around the PMOS is not really necessary, but it helps to fight SEL (it is a robust n-well contact).

Slide 91
To summarize, we can obtain Total Ionizing Dose (TID) tolerance using a deep submicron process (<= 0.25 um) and ELTs and guard rings.
This is a much more efficient way of making radiation tolerant ICs compared to using a rad-hard technology: with deep submicron technologies we also gain in speed, power consumption, density, cost and yield!
This is because all the rad-hard processes available today are much less developed compared to a modern CMOS process.

Slide 92
What about tolerance to Single Event Effects (SEE)?
- Single Event Latch-up (SEL): As already mentioned, guard rings are quite effective in preventing SEL
- Single Event Gate Rupture (SEGR): In many circuits we have done in a 0.25 um CMOS process, we have never had this problem. Results found in the literature are also quite positive.
- Single Event Upset (SEU) is definitely something we have to deal with. The use of ELTs increases a bit the node capacitance. Minimum size ELTs are bigger than the standard minimum size transistors, and therefore their gate capacitance is larger. This decreases a bit the sensitivity to SEU, but still in many occasions this is not enough. So one needs to investigate other solutions at the circuit level (redundancy, special memory cell architectures, …)

Slide 93
This slide and the next one are on reliability. This is a huge, complex topic. Very far from treating it in detail, we would like here just to mention a few key points.
Why are some words in red in the definition?
- Probability: the definition involves statistics: we have to accept that a few devices may fail (how many???)
- Required function: we have to define the failure
- Stated conditions: we have to define them all
- Period of time: involves the concept of lifetime of a circuit

Slide 94
A few more comments:
- Oxide breakdown: Prevention: improve oxide quality, alternative dielectrics, screening
- Hot carriers: non destructive. High energy electrons injected into the oxide. Shifts the Id-Vg curve. Prevention: Special Drain Layout, increase oxide quality, decrease V_{DD}
- Electromigration: due to high current densities. Prevention: design rules, new materials.
- ESD: Prevention: antistatic measures, ESD protection circuits

Section: Mixed-signal circuits

Slide 95
The last part of the presentation addresses some of the issues that we encounter when designing a mixed-mode (or mixed-signal) circuit, i.e. a circuit in which analog and digital blocks coexist.

Slide 96
Using a standard digital CMOS process (I mean with this a CMOS process which is not optimized also for analog design) has some advantages and some complications: we will see some of the points in more details in the following slides.
Some more comments to this slide:

- Price: CMOS processes used to be relatively inexpensive. This is not quite true anymore: the price of a mask set for a 130 nm CMOS process is around 400K USD! Anyway, adding bipolars to that would make it even more expensive!
- Modern CMOS processes allow making very dense and high performance digital circuits. One has to take this into account when designing the chip at the system level, as the partition line between analog and digital might move more towards having more done with digital and less with analog. This means that some of the signal processing can be done more efficiently with digital circuits.
- Lower power consumption: we have already seen that this is true mainly for digital!
- Low power supplies are very good for the digital power consumption, but can create problems in the choice of the analog circuit architectures.
- The lack of analog components and the inadequate modeling of a few analog parameters are true especially in the early stages of development of a process. When a process is more mature these problems are normally at least partially solved. Also, as we have already mentioned, the lack of analog components is not really a problem anymore. The push to make System on Chip (SOC) circuits and RF CMOS circuits made these processes evolve, so that today we do have a quite large choice of analog components.

Slide 99
The plot shows that in modern submicron processes the wire thickness is becoming bigger than the distance between wires. This means that the parasitic capacitance between wires at minimum distance is increasing.
Making a capacitor exploiting the parasitic capacitance between different metal levels (eventually using several levels) it therefore a not very efficient way of doing, as the thickness of the oxide between levels is bigger than the minimum distance between two wires of the same metal level. A better way would then be using the parasitic capacitance between wires.

Slide 100
These are a few examples of how we can make dense capacitors exploiting the parasitic capacitance between wires. More details can be found in the references. The fractal capacitor is an interesting solution as it maximizes the perimeter to area ratio, but I am afraid it is going to be slightly difficult to design…

Slide 101
Another interesting option to make dense capacitors is to exploit the high specific capacitance of the gate oxide. There are 4 ways of doing this, shown in this slide and in the next one.
The left structure in this slide has the advantage that it can be shielded from the substrate noise by applying a fixed bias to the n-well.

Slide 102
Here we have the two other possibilities. The structure on the right has the disadvantage that the bottom plate of the capacitor is AC grounded.

Slide 103
These MOS capacitors can go to very high density values. In the case of a 0.13 um technology, for example, we have around 10 fF/um^2. But the big problem with them is that they are highly non linear, and therefore they are unusable in many analog applications. This is a measurement we have done on two of the four structures we have illustrated, designed in a 0.25 um technology. The bottom plate here was grounded and the top plate was swept between -2.5 V and 2.5 V.
We can see that these structures are really highly non linear.
What is interesting for us is the part of the curve for $V > 0$. We can see that the NMOS in the N well capacitor is the best option, as it is the more linear of the two.

Slide 104
The noise from the substrate affects the devices through the bulk transconductance and the parasitics between the substrate and the gate, source and drain.
As explained in the slide, there are three ways through which the noise originated by the digital circuits can disturb the analog circuits:

- The power and ground lines
- The parasitic capacitances between interconnection lines
- The common substrate

To solve the first problem, we can keep separated the bias lines for analog and digital circuits. This is quite effective.

For the second problem, we have to make the layout with extreme care.
For the third problem… see next slides!

Slide 105
Also important is the position of the pins in the package!