P. Bruschi

Lecture Notes on Analog Circuit Design

Part 1: Design flow and Devices

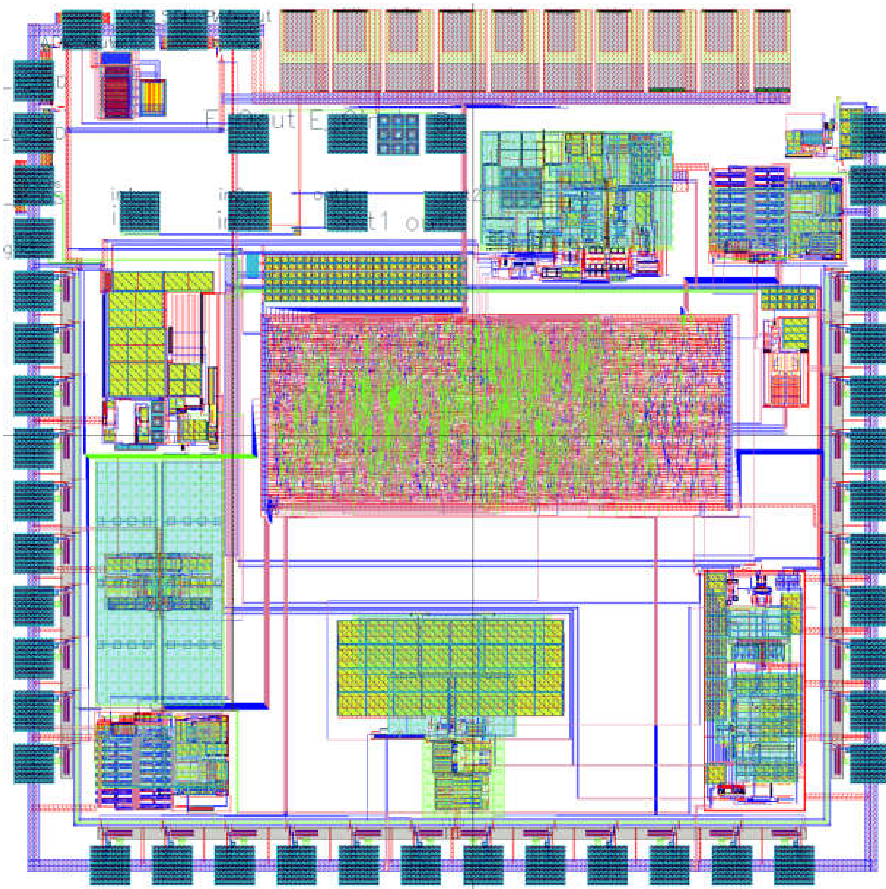

Table of Contents

1 Fundamentals of Integrated Circuit Design 5
1.1 Definitions 5
Integrated circuits 5
Design flows. 6
1.2 Analog Design Flow 8
1.3 Common features of a CAD environment 11
Structure of a project: libraries 11
Structure of a project: instances and hierarchy 12
Schematic editor 13
Layout editor 15
Layout rules 17
1.4 Example of technology: the CMOS process 20
General considerations 20
Simplified process flow and layout elements of a standard 1P2M n.Well CMOS process 20
Bonding pads 24
Possible variants present in commercial CMOS processes 25
1.5 Brief list of technologies alternative to CMOS 27
1.6 Resistances and capacitances in Integrated Circuits 28
Resistances 28
Capacitances 28
1.7 References 29
2 Passive Components in Integrated Circuits. 30
2.1 Resistors 30
General considerations. 30
Polysilicon resistors 31
Diffusion resistors 32
2.2 Capacitors 34
General considerations 34
Metal-Metal capacitors 35
Polysilicon capacitors 37

Junction capacitors. ... 38
2.3 Integrated inductors ... 39
3 Active device models and layouts ... 40
3.1 MOSFET layouts ... 40
Layout description ... 40
Designer options ... 42
3.2 Mosfet models ... 43
Control voltages and operating regions. ... 43
Drain current equations in strong inversion ... 45
Drain current in weak inversion ... 45
Moderate inversion ... 46
Saturation voltage, V_{DSAT} ... 46
Junction currents ... 46
Temperature effects ... 47
Small signal model and parameters ... 47
Transconductance models in saturation region ... 50
Capacitance models ... 51
Matching parameters ... 53
3.3 Bipolar Transistor Layouts. ... 54
Layout descriptions ... 54
Designer options for BJTs ... 55
3.4 Bipolar transistor models ... 56
Large signal model ... 56
BJT small signal model ... 58
3.5 References ... 60
4 Process errors ... 61
4.1 General definitions ... 61
4.2 Fabrication errors in a microelectronic process: global and local errors. ... 63
4.3 Matching errors. ... 65
4.4 Local granularity: The Pelgrom Model ... 66
4.5 Gradients ... 68
4.6 General rules for matching components ... 72
4.7 Rules for accurate ratios. ... 73

4.8 Error propagation elements applied to matching errors 77
One-dimensional case 77
Multi-dimensional case 79
Useful examples of relationships that occurs frequently 79
4.9 References 80

1 Fundamentals of Integrated Circuit Design

1.1 Definitions

Integrated circuits

An integrated circuit (IC) is formed by components and interconnections that are fabricated on a single silicon piece of semiconductor, typically indicated as "chip" or "die". An integrated circuit can be divided into several sub-circuits, performing different functions, typically indicated as "cells". A whole chip can also be regarded as a cell (top-cell).

Each cell includes different views (cell-views), consisting in different ways to represent the cell. Different cell-views are used at the various steps of the design flow. We will examine this concept later. Now, let us focus on two different cell-views:

1) Schematic cell-view

2) Layout cell-view

The schematic view is the schematic diagram of the electrical circuit, representing the cell as a set of simpler parts, connected by means of their terminals.

The layout view is a geometrical representation of the cell. The object is contained in a 2-dimensional region (cell-outline), which define the substrate area occupied by the cell. The layout is a collection of 2-D shapes (e.g. rectangles), which tell the factory (called silicon foundry) the exact areas, within the cell outline, where the various process steps have to be carried out. Each shape represents an object that will be fabricated on the substrate. The particular type of object (e.g. polysilicon, metal, doping implantation) is represented by a property of the shape, which is called "layer".

Figure 1.1 shows a possible schematic and layout view for a CMOS nand gate. The available layers in a CMOS process are collected in the so called "layer palette". Note that modern IC fabrication processes include a much greater number of layers. The layer set is completely defined by the process and the designer has no control on it. The layout designer decides where the various layers (roughly corresponding to process steps) have to be applied. For example, referring to Fig.1.1, the areas that will be covered by polysilicon in the fabricated cell are shown in Fig.1.2. Note that, to improve readability, the layout editor shows different color shades where two or more layers cross each other.

The actual project of an integrated circuit generally uses a much wider set of views. For example, complex logical cells are often represented by means of hardware description languages (such as VHDL and Verilog). Therefore, the views "vhdl" or "verilog" may be necessary. We will introduce another type of view, the "symbol" view, later in this document.

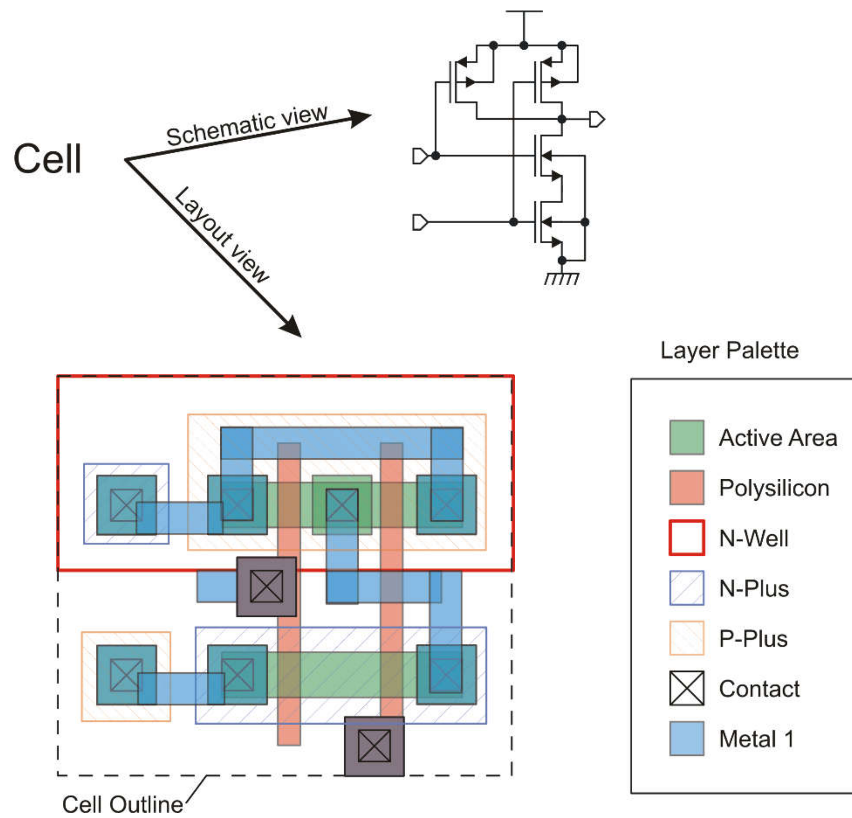

Fig.1.1 Schematic and layout views of a cell

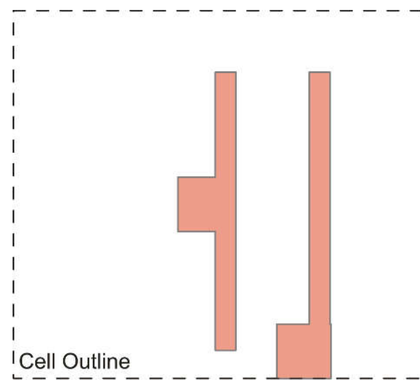

Fig.1.2. Areas covered by polysilicon in the cell to be fabricated from the layout of Fig.1.1.

Design flows.

A Design Flow is the collection of all steps that are necessary to design the IC, from the specifications to the layout. The three main design flows that are currently followed are:

- Analog Design Flow
- Digital Design Flow

- Mixed-Signal Design Flow

The various design flows with their typical products and dependencies are shown in Fig.1.3. The analog design flow is optimized to produce analog cells. Generally, at the core of the design flow there is the schematic view, which represents the electrical network (simpler parts, connected through nodes), generally indicated with “netlist”. The design process is mainly manual, with very low standardization (the designer “style” still counts). On the other hand, the Digital Design Flow is founded on an HDL description, which is often behavioral (that is, not based on connection of actual components). The Digital design flow is used to produce complex digital architectures, such as finite-state machines, microprocessors, digital filters. The design flow from the HDL description to the layout is mostly automated. In many cases, the designer even does not need to inspect the final layout. Note that the Digital Design Flow requires the availability of a library of simple standard cells (logical gates, flip-flops, multiplexers etc.) which should be produced with the analog flow (see the figure).

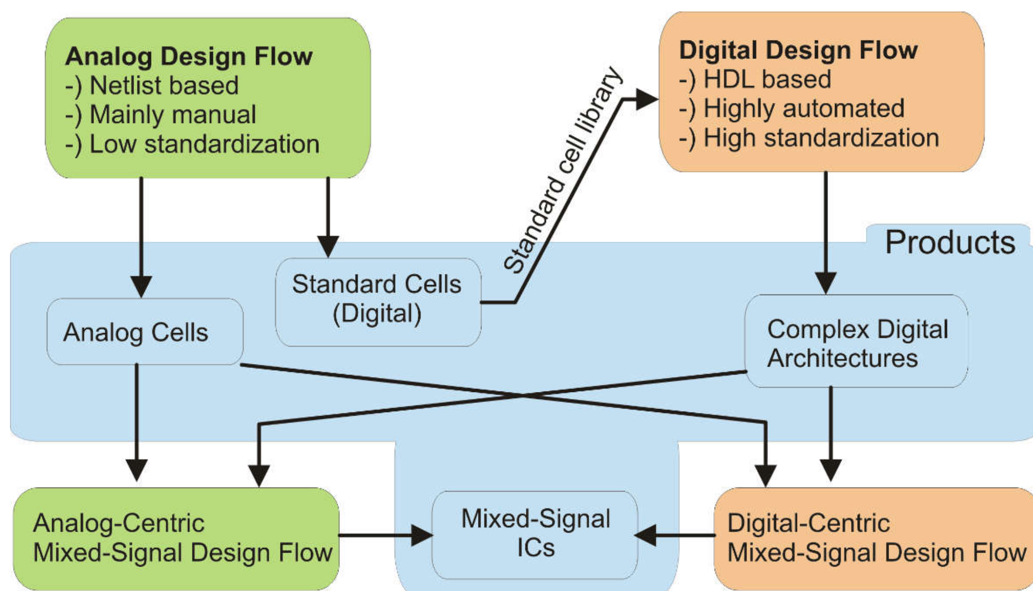

Fig.1.3. Integrated Circuit Design Flows

Mixed-Signal circuits, i.e. circuits including both digital and analog blocks, can be designed with either an “Analog-Centric” or “Digital Centric” flow. In both cases, the analog and digital cells are designed with their proper flows. In the Analog-Centric flow, the digital cells are combined with the analog ones using the same tools of the analog flow. In the Digital-Centric flow, the analog cells are combined with the digital ones using the Digital-flow approach. Generally, the analog-centric flow is preferable when the analog subsection is dominant and/or when the analog and digital cells cannot be simulated separately. Alternative ways to indicate the Analog-Centric and Digital-Centric flows are Analog-Top or Digital-Top flows, respectively, with reference to the environment of the final design composition.

1.2 Analog Design Flow

A simplified diagram showing the various steps of the Analog Design Flow are shown in Fig.1.4. The figure is formed by three columns. The central column, enclosed into green boxes, includes the main design steps. The specific CAD tool, when applicable, are indicated in italic, between round parentheses.

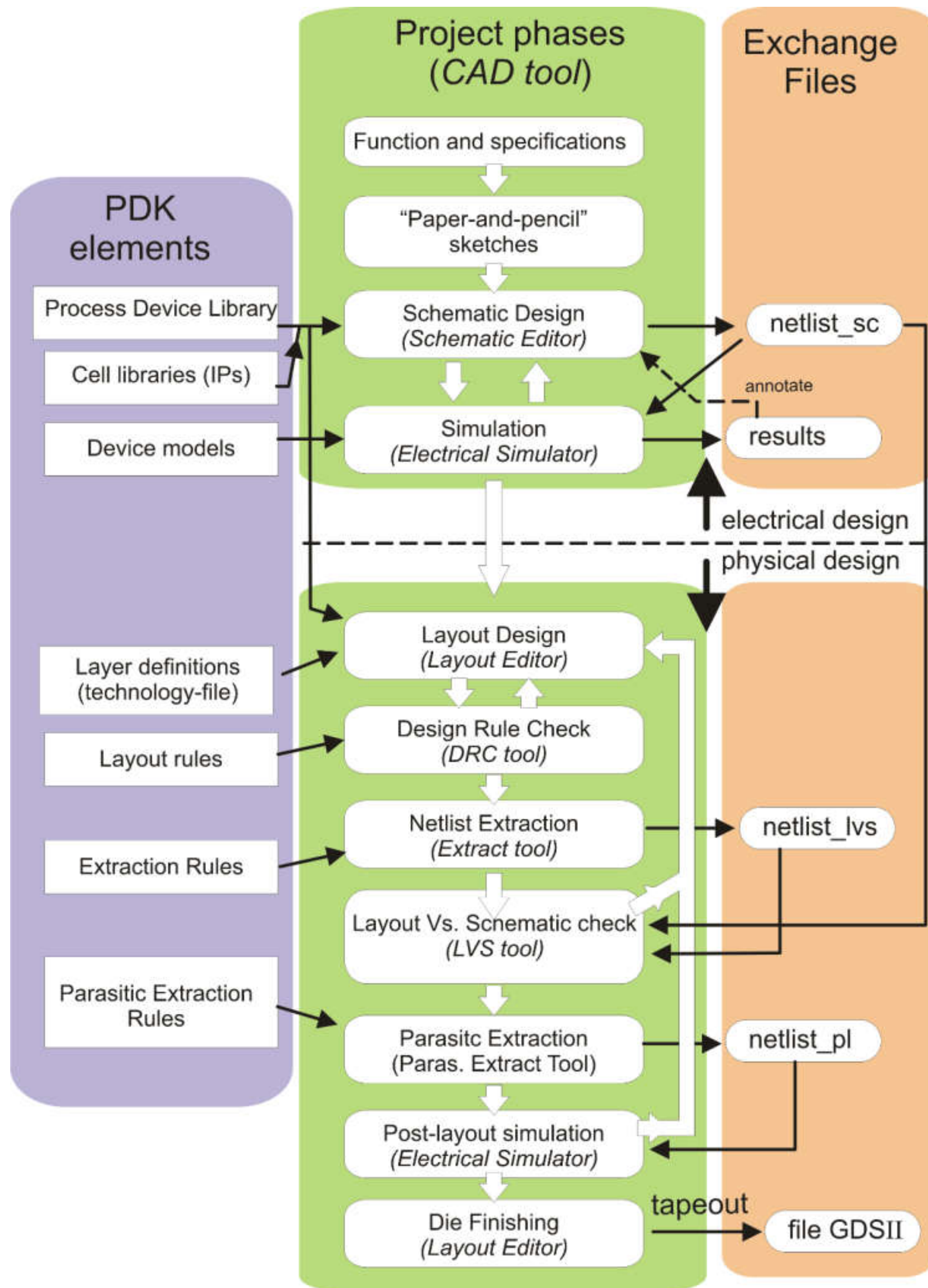

Fig.1.4. Analog Design Flow. The required PDK (Process Design Kit) files, the CAD tools and common exchange files are indicated

The CAD tools are generally available within a single CAD platform. The platform allows the various tools to be executed from the same user-interface and simplifies data exchange between the tools. A given CAD platform may include tools of different vendors.

The CAD platform should be customized with a series of files that contain information about the fabrication process. This collection of files forms the so-called Process Design Kit (PDK). A PDK includes files that are required at different stages of the project. The main PDK elements are collected inside the violet box in Fig.1.4. Most of the data included in the PDK are detailed in human readable form in the DRM (Design Rule Manual).

Finally, the orange boxes collect the main files that are generated and exchanged between the tools.

The process flow of Fig.1.4 is applicable to the design of a complete chip or of a single cell, with minor differences. The project is divided into the two main phases:

- Electrical (or Schematic) design. In this phase the schematic view of the circuit is developed.
- Physical design. This phase includes all steps necessary to design a layout that implements the schematic view designed in previous step.

At the beginning of the design flow, there is the description of the function that the circuit must implement and a series of specification that define the desired performances of the circuit. For example, we may have to design an operational amplifier (= function) with at least 1 MHz Gain-Bandwidth Product and a static gain greater than 80 db (= specifications).

The first phase of the design is the most creative and generally consists of drawing sketches by "pencil and paper" and making approximate, first order calculations. The designer uses his/her experience to find the best topology that meets the requirements. If a suitable topology is not available in textbooks and in scientific articles, the designer should try to develop an original topology, using transistor-level design skills. In this phase, it is very important to be able to perform simplified circuit analysis to obtain a rough estimate of the circuit performances and perform transistor sizing in order to obtain a first version of the circuit to start with. In this preliminary phase it is very useful to divide complex circuits into simpler ones (design partitioning).

The circuit is then drawn using a *schematic editor*. The designer will use the PDK device library, which includes all the devices that can be actually implemented with the chosen technological process. If available, the designer can use also macro-cells (e.g. operational amplifiers, comparators, etc.) that the foundry or other designers of the group have previously created. These cells are grouped into independent libraries and are generally referred to as "Intellectual Properties (IP).

When the circuit or part of it has been drawn, then a series of simulations are performed by means of an *electric simulator*, in order to estimate the circuit performances. Currently, several different electrical simulators are available, but all of them represent further developments of the SPICE (Simulation Program with Integrated Circuit Emphasis) program. In order to allow a circuit to be simulated, it is necessary to extract the netlist ("netlist_sc" in Fig.1.4), which is a textual file (ASCII characters) that lists all components and connections present on the circuit. The netlist should adhere to the typical SPICE format. For the simulations to be executed, it is necessary to provide the simulator with the so-called "device models". These files, which are part of the PDK, contain a set of parameters that are required to represent the behavior of all available devices. Device models generally include a very large number of parameters for each device, which are estimated by the foundry by means of a long series of delicate

measurements. The simulations produce data that can be plotted as graphs (e.g. voltage vs time plots), typically referred to as "waveforms" or simple collections of node voltages and branch currents, as in the case of the rest point. The waveforms are saved in binary files that are displayed with proper programs, which generally are part of the simulator package. Single data sets, such as rest-point data, can even be displayed directly in the schematic view, with an operation called "annotate", facilitating interpretation. Analysis of the simulation data gives information to the designer on the changes that should be applied in order to approach the target performances. Generally, a very large number of cycles between the schematic editor and the simulator are necessary to get the desired results. As the design gets refined, simulations become more and more sophisticated and start regarding the analysis of "manufacturability", that is the impact of process variations on the performances of the circuit. A good design should guarantee that most of the dies that will be fabricated satisfy the initial specifications, regardless of all possible variations and different operating conditions (supply voltage, temperature etc.).

When the final version of the schematic design is reached, then the physical design phase can be initiated. The physical design is carried out using the *layout editor* which is a graphical tool that allows insertion of shapes (rectangles, paths, polygons etc.) into a design area, where each point of the plane (i.e. substrate surface) is represented by means of orthogonal coordinates, typically expressed in microns. For the layout to be consistent with the chosen process, it is strictly necessary to use only the layer-set provided by the foundry within the PDK. The layer definitions are included in a file called "technology file", or "tech-file". Depending on the available tools and type of PDK, the layout creation is widely assisted by automated procedures. Generally, it is not necessary for the layout designer to draw the devices, since the latter are generally available as cells (p-cells, as we will see later). In the analog design flow, the designer should place and size the devices and draw the interconnections. The exception is given by the design of elementary digital gates (inverter, nand gates etc.) where, in order to optimize area occupation and speed, the designer defines also the shape of the individual devices.

In order to guarantee that the layout can be actually fabricated, it is necessary to respect a series of design rules, defined in the DRM and included into the PDK. A particular tool, the DRC (Design Rule Checker), is used to check whether the layout meets all the design rules. During the creation of the layout, it is necessary to launch the DRC several times and to apply corrections according to indications contained in the DRC report. The fact that the layout does not contain violations of the design rules is not sufficient to assert its correctness. Indeed, it is necessary to check also whether the layout implements the same electrical network as the schematic. The operation of checking the correspondence between schematic and layout is called LVS (Layout *Vs* Schematic). To perform this operation, the netlist actually implemented by the layout is extracted (netlist_lvs in Fig.1.4). Then, the LVS tool performs comparison between the layout netlist and the schematic netlist, producing a report that details all discrepancies between the two networks. Clearly, if the LVS detects errors, it is necessary to correct the layout and repeat all verifications (DRC and LVS). Netlist extraction from the layout is controlled by proper "extraction rules" included into the PDK.

For critical designs, where the parasitic components introduced by the interconnections can significantly affect the circuit behavior, it is necessary to perform a different kind of extraction, called "parasitic extraction". This operation extracts a netlist (netlist_pI in Fig.1.4) from the layout, including also interconnect parasitic components (capacitances and resistances). This netlist is then simulated (post-layout simulation), obtaining an estimation of the circuit performances which closely match the physical design. Due to the presence of parasitic components, the performances estimated at this stage may significantly differ from those obtained from the simulation of the schematic design. Note that the

parasitic components of most devices (Mosfets, BJTs etc) are already well represented by the device model. Thus, simulations performed in the electrical (schematic) design phase is already close to the actual behavior of the circuit, even regarding dynamic parameters (such as gain-bandwidth product). The effect of interconnect parasitics can be critical only in particular cases, such as switched capacitor circuits, complex digital architectures with particularly long interconnections or radio-frequency (RF) circuits. Clearly, parasitic extractions needs an extended set of extraction rules ("Parasitic extraction rules" in Fig.1.4) with respect to the simpler case of the extractions made for the LVS purpose. If the performances estimated with the post-layout simulations are not satisfactory, it is necessary to find the critical interconnects and modify the layout to reduce their impact on the circuit behavior.

When a layout that meets all specifications is obtained, it can be released for production. If we are dealing with a single cell (an "IP"), then no other steps are necessary. The cell will be stored for further use and a brief report that summarizes the cell performances will be created. In the case that the circuit that we are creating is a complete chip, then it is necessary to perform some further steps that are generally referred to as "die finishing". These steps, which are well documented in the DRM and automated by CAD routines, vary according to the foundry and the particular technological process. An example of die-finishing is the placement of "planarization dummies", which are generally metal patches placed in empty areas to obtain a nearly constant density across the chip for all metal levels, in order to facilitate planarization procedures. When the final layout is ready, it is exported in a universal format. The current format for layouts is the GDSII format (GDS means Graphic Database System while "II "stands for the roman number "2"). The GDSII is then sent to the foundry for fabrication. This final step is called "tapeout", since in earlier times, the GDSII file was saved on a magnetic tape that was physically sent to the foundry. Nowadays, file transfer is performed over an Internet connection.

1.3 Common features of a CAD environment

Structure of a project: libraries

In most CAD platforms, projects are organized as a set of libraries. Libraries are simply a collection of cells, which, in turn, are described by a series of views, as shown in Fig.1.5.

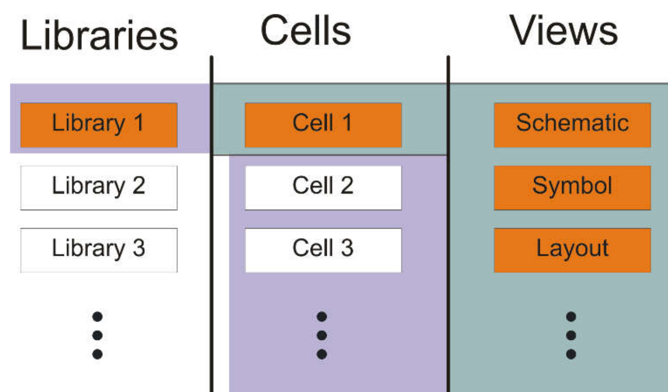

Fig.1.5. Project organization as a collection of libraries cells and views.

Libraries may have several purposes. The current design generally consists of one or more libraries, which includes the cells that are being developed. All devices that are available in the chosen process are given as a series of cells included in the so-called "process library". Cells that were designed by other

employees or by different companies are organized also in specific libraries. A common library is the standard-cell library, which includes all elementary digital cells to be used to synthesize digital sub-circuits (generally with a digital design flow). Finally, auxiliary libraries can also be present. These libraries includes ideal components (such as resistors capacitors, voltage sources etc.) to be used to simulate the circuits that we are developing (e.g. provide the supply voltage and input signals, represent a certain load condition etc.). These ideal components should clearly not be present in the cell being designed, since they would alter the result of the LVS.

Structure of a project: instances and hierarchy

A project generally consist of a top cell, which is the target cell that should be produced in the end. The top-cell can be either a functional block (e.g. an IP) or an integrated circuit. The top-cell is composed of simpler cells, which are connected each other by means of proper nodes called terminals. In the example of Fig.1.6, cell *C* includes as sub-parts cells *A*, *B* and *K*. A cell may appear into another an arbitrary number of times. For example, cell *A* appears two times into cell *C*. Every time we use a cell inside another one, we create an “instance”. All instances are independent objects, even if they refer to the same original cell (as the two instances of cell *A* into *C*). This corresponds to the physical structure of the final object: cell *A* is a circuit with electron devices inside. The two instances correspond to two copies of the same cell placed into two distinct areas of the chip. The electron devices of the two instances will be perfectly independent objects, with their own voltages and currents.

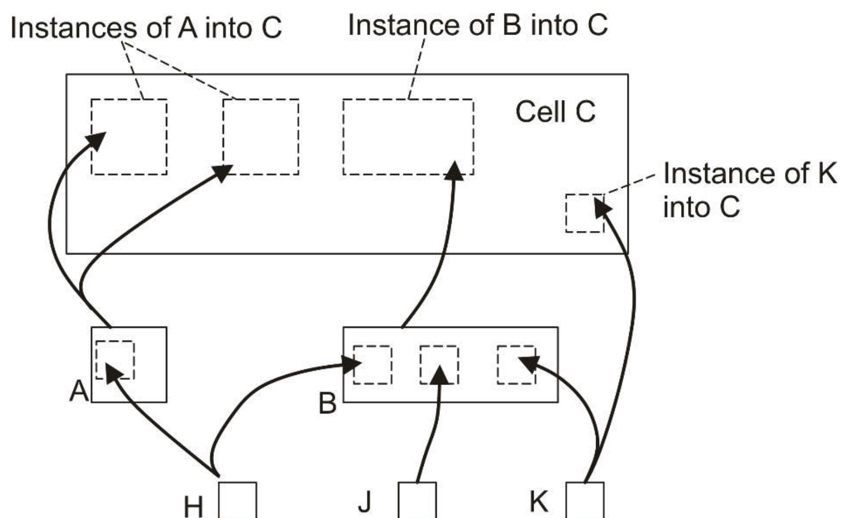

Fig.1.6. Example of hierarchical organization of a project

The fact that the two (or more) instances are linked to the same cell, is clearly used to optimize the design flow, since one needs to draw the cell only one time and then he/her can place the cell thousands of times with no need of redraw it. Note that if we change the original cell, all instances are changed: this is a powerful property, that allows us to update a project in a very fast way, but the designer should be well aware of it, since it may cause unwanted effects. For example, if we want to change the properties of only one instances, we cannot change the original cell but we have to create a new one and to replace the instances with it.

If we expect to make many instances of the same cell that differ by some property (e.g. a resistor value, or a MOSFET width), it is convenient to create a so-called P-Cell (Parametric Cell). A P-Cell is

characterized by a series of parameters that can be set when the cell is instantiated. In this way, it is possible to personalize the instances and it is not required to create a different cell for each different value of the parameter(s).

The fact of describing a cell as a combination of instances of simpler cells implicitly create a hierarchy, simply defined by the following law: *if cell Y includes at least an instance of cell X, then the hierarchical level of Y is higher than X one.* The transitive relation applies to hierarchy: if, in terms of hierarchy, *Y* is higher than *X* and *X* is higher than *Z*, then *Y* is higher than *Z*. Considering Fig.1.6, *C* is clearly at an higher hierarchical level than *A*, *B* and *K*, but, for the transitive relation, also than *J* and *H*. On the other hand, *J* and *K* are at the same level as *A* even if *J* is instantiated into *B*, which is at the same level as *A*.

A cell can contains instances of simpler cells from different libraries, which can either be custom libraries, created by the user, or PDK libraries, including all available devices or IPs created by the foundry or third parties.

Schematic editor

There are two main cell views that are involved in the schematic design:

- Schematic view
- Symbol view

The schematic view is a way to represent a cell as the connection of simpler ones. Connection occurs through terminals. A set of terminals connected together forms a node (or “net” in the CAD jargon). Then a schematic view includes instances of simpler cells and nodes. The schematic view is conveniently represented in a graphical way, where lines (wires) are used to connect the terminals and instances are represented with simple shapes (e.g. rectangles). Terminals are generally represented as small dots or short straight lines. The object that is used to represent the instance of a cell is the symbol view. Terminals are selected nodes of a cell that we intend to use in order to connect the instances of cell to other instances. For general purpose cells performing universally known functions (like logical gates, amplifiers, electron devices), a picture recalling the standard symbol is used for the symbol view (e.g. a triangle for an amplifier, classical symbols for MOSFETS, resistors, capacitors, etc.). For custom cells created for the user, a generic rectangle is preferred.

Figure 1.7 shows the hierarchical schematic design of a 3-input nand-gate (Nand3). The schematic view includes two instances of a 2-input nand gate (Nand2) and one instance of an inverter. The symbol view of the Nand3 includes the three input terminals and the output one. The Nand2 cell also owns a schematic and a symbol view. Let us focus on the schematic view to define all passible node types:

- **Terminals** (in1, in2, out). The purpose of terminals is to allow the cell to be connected to other ones when it is instantiated into a higher-level cell. Terminals should be defined first in the schematic view by connecting the selected nodes to proper objects (generally called “pins” or “ports”. Different types of pins are possible, with “input”, “output” and “bidirectional” being the most common. The symbol type is used only for checking purposes, for example to issue a warning when two output pins are connected together. In an analog design, “bidirectional” pins are often used since “input” and “output” categories are often not applicable for analog nodes.
- **Global nodes** (gnd, Vdd). Global nodes are used to connect nodes of a large number of instances together, with no need of defining a terminal for that purpose. For example, in the Nand3 cell of

Fig.1.7, the *gnd* nodes of all cells are connected together. This obvious condition recurs in practically all circuits. Use of dedicated terminals to indicate an obvious condition would reduce readability of the schematic representation. The same situation applies to the supply voltage connection (Vdd). For this reason, it is possible to make a node global. This can be done using a conventional name for that node or, preferably, connecting the node to a dedicated object that represent the global node (e.g. the gnd symbol). A set of global nodes (e.g. gnd, Vdd, Vcc, AVdd, DVdd etc) are pre-defined in the schematic editor, but custom global nodes can be created by the user, for example to bring a clock signal to a large number of cells with no need of adding a dedicated terminal.

- **Internal nodes** (N1). All nodes that are neither terminals nor global nodes are internal nodes. Internal nodes cannot be used for connecting instances together.

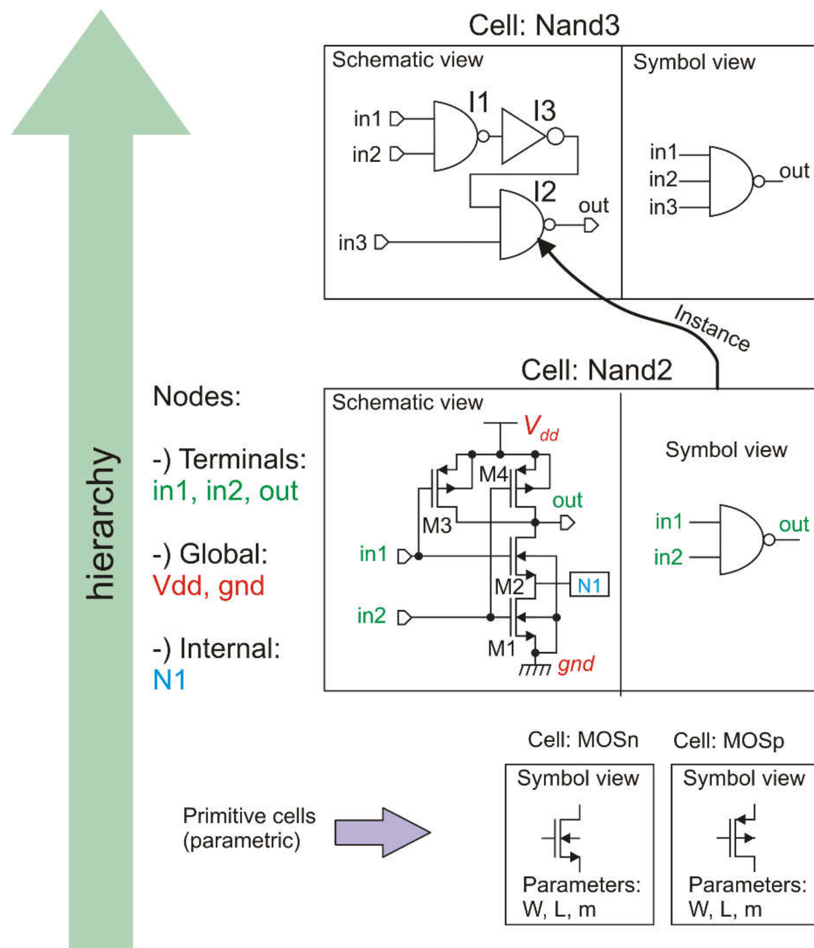

Fig.1.7. Elements and structure of a schematic editor.

When we create a terminal, we have to choose a name for it. Distinct terminals should have different names. When we connect a node to a symbol pin, in order to give that node the status of "terminal", we give the terminal name to that node. It is possible to give a name also to an internal node by connecting a "label" to the node. This operation can be used to connect several nodes together (forming a single node) without using wires. In fact, in complex circuits, most connections are made by label, since wires

would be too tangled to provide a fast view of the circuit architecture. If we give an internal node the same name as a pin, we connect it to the pin.

Note that there are cells that cannot be decomposed into simpler ones. These are, for example elementary devices such as MOSFETS, resistors BJTs and so on. For these cells, a schematic view is not applicable and only the symbol view is present. These cells are called "primitive cells". Primitive cells are generally P-Cells, since one or more parameters are necessary to define the properties of the cell. For example, a resistor cell will have at least a "resistance" parameter. MOSFETs will require at least a "W" (channel width) and "L" ("channel length") parameter. In most schematic editors, parameter values can be associated to a given instance by means of a graphical interface (e.g. a pop-up menu).

Even if internal nodes are not accessible for connection purposes, they actually exist, since when we create an instance of a cell, we generate an object that contains all the elements of the cell. Internal nodes are still accessible when we perform simulations or verifications (e.g. LVS). For example, if we want to plot simulation data for node N1 of instance I1 in the Nand3 cell, we have to refer to that node with the hierarchical expression I1.N1, where I1 is the name of the instance. With the same syntax, we can access an instance of a cell included into another instance. For example, device M1 (instance of a MOSFET cell) inside instance I1 in Nand3 will be I1.M1. Fig. 1.8 shows a few example of this way to indicate nodes and instances that are present at lower hierarchical levels. This syntax can be nested to access the lowest level of hierarchy from the current level.

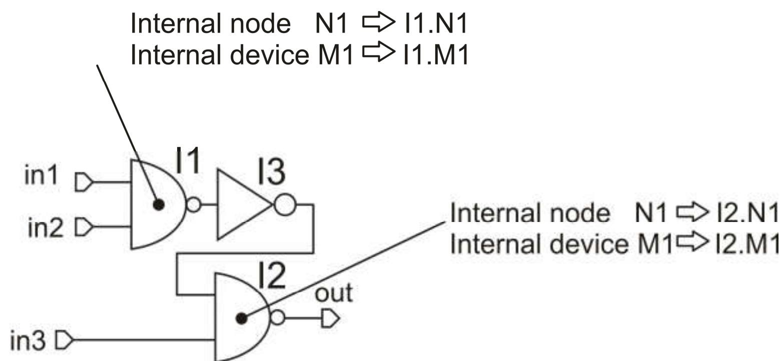

Fig.1.8 Typical convention used to access internal nodes and devices.

The netlist extracted from the layout generally maintains the hierarchical structure of the

Layout editor

The hierarchical organization of the schematic design is reflected into the layout design. Complex cells are formed by instances of simpler ones. However, instances appear in a different way with respect to the schematic design, where the symbol view conceals the complexity of the schematic view, improving readability. In the layout design, when an instance is created, the whole layout of the cell being placed is visible. This is necessary because, in the layout design, geometrical coordinates are of primary importance. Thus, it is necessary to have visual control on what we are placing into a layout in order to avoid that the cell being instantiated conflict with other elements already present in the layout. Furhermore, while placing (e.g. instantiating) a cell into a layout, we can immediately orientate it (rotate, flip) to facilitate interconnections with the surrounding cells, or to position it in such a way to optimize

silicon area occupation. However, the instance operation is different from a simple "copy and paste" action, since instances are selected, moved, rotated and flipped as a whole, with no risk to break them into parts. As a counterpart, we cannot select or modify any part of an instance, even if we can see all individual shapes that form it. To modify an instance we have to modify the original cell, but as already stated, the change is applied to all instances. To modify a single instance, we have to duplicate the original cell, modify it, and use it to replace the instance of interest. As for the schematic view, we can also define the original cell as a P-Cell. In the layout view, it is possible to use a specialized language to change the shape of the cell according to one or more parameters.

The instance mechanism that we have detailed above is not desirable when we have to conceal the layout of a cell. This occurs when a company designs an IP and sells it to clients, allowing them to use it in their projects, but not to analyze the layout. In this case, it is possible to use a special view called "**abstract**" view. Using this view, it is possible to show only the outline of the cell and the points (pins) where the users have to make the connections. The final layout is not made by the user, but by the foundry, where the company that designed the cell has deposited the full layout. The foundry simply replaces the abstract view with the corresponding layout view. Clearly, use of abstract views in a layout prevents DRC and LVS from being applied to the whole chip.

As shown in Fig.1.4, the PDK provides the layout editor with a proper set of layers. Layers can be divided into three types, according to the following list:

- **Technological or "tooling" layers**. These layers correspond to objects that can be actually fabricated on the chip with the chosen process.
- **Derived layers.** These layers are obtained with logical operations performed on other layers. For example, we can define a layer "GATE" as the intersection (= logical and) of the POLY and ACTIVE tooling layers. In that case, a GATE layer will be generated (when required) in any region of the layout where POLY and ACTIVE shapes overlap. Derived layers are mainly used to facilitate writing of the DRC and Extraction rules.
- **Service Layers.** These layers are used to add information on the layout and have no direct correspondence to objects to be fabricated on the chip. Examples are the layers to be used to mark selected interconnecting lines (pieces of metal) as pins or layers used to inform the foundry that there are areas where we have intentionally violated the DRC rules to design experimental objects.

Only the tooling layers affect the creation of the set of photomasks by which the chip is fabricated. Nevertheless, there is not a 1:1 correspondence between a layer and a photomask. This is due to several reasons listed below:

- Dimensions of the shapes drawn in the layout are those of the final object that will be fabricated on the chip. Due to several non-ideal effects occurring during the fabrication process, the patterns drawn in the photomasks are altered. Phenomena involved in this process are proximity and interference effects during UV exposure, under-etch and lateral diffusion of dopant. For this reason, masks are properly modified to take into account these effects and make sure that the final object is as close as possible to the designed one. This operation is transparent to the user and is performed by the foundry in the so-called "Mask Data Preparation" (MDP) phase.

- In order to produce an object, often several photolithographic steps are required, each one involving an individual photomask. The designer does not have generally control over the single steps, but only on the final object, for which, he/she will use only a single layer.
- Sometimes, an operation performed on the chip (e.g. doping) is frequently done only in regions that depends on other process steps. For example, in a CMOS process, n-plus doping is applied over areas where p-plus is not applied. Then we can save design time by obtain the n-plus mask by means of a complement operation (logical NOT) applied to the p-plus mask. In these cases, there will not be an n-plus layer, but only a special mask, to be used in the rare case that we do not want neither a p-plus nor an n-plus doping. The choice of the particular layer set depends on the foundry. For the example given above, a foundry may choose to provide the designer of both the n-plus and p-plus layer.

Note that, depending on the type of process step involved, the shape created in a certain layer may need to be inverted before creating the mask. In the case that the mask is a positive copy of the layer (after the adjustments detailed above, during MDP), then the layer association is marked as "clear field". If an inversion is required, then the layer is marked "dark field". Fig. 1.9 illustrate this convention.

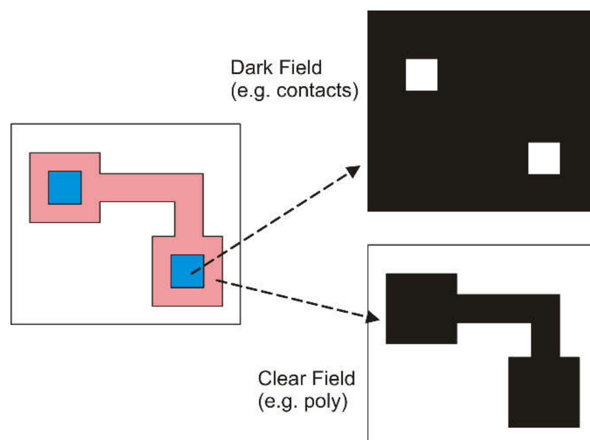

Fig.1.9. Layer to mask conversion in the case of clear and dark field.

Layout rules

Layout rules (sometimes referred to as "Topological Layout Rules" – TLR) guarantee that the objects that have been drawn by the designer can be actually fabricated. For example, if we draw a long thin line of metal that we want to use to carry a signal across a region of the layout, we need that the line is continuous. If the line width is too small, than there will be the risk that the line will be interrupted in some points. Then, to guarantee that the line is continuous on the chip, we need to design it with a width greater than a minimum value, reported in the DRM. Examples of design rules are illustrated in Fig. 1.10. Spacing rules guarantee that two object that we draw as distinct entities, actually have no interactions that can alter normal operation. This concept covers the trivial case of two conducting lines that we want to be isolated and that, if the minimum spacing is not respected, risk to come into contact. Spacing applies also to more complex cases. For example, a minimum distance can be applied to two diffusions whose depletion regions risk coming into contact, producing a punch-trough effect. The extension rule is

introduced to guarantee that a shape of layer "B " (pink in the figure) crosses a shape of layer "A" (light blue) dividing the latter into two separate parts. This is, for example, the case of planar MOSFET layout, where layer "A" is active area and "B" polysilicon.

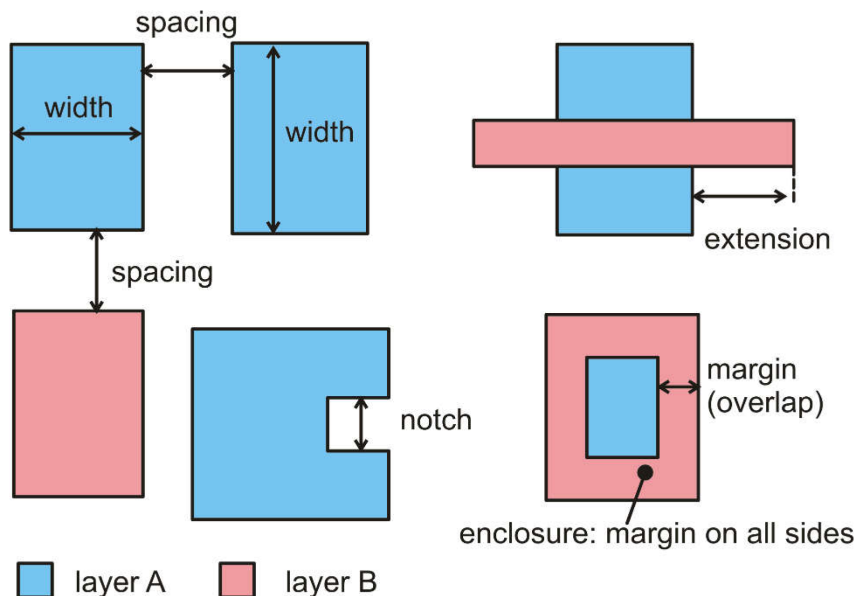

Fig.1.10. Example of layout rules. Notches are generally treated as spacing.

The margin rule guarantees that object in layer "A", which is drawn over object in layer "B" stays at sufficient distance from the border of layer "B". The enclosure rule commands that a margin is respected from all edges of the shape in layer "B". The margin rule is sometimes indicated as "overlap", with identical geometrical meaning.

In most cases, layout rules are expressed as minimum values. There are rare, albeit important, cases in which the rule is an exact value. This applies, for example, to the width of contacts and vias. Cases of maximum value are also possible.

An important rule, which is derived from the minimum width and minimum spacing the pitch, defined by:

$$pitch = W + S \tag{1.1}$$

where W and S are the minimum width and spacing, respectively. As Fig. 1.11 clearly illustrates, the pitch is minimum step by which objects in a given layer can be repeated. If we have to create a required BUS of N parallel lines, we need at least an N*pitch space. The pitch is the parameter that is more frequently used to indicate how the interconnection layers of a given process are efficient. Pitch applies to every element that have to be arranged in regular arrays, such as memory cells and pixels of a image sensors.

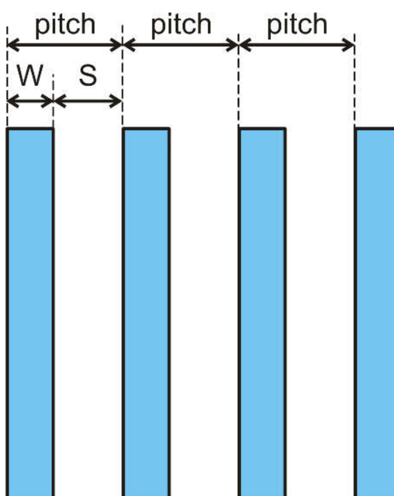

Fig.1.11. Definition of pitch as the sum of the minimum width (W) and minimum spacing (S).

Layout rules can be expressed in two different ways:

- Micron (or absolute) rules
- Lambda-based (or scalable) rules

Micron rules are expressed with their absolute value, typically given in micron or, more recently, in nanometers.

Lambda-based rules are given as multiples of a *length unit* called "lambda" (λ), which is given in micron and is a property of the process. The minimum dimension in a lambda-based rule set is 2λ. Other rules can be 3λ, 4λ and so on. Lambda rules were popular when scaling down of circuit size was mainly limited by the resolution of photolithography. Then, every time photolithography was improved, all rules of the process could be scaled down uniformly. Therefore, a layout designed for a given version of the process was still usable when the process was improved. The only required operation was changing (i.e. reducing) the value of lambda. The idea of lambda-based rules was introduced by Meade and Conway [1] around the end of the '70s and was aimed to facilitate migration of a design across different technological processes. Lambda rules turned out to be no more convenient when photolithography resolution got better than 1.0 μm and other physical phenomena started to contribute to limit device scaling. The design rules of modern processes do not scale down uniformly, so that they must be re-written when a technology is improved. For this reason, the design rules of practically all available IC fabrication technologies are micron rules.

There are two notable exception. The first one is constituted by the so-called MOSIS scalable rules. MOSIS is a service, based in the United States, which takes processes of international silicon foundries and re-elaborate the design rules (micron rules) to derive a set of lambda-rules. To do this and maintain manufacturability, a few rules should be properly increased, reducing the competitiveness of the process. In order to reduce the value of lambda and allow the layouts developed with a set of rules to be fabricated with an improved version of the process, it is necessary to wait until all rules are scaled down significantly. Even in that case, it is the rule that has been scaled-down by the worst factor to determine the scale-down factor for the lambda value, and then for all other rules. Layouts designed with these sets of rules are still correct, but are also somewhat larger than the process could allow if the design were

done following the actual (micron) rules provided by the foundry. The MOSIS service is mainly targeted to University or research institutions, for which the possibility of re-using older layouts and educational documents (made possible by scalability of the lambda rules) is more important than competitiveness.

Another example of scalability is represented by the so-called “shrunk” processes. These processes are upgraded versions of an older process for which it is possible to apply a uniform scaling of all rules. The scaling factor is generally modest (e.g. 10 %), so that the “shrunk” process cannot be considered as a new technology. However, this operation allow extending the competitiveness of the process, which, not being the leading technology any more, can also be offered at a discounted price. Generally, in order to save time, a new set of rules is not created and the design kit remain the same. The only difference is that all geometries are simply scaled down before manufacturing. The same scale factor is also automatically applied to all drawn geometries (i.e. MOSFET lengths and widths) before every simulations is launched.

1.4 Example of technology: the CMOS process

General considerations

A first example of complementary MOSFET p-n logic was proposed by F. Wanlass of Fairchild R & D Laboratory in 1963 [2]. After being limited only to ultra-low-power, low performance circuits (such as digital watches), CMOS integrated circuits progressively supplanted all other type of digital circuits. At present, CMOS technologies are by far the most used process for fabrication of all kind of digital circuits and constitute the preferred choice for mixed-signal circuits. For these reason, a standard CMOS process will be briefly recalled in this section.

The purpose of this description is showing which elementary objects (e.g. doped areas, metal shapes, etc.) can be fabricated on the chip with a typical CMOS process and highlighting the relationship of these objects with the layers that are generally available in the PDK. The way the object are fabricated (i.e. the actual technological procedures) is not the subject of this document.

Simplified process flow and layout elements of a standard 1P2M n.Well CMOS process

We will refer to a standard n-well CMOS process with one polysilicon and two metal levels (1P2M). N-well processes are used more frequently than P-well ones. Many CMOS process flows starts from a heavily doped substrate (p doping, order of 10^{19} cm^{-3}) on top of which a lightly doped epitaxial layer (p doping) is grown. Selected areas of the epitaxial layer are either n-doped (forming the n-wells) or p-doped (forming the p-wells). In this way, it is possible to set the correct doping for the n-wells and p-wells independently of the initial substrate doping (twin-tub processes), obtaining the desired threshold voltage for both the n-MOSFETs (placed inside the p-wells) and p-MOSFETs (placed into the n-wells).

An important distinction should be made at this point: since the substrate is p-type, all p-wells are electrically connected by ohmic paths. Thus, p-wells cannot be electrically independent On the contrary, n-wells forms a p-n junction with the substrate and can be isolated from the latter (and each other) by providing a proper reverse bias to the junctions. This fact has an important impact on the degree of freedom that the designer has in connecting the substrates (bodies) of MOSFETs. To make sure that the well-substrate junctions and drain/source-substrate junctions are reverse biased, the p-wells (and then the substrate) are connected to the lower supply voltage node. This node, conventionally indicated with Vss, coincides with ground (gnd) in single supply circuits. The heavily doped substrate placed under the epitaxial layer works as a ground-plane that improve potential uniformity of the p-wells, reducing the so-

called substrate noise and the risk of latch-up. The lightly doped epitaxial layer, instead, is ideal for device fabrication, due to the initial low impurity density.

CMOS processes are classified by the minimum allowed channel length for both the n and p MOSFETs. In a polysilicon-gate process, the minimum channel length coincides with the minimum width of the polysilicon lines.

Figure 1.12 (a) shows the situation after p-well and n-well definition. In the following step, represented in Fig. 1.12 (b) creation of the active areas is depicted. At this stage, the silicon surface is covered by a thick oxide layer (FOX – Field-oxide) except for the active areas, where the crystalline silicon is exposed. In modern processes (0.25 μm channel length and below), active areas are separated by shallow trenches (STI: Shallow Trench Isolation) and the FOX is the oxide used to fill the trenches. The designer decide the position and shape of the n-wells by means of the N-Well layer and the active areas by means of the Active layer. An example of layout corresponding to the cross section of Fig.1.12 (b) is shown in Fig.1.12 (c).

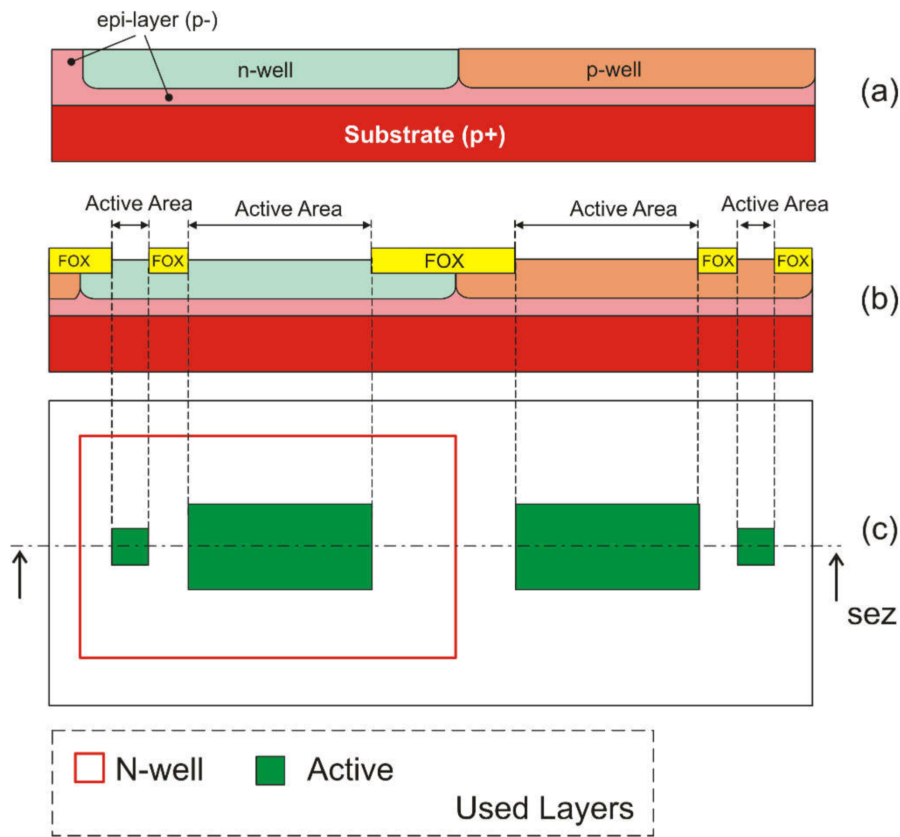

Fig.1.12.(a): Formation of the n-well and p-well. Note that the p-well electrically short-circuited to the substrate. (b) Formation of the active areas. (c): Top-view (layout) of the circuit at stage (b).

Note that four active areas have been represented in the example of Fig.1.1. The two bigger areas will be used to design an n-MOSFET and a p-MOSFET, while the two smaller areas will be used to create contacts for the p-well and n-well.

The next step is the growth (or deposition) of the thin gate oxide inside all active areas. After this step, the entire wafer surface is covered by the polysilicon layer, which is then patterned. The gate oxide is

removed together with polysilicon. Therefore, active areas not covered by polysilicon are free from gate oxide. The situation after polysilicon patterning is shown in Fig.1.13 (a). The following step is doping of the active areas with either n-plus or p-plus implantation. Note that four cases are possible, as shown in table 1.1:

Table 1.1 – Possible types of active areas

Active area position	Type of doping	Function of the active area
Substrate (p-Well)	n-plus	n-MOSFET drain /source, diode
	p-plus	Substrate contact ("Substrate Tap")
N-Well	p-plus	p-MOSFET drain /source, diode
	n-plus	Well contact ("Well Tap")

Finally, both the active areas and polysilicon are silicided, i.e. their surface is transformed into a metal silicide (e.g. Titanium or Cobalt silicide). The process, which is also referred as "salicide" (Self-Aligned silicide), is used to lower the sheet resistance of active areas and polysilicon. The situation after these steps is shown in Fig.1.13 (b), while the corresponding layout is shown in Fig.1.13 (c). Only three new layers, namely Poly, N-plus and P-plus are required. Note that n-plus and p-plus doping occurs only in active areas, since it is stopped by the FOX. Therefore, the N-plus and P-plus layers can be much larger than the active areas, since they have no effect outside of the latter. Active areas must be completely covered by either the n-plus and p-plus doping. The active area portions that are covered by polysilicon are also not affected by n-plus or p-plus doping, since dopant is stopped by the polysilicon layer. This fact is essential to form the MOS transistors (self-aligned drain and source).

Polysilicon that intersects an active area forms a MOSFET gate. Fig. 1.13 (b) and the corresponding layout of Fig.1.13 (c) clearly shows that the bigger active areas are split into three parts by the polysilicon line that crosses them. The two areas that are not covered by polysilicon are the drain and source, which have opposite doping with respect to the underlying silicon (p-well or n-well). The part covered by polysilicon has the same doping than the substrate and forms the channel of the MOSFET. The portions of active area that have received the n-plus or p-plus doping are generally indicated as "diffusions".

Polysilicon is doped by the same n-plus or p-plus implantation used for the drain and source (once again, note that the doping layer covers the entire active areas). Then polysilicon is p-doped when it forms the gate of p-MOSFETs and is n-doped for n-MOSFETS. This is required to obtain complementary threshold voltages. Generally, it is not necessary to provide doping of polysilicon outside active areas, where it serves as an interconnect layer, since silicide is sufficient to provide a low sheet resistance.

At the stage depicted by Fig.1.13 (b) active devices have been already created. The steps required to get to this point are indicated with FEOL (Front-End Of the Line). The following steps, which are required to provide the main interconnections, are called BEOL (Back-End Of the Line).

In order to create the interconnections, the devices and polysilicon layer should be isolated from the metal layer that will be deposited on top of them. Therefore, an insulator is deposited over the whole wafer and then holes are opened through this layer only at the points to be connected. These holes are defined by a layer called "Contacts". The situation after contact opening is shown in Fig. 1.14 (a). Holes are filled with tungsten to provide electrical contact (tungsten plugs). After planarization, the first metal level (metal 1) is deposited and patterned. An insulating layer is then deposited and holes ("vias") are opened

through it where a contact between the metal 1 and metal 2 (second metal level) is required. Again, vias are filled with tungsten plugs, a planarization step is applied and a second metal layer is deposited. We will stop the process here, since we are considering a CMOS process with a dual metal layer. Finally, a passivation layer (dielectric) is deposited to protect the metal-2 layer from humidity and accidental scratches.

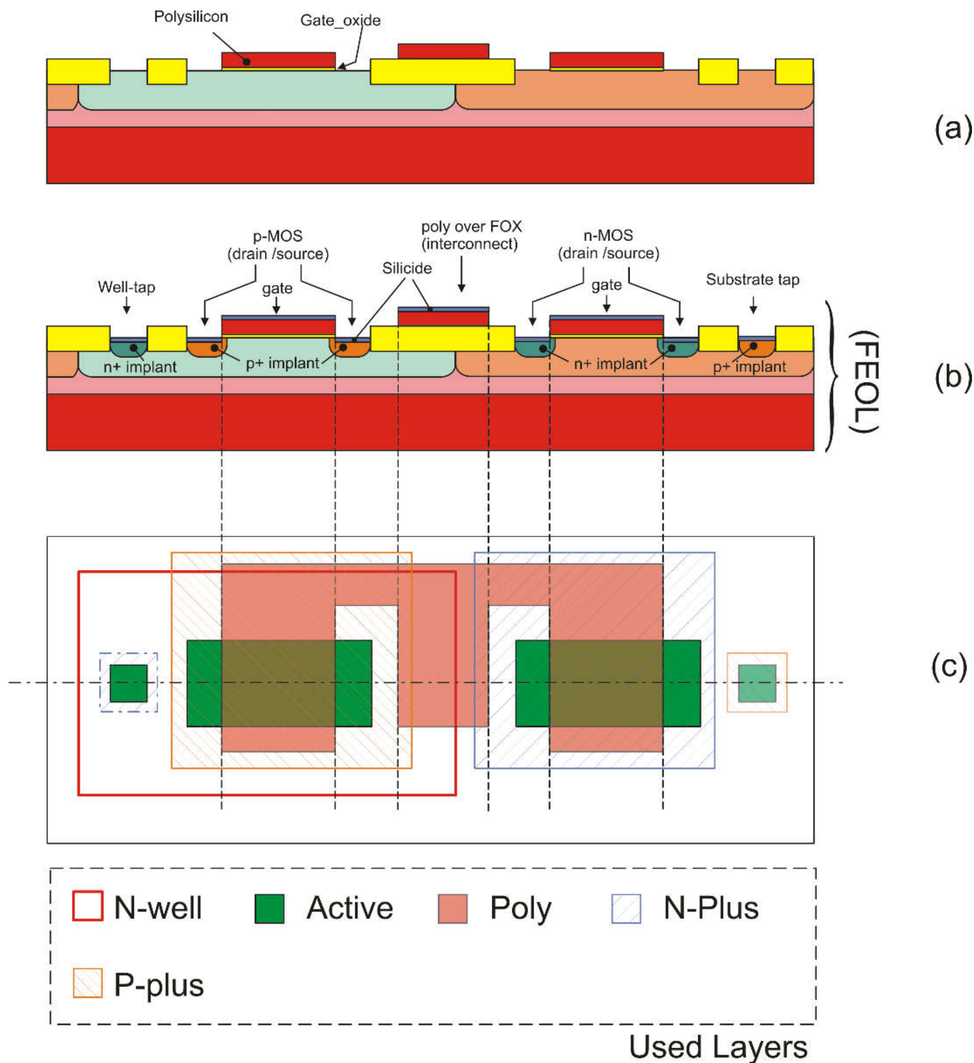

Fig.1.13. (a) Situation after patterning of the polysilicon layer. (b) Complete FEOL, including active area doping and Self Aligned Silicide (Salicide) growth over polysilicon and diffusions..

The new layers required to define the interconnection layers are the following: contact, metal 1, via and metal 2. Note that the contact layer is used to connect a metal 1 shape to both an active area or a polysilicon shape, depending on where the contact is placed. Generally, contacts and vias cannot be drawn with arbitrary dimensions, but they should be drawn as squares of fixed side length, defined in the DRM. This is an example of "exact" design rule. The DRM gives also indication on the maximum current that can be carried by a single contact. To allow a connection between two layers to carry more current than a single contact (or via), the connection will be implemented using a number of contact sufficient

to meet the requirement. Note that three contacts are used to connect the drain /source areas of the MOSFETS to metal-1 lines.

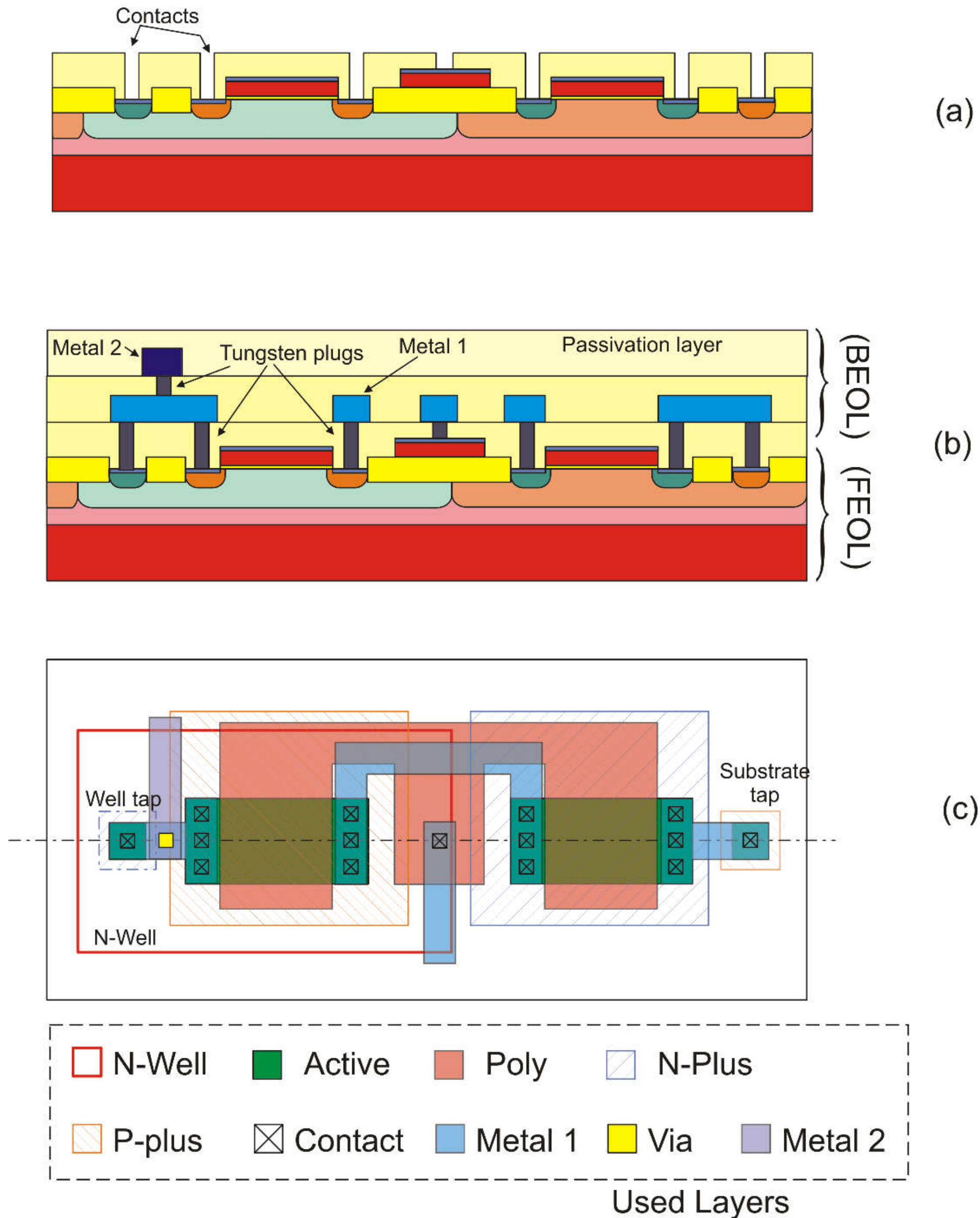

Fig.1.14. BEOL for a dual metal layer CMOS process. (a) Contact opening through the dielectric layer. (b) Situation after deposition and patterning of the second metal level. (c) A possible layout corresponding to the cross section (b).

Bonding pads

The fabrication process of an integrated circuit is concluded with the packaging phase. This consists in gluing the die into a package and then connecting the terminals of the chip to the package pins. Bonding may be accomplished in several ways that depend on the type of integrated circuit, the number of terminals to be connected and the chip performances. In all the cases, it is necessary to create special structures called "bonding pads" (or simply "pads") on the die. Pads are metal areas (typically squares)

as large as several tens of microns (e.g. 60 μm × 60 μm), which are suitable to be connected to similar structures that are present in the chamber of the package where the die is placed. "Wire bonding" is the most common technique to connect the bonding pads of the chip to the pads of the package. Figure 1.15 (a) shows packaging of a chip into a surface mount case, with connections made by means of wire bonding. The structure of a bonding pad for a dual metal process is shown in Fig. 1.15 (b). Opening of the passivation over the metal 2 later is defined by means of the Passivation opening layer (often simply called "Pass", or "Passivation"). Generally more than a metal layer is stacked (metal 2 and metal 1 in the figure) to improve robustness and allow connection with several different metal layers.

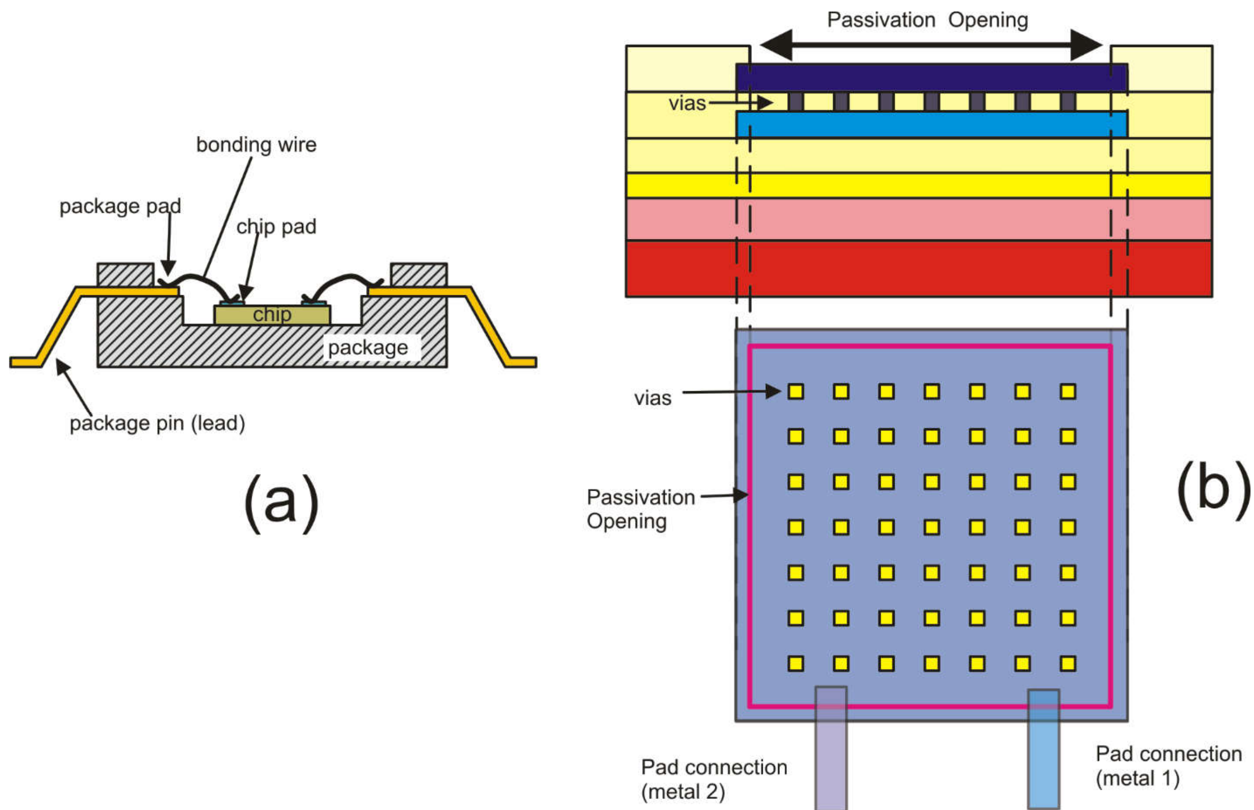

Fig.1.15. (a) simplified view of chip packaging and wire bonding (b) structure of a bonding pad in a dual metal process.

Possible variants present in commercial CMOS processes

Modern CMOS processes may present additional characteristics that make them more performant or more versatile. A non-exhaustive list of additional features is the following:

- More metal levels (even 6-8 metal levels). A large number of interconnection layers allows more dense integration of digital circuits and may facilitate carrying large currents by simply placing lines of different layers in parallel. Note that in fast microprocessors the power lines may carry several amperes,

- Thick metal layer. Normal metal layers are generally nearly 0.5 μm thick. This value allow very small line widths to be achieved. The upper metal level is generally thicker (up to 3 μm), in order to carry more current with reduced sheet resistances.
- Native MOSFETs (ZVT devices). These devices are n-MOSFETs fabricated on the original lightly doped substrate. In this way, their threshold is much lower than the one of regular n-MOSFETs, often close to zero. Native MOSFESs are not used in digital circuits but may be very useful for analog blocks. Generally, Native MOSFETs do not require additional masks or process steps.
- Low threshold MOSFETs (LVT devices). These devices are fabricated in special n-wells or p-wells, where the doping is adjusted to obtain a lower threshold voltage. LVT devices can be either n or p MOSFETs and are used for analog applications or ultra-low voltage logic circuits.
- MOSFET families with different voltage ratings. Deep submicron process feature MOSFETs that are damaged by supply voltages higher than 2 V. The weakest point is generally the oxide gate, which imposes strict limits to the gate-source and gate-drain voltage. Unfortunately, communication between different chips on a printed circuit board still occurs at higher voltages (e.g. 3.3 V). To allow the chip to communicate with such voltage levels, the process generally includes an additional CMOS family (n and p MOSFET pair) that can stand higher voltage thanks to a thicker gate oxide. These devices are larger and slower than the so-called "CORE" devices of the process and, for this reason, are used only to build I/O buffers or analog blocks. Clearly, additional masks are required to decide in which active areas the thicker gate oxide has to be deposited.
- Passive devices. These devices (resistors, capacitors and, less frequently, inductors) are practically indispensable for precision analog circuits. Generally, additional process steps and/or masks are required. For example, polysilicon resistors require a mask that prevents silicide to be grown onto selected polysilicon stripes that have to work as resistors.
- Triple well option. In the twin-well (twin-tub) process described above, all p-wells are connected to the substrate, which, in turn, is connected to the lowest supply voltage. As a result, it is not possible to connect the body of selected n-MOSFETs to their sources to avoid the body effect. This is possible for p-MOSFETs, since their bodies are the n-wells, which are insulated from the substrate. The fact that all p-wells (i.e. the bodies of n-MOSFETs) are connected together may introduce unwanted coupling between different sub-circuits (substrate noise). Triple-well processes allow creation of p-wells that are insulated from the substrate. The principle is illustrated in Fig.1.16. The key element is a buried well, which is obtained by means of high-energy ion implantation, or simply as a buried layer created before epitaxial layer growth. The p-well to be insulated is closed into a box, whose side walls are obtained with a ring made with the conventional n-well. The bottom of the box is closed by the buried n-well. Triple well process, beside allowing independent biasing of both p-wells and n-wells, are also suitable for space applications since they are less prone to the latch-up phenomenon, which is much more severe in environments exposed to high energy particles.

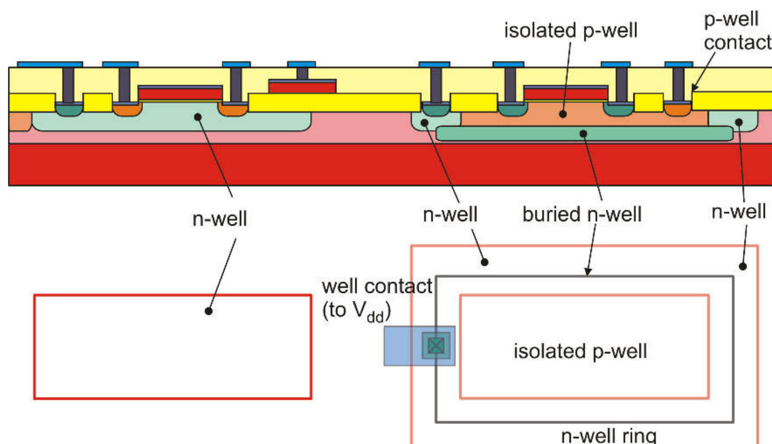

Fig.1.16. Cross-sectional view and layout of an insulated p-well obtained with a triple-well process.

1.5 Brief list of technologies alternative to CMOS

CMOS technology is used to fabricate the largest part of integrated circuits produced every year in the world. Currently there is a very large number of CMOS processes available in the market, covering most cost/performance combinations. Speed improvements has recently allowed CMOS technologies to compete with Bipolar and BiCMOS processes in RF applications. The refinement of switched techniques for offset and flicker noise reduction is making CMOS analog circuits match the precision performances of Bipolar amplifiers.

Nevertheless, there are fields where different technologies may still represent a better choice. Table 1.2 includes significant alternatives to the CMOS technology, including both long established technologies and emerging materials.

Table 1.2. Alternative technologies to pure CMOS processes

Technology	Available Devices	Notes
Bipolar	Vertical NPN, Lateral PNP	Used for precision and/or fast amplifier. Si-Ge versions for RF applications
Complementary Bipolar	Vertical NPN, Vertical PNP	
BiFet	BJTs and JFETs	Used for precision / low bias current amplifiers
BiCMOS	CMOS + BJTs	Especially relevant for Mixed Signal System on a Chip including RF links or high speed digital communication interfaces.
BCD	Bipolar, CMOS, DMOS	Invented by STMicroelectronics, BCD technologies are nowadays the best choice for smart power applications, due to the high voltage / high power capability of Double Diffused MOSFETs (DMOS)
SOI Silicon on Insulator.	Depends on the process from which it is derived (CMOS, BiCMOS or BCD)	High voltage applications. Space applications, due to resistance against latch-up
GaN, SiC	BJTs, MESFETs, HEMT	Use of wide-gap semiconductor is particularly promising for high power devices and RF applications.

1.6 Resistances and capacitances in Integrated Circuits

Every shape designed using a conductive layer (metal, polysilicon, doped active area etc) exhibits resistances and capacitances. As we have seen, simulations that take into account all interconnect parasitic components are called “post-layout” simulation and require a preliminary phase of parasitic extraction.

Resistances

The example of Fig.1.17 shows a rectangular interconnect line. The resistance between the two indicated terminals is simply given by:

$$R = R_S \frac{L}{W} \tag{1.2}$$

where R_S is the sheet resistance which is measured in Ohm, or in Ohm/squares (Ohm/sq). The latter unit is still equivalent to Ohm, since “squares” stands for “number of squares”, meaning the number of squares of side W that can be lined-up along L. Notice that number-of-squares=L/W.

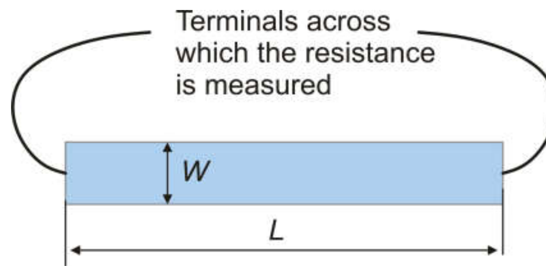

Fig.1.17. Length and width of an interconnection line.

Capacitances

The typical capacitances that affects interconnection lines are shown in Fig. 1.18 (a). Capacitances can be of vertical (C_V in Fig.1.18) or lateral (C_O in Fig.1.18) type. Vertical capacitances are created every time two interconnecting lines of two different layers intersect. The intersection capacitance is given by:

$$C_V = k_A A + k_P P \quad \text{with} \quad A = W_A \cdot W_B \ \text{ and } \ P = 2W_A + 2W_B \tag{1.3}$$

where A e P are the area and the perimeter of the intersection, respectively, while k_A and k_p are the capacitance per unit area and unit perimeter, respectively. The capacitance proportional to the intersection perimeter is due to the electric field line that, as shown in Fig.1.18 (a), extend from the edges out of the intersection area. Perimeter capacitances (also called “fringe” or “edge” capacitances) give a significant contribution when one or both sides are very small. If both sides are large (typically much larger than the minimum width) then the area contribution dominates.

The lateral capacitance occurs between two lines that are on the same layer. A rough approximation that does not take into account boundary (fringe) effects is the parallel plate formula:

$$C_L = \varepsilon_d \frac{L \cdot t_m}{d} \qquad (1.4)$$

where ε_d is the permittivity of the dielectric between the lines, t_m the thickness of the lines and L the length of the segments that runs parallel to each other. For lines spaced 0.25 µm, the lateral capacitance can be of the order of 0.1 fF/µm. This seems a very small value. However, for lines that runs parallel for several hundred microns, the lateral capacitance can be as large as tens of fF. Since C_L introduces coupling between two otherwise independent lines (i.e. a cross-talk), the circuit performances may be strongly degraded. The same consideration apply to polysilicon lines.

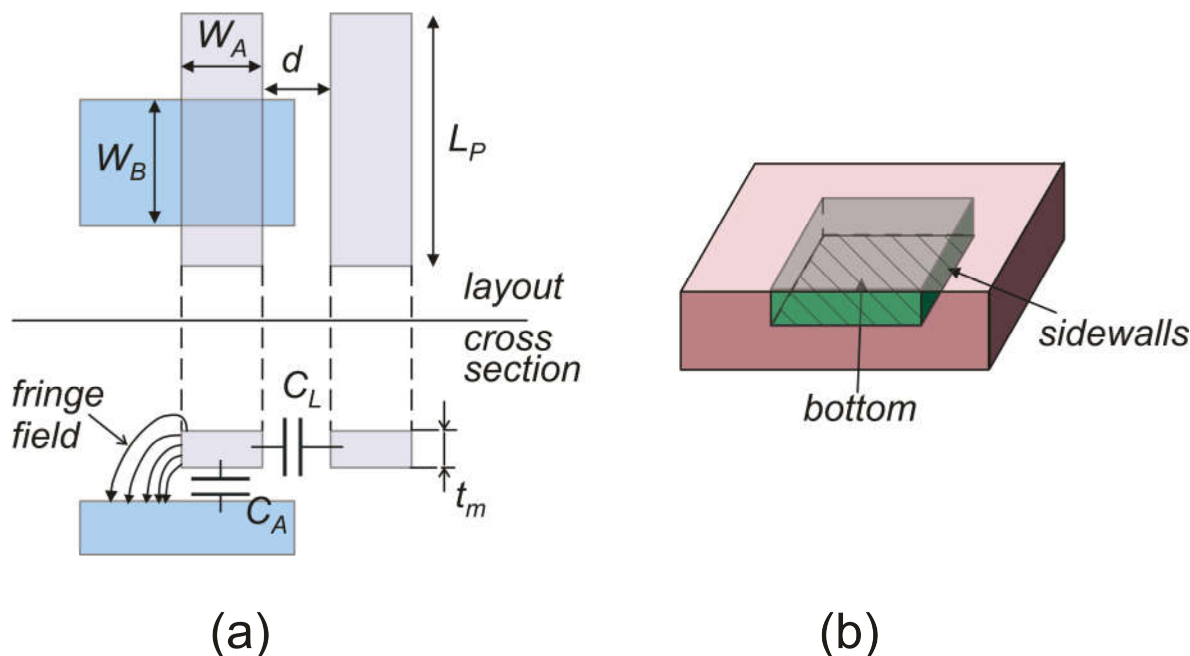

Fig.1.18. Layout and cross-section of metal lines coupled by parasitic capacitances (a); simplified view of a diffusion embedded into a substrate (b).

If the conducting layer is a doped active area (diffusion), then it will be insulated from the substrate by a reverse-biased junction. Then the boundary of the diffused layer will consist of a bottom surface and four sidewalls. Since the junction depth is fixed, the total sidewall area is clearly proportional to the perimeter. Therefore, the capacitance will be the sum of an area and a perimeter contribution, just as in (1.3). Since junction capacitances decrease as the reverse bias is increased, a worst-case estimation will be based on zero-bias capacitances.

1.7 References

[1] Carver Mead and Lynn Conway, “Introduction to Vlsi Systems” Addison-Wesley, Reading (MA), 1980).
[2] Wanlass, F. M. and Sah, C.T. "Nanowatt Logic Using Field-Effect Metal-Oxide Semiconductor Triodes," ISSCC Digest of Technical Papers, February 20, 1963, pp. 32-33.

2 Passive Components in Integrated Circuits.

2.1 Resistors

General considerations.

The generic layout of an integrated resistor is shown in Fig.2.1 (a). We have a resistive layer, which is properly shaped to obtain the required resistance. Contacts to an interconnect layer (metal 1 in the example) are placed at both end of the resistive shape, creating the two resistor terminals. The rectangular shape that is included between the contacts is the resistor section, indicated with resistor body.

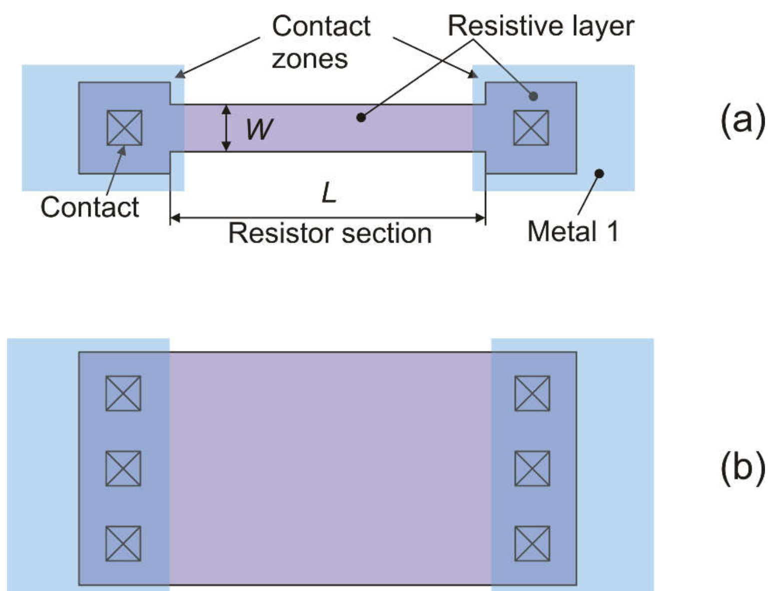

Fig.2.1. Generic resistor layouts.

The resistance of the resistor body is given by:

$$R = R_S \frac{L}{W} \quad with \quad R_S = \frac{\rho}{t} \tag{2.1}$$

where R_S is the so called sheet resistance, ρ and t are the resistivity and thickness of the resistive layer. In order to obtain relatively large resistances, generally the L/W ratio should be made large. To keep dimensions small, W is then set to the minimum value allowed by the design rules. Frequently, this width is not large enough to accommodate a contact. For this reason, the resistor shape is enlarged at both ends, as shown in Fig.2.1(a), introducing a sufficient overlap of the contact object on all sides. If W has to be made much larger than the minimum width (for example to fabricate resistances of low value or to make it withstand large currents) then there is no need to enlarge the resistor ends. For large widths, it is

possible to place more than one contact at both ends, in order to reduce the contact series resistance and increase the maximum current capability of the device.

A process may offer different kind of resistors, which are distinguished by the type of resistive layer. The criteria that are used to choose the best type of resistor are the following:

- **Sheet resistance (R_S)**. To obtain large resistance values, it is necessary to choose the resistor type with the larger R_S. For particularly low resistance values, layers with exceptionally low R_S (such as the metal layers) can be used.
- **Voltage dependence**. Some type of resistors exhibit a large deviation from the ideal V=I/R law. This non-ideality is expressed as a dependence of the resistance on the applied voltage V. A typical expression that is used to model this effect is:

$$R(V) = R(0)[1 + \alpha_{V1}V + \alpha_{V2}V^2] \tag{2.2}$$

  where $R(0)$ is the resistance measured at very low applied voltage while α_{V1} and α_{V2} are empirical coefficients. For applications requiring high precisions, resistors should have negligible dependence on voltage.
- **Temperature dependence**. Temperature dependence is generally expressed with the following formula::

$$R(T) = R(T_0)[1 + \alpha_1 \cdot (T - T_0) + \alpha_2 \cdot (T - T_0)^2] \tag{2.3}$$

  where $R(T_0)$ is the resistance measured at the nominal temperature T_0 (typically 300 K), while α_1 and α_2 are the first order and second order temperature coefficient, respectively. In particular, coefficient α_1, also named TCR (Temperature Coefficient of Resistance) is given by:

$$a_1 = TCR = \frac{1}{R}\frac{dR}{dT} \tag{2.4}$$

- **Parasitic components**. Resistors, as all other integrated devices, are in close contact with the substrate. As a result, there will be parasitic capacitances between the resistor body and the substrate. In the case of diffused resistors, parasitic components include reverse biased junctions between the resistor body and the substrate. Parasitic components contribute to make the resistor behavior non-ideal.

Polysilicon resistors

The cross-section and layout of a polysilicon resistor is shown in Fig.2.2. The polysilicon layer is placed over the field oxide (FOX), providing effective isolation from the substrate. Since polysilicon is salicided by default in modern CMOS processes, if a unsalicided resistor is needed, then the resistor body should be protected from salicide generation by means of a proper layer (Si-Prot in Fig.2.2). Salicide is left on the contact areas in order to reduce the contact series resistance.

Depending on the process, several different types of polysilicon resistors can be available. Table 2.1 includes a few common cases, with the typical sheet resistances and TCRs.

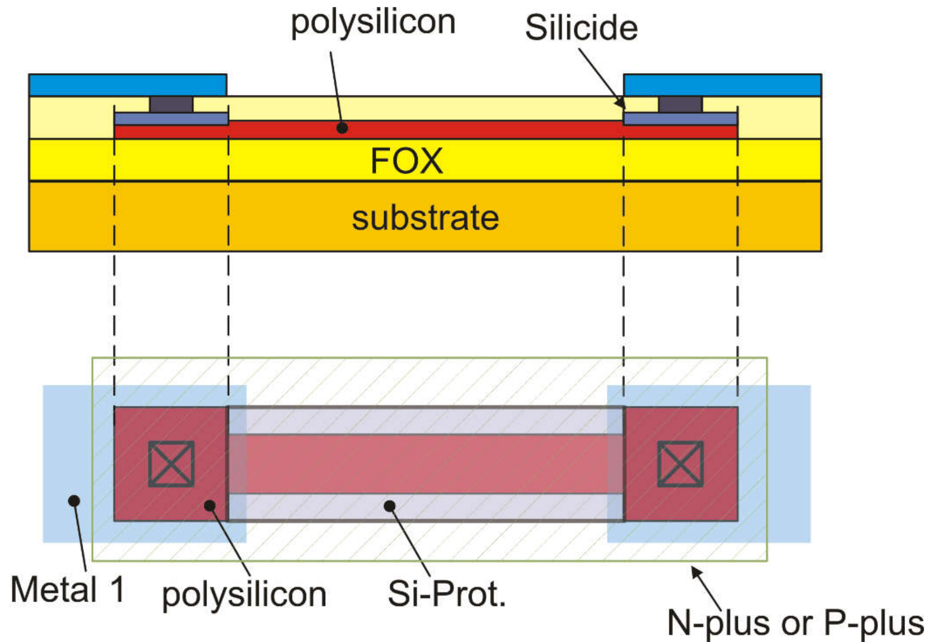

Fig.2.2. Cross-section and layout of a polysilicon resistor.

High-res polysilicon resistors are formed by lightly doped polysilicon. An additional layer (i.e. photomask) is generally required to indicate that the resistor body should not be highly doped as standard polysilicon. N-plus and p-plus polysilicon doping is generally performed with the same implant used to for source/drain doping. Finally, salicided polysilicon resistors, when available, are used when low resistance values are required. All type of polysilicon resistors are marked by very low dependence on voltage, i.e. their current vs voltage characteristics is highly linear.

Table 2.1. Commonly available polysilicon resistors in CMOS processes

Resistor Type	Sheet resistance	TCR	Non linearity (α_{V1})
n-plus polysilicon	30-150 Ω	100-500 ppm/°C	50 ppm / V
p-plus polysilicon	50-400 Ω	250-1000 ppm/°C	-50 ppm / V
high-res polysilicon	400-4000 Ω	–1000 ppm/°C, –3000 ppm/°C	100 ppm / V
Salicided polysilicon	5-10 Ω	2500-3500 ppm/C	-

Diffusion resistors

The resistive layer of a diffused resistor consists of a portion of the crystalline silicon whose doping is opposed to that of the surrounding substrate. In this way, a junction is created between the resistive layer and the substrate. This junction should be properly reverse-biased to provide isolation between the

substrate and the resistive layer. The cross-section and layout of an n-plus resistor implanted into the p-substrate is shown in Fig. 2.3. As in the case of polysilicon, the resistor body should be protected from salicide formation in order to obtain large R_S values. Other possible diffused resistors are p-plus diffusion in n-Well and n-Well diffusion in substrate.

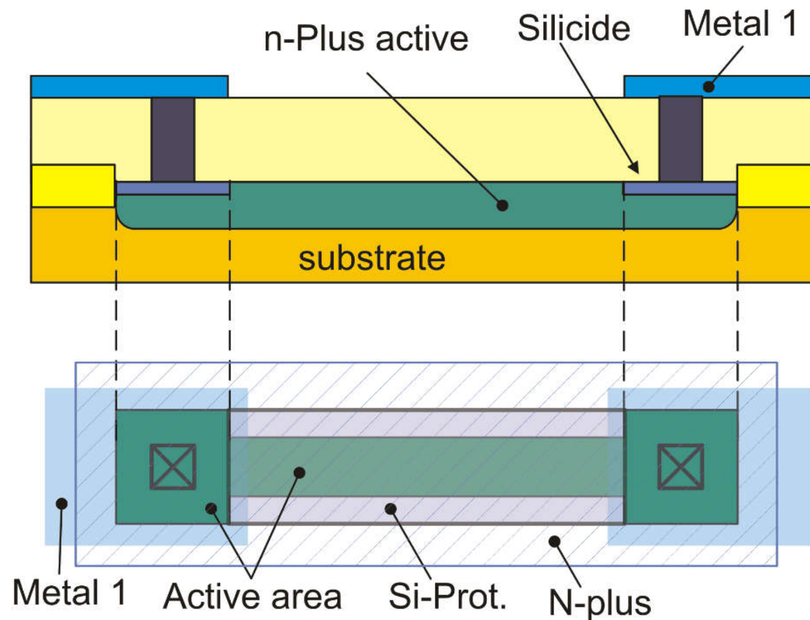

Fig.2.3. Diffused (implanted) resistor: n-plus implant in substrate.

Diffused resistors generally present a relatively large dependence on the applied voltage. The reason is illustrated in Fig.2.4:

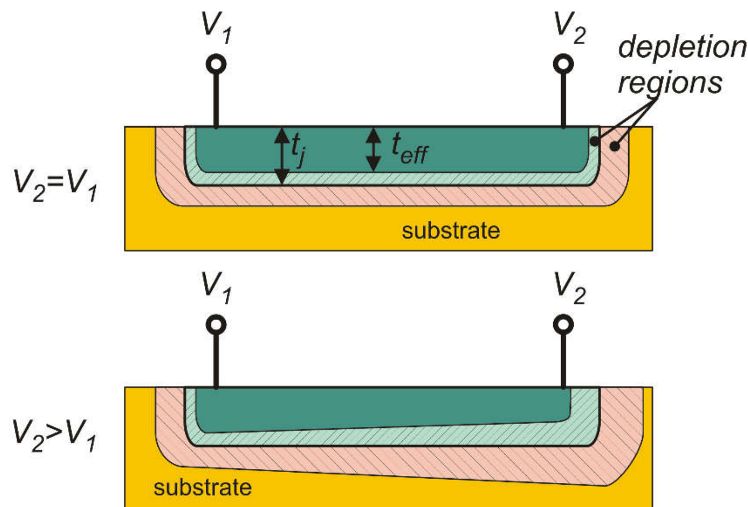

Fig.2.4. Depletion regions in a diffused resistor in the case of null applied voltage (top) and positive applied voltage (bottom).

The problem arises from the depletion regions that form at the boundary between the implant and the substrate. These regions extend also into the resistor body, reducing the effective thickness (t_{eff}) with respect to the thickness of the implant (t_j). The thickness reduction produces a resistance increase, due to expression (2.1). Since the depletion region thickness depends on the voltage difference between the substrate and the implant, the total resistance will also depend on voltage. Note that both the differential voltage across the terminals (V_2-V_1) and the common mode voltage (V_1+V_2)/2 affect the resistance. A change in the common mode voltage with null differential component (Fig.2.4, top) produces a uniform enlargement or reduction of the effective thickness. Application of a differential voltage (Fig.2.4, bottom) produces a restriction of the effective thickness that is maximum at the terminal with higher voltage (V_2 in Fig.2.4). As the applied voltage V2-V1 progressively increases, the effective thickness decreases, producing an increase of the resistance, i.e. a non-linearity in the current-voltage curve. For this reason, diffused resistors should be avoided when high linearity have to be achieved in the presence of large applied voltages. Modern electrical simulators offer resistor model that takes into account both the common mode and differential mode voltages applied to the resistors ("three terminal resistors") using expressions more complicated than expression (2.2).

Table 2.2 summarizes the characteristics of a few diffused resistor types.

Table 2.2. Parameters of diffused resistors in a CMOS process

Resistor Type	Sheet resistance	TCR	Non linearity (α_{V1})
n-plus on substrate	30-80 Ω	1000-1500 ppm/°C	400 ppm / V
p-plus on n-well	50-150 Ω	1000-1500 ppm/°C	400 ppm / V
n-Well on substrate	400-4000 Ω	2000 -3000 ppm/°C	3000 ppm / V

2.2 Capacitors

General considerations

Capacitances consists of two conductors separated by a thin electrically insulating layer. As for resistors, several parameters have to be taken into account when choosing a capacitor type among the available ones.

- **Capacitance per unit area**: this figure is extremely critical when large capacitors have to be integrated into the chip. Typical values ranges between 1 fF/μm^2 to 8 fF/μm^2.
- **Linearity**: this parameter is related to the dependence of capacitance on the applied voltage. A linear capacitance should be voltage independent. In some cases, the fact that a capacitance depends on voltage can be exploited to obtain tunable resonant circuits.
- **Temperature dependence.** Integrated capacitors exhibit a very low temperature dependence (a few tens of ppm/°C. Thus, this is seldom a critical point.
- **Parasitic components**: as for resistors, capacitors may be affected by parasitic components, typically capacitances to the substrate and parasitic diodes, when insulation is obtained by means of reverse-biased junctions.

Metal-Metal capacitors

Metal-metal capacitors can be divided into two different categories:

- MIM (Metal – Insulator –Metal) capacitors
- MOM (Metal – Oxide – Meta) or “flux” capacitors.

MIM capacitors are simply two stacked metal surfaces separated by a thin dielectric. The latter can be silicon dioxide, a different inorganic insulator or a polymeric insulator. The structure of a MIM capacitor is shown in Fig. 2.5. The capacitance is created between a metal layer and the successive, indicated with “metal (n-1)” and “metal n” in the figure. Due to planarization requirements, it is not convenient to use these two interconnect layers to build the two plates of the capacitor. Instead, an auxiliary metal layer, indicated with “metal (n-1) aux” in the figure is used. This metal layer forms a parallel plate capacitor with the metal (n-1). The two plates are separated by a thin dielectric layer, which is deposited on top of the metal (n-1). The upper plate of the capacitor, made of the auxiliary metal, is connected to the metaln layer by means of standard vias.

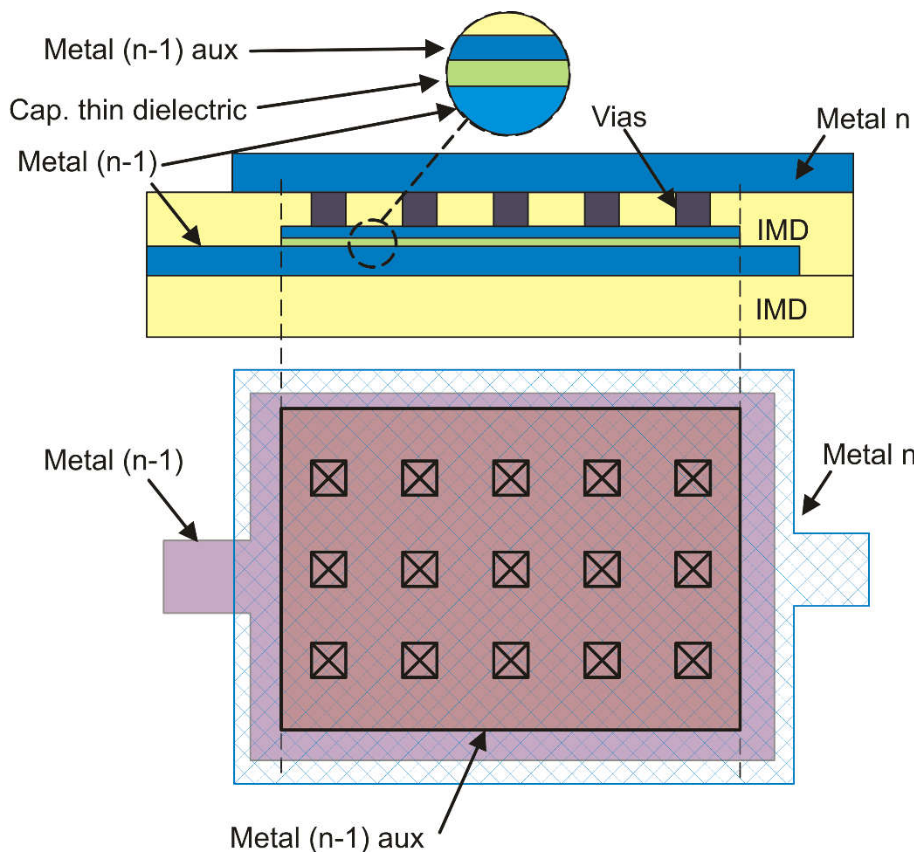

Fig.2.5. Cross-section and layout of a MIM capacitor.

The MIM capacitor layout, shown in Fig.2.5 (bottom) uses only an additional metal layer, the metal (n-1) aux layer, that define the shape of the corresponding metal plate and of the thin dielectric above it.

A MOM capacitor (flux capacitor) is formed by metal lines that are opposed laterally, instead of vertically as in the MIM device. The structure of a MOM capacitor is sketched in Fig.2.6: several parallel lines of a selected metal are alternatively connected to the two terminals of the capacitor, forming a fingered structure. The total capacitance is given by $C_{tot}=(N-1)C_1$, where N is the number of parallel lines and C_1 is the capacitance between two lines.

The advantage of the MOM capacitor is that it does not require additional process steps, because it can be obtained using only standard metal layers. Furthermore, MOM capacitors are generally capable of withstanding much higher voltages.

The main drawback is that, for the same used area, the capacitances that can be obtained are smaller than allowed by MIM capacitors. This problem can be overcome using multi-layer MOM capacitors, which involve several metal layers connected by means of vias, as shown in Fig. 2.7.

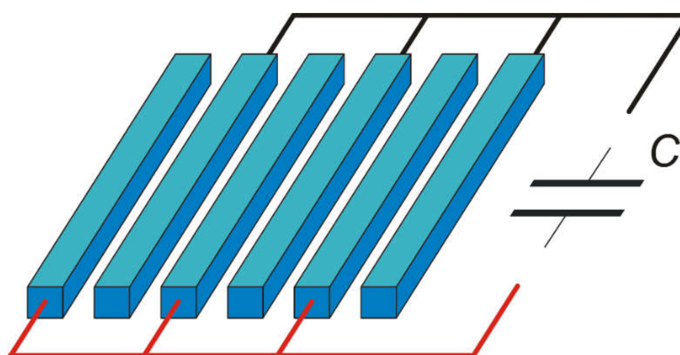

Fig. 2.6. MOM capacitor designed using a single metal layer.

With a MOM capacitor designed with several metal layers, it is possible to match the capacitance-per-unit area of the MIM capacitors. The price to pay is creation of a region that cannot be crossed by interconnections, since all metal layers are used by the capacitor.

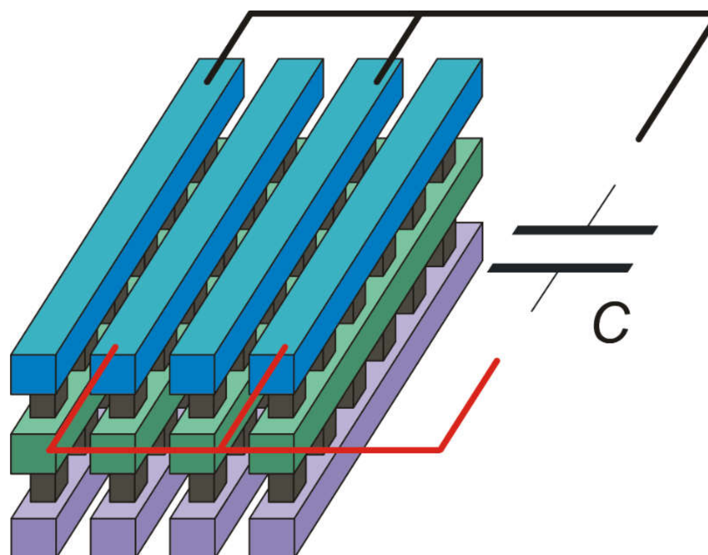

Fig.2.7. Multi-layer MOM capacitor.

On the other hand, the MIM capacitor uses only two of the standard interconnection layers. Frequently, the metal n is the upper metal layer and this allow to put the capacitor on top of active circuits, saving enough lower metal levels to provide basic interconnections.

Polysilicon capacitors

Another common capacitor type is the poly-poly capacitor. This device require two polysilicon levels. Such a feature is frequently available in IC fabrication processes, since it is required to fabricate the floating gates of EEPROMs. The two polysilicon levels are separated by a thin oxide layer, which can be grown by means of thermal oxidation just as the gate oxide. This allows tight control of the dielectric thickness, obtaining high capacitance-per-unit area values. The structure of a poly-poly capacitor is represented in Fig.2.8.

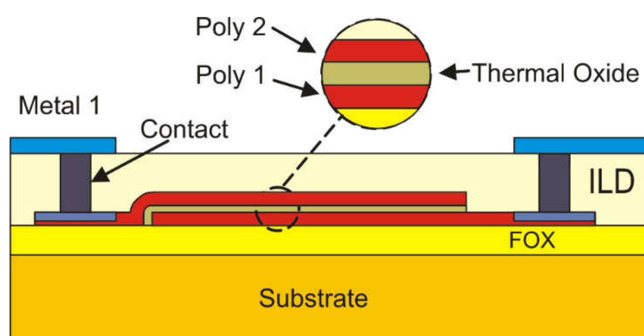

Fig.2.8. Cross-section of a poly-poly capacitor.

When two poly layers are not available, it is possible to create capacitance with a similar high capacitance density (cap. per unit area) using the single poly layer and an n+ diffusion created in the substrate. The structure is shown in Fig.2.9. The dielectric is just the gate oxide of the MOSFETs. The n+ diffusion (indicated with cap-implant in the figure) cannot be obtained using the n-plus layer of the n-MOSFET source and drains, since the n-plus is masked by the poly layer (see the CMOS process flow). Thus, a special diffusion (additional process step and photomask) is required to dope the active area before gate oxide growth and poly deposition.

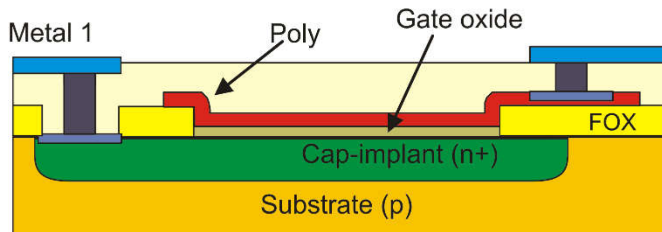

Fig.2.9. Poly-diffusion capacitor.

The main drawback of the poly-diffusion capacitor with respect to the poly-poly capacitors is the parasitic diode between the bottom plate (cap-implant) and the substrate. This diode introduce a leakage current (diode reverse saturation current) and a voltage dependent parasitic capacitance. The bottom plate should be connected only to nodes that are not sensitive to these problems. The top plate (polysilicon) is not affected by significant parasitic components, since it is insulated from the bottom plate by the gate oxide.

Junction capacitors.

These devices consists of a reverse-biased junction. A possible layout is shown in Fig.2.10. The bottom plate is an n-Well, while the top plate is a p+ diffusion. In order to keep the junction reverse biased, the polarity indicated in the figure should be respected. Junction capacitors are strongly nonlinear (voltage dependent) and a large parasitic diode is present from the bottom-plate to the substrate. These devices are used for frequency compensation purposes when no other capacitor types are available or when the dependence on bias voltage is desirable, as in voltage-controlled capacitors (varactors or varicap diodes) for tuning of resonant circuit.

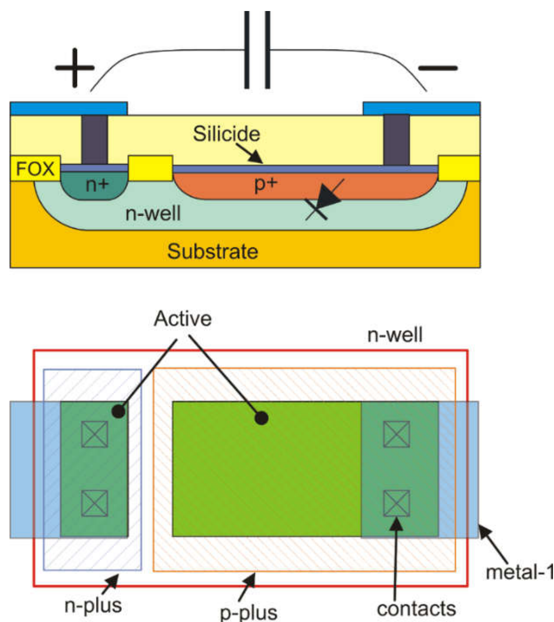

Fig.2.10. Junction capacitor

The properties of frequently used integrated capacitors are summarized in Table 2.3.

Table 2.3. Typical parameters of integrated capacitors

Capacitor Type	Capacitance per unit area	Linearity (voltage dependence)	Parasitic components
MIM	1 fF /μm^2	< 100 ppm/ V	Capacitance to substrate (bottom plate only)
MOM (fkux) 1 metal layer	0.1 fF /μm^2	< 100 ppm/ V	Capacitance to substrate (both terminals only)
MOM (flux) 6 metal layer	1 fF /μm^2	< 100 ppm/ V	Capacitance to substrate (both terminals only)
poly-poly	6 fF /μm^2	100 – 1000 ppm/V	Capacitance to substrate (bottom plate only)
poly-diffusion	6 fF /μm^2	100 – 1000 ppm/V	Diode to Substrate (bottom plate only)
junction	1 fF /μm^2	very –high (up to 30 % / V)	Diode to Substrate (bottom plate). Diode between terminals.

2.3 Integrated inductors

Inductors are the type of passive device that finds more difficulties to be integrated. The main problem is the small inductance values that can be obtained. In practice, due to the small available areas, only inductance values of a few nH can be obtained with integrated inductors. For an inductor to be useful, the magnitude of its impedance should be much greater than interconnect resistances. Since the magnitude of an inductor of inductance L at frequency f is given by $2\pi fL$, in order to have impedances of the order of a few ohm with L in the nano-Henry range, frequencies should be in the GHz range. As a result, integrated inductors can be used only for radio front-ends operating at very high frequencies. The layout of an integrated spiral inductor is shown in Fig.2.11. Notice that at least two metal levels are required. Another problem of integrated inductor is the close distance to the conductive substrate, which favors the development of induced currents. These currents dissipate power, reducing the quality factor of the integrated inductors.

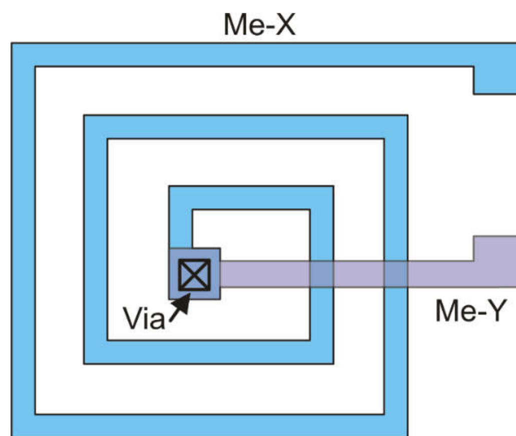

Fig. 2.11. Integrated inductor

3 Active device models and layouts

3.1 MOSFET layouts

Layout description

The layout of a planar MOSFET including drain and source contacts is shown in the bottom-left corner of Fig.3.1. The longitudinal (AA’) and transverse (BB’) cross sections of the device are shown on the top and bottom-left corners, respectively. These pictures are representative of both the p and n devices. For this reason, the type of doping is not specified. Clearly, doping of the drain/source regions are of opposite type with respect of the body region.

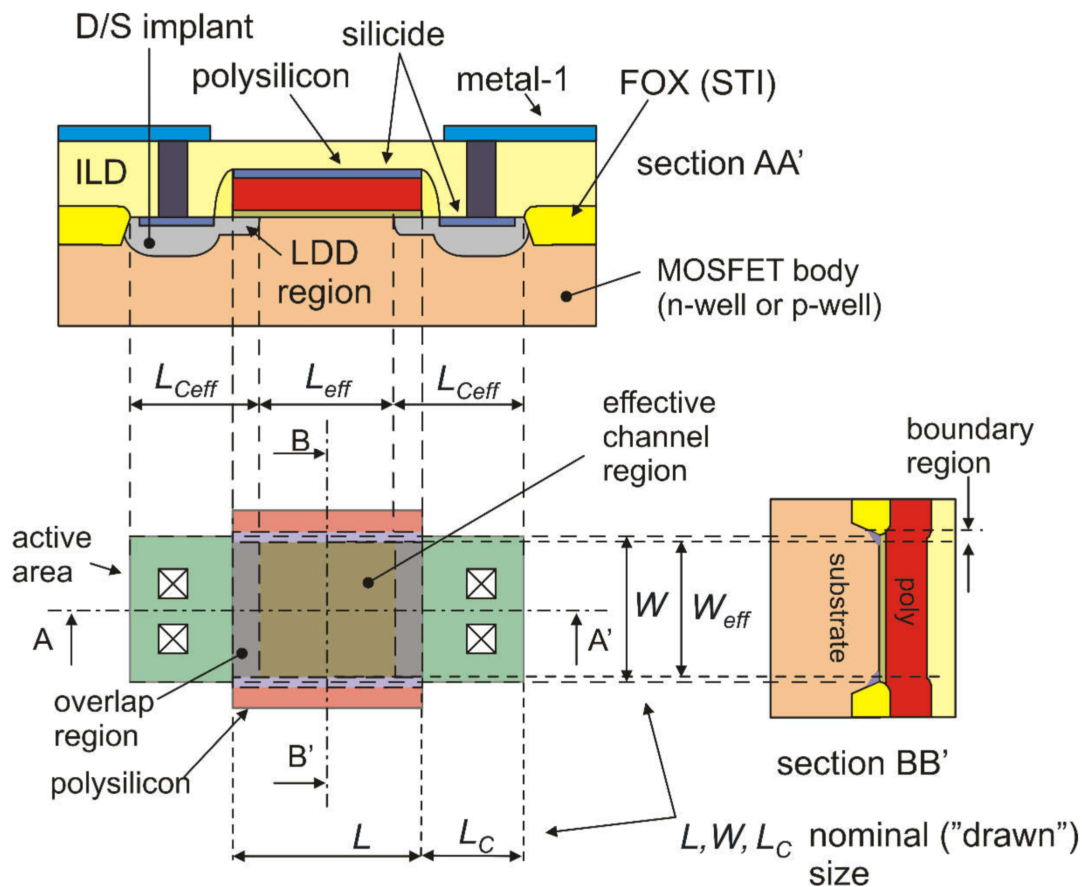

Fig. 3.1. Planar MOSFET layout with longitudinal (AA’) and transverse (BB’) cross-sections. Metal 1 and D/S implant layers are not shown in the layout for simplicity.

Ideally, a MOSFET is formed when an active area is crossed by a polysilicon line. The width of the polysilicon line determine the channel length (L), while the width of the active area is the channel width (W). The source and drain areas coincides with the two portions of the active area that are not covered by polysilicon and face each other across the channel. The drain/source regions are nominally rectangles of L_C size along the longitudinal direction and W size along the transverse one. Dimension L_C can be indicated as drain/source length. Note that, generally, L_C is set to the minimum value allowed by the design rules, in order to minimize parasitic junction capacitances. The drain/source areas are filled with the maximum number of contacts allowed by the design rules in order to reduce the drain/source series resistance.

In a real device, extension of the drain-source doped regions under the gate reduces the effective channel length (L_{eff}) with respect to the nominal (drawn) value (L). For the same reason, the actual drain/source lengths (L_{Ceff}) are longer than L_C. Note that the drain/source doped areas extends under the gate mainly with their LDD (Light Doping) portions.

Similarly, the effective width of the MOSFET (W_{eff}) is slightly different from the drawn geometry (W). The reason is less intuitive than for the channel length reduction. The effect is caused by the presence of boundary regions located at the interface of the active area with the field oxide. In these points, the distance between the polysilicon gate and the MOSFET body gradually increases as the polysilicon line goes out of the active area. As a result, there are channel portions (within the drawn width W) where inversion is less effective (smaller density of mobile charge) and zones, beyond the nominal width, where the gate induces depletion charges into the substrate (fixed charge). The presence of these boundary regions generally result in reduction of the effective width ($W_{\rm eff} < W$).

Generally, it is not necessary for the layout designer to draw all the masks required to complete the MOSFET fabrication. For example, in many process PDK, the spacer oxide and the LDD doping is automatically drawn when the polysilicon, active area and drain / source doping layers are drawn. Simplified layout for the n-MOSFET and p-MOSFET are shown in Fig. 3.2 and Fig.3.3, respectively.

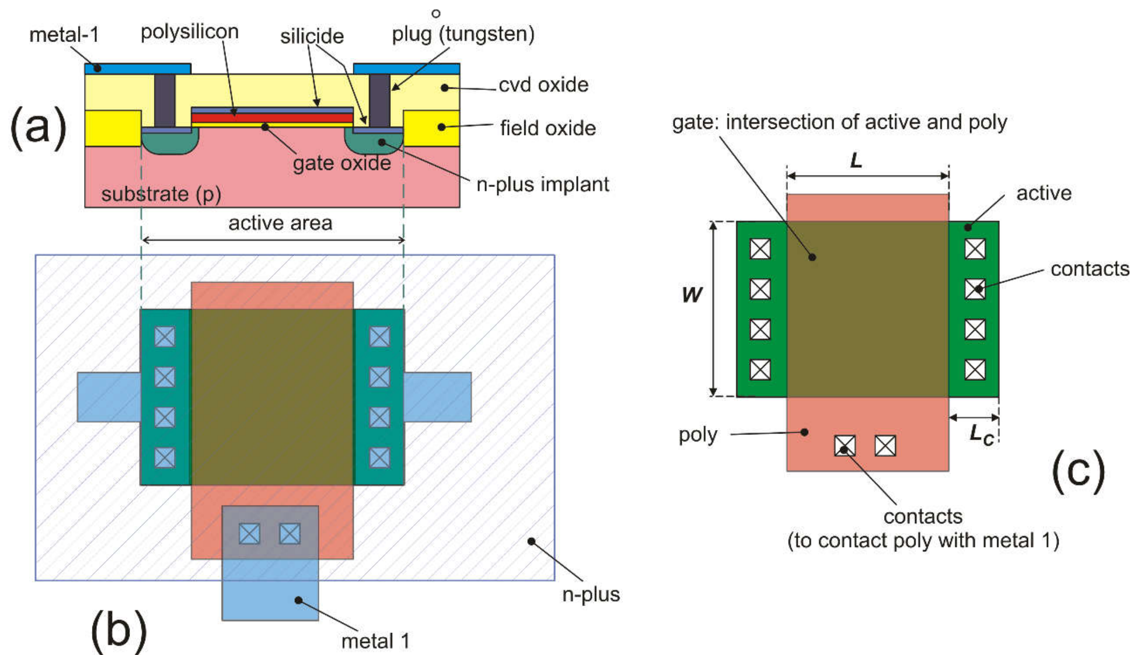

Fig.3.2. Simplified cross-section (a) and layout (b) of an n-MOSFET. An extract of the layout showing only active, poly and contact layer is proposed in (c), with indication of the most relevant lengths.

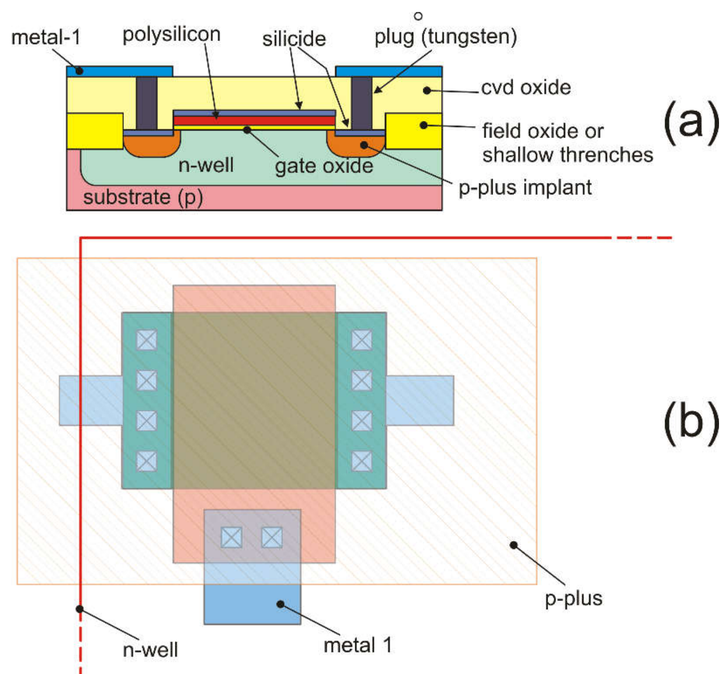

Fig.3.3. Simplified cross-section (a) and layout (b) of a p-MOSFET. The difference with respect to the n-MOSFET layout is the p-plus doping (in place of n-plus) and enclosure into an n-Well.

Giving the designer the minimum number of layers that are necessary to define a standard layout is not the only possible choice. There are foundries that give more control to the designer, including also layers that refer to process steps that can be generated automatically.

In a PDK oriented to analog design, MOSFETs are placed into the layout by creating an instance of the corresponding p-cell, which coincides with the layout view of the device. Through a graphical interface, it is possible to choose the W and L of the device being placed. This procedure generates all required layers, including contacts. The role of the layout designer is then arranging the devices inside the assigned area and making all connections.

Designer options

Processes may offer different MOSFET families as an option: for example, low threshold devices, or devices with increased oxide thickness for handling higher logical levels may be available as an alternative to the regular complementary (n and p) devices. For any MOSFETs that the designer places into the circuit, the width (W) and length (L) should be specified. Another degree of freedom that is under control of the designer is the type of layout. For example, instead of the simple layouts shown in Fig.3.2 and 3.3, it is possible to use fingered layouts, which are particularly convenient for MOSFETs with large W values.

3.2 Mosfet models

In this part, we will consider an n-MOSFET (enhancement type), since most quantities (currents and voltages) are positive for this type of device. Type-p devices will be briefly reviewed at the end of this document. The large signal model of an n-MOSFET is shown in Fig.3.4. The core of the device is the I_{DS} controlled source. Resistors R_S and R_D, connect the edge of the channel, S' and D', to the actual source and drain contacts, S and D, respectively. Diodes D_{BS} and D_{BD} represent the junctions that isolate the source and drain from the body (B). The various capacitances are generally characterized by a non-linear relationship between voltage and charge. In the following part of this document, the series resistance will be neglected for simplicity. Then, we will consider that S≡S' and D≡D'.

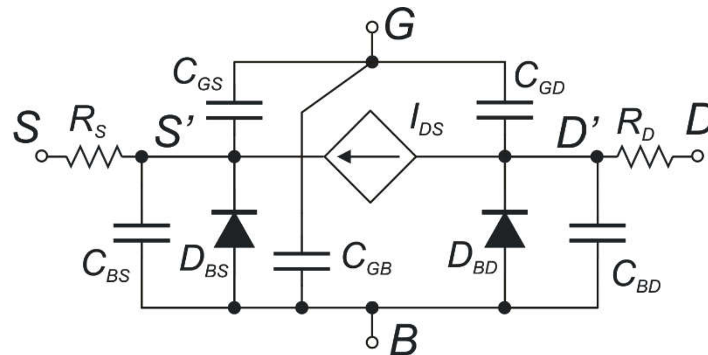

Fig. 3.4. Large signal equivalent model of an n-MOSFET

Control voltages and operating regions.

The drain-current I_{DS} depends on the voltages across the MOSFET terminals. There are two possible ways to define these voltages: source-referred or body-referred. Source referred potentials are more commonly used and consists of V_{GS}, V_{BS} and V_{DS}. Body-referred voltages are V_{GB}, V_{SB} and V_{DB}. Obviously, the two approaches are equivalent, since it is always possible to transform body-referred potentials into source referred ones and vice versa. However, source-referred potentials are simpler and more intuitive. The only real drawback of source-referred potentials is that, for a symmetrical MOSFET, it is necessary to determine which terminal is working as the source before writing down the current equations. We have to consider the voltages of the two terminals that are candidate to be the source or the drain and apply the following rule:

- in an n-channel device, the source is the terminal that has the lower potential;
- in a p-channel device, the source is the terminal that has the higher potential.

The voltage V_{BS} should be such that both the body-source and body-drain diodes do not carry a significant current. The ideal situation is to make sure that both diodes are reverse-biased or, at least zero-biased. For an n-MOSFET, this means:

$$V_{BS} \leq 0 \Rightarrow V_S > V_B \tag{3.1}$$

Voltage V_{BS} affects the threshold voltage V_t through the body effect [1]:

$$V_t = V_{t0} + \gamma\left(\sqrt{\phi_s - V_{BS}} - \sqrt{\phi_s}\right) \tag{3.2}$$

where V_{t0} is the threshold voltage for V_{BS}=0, whereas γ and ϕ_s are equal to:

$$\phi_s = 2\psi_B = 2\frac{k_B T}{q}\ln\left(\frac{N_A}{n_i}\right); \qquad \gamma = \frac{\sqrt{2q\varepsilon_{Si}N_A}}{C_{ox}} \tag{3.3}$$

and: k_B=Boltzmann constant, T=absolute temperature, q = electron charge, N_A= dopant concentration of the body, e_{Si} the silicon permittivity and C_{ox} the gate capacitance-per-unit area.

It is useful to calculate the sensitivity of V_t with respect to V_{BS}. Due to (3.1), it is convenient to express the sensitivity with respect to $V_{SB}= -V_{BS}$, obtaining [2]

$$\frac{dV_t}{dV_{SB}} = m-1; \quad \text{with} \qquad m = 1 + \frac{C_{dm}}{C_{ox}} \tag{3.4}$$

where m is the so-called subthreshold slope factor and C_{dm} is the depletion layer capacitance given by:

$$C_{dm} = \sqrt{\frac{q\varepsilon_{Si}N_A}{2\left(2\phi_f + V_{SB}\right)}} \tag{3.5}$$

The value assumed by m for V_{BS}=0 varies from 1.2 to 1.3, so that dV_t/dV_{SB} is in the range 0.2-0.3. The threshold voltage to body bias sensitivity decreases for increasing reverse voltages.

As far as V_{GS} and V_{DS} are concerned, six regions can be roughly distinguished, as shown in Table 3.1, where V_T is k_BT/q. Note that V_{GS} appears through the quantity (V_{GS}-V_t) which is called "overdrive voltage"

Table 3.1. MOSFET operating regions on the basis of V_{GS} and V_{DS}.

	$V_{GS} - V_t \le 0$	$0 \le V_{GS} - V_t \le 4V_T$	$V_{GS} - V_t \ge 4V_T$
$V_{DS} \le V_{DSAT}$	Triode – Weak Inversion	Triode – Moderate Inversion	Triode – Strong Inversion
$V_{DS} \ge V_{DSAT}$	Saturation – Weak Inversion	Saturation – Moderate Inversion	Saturation – Strong Inversion

Independently of the condition of strong or weak inversion, saturation region is characterized by reduced dependence of the I_{DS} on the V_{DS}. On the contrary, in triode region, the drain current shows a strong dependence on V_{DS}. For very small V_{DS} values ($V_{DS} \ll V_{DSAT}$) the I_{DS} vs V_{DS} dependence is linear.

Strong inversion is a region where the inversion layer is well formed and the depletion layer in the MOSFET body is not affected by V_{GS}. In strong inversion, the dependence of I_{DS} on V_{GS} and V_{DS} can be approximated by square laws (MOSFET parabolic equations). Weak inversion, also indicated with "subthreshold region", is marked by exponential dependence of I_{DS} vs V_{GS}. Moderate inversion is the transition region between weak and strong inversion.

For $V_{GS} << V_t$, I_{DS} becomes so small that the MOSFET can be considered turned off (off state). This occurs when the I_{DS} current becomes of the same order of the junction leakage currents.

Drain current equations in strong inversion

In strong inversion, the usual equations used in triode and saturation regions are:

$$I_{DS} = \beta_n \left(V_{GS} - V_t - \frac{V_{DS}}{2} \right) V_{DS} \qquad \text{(Triode)} \tag{3.6}$$

$$I_{DS} = \beta_n \frac{\left(V_{GS} - V_t\right)^2}{2} \left[1 + \lambda\left(V_{DS} - V_{DSAT}\right)\right] \qquad \text{(Saturation)} \tag{3.7}$$

where:

$$\beta_n = \mu_n C_{OX} \frac{W_{eff}}{L_{eff}} . \quad W_{eff}=W-2W_D, \quad L_{eff}=L-2L_D \, , \tag{3.8}$$

and μ_n is the electron mobility in the channel. Parameters W_D and L_D are the reduction that the channel width and length, respectively, suffer from each side of the gate with respect to the corresponding drawn geometries W and L.

Lambda (λ) is an important parameter that determine the dependence of I_{DS} on V_{DS} in saturation region. This dependence is due to multiple phenomenon, such as channel length modulation or drain-induced barrier lowering (DIBL). The ideal situation would be λ=0, i.e. no dependence on V_{DS}. In practice, it is difficult to obtain $\lambda < 0.01\ V^{-1}$. The most important parameter that affects λ is the channel length. As a first order approximation it possible to express the inverse of lambda on channel length with a linear relationship, valid for lengths somewhat larger than the minimum L of the process:

$$\lambda^{-1} \cong k_\lambda L_{eff} \tag{3.9}$$

where k_λ (V/µm) is a constant.

Drain current in weak inversion

Weak inversion is also indicated as sub-threshold region. Strictly speaking, sub-threshold region would require that $V_{GS}-V_t<0$. In practice, the terms weak inversion and sub-threshold are both used to indicate a region where the I_{DS} dependence on both V_{GS} and V_{DS} becomes exponential, according to the equation:

$$I_{DS} = I_{SM} e^{\frac{V_{GS}-V_t}{mV_T}} \left(1 - e^{\frac{-V_{DS}}{V_T}} \right) \left[1 + \lambda\left(V_{DS} - V_{DSAT}\right)\right] \tag{3.10}$$

where I_{SM} is given by:

$$I_{SM} = \mu_n C_{dm} \frac{W_{eff}}{L_{eff}} V_T^2 = \mu_n C_{ox} (m-1) V_T^2 \frac{W_{eff}}{L_{eff}} \tag{3.11}$$

When V_{DS} is high enough, the exponential term exp($-V_{DS}/V_T$) can be neglected with respect to one and the drain current shows a reduced dependence on V_{DS} (limited to the λV_{DS} term as in strong inversion. In this condition, the MOSFET is in saturation and (3.10) reduces to:

$$I_{DS} = I_{SM} e^{\frac{V_{GS}-V_t}{mV_T}} \left[1 + \lambda \left(V_{DS} - V_{DSAT}\right)\right] \tag{3.12}$$

For low V_{DS} value, the exponential term in the round parentheses is no more negligible and the current begins to show a strong dependence on V_{DS}. This is the analogue of the triode region defined for the strong inversion.

Moderate inversion

In moderate inversion, I_{DS} dependence progressively changes from parabolic to exponential. Simple equations for this region do not exist. However, the EKV (Enz-Krummenacher-Vittoz) model [3] consists of a single I_{DS} equation for all three regions (weak-moderate-strong inversion). This equation is rather complex for hand calculation and require the voltage to be expressed with the body-referred method.

Saturation voltage, V_{DSAT}

The saturation voltage V_{DSAT} can be approximated by the following expressions:

$$V_{DSAT} \cong \begin{cases} (V_{GS} - V_t) & \text{in strong inversion} \\ 4V_T \ \ (100 \ \text{mV}) & \text{in moderate and weak inversion} \end{cases} \tag{3.13}$$

These, are empirical expressions that set a lower limit to V_{DS} for having a small dependence of I_{DS} on V_{DS}. Definitions that are more related to physical phenomena can be also used. For example, textbooks on electron devices [4] often propose the expression (V_{GS}-V_t)/m for V_{DSAT} in strong inversion. Since m is slightly greater than one, using this definition lead to a lower V_{DSAT} value. On the other hand, the definitions given in (3.13) are simpler to use for design purposes and give a good representation of the experimental MOSFET behavior.

Junction currents

The source-body and drain-body junctions are normally reverse biased. As a result, these junctions, represented by the D_{BS} abd D_{BD} diodes in Fig.3.4, carry only the inverse saturation current, which is given by:

$$I_J = A_J J_S \tag{3.14}$$

where I_J is the current flowing through the junction, A_J is the area of the junction (source or drain area) and J_S the inverse saturation current density. Considering the layout of Fig.3.4, A_D and A_S (drain a source areas, respectively) are both given by $W \cdot L_C$. Typical values of J_S at room temperature are of the order of 0.1 fA/μm^2.

Temperature effects

The effect of temperature on the MOSFET dc characteristics can be represented by the temperature dependence of β_n and V_t. Temperature affect b through the mobility, which generally decreases with temperature. The following equation can be used to represent the temperature dependence of β_n:

$$\beta_n(T) = \beta_n(T_0)\left(\frac{T}{T_0}\right)^{-\alpha_\mu} \tag{3.15}$$

where T_0 is a reference temperature (for example 300 K) and α_μ a process-dependent constant that can vary in the 1.2-2 interval.

The temperature dependence of the threshold voltage is often approximated with a linear expression:

$$V_t(T) = V_t(T_0) - \alpha_{VT}\left(T - T_0\right) \tag{3.16}$$

where α_{VT} is of the order of 1 mV/K. As temperature increases, both V_t and β_n decreases producing opposite effects on I_{DS}: the V_t reduction increases I_{DS} while the β_n reduction causes a proportional decrease of I_{DS}. These effects generally do not compensate each other:

- At low V_{GS}-V_t, the threshold voltage dominates and the current increases with temperature.
- At high V_{GS}-V_t, it is β_n to dominate and the current increases with temperature.

Compensation of the two effects occurs only for a particular value of V_{GS}-V_t., which typically falls well into the strong inversion range. The presence of an operating point with low temperature sensitivity (ZTC, zero temperature coefficient) can be used for reference voltage generation [5].

The behavior of p-channel devices is similar to n-channel ones. Expression (3.15) can be used also for $\beta_p=\mu_p C_{ox} W_{eff}/L_{eff}$, while (3.16) is applicable when the absolute value of p-channel threshold voltage is considered. Then, for p-channel devices, the absolute value of V_t decreases with temperature, increasing the magnitude of I_{DS}, just as for n-channel devices [5].

As far as leakage is concerned, it should be observed that the junction saturation currents of all junctions (e.g. body-drain, body-source junctions) roughly doubles for every temperature increment of 10 °C. This means that these currents are multiplied by about a factor of 1000 for a temperature increment of 100 °C.

Small signal model and parameters

The small signal model of the MOSFET is shown in Fig. 3.5. Notice that, for simplicity, the series resistances R_S and R_D have been neglected in the small signal circuit. The reduced circuit for dc signals is obtained from the circuit of Fig. 3.5 by removing all parasitic capacitors. The result is shown in Fig.3.6.

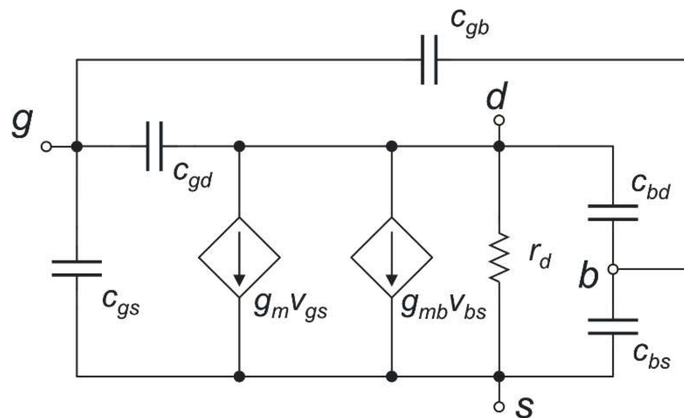

Fig.3.5. MOSFET small signal model.

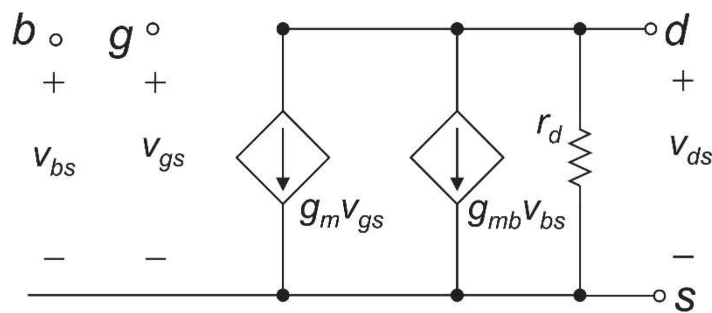

Fig.3.6. MOSFET small signal model for dc signals.

Let us start from the dc equivalent circuit. It is important to study how the parameters vary with the operating point. We will focus on g_m and r_d, since g_{mb} is related to g_m by the following relationship:

$$g_{mb} = (m-1)g_m \quad \cong \quad 0.2 g_m \tag{3.17}$$

which is a direct consequence of (3.4).

It is interesting to consider the following two aspects:

- How g_m and r_d varies as a function of V_{DS} for a fixed overdrive voltage (V_{GS}-V_t).
- How g_m varies in saturation region as a function of V_{GS}-V_t.

The first point will be analyzed considering (V_{GS}-V_t)>4V_t, i.e. strong inversion. Then, considering (3.6), we can calculate the values of g_m and g_{ds}=1/r_d in triode region:

$$g_m \equiv \left(\frac{\partial I_{DS}}{\partial V_{GS}}\right)_{V_{DS},V_{BS}=const} = \beta_n V_{DS} \tag{3.18}$$

$$\frac{1}{r_d} = g_{ds} \equiv \left(\frac{\partial I_{DS}}{\partial V_{DS}} \right)_{V_{GS}, V_{BS} = const} = \beta_n \left(V_{GS} - V_t - V_{DS} \right) \tag{3.19}$$

In <u>saturation</u> region, the same parameters can be find using (3.7), and neglecting the dependence of V_{DSAT} on V_{GS}. :

$$g_m \equiv \left(\frac{\partial I_{DS}}{\partial V_{GS}} \right)_{V_{DS}, V_{BS} = const} = \beta_n \left(V_{GS} - V_t \right) \left[1 + \lambda \left(V_{DS} - V_{DSAT} \right) \right] \cong \beta_n \left(V_{GS} - V_t \right) \tag{3.20}$$

$$\frac{1}{r_d} = g_{ds} \equiv \left(\frac{\partial I_{DS}}{\partial V_{DS}} \right)_{V_{GS}, V_{BS} = const} = \lambda \frac{\beta_n}{2} \left(V_{GS} - V_t \right)^2 \cong \lambda I_{DS} \tag{3.21}$$

For $V_{DS}=V_{DSAT}=V_{GS}-V_t$, triode expression (3.19) yields $g_{ds}=0$, while, for the same V_{DS}, the saturation formula (3.21) gives $g_{ds}=\lambda I_D$. The reason is the discontinuity (of both I_{DS} and its first derivative) obtained for $V_{DS}=V_{DSAT}$ using (3.6) and (3.7). Assuming that the correct value of g_{ds} for $V_{DS}=V_{DSAT}$ is given by the saturation equations, the simplified behavior of gm and g_{ds} across both triode and saturation region are represented in Fig.3.7.

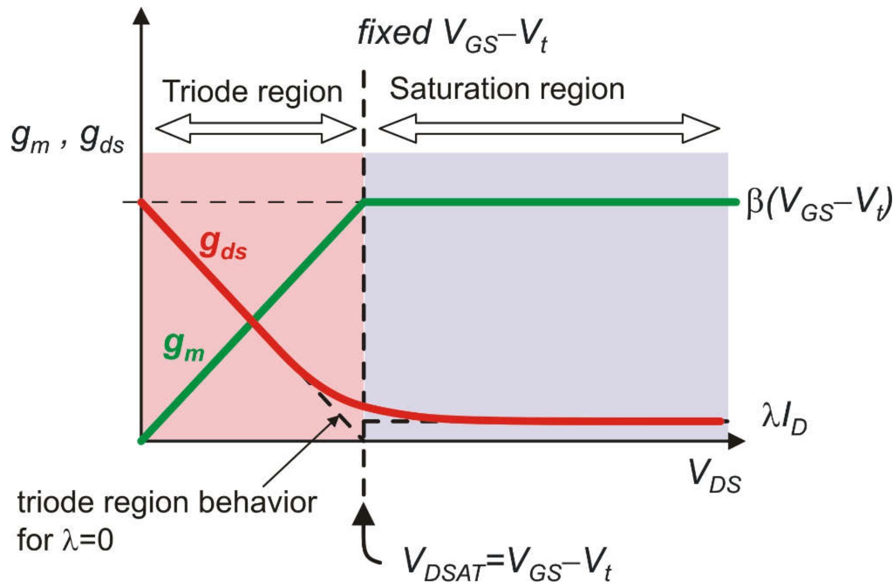

Fig. 3.7. Simplified plots of g_m and g_{ds} as a function of V_{DS} for constant $V_{GS}-V_t$, in strong inversion.

A similar behavior can be derived for $V_{GS}-V_t$ values that set the device in weak or moderate inversion, but with different values for g_m, g_{ds} and V_{DSAT}.

In all cases, it is important to observe that in saturation both g_m and $r_d=1/g_{ds}$ assume their maximum value in saturation. When V_{DS} gets lower than V_{DSAT} and the device gets into triode region, both g_m and r_d start

decreasing progressively, down to their minimum values that are reached for V_{DS}=0. In particular, for V_{DS}=0 g_m is zero.

Transconductance models in saturation region

The second point can be analyzed considering three different equivalent expressions for g_m in strong inversion:

$$g_m = \beta_n (V_{GS} - V_t) \tag{3.22}$$

$$g_m = \sqrt{2\beta_n I_D} \tag{3.23}$$

$$g_m = \frac{2I_D}{(V_{GS} - V_t)} \tag{3.24}$$

The first one simply coincides with (3.20), The second one can be obtained from (3.22) considering that, neglecting the λV_{DS} term in (3.7), then the overdrive voltage in strong inversion is given by:

$$V_{GS} - V_t = \sqrt{\frac{2I_D}{\beta_n}} \tag{3.25}$$

The third expression, (3.24), can be obtained considering that, neglecting λV_{DS} in (3.7), $V_{GS}-V_t$ is equal to $2I_D/\beta_n(V_{GS}-V_t)$.

In weak inversion, using (3.12), it is possible to find the following expression for g_m:

$$g_m = \frac{I_D}{mV_T} \tag{3.26}$$

Considering that for a BJT $g_m=I_C/V_T$, it is possible to use the following g_m expression for a MOSFET in strong, moderate and weak inversion and extend it to the BJT:

$$g_m = \frac{I_D}{V_{TE}} \tag{3.27}$$

where V_{TE} (equivalent V_T) is a voltage, defined just by equation (3.27), which assumes the following value:

$$V_{TE} = \begin{cases} (V_{GS} - V_t)/2 & \text{MOSFET in strong inversion} \\ mV_T & \text{MOSFET in weak inversion} \\ V_T & \text{BJT} \end{cases} \quad (3.28)$$

The typical behavior of V_{TE} is represented in Fig.3.8

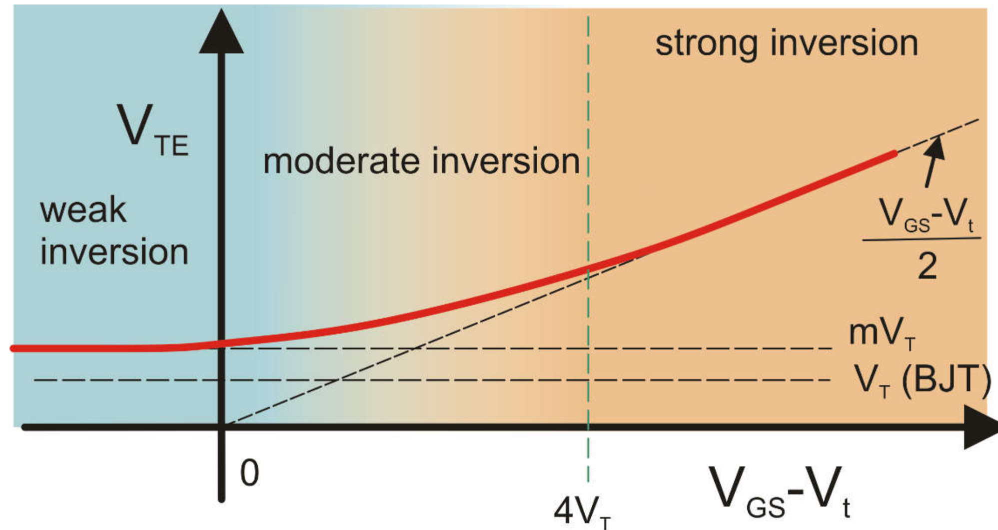

Fig. 3.8. Behavior of the V_{TE} parameter as a function of V_{GS}-V_t, compared with the BJT value.

A parameter that is sometimes quoted in scientific articles on analog design [6] is g_m/I_D, equal to $1/V_{TE}$. This parameter is sometimes referred to as "transconductance efficiency", since a high g_m/I_D value means that it is possible to obtain large transconductances with relatively small I_D. Notice that it is often necessary to obtain a large g_m from selected devices in a circuit, in order to achieve low noise and/or wide bandwidth performances. With high values of g_m/I_D it is possible to meet noise and bandwidth specifications with lower current consumption. Considering that V_{TE} is the inverse of g_m/I_D, Fig.3.8 means that, for a MOSFET, the best efficiency is found at the lowest overdrive voltages, ideally in weak inversion. Due to the factor m>1, a BJT achieves a V_{TE} lower than a MOSFET in weak inversion. Then BJT are superior to MOSFETs in this respect.

Capacitance models

Parasitic capacitances shown in the circuit of Fig. 3.5 are the small signal equivalent of non-linear capacitors shown in Fig. 3.4.

Each one of the capacitances that involve the gate, (c_{gs}, c_{gd}, c_{gb}) can be divided into two components, named "overlap" and "intrinsic" capacitances, according to:

$$\begin{aligned} c_{gs} &= c_{gs}^{(ov)} + c_{gs}^{(i)} \\ c_{gd} &= c_{gd}^{(ov)} + c_{gd}^{(i)} \\ c_{gb} &= c_{gb}^{(ov)} + c_{gb}^{(i)} \end{aligned} \quad (3.29)$$

where the apex (ov) stands for "overlap" and (i) for "intrinsic". Overlap capacitances derive from the gate partial superimposition on the other electrodes. For example, examining Fig.3.1, it is possible to recognize the gate-source and gate-drain overlap areas. Gate-body overlap is due to the boundary regions, where the gate layer (polysilicon in Fig.3.1) progressively gets away from the substrate. In these regions, the gate is still close enough to the substrate to create a significant capacitance, but not so close to be able to invert the substrate surface and create the channel. Overlap capacitances are practically independent of the state of the channel (i.e. depletion, inversion, etc.) and do not depend on the MOSFET terminal voltage. Their value is given by:

$$\text{overlap capacitances:} \begin{cases} c_{gs}^{(ov)} = c_{gso} \cdot W \\ c_{gd}^{(ov)} = c_{gdo} \cdot W \\ c_{gb}^{(ov)} = c_{gbo} \cdot L \end{cases} \tag{3.30}$$

where c_{gso}, c_{gdo}, c_{gbo} are constant capacitance-per-unit length coefficients (unity: F/m). Inspection of Fig.3.1 shows that the gate-source and gate-drain overlap capacitances are located along the gate width , whereas the gate-body capacitance is distributed along the gate length. For this reason the overlap components of c_{gs}, c_{gd}, are proportional to W, while the gate-body one is proportional to L.

The intrinsic capacitances are strictly related to the charge accumulation in the channel. These capacitances are strongly non-linear. Simplified expressions, which are often used for hand calculations consist in the so-called Meyer model. According to this model the first order approximation of the three intrinsic capacitances in off-region ($V_{GS}<<V_t$), triode and saturation regions are given in table 3.2.

Table 3.2. Meyer Model for the mosfet intrinsic capacitances

	Off –state	Triode	Saturation
$c_{gs}^{(i)}$	0	$\frac{1}{2}C_{ox}WL$	$\frac{2}{3}C_{ox}WL$
$c_{gd}^{(i)}$	0	$\frac{1}{2}C_{ox}WL$	0
$c_{gb}^{(i)}$	$\left(\frac{1}{C_{ox}WL}+\frac{1}{C_{dm}}\right)^{-1}$	0	0

A major drawback of the Meyer model is that it does not respect charge conservation. Such a model fails to represent transient phenomena where a MOSFET crosses different operating region, as occurs with logical gates or switches (pass – transistors). A more accurate description requires that the intrinsic capacitances be replaced by "capacitance coefficients" generically defined by:

$$c_{ij} \equiv \frac{\partial Q_i}{\partial V_j} \tag{3.31}$$

where Q_i is the charge accumulated on terminal "i" and V_j is the voltage of terminal "j". The terminal taken into consideration are drain, gate and source. The body is used as a reference for the voltage of the other electrodes. Then we have three self-capacitance coefficients (c_{dd}, c_{gg}, c_{ss}) which are all positive, and six mutual capacitance coefficients (c_{gd}, c_{dg}, c_{gs}, c_{sg}, c_{sd}, c_{ds}), which are negative, since the charge displaced on an electrode by a positive voltage change applied to a different electrode is negative (as happens even in an ideal parallel plate capacitor). Capacitances c_{sd} and c_{ds} are generally negligible. Finally, it should be observed that, generally, c_{ij} is different from c_{ji}, due to the nonlinear behavior of the MOSFET. As a result, the mutual charge induction between the gate and the other two electrode cannot be represented by a simple capacitance, but two distinct coefficients are required. This model is called charge-oriented model and was introduced by Dutton and Ward [7]. All modern simulation model adopt the charge-oriented model for the intrinsic capacitances.

Finally, c_{bd} and c_{ds} are junction capacitances, marked by a strong dependence on voltage. Normally, the drain-body and source-body junctions are reverse biased. The larger the (reverse) bias, the smaller the capacitance. Commonly used expressions for these capacitances are the following:

$$c_{bs} = \frac{C_J A_S}{\left(1+\frac{V_{SB}}{V_0}\right)^{m_j}} + \frac{C_{JSW} P_S}{\left(1+\frac{V_{SB}}{V_0}\right)^{m_{jsw}}} \tag{3.32}$$

$$c_{bd} = \frac{C_J A_D}{\left(1+\frac{V_{SB}}{V_0}\right)^{m_j}} + \frac{C_{JSW} P_D}{\left(1+\frac{V_{SB}}{V_0}\right)^{m_{jsw}}} \tag{3.33}$$

where: A_S=source area, P_S= source perimeter, A_D=drain area, P_D= drain perimeter. Parameters C_J and C_{JSW} are the (zero-bias) capacitances per unit area and unit perimeter, respectively. V_0 is the built-in potential of the junction whereas m_j and m_{jsw} exponents, called "grading coefficients", depend on the doping profiles across the bottom and sidewalls of the source/drain implants, respectively. These exponents are generally in the range 0.33-0.5.

The overlap capacitance and the junction capacitances are classified as extrinsic capacitances.

Matching parameters

The most frequently used expressions for the standard deviation of the threshold voltage and transconductance factor (β) matching errors of two matched MOSFETs, are the following:

$$\sigma_{\frac{\Delta\beta}{\beta}} = \frac{C_\beta}{\sqrt{WL}}; \qquad \sigma_{Vt} = \frac{C_{Vt}}{\sqrt{WL}} \tag{3.34}$$

3.3 Bipolar Transistor Layouts.

Layout descriptions

Typical BJTs that are available in a standard bipolar or BiCMOS process are the vertical NPN and the lateral PNP transistors. Both devices are created inside a moderately n-doped region, which is isolated from the surrounding p-doped substrate. These isolated n-type areas are generally called “pockets”. In earlier pure bipolar process, the pockets were obtained from an n-doped epitaxial layer, grown onto a p-substrate. The epi-layer was divided into isolated pockets (epi-pockets) by means of isolation p+ implants, or trenches filled by silicon dioxide. Modern BiCMOS process are generally derived by simpler CMOS process. Therefore, the n-pockets are simply n-wells diffused through a p^- epitaxial layer. In order to reduce the resistivity of the pocket, without altering the superficial moderately doped region, a highly n-doped buried layer (or buried well) is created on the bottom of pocket. A possible cross-section and layout of a vertical NPN BJT built inside an n-well are shown in Fig. 3.9. The n-pocket coincides with the collector. Creation of the base region requires a special doping (p-base), which is normally not available in a standard CMOS process. The emitter can be obtained with an n+ implant, or by simply putting an n-doped polysilicon strip in contact with the p-base diffusion [2].

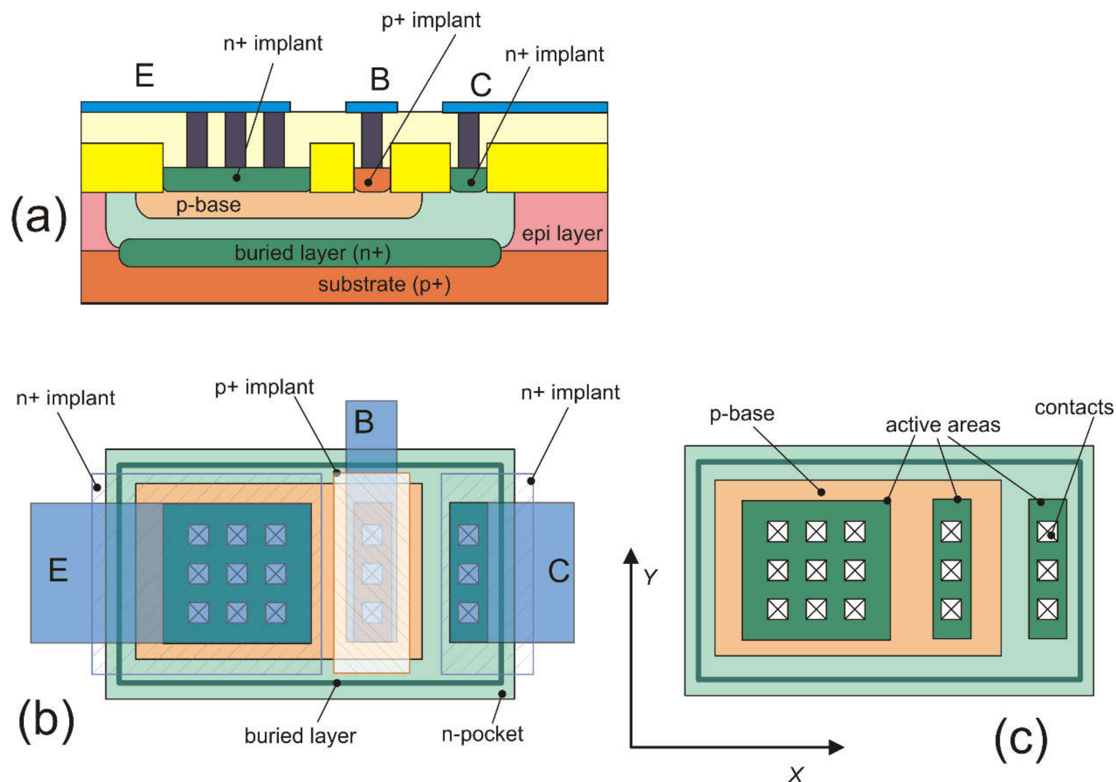

Fig. 3.9. Cross section (a) and Layout (b) of a vertical npn BJT. A layout excluding n-plus, p-plus and metal layers is shown in (c).

The cross-section and layout of a lateral PNP BJT are shown in Fig. 3.10 (a). The pocket is the base, while the emitter and collector are p-diffusions (p-base in the example of Fig.3.10). Transport of holes from the emitter to the collector occurs laterally, under control of the thin n-well (n-pocket) layer that separates them and acts as the base. In early Bipolar processes, lateral PNP BJTs had a very low beta,

due to the difficulty of reducing the base length. With modern photolithography resolutions, this no more a limiting factor and PNP transistors with beta of the order of one hundred or more can be easily obtained. One of the major limitations of lateral PNP transistors is the possibility to control the doping profiles of the emitter and collector independently.

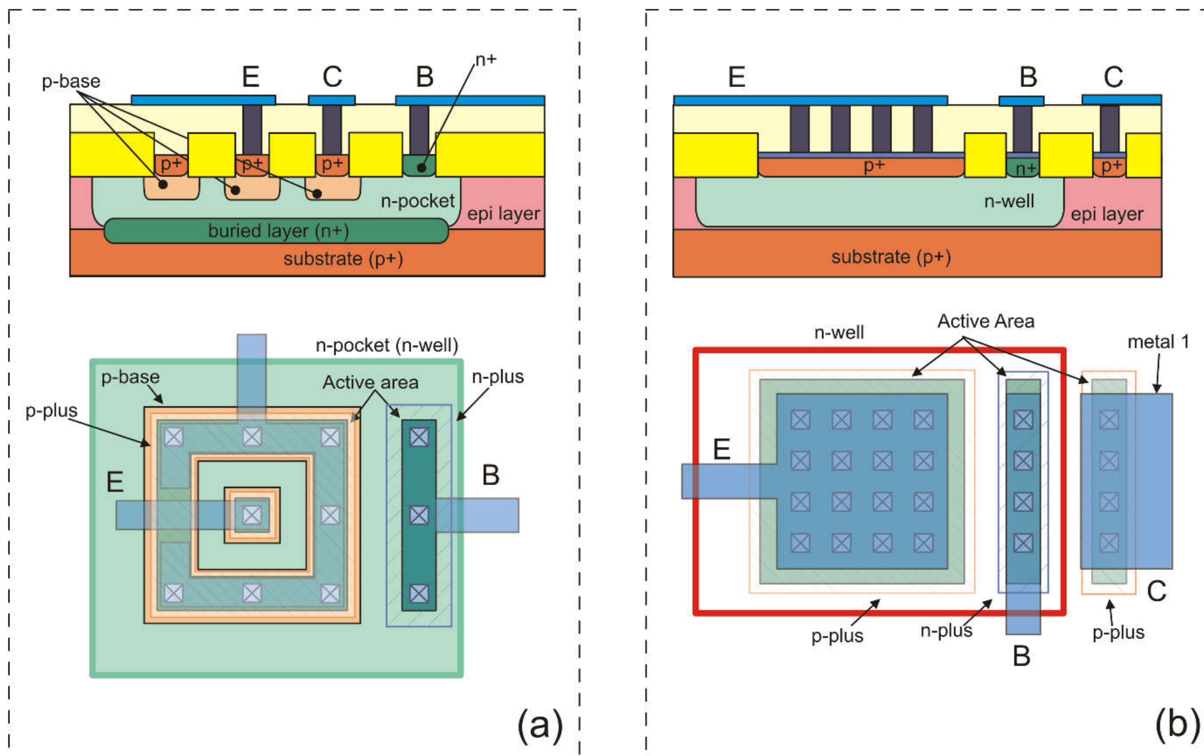

Fig.3.10. Cross-section and layout of a lateral BJT (a) and of a substrate PNP (b).

The high resistivity of the base (n-pocket) introduce a parasitic base resistance that adversely affect noise and high frequency performances. In spite of these limitations, lateral PNP transistors are widely used in general-purpose analog circuits, such as operational and instrumentation amplifiers. Bipolar analog circuit requiring NPN and PNP transistors with similar characteristics should be based on both NPN and PNP vertical transistors (Complementary Bipolar Processes).

A vertical PNP BJT, which is available even in standard n-Well CMOS technology with no need of additional masks or process steps, is shown in Fig.3.10 (b). The base is formed by an n-Well, while the emitter is a p+ region (doped active area) created into the n-Well. The emitter doping is introduced with the same step used to create the drain/source of p-MOSFETs. A major limitation of this kind of device is that the collector the substrate, thus it should be mandatorily fixed to the smallest supply voltage. Substrate BJTs in CMOS processes are typically used to build reference voltage circuit based on the Band-Gap principle.

Designer options for BJTs

Process design kits (PDK) generally give the designer less freedom to customize bipolar transistors with respect to MOSFEts. We have seen that the designer may assign both the width and length of a MOSFET within a very wide interval of values. The PDK offers a set of BJTs both npn and pnp, that differ for the

type of layout. These devices are called elementary transistors. The designer can place an elementary device directly into the circuit or personalize it through a dimensionless parameter called *area*. The area acts as a multiplier for many parameters of the elementary device. The saturation current (I_S) and parasitic capacitances are multiplied by the *area* factor, while the series resistances are divided by *area*. These transformations of parameters are taken into account by the electrical simulator. The parameter area must be greater or equal to one, meaning that we can modify an elementary transistor only making it bigger, not smaller. As far as the layout is concerned, the parameter *area* can be implemented in different ways. The simplest method, which is always available, is placing a number N of elementary devices in parallel. In this way *area*=N. The main limitation is that area must be an integer. A different approach that allows also factionary area values (but still *area* ≥ 1) is stretching the layout of the elementary transistor along a selected direction. For example, for the vertical transistor of Fig.3.9, it is possible to stretch the layout by a factor equal to *area* along the Y direction. This increases the effective emitter area, proportionally increasing I_S . Layout stretching is not possible for the lateral PNP of Fig.3.10 (a), thus the only option for this type of device is paralleling elementary transistors (integer areas only).

3.4 Bipolar transistor models

Large signal model

Large signal dc analysis of bipolar transistors is performed using the Ebers Moll equivalent circuit, properly modified to take into account the Early effect. As far as large-signal transient analysis is concerned, the charge-control model is generally used [3]. Here, we will limit to recall the simplified collector current equation in the forward-active region, which is used by SPICE-like simulators [8]:

$$I_C = I_S e^{\frac{V_{BE}}{V_T}} \left(1 + \frac{V_{CB}}{V_A}\right) \tag{3.35}$$

where I_S is the saturation current, $V_T = k_B T/q$ and V_A is the Early voltage.

This equation is often simplified by considering that $V_{CB}=V_{CE}-V_{BE}$ and that V_{BE} is nearly constant in the forward-active region. By this simplification, this leads the following equation, which is generally referred to in most textbook on electronic design:

$$I_C \cong I_S e^{\frac{V_{BE}}{V_T}} \left(1 + \frac{V_{CE}}{V_A}\right) \tag{3.36}$$

The base current can be estimated by means of the well-known relationship:

$$I_B \cong \frac{I_C}{\beta} \tag{3.37}$$

where β is the dc current gain.

The active region is characterized by $V_{CE} \geq V_{CESAT}$. Textbooks on electron devices define saturation as a condition where both the BE and BC junctions are forward-biased. From this definition, V_{CESAT} would

be such that V_{BC}=0, then V_{CESAT}= V_{BE}. For practical purposes, saturation is considered the region where I_C starts to exhibit a strong dependence on V_{CE} and the base current gets much larger than the value given by (3.37). Typical values of V_{CESAT} that correspond to the practical definition are in the range 100-200 mV.

Equations (3.36) and (3.37) represent a good approximation of I_C and I_B over a large interval of collector currents. In particular, the exponential behavior represented by (3.36) is generally maintained across an I_C range of several decades. Outside this range, the collector current deviates from the exponential behavior. This is well represented by the so-called Gummel plot, which is generally included in the process manuals for any elementary device. An example of Gummel plot is shown in Fig. 3.11. The plot gives a representation of the collector and base currents as a function of the base-emitter voltage, for a fixed collector-emitter voltage. The latter is chosen to keep the device in forward-active region. Considering the semi-logarithmic scale (linear voltage, logarithmic currents), an exponential behavior turns out into a straight line.

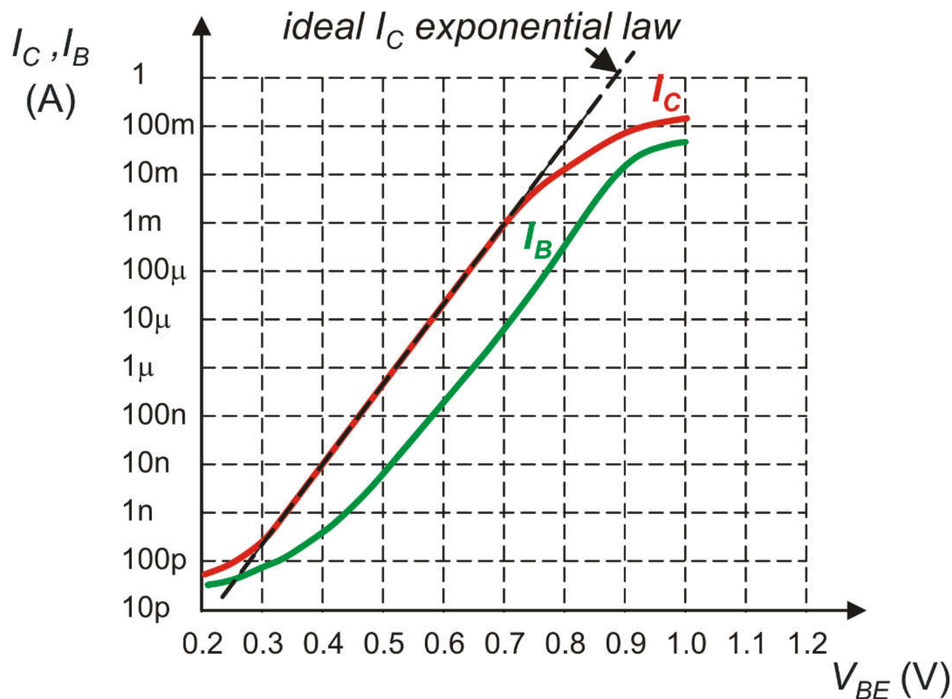

Fig.3.11. Example of Gummel plot.

Equation (3.36) holds true region where the I_C curve is well approximated by a straight line. Deviation from the straight-line at low currents is mainly due to leakage currents (e.g. currents from the reverse-biased CB junction). At high currents, the discrepancy is due to high-injection effects and to the base series resistance. Due to the logarithmic scale, the distance between the I_C and I_B curve is proportional to log(β). At low and high collector currents, the I_C and I_B curves get closer, meaning a reduction of β. This is well represented by the curve of β (see Fig.3.12), which is also generally included into the process manual.

The fall of beta at low currents is mainly due to generation-recombination and tunneling current associated to the base-emitter junction. The beta decrease at high currents is due to base widening effects occurring at high current densities (e.g. Kirk effect) [2].

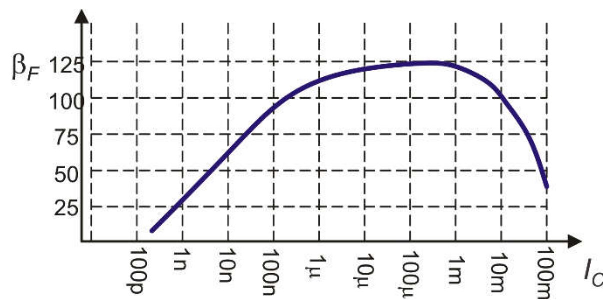

Fig.3.12 . Beta as a function of I_C.

For a transistor having *area* > 1, the same Gummel plot (or same beta plot) of the corresponding elementary device can be used by simply dividing the actual device current by the *area* factor. As an example, let us consider an elementary BJT having the Gummel plot of Fig. 3.11. If we bias the elementary device (*area*=1) with a current of 10 mA, then we are out of the region where the exponential equations hold. This may have several adverse consequences. For example, the V_{BE} cannot be considered as a quasi-constant voltage ($V\gamma$) anymore since its increases with I_C gets larger than in the exponential region. Furthermore, there are circuits that rely on the exponential behavior, such as band-gap voltage references or analog multipliers. If we cannot reduce the bias current (e.g., due to noise or bandwidth constraints), we can still set the area parameter to a value greater than one. Using *area*=10, the equivalent I_C current to be used in the Gummel plot of the elementary device is the actual I_C current (10 mA) by 10. The resulting value (1 mA) is well within the exponential region. Obviously, the maximum collector current of a device is *area* times as large as the maximum current of the elementary transistor.

BJT small signal model

A simplified small-signal equivalent circuit of a vertical BJT is shown in Fig.3.13. This circuit is similar to the MOSFET equivalent circuit (with the exception of the input resistance r_{be}), allowing straightforward porting of circuit analysis results across the two device types. For this reasons the circuit of Fig.3.13 is more frequently used than h-parameter one for integrated circuit design.

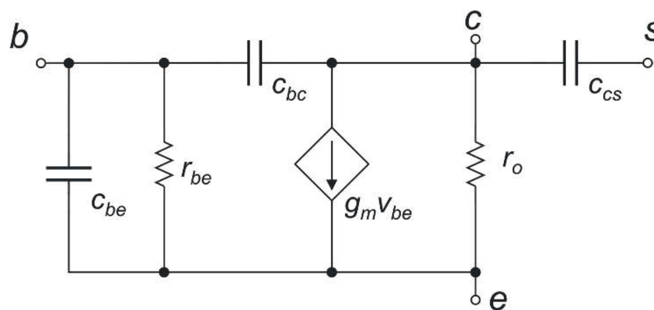

Fig.3.13. BJT small signal equivalent circuit

Terminal "s" represents the substrate. In a vertical npn BJT (see Fig.3.9), the substrate (p-doped) forms a junction with the collector (n-doped). This junction is kept reverse-biased by setting the substrate to the smaller supply voltage. Nevertheless, due to the junction, a capacitance (c_{cs} in Fig.3.13) is present between the collector and the substrate. Due to the large extension of the collector body, this capacitance can be relatively large, affecting the performances of circuit that use npn BJTs. In the case of a lateral

structure such as the pnp BJT of Fig. 3.10 (a), it is the base to be coupled to the substrate with a junction capacitance.

The dc parameters, namely g_m, r_0 and r_{be} can be derived from (3.36) and (3.37), obtaining the following expressions:

$$g_m \equiv \left(\frac{\partial I_C}{\partial V_{BE}}\right)_{V_{CE=const}} = \frac{I_C}{V_T} \tag{3.38}$$

$$r_0 \equiv \frac{1}{\left(\frac{\partial I_C}{\partial V_{CE}}\right)_{V_{BE=const}}} = \frac{1}{\frac{1}{V_A} \cdot I_S e^{\frac{V_{BE}}{V_T}}} \cong \frac{V_A}{I_C} \tag{3.39}$$

$$r_{be} \equiv \left(\frac{\partial V_{BE}}{\partial I_B}\right)_{V_{CE=const}} = \left(\frac{\partial I_B}{\partial V_{BE}}\right)^{-1}_{V_{CE=const}} = \left[\frac{1}{\beta}\left(\frac{\partial I_C}{\partial V_{BE}}\right)_{V_{CE=const}}\right]^{-1} = \beta \frac{1}{g_m} \tag{3.40}$$

Notice that expression (3.40) is an approximation, which is acceptable only if β can be considered independent of I_C. The dc version of circuit in Fig. 3.13 is equivalent to the well-known hybrid parameter circuit with h_{re}=0. Equivalence with the remaining h-parameters are: $h_{ie}=r_{be}$, $h_{oe}=1/r_o$, $h_{fe}=g_m r_{be}$.

The capacitances that appear in the equivalent circuit are all due to junctions, and then are strongly voltage dependent. In active-forward region, the base-collector and collector-substrate junctions are generally reverse biased. The only exception occurs for $V_{CESAT} \le V_{CE} \le V_{BE}$, where the base-collector is be weakly forward biased. In all cases, c_{bc} and c_{cs} are dominated by the depletion-layer capacitance, given by:

$$c_{bc} = \frac{C_{JC}}{\left(1+\frac{V_{CB}}{V_{JC}}\right)^{m_{jc}}}, \qquad c_{cs} = \frac{C_{JS}}{\left(1+\frac{V_{CS}}{V_{JS}}\right)^{m_{js}}} \tag{3.41}$$

where C_{JC} and C_{JS} are the corresponding zero-bias capacitances, V_{JC} and V_{JS} are the built-in potentials of the two junctions and m_{jc}, m_{js} the corresponding grading coefficients.

On the other hand, the base-emitter junction is forward biased in the forward-active region. Therefore, both the depletion-layer (c_{te}) and diffusion (c_{de}) capacitances must be considered. Then, $c_{be}=c_{te}+c_{de}$, with:

$$c_{te} = \frac{C_{JE}}{\left(1-\frac{V_{BE}}{V_{JE}}\right)^{m_{je}}} \tag{3.42}$$

$$c_{de} = \tau_F g_m \tag{3.43}$$

where τ_F is the forward transit time [2]. At sufficiently high I_C values, c_{de} is much larger than c_{te} and c_{bc}. In these conditions, the forward transit time can be related to the transition frequency (f_T) of the bipolar transistor through the following equation:

$$f_T = \frac{1}{2\pi\tau_F} \tag{3.44}$$

3.5 References

[1] P.R. Gray, P. J. Hurst, S. H. Lewis, R. G. Meyer, "Analysis and Design of Analog Integrated Circuits", 4th edition, John Wiley & Sons, New York, 2001.

[2] Y. Taur, T.H. Ning, "Fundamentals of Modern VLSI Devices", Cambridge University Press, 2nd edition, 2009.

[3] C. C. Enz, F. Krummenacher and E. A. Vittoz. "An Analytical MOS Transistor Model Valid in All Regions of Operation and Dedicated to Low-Voltage and Low-Current Applications", Analog Integrated Circuits and Signal Processing, vol. 8, pp. 83-114, 1995.

[4] Y. Tsividis, "Operation and Modeling of the MOS Transistor", 2nd edition, Oxford University Press, New York, 1998.

[5] I. M. Filanovsky, A. Allam "A Mutual compensation of mobility and threshold voltage temperature effects with applications in CMOS circuits". IEEE Trans Circuits and Syst I, vol. 48, pp. 876–884, 2001.

[6] F. Silveira, D. Flandre, and P. G. A. Jespers, "A g_m/I_D Based Methodology for the Design of CMOS Analog Circuits and Its Application to the Synthesis of a Silicon-on-Insulator Micropower OTA", IEEE J. Solid State Circuits, vol. 31, pp. 1314-1319, September 1996

[7] D. E. Ward and R.W. Dutton, " A Charge-Oriented Model for MOS Transistor Capacitances", IEEE J. Solid State Circuits, vol. SC-13, No 5, pp. 703-708, October 1978.

[8] A.S. Sedra, K.C. Smith "Microelectronic Circuits", 7th edition, - 2016 - Oxford University Press, appendix B: SPICE Device Models And Design Simulation Examples Using Pspice And Multisim, available on line: https://global.oup.com/us/companion.websites/9780199339136/student/app/

4 Process errors

4.1 General definitions

Fabrication of an integrated circuit is subjected to errors that make the final product different from the designed device. This problem, which is clearly typical of all industrial processes, needs to be well characterized in order to estimate the actual deviation that can be expected to occur from the ideal case. Let us start from very common definitions. We will focus on a component (e.g. a resistor) integrated on a silicon chip. Of that component, will consider a particular quantity (e.g. its resistance) that we will generically indicate with "*A*". The value of *A* assigned to the given component in the design phase is indicated as "nominal" value (A_N). Due to process errors, components integrated in the fabricated chips will show a value of *A* that differs from the nominal value. In addition, different realizations of the same component will show different value of *A*. The best way to represent the variability of the fabricated values (also indicated as "process spread") is using a histogram.

To build a histogram, we need to consider a large number of different specimens of the same component. Let us indicate the number of different samples with "*n*". Among this set, the quantity A assumes a minimum and maximum value. We divide the interval between the minimum and maximum into a series of uniformly sized sub-interval, called "bins", of width ΔA. For each bin, we count the number of samples whose quantity *A* falls into it. A graphical representation of a histogram is shown in Fig.4.1, where the quantity represented in the *y*-axis is the fractional number ($\Delta n/n$) of samples included in each bin.

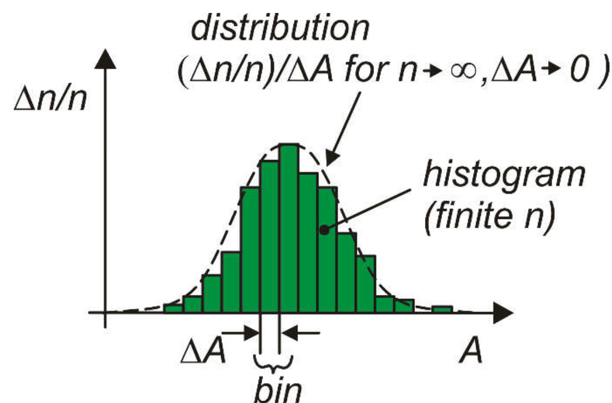

Fig.4.1. Example of histogram.

If we imagine to progressively increase the number of samples and, at the same time, increase the number of bins (reducing the width of each bin), the histogram tends to the ideal distribution that characterize the errors for the given fabrication process. To be more precise, the distribution is obtained by dividing the height of each bar in Fig. 4.1 ($\Delta n/n$) by the width of the bins (ΔA). Since *A* is a continuous variable, the distribution coincides with the probability density function.

The elements of the distribution that are of interest for the production process are illustrated in Fig. 4.2. These elements are summarized below:

A_N: this is the nominal value, defined in the design phase.

A_i: The value of quantity A for a generic i-th component.

$\langle A \rangle$: the mean of the distribution.

e_S: The systematic error = $\langle A \rangle - A_N$

e_R: Random error for the i-th component = $A_i - \langle A \rangle$.

The mean of the process can be estimated by taking the mean of A over a large set of components. The actual values of A taken on different components tends to group around the mean value. Differences from the mean value constitute the random error. The difference of the mean with respect to the nominal value is the systematic error. In a correct design, the systematic error should be negligible with respect to random errors. The presence of a non-negligible systematic error can be due to design errors, inaccurate or faulty fabrication process or from inaccuracy of the models used to represent the component behavior. For example, an excess systematic error may derive from neglecting the contact resistance of integrated resistors. In this case, the resistance of the fabricated resistors will be on average larger than the value set by design (nominal value).

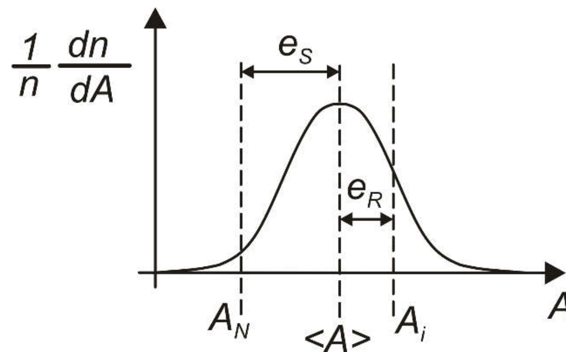

Fig. 4.2. Elements of the distribution.

The magnitude of random errors is well represented by the standard deviation (or standard error), which is the square root of the mean square value of the deviation from the mean. It is defined by:

$$\sigma_A = \sqrt{\left\langle \left(A - \langle A \rangle \right)^2 \right\rangle} \tag{4.1}$$

If we have a finite set of data (finite sample N dada), the best estimate (unbiased estimate) of the standard deviation of the whole fabrication process is given by:

$$\sigma_A = \sqrt{\frac{\sum_{i=1}^{N}\left(A_i - \mu_{A,N}\right)^2}{N-1}} \tag{4.2}$$

where $\mu_{A,N}$ is the mean calculated over the finite sample of N data. The square of the standard deviation is the variance.

The knowledge of the standard deviation is particularly important when the type of distribution is given, since it allows determining the fraction of data that fall with a given interval around the mean. Note that in most cases of interest for a fabrication process, the distribution is Gaussian. This occurs because fabrication process involve a large number of phenomena that contribute to the total random error. Generally, these phenomena are independent, so that the final distribution tends to a Gaussian even if the single distributions are not Gaussian (central limit theorem). A Gaussian distribution is perfectly determined when its man and standard deviation are given. The fraction of data that falls within an interval centered around the mean is given in the table 4.1:

Max deviation from the mean	$\pm\sigma$	$\pm 2\sigma$	$\pm 3\sigma$	$\pm 4\sigma$
Fraction of data within the interval	68.3 %	95.4 %	99.7 %	99.994 %
Fraction of data outside the interval	31.7 %	4.6 %	0.3 %	0.006 %

Table 4.1: Fraction of data that fall inside or outside an interval around the mean for a Gaussian distribution as a function of the maximum deviation from the mean.

4.2 Fabrication errors in a microelectronic process: global and local errors.

Figure 4.3 depicts the different scales of an integrated circuit (IC) fabrication process. At the smallest level there is the chip. At this stage, if we place several identical copies of the same component (nominally identical components) the differences among them are very small. For example, if we design a chip with different copies (instances) of a 1000 Ω resistor, we have good chances to get components that differ from each other by less than a few Ohms. At the second level of the fabrication process, there is the wafer, which collects hundreds or even thousands of dies (chips). The uniformity of process geometrical or physical parameters over a large wafer is much worse than over a single chip. Therefore, if we consider the set of components fabricated on the chip of the whole wafer, differences between these components begin to get significantly larger. Differences gets larger and larger as we consider the successive scale levels, that is the batch of wafers fabricated in a single run and, finally different runs. Differences between components fabricated in different runs can be very large, reaching even ± 20 %. If we consider again a resistor that is designed to have a resistance of 1000 Ω, we can likely get resistors of 800 Ω and 1200 Ω in different runs.

It is useful to introduce two new quantities:

$<A>_{chip}$: The mean performed on all components integrated on a given chip. This value will change from one chip to another. Even if we cannot place an infinite number of copies of the same component on the same chip, we can imagine being able to reproduce the fabrication of that chip perfectly just in terms of mean values of all parameters (doping levels, oxide thickness, etc.). By this expedient, it is possible to justify the introduction of a mean, which is a property of a hypothetical process that led to the fabrication of that particular chip, and then refer to an infinite number of components.

$<A>_{process}$ The mean performed over the totality of components fabricated by that process. Clearly, $<A>_{\mathrm{process}}$ is also the mean of $<A>_{chip}$ calculated over all chips produced by that process.

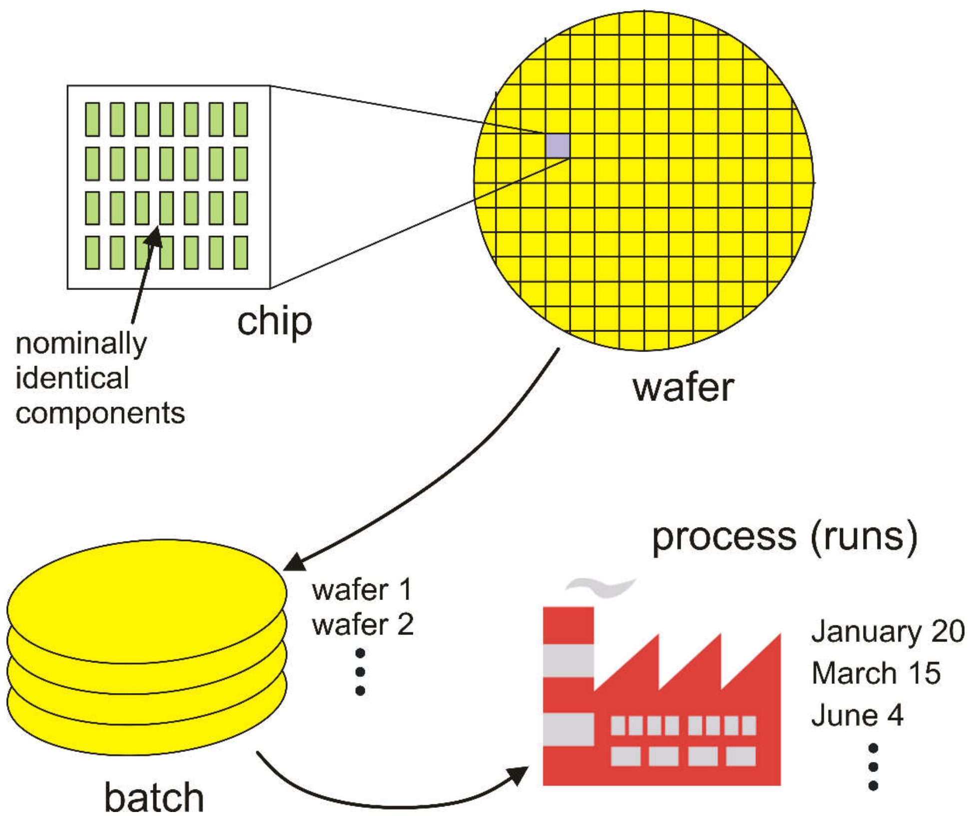

Fig. 4.3. Different scales of the fabrication process.

We can now divide the random errors into two different contributions:

-) **Local errors**, given by the difference between the value of the quantity of interest (A) assumed by a component with respect to the mean of the chip where it is located. Considering the discussion at the beginning of this paragraph, there is generally a good uniformity of parameters across a single chip, and then all components in that chip will exhibit values of A very close to $<A>_{\mathrm{chip}}$. In other word, local errors are generally very small. Symbolically, the local error for component i-th is given by:

$$e_{local} = A_i - \langle A \rangle_{chip} \tag{4.3}$$

where $<A>_{chip}$ refers to the chip where component i-th is placed .

-) Global errors: given by the difference of the mean of a given chip with respect to the mean of the process. This error can be very large, since process parameters can vary much depending on; (i) the position of the chip in the wafer, (ii) the position of the wafer in the batch and, most importantly, (iii) the run the batch belongs to. (see Fig. 4.3). Symbolically, the global error for a given chip is given by:

$$e_{global} = \langle A \rangle_{chip} - \langle A \rangle_{process} \tag{4.4}$$

Figure 4.4 shows a graphical representation of the various error components. The random error is decomposed into a local and global error. The mean of single chips is distributed according to the global distribution shown at the bottom. The local distributions of two distinct chips ($chip_1$ and $chip_2$) are shown at the top of the figure. Decomposition of the random error is shown for a component belonging to $chip_1$.

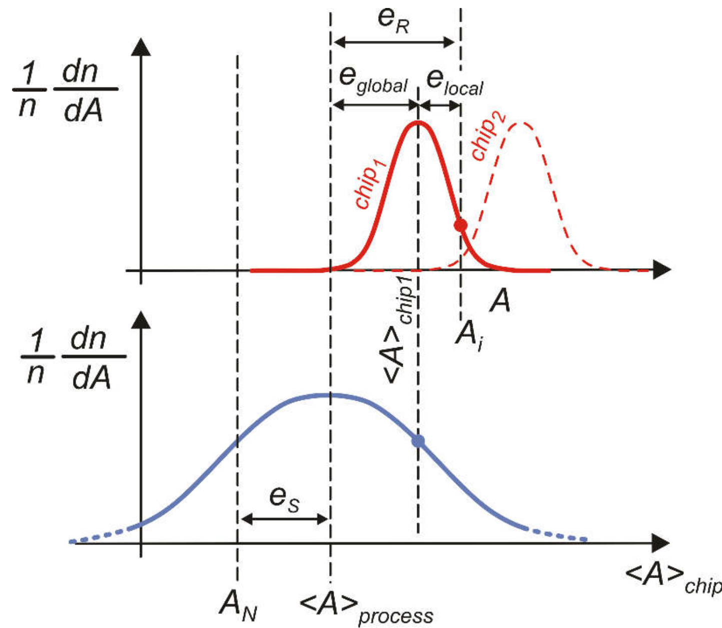

Fig. 4.4. Local (top) and global (bottom) errors. The width of local error distribution is comparatively much smaller than swon in the figure, where it has been artificially enlarged for visibility purpose.

Global and local errors are represented by distinct distributions, characterized by two distinct standard deviations, σ_{global} and σ_{local}, respectively. Different chips are characterized by different local means ($<A>_{chip}$), but all chips have the same standard deviation. This mean that distribution from different chips are simply shifted along the A axis, as shown in Fig. 4.4, but maintain the same shape and width. For the considerations made about the magnitude of global and local errors, we have:

$$\sigma_{global} >> \sigma_{local} \tag{4.5}$$

4.3 Matching errors.

A matching error is defined as the difference assumed by quantity *A* between two nominally identical components. In microelectronics, matching errors are considered only between components that are placed on the same chip. Therefore, matching errors are the consequence of local errors. If consider two component, identical by design, and indicate with A_1 and A_2 the value assumed by A on component 1 and component 2, respectively, then we can define the two quantities:

$$\begin{cases} \Delta A = A_1 - A_2 \\ \overline{A} = \dfrac{A_1 + A_2}{2} \end{cases} \tag{4.6}$$

Where ΔA is the matching error, while $\overline{A}$ is the midpoint value. Equations (4.6) can be solved to express A_1 and A_2 as a function of the matching error and midpoint value:

$$\begin{cases} A_1 = \overline{A} + \dfrac{\Delta A}{2} \\ A_2 = \overline{A} - \dfrac{\Delta A}{2} \end{cases} \tag{4.7}$$

There are two main causes of matching errors:

- Local granularity
- Gradients

4.4 Local granularity: The Pelgrom Model

Matching errors between identical components that are placed very close to each other into the same die are due to local non-uniformity ("granularity") of the material properties. To understand this, let us consider doping: dopant atoms are randomly distributed over the substrate and the number of dopant atoms that are present inside a given component will obviously vary, depending on the component location.

This phenomenon is clearly illustrated in Fig. 4.5, where the rectangle shows a portion of the chip area and the red crosses are dopant atoms. The yellow and green rectangles represent the area occupied by two nominally identical devices. Three possible placement for the two components are proposed. N is the total number of dopant atoms that fall inside the two components in a given location, while ΔN is the difference between the number atoms inside the yellow component and the number inside the green one. Note that the fluctuation occurring from one location to another is very large, reaching 38 %.

Repeating the experiment with larger components the relative fluctuation of the number of atoms is significantly reduced. This is due to the averaging effect that large areas operate on the local irregularity. Figure 4.6 represents a case in which the component width and height have been doubled, showing the considerable reduction of $\Delta N/N$. The same effect applies to other quantities, such as the gate oxide thickness, which exhibits local variations due to the unavoidable surface roughness.

As the examples in Figs. 4.5 and 4.6 clearly show, on large area devices, these short-length variations tend to have a smaller relative impact, since the device will include areas with both minimum and maximum levels of the physical quantities of interest, producing a sort of compensation.

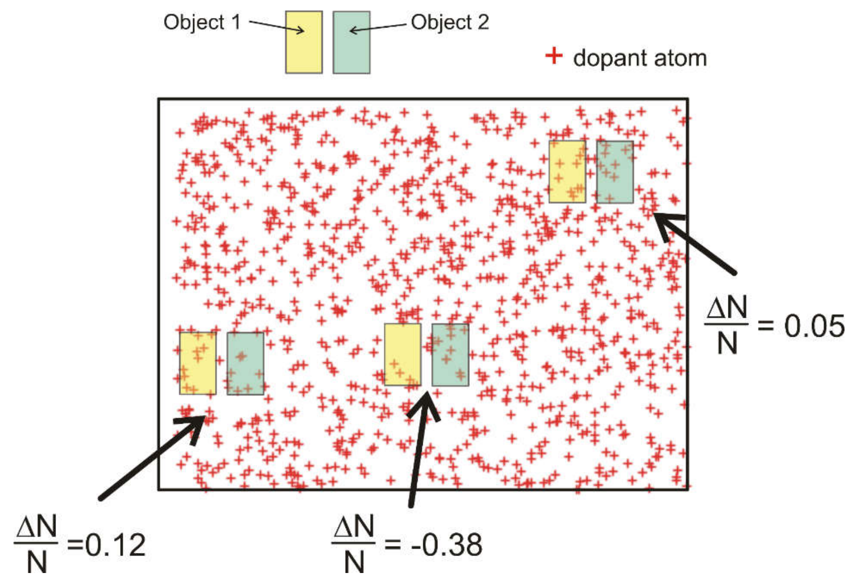

Fig.4.5. The red crosses represent dopant atoms, while the green and yellow rectangle represent the area occupied by two identical components. Different placements result in different atom distributions within the component areas.

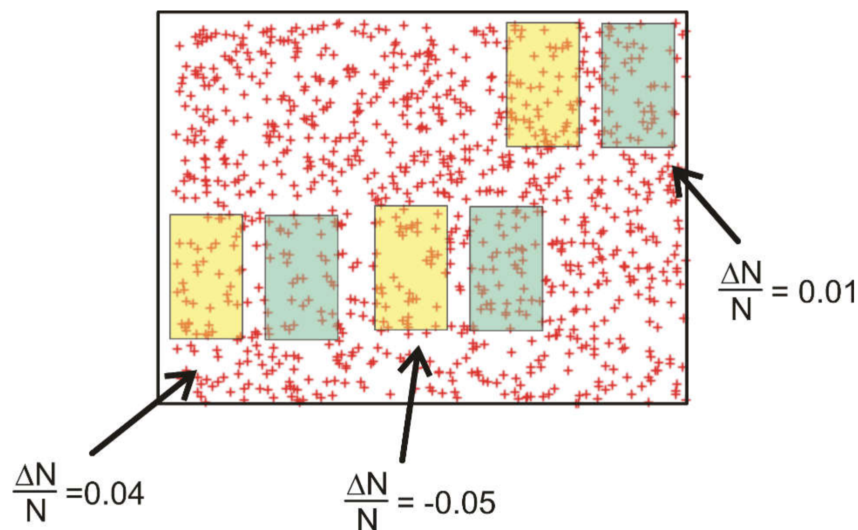

Fig.4.6: Same atom distribution as in Fig. 4.5, but with components of larger area. Note the relative fluctuation is much smaller than in the case of small components.

For this mechanism, matching errors will be smaller in large area devices. This intuitive idea is well represented in a quantitative way by the Pelgrom model [1] that express the standard deviation of the MOSFETS parameters as a function of the device gate area (WL) in the following way:

$$\begin{cases} \sigma_{\Delta V_t} = \dfrac{C_{Vt}}{\sqrt{WL}} \\ \sigma_{\frac{\Delta\beta}{\beta}} = \dfrac{C_{\beta}}{\sqrt{WL}} \end{cases} \tag{4.8}$$

where C_{Vt} and C_β are constant parameters that are typical of the fabrication process. These matching parameters can be found in the process DRM (Design Rule Manual). This model is generally valid also for other kind of devices, such as resistors, capacitors and bipolar transistors. For example, the standard deviation of the relative matching error of integrated resistors can be expressed by:

$$\sigma_{\frac{\Delta R}{R}} = \frac{C_R}{\sqrt{WL}} \tag{4.9}$$

where C_R is constant that depends on the process and on the type of resistor (polysilicon, high-resistivity polysilicon, diffusion etc.). Constants C_β and C_R are expressed in mm, while C_{Vt} is typically in mV·µm, so that W and L should be expressed in µm in expressions (4.8) and (4.9).

4.5 Gradients

Gradients indicate that important quantities that affect properties of devices are not uniformly distributed on a macroscopic scale. This means that these quantities gradually varies across the chip area.

Quantities of interest can be, for example:

-) Dopant density

-) Oxide thickness

-) Geometrical process biases (e.g. etching undercut)

-) Temperature (e.g. due to power devices present on the chip)

-) Mechanical stress (mainly due to the packaging process)

The effect of the gradient of a given quantity "A" (can be one of the list above) on the matching of two components is shown in Fig. 4.7

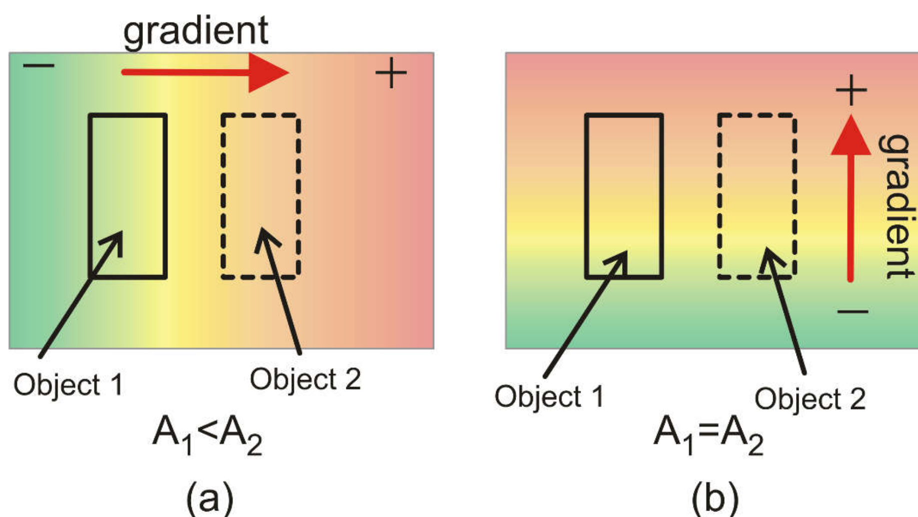

Fig. 4.7. Effect of gradients on device matching. Different colors represent different values of a given quantity "A". In (a) the average of quantity A is larger for component 2, while in (b) the two components receive the same average of A.

The effect can be different depending on the gradient orientation with respect to the line that join the two component locations. If the gradient and the joining line are parallel, as in the case of Fig. 4.7(a), the mismatch will be maximum; if the gradient and the joining line are orthogonal, the gradient does not cause mismatch, as shown in Fig. Fig. 4.7(b). Unfortunately, generally it is not possible to predict the gradient direction, since it will vary according to the position of the chip in the wafer, the wafer in the batch and so on. Only in the case of mechanical stress and temperature distribution, it is possible to have an idea of the gradients from the way the chip is mounted on the package and from the position on the chip of power devices, that can be important heat sources. However, even in these cases the prediction is fairly inaccurate, so that gradients are likely to produce mismatch.

An effective solution is offered by the so-called common centroid configurations. The two devices that should match are split into different identical parts that are then placed in such a way that parts from object 1 are interleaved with parts from object 2. The requirement is that the centroids of the two devices coincides. This method is illustrated in Fig. 4.8: Component A_1 is divided into the two identical parts $A_{1,1}$ and $A_{1,2}$, while A_2 is divided into $A_{2,1}$ and $A_{2,2}$. The centroid of the two devices is indicated with C. Note that by splitting each device into two parts, the centroid is allowed to lie outside each convex shape that form the device (rectangles in the example). In this way, we can make the centroid to coincide even if the two devices does not overlap in any point.

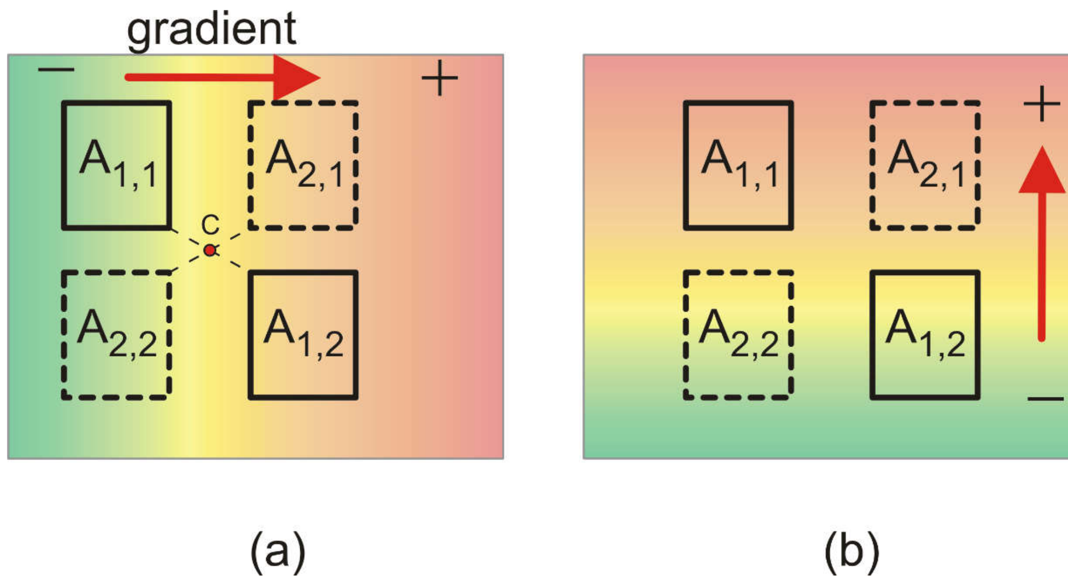

Fig. 4.8. Common centroid configuration in the case of two different gradient orientations.

From Figure 4.8 (a) and 4.8 (b) it is clear that now the two devices are affected by the quantity of interest (to which the gradient refers) in exactly the same way, independently of the gradient direction. In both cases depicted in Fig. 4.8 (a) and 4.8 (b), each device has a part that receives a larger value of the quantity while the other part receive a smaller value. On average, both devices receive the same value.

Note that Fig. 4.8 (a) and (b) represent two particular cases. Fig.4.9. represent the case of an oblique gradient. Object A_1, formed by parts $A_{1,1}$ and $A_{1,2}$ gets an intermediate value of the quantity. Object A_2 receives an higher value (part $A_{2,1}$) and a lower value ($A_{2,2}$). Again, we have a compensation and on average the two components A_1 and A_2 receive the same effective value of the quantity of interest. Differently from the cases depicted in Fig. 4.8 (a) and (b), the symmetry is not perfect for oblique

gradients and compensation is perfect only if the gradient is constant across the whole area occupied by the two components.

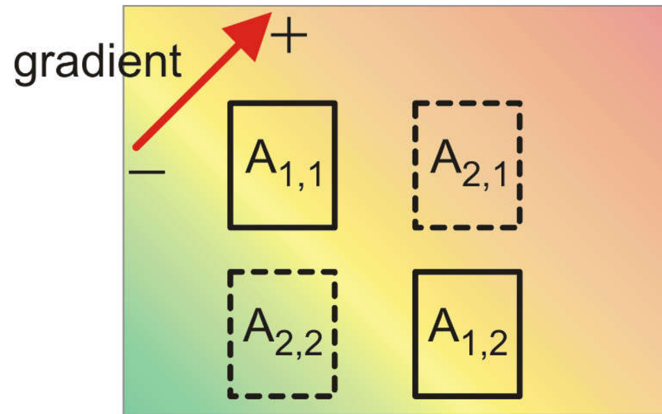

Fig. 4.9. Common centroid configuration in the case of oblique gradient.

In order to apply the common centroid approach, it is necessary to split each component into two parts that, properly connected, must still behave like the original component. The way components can be split and re-connected depends on the type of device. Figure 4.10 show the two options that can be adopted when the two components to match are resistors (R_1 and R_2, of nominal value R).

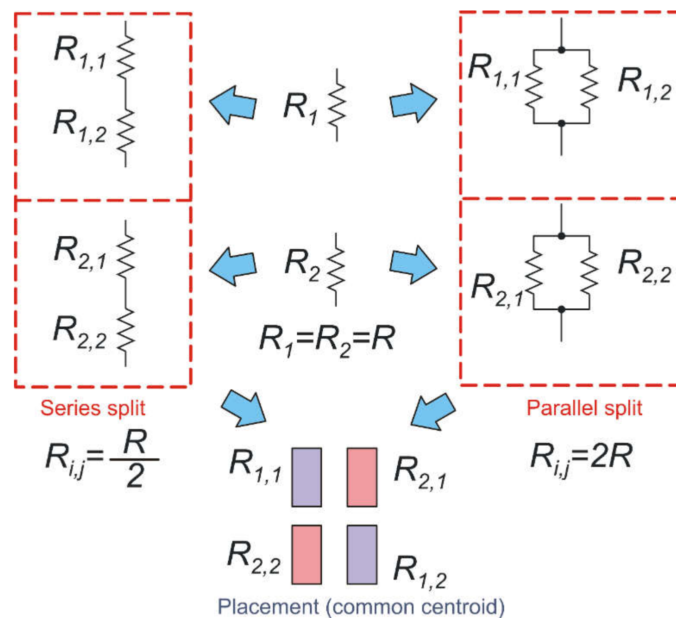

Fig. 4.10. Common centroid configuration of resistors implementd with series split and parallel split.

On the left, each one of the two resistors is split into two parts that are then reconnected in series. In order to maintain the original value, the individual parts should have half of the original value, i.e. their resistance should be R/2. On the right, the parts are connected in parallel. Then, to maintain the resistance of the original components, each part should have a resistance 2R. The placement is the same for the series and parallel split. The series split is advantageous in the case that R is large, resulting in particularly long resistors. The parallel split has to be preferred only in the case of small resistance (short resistors).

In the case of active devices, such as MOSFETs or BJTs, the only possible way to split the original components is the parallel split. Figure 4.11 illustrate the case of common centroid applied to MOSFETs. If the β ($=\mu_n C_{ox} W/L$) of the The original devices, M1 and M2, have the same nominal parameter β ($=\mu_n C_{ox} W/L$). Then the parts in which they are split are characterized by β/2. In practice, the parts have half the width (W) of the original MOSFETs. In a parallel connection of MOSFETs, the effective beta is the sum of the betas of the devices that form the parallel. The actual arrangement of the single parts is shown in Fig. 4.11, on the right. Series split configurations are not applicable to common centroid arrangement of MOSFETs. The reason is that in a series of MOSFETs, the two parts do not contribute in the same way the property of the composite device. The same applies to BJTs, for which the only possible split is the parallel one.

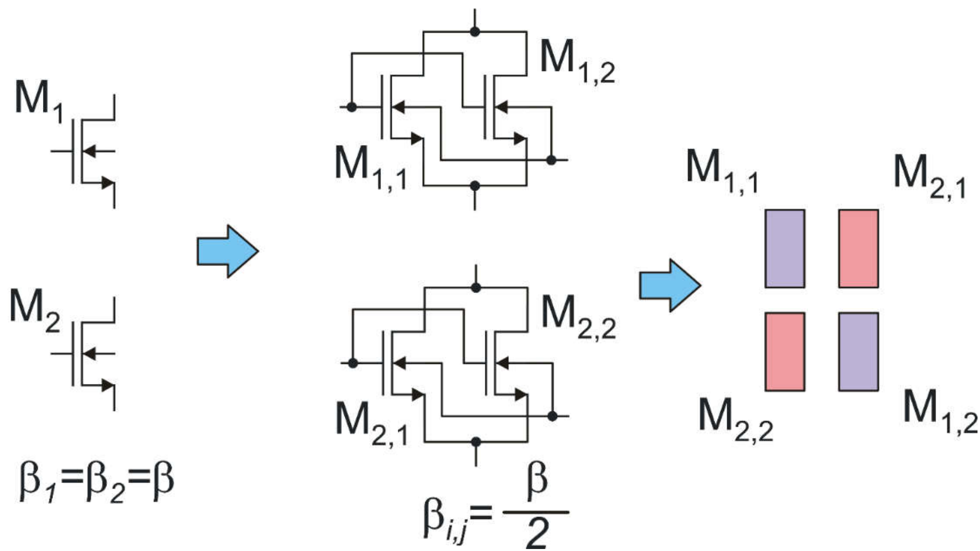

Fig. 4.11. Common centroid applied to MOSFETs.

Finally, common centroid configurations are widely used also for capacitors. A good matching between capacitors is a key element for the accuracy of switched capacitors circuits and in particular of charge-redistribution analog to digital converters. In principle, common centroid configurations can be applied to capacitors using both the series and the parallel split, just as for resistors. In practice, series connection of integrated capacitors has to be discouraged, because the dc voltage of intermediate nodes in a series of capacitors cannot be easily controlled. As a result, common centroid schemes are applied to capacitors using mainly the parallel split approach.

4.6 General rules for matching components

In addition to the rules introduced in paragraphs 4.4 (area of devices) and 4.5 (common centroid arrangement), there are also important rules that have to be mandatorily or optionally followed to reduce matching errors between component pairs. Figure 4.12 shows two mandatory rules. On the left, two objects with identical W/L ratios (e.g. two resistors or two MOSFETs) are shown. Setting the aspect ratios to be equal is not sufficient to obtain a good matching, even in the case that the expressions of the quantities of interest (e.g. resistance) include only the W/L ratios. The reason is that the properties of the materials that compose the devices tends to be different in the proximity of the boundaries of the device area (borders). For example, the resistivity of a conducting layer may be higher close to the borders due to reduction in dopant concentration or to increased scattering mechanisms. Since the extensions of the borders does not depend on the device dimensions, border-related effects will have a greater relative impact on the smaller device. For this reason, matched devices should be identical (same width and length). Note that the matching errors introduced by different device areas are systematic. Figure 4.12 shows two identical object (same lenghts and widths) that are placed along orthogonal directions. This may lead to poor matching since material properties can be anisotropic. The typica cause is mechanical stress due to the packaging procedure: packaging often occurs at a temperature that can exceed one hundred degrees. Successive cooling down produces mechanical stress through the different thermal expansion coefficients of the chip and package materials. Mechanical stress has generally a prevalent direction and this results in mentioned anysotropy. In addition, also the device dimensions are unevenly modified by the stress. A resistor subjected to mechanical stress that having a prevalent axis parallel to resistor length, will become longer and narrower than the original device. The opposite occurs if the stress is orthogonal to the resistor length. The stress will then cause different changes in the W/L ratios of devices width different orientations. As a result, matched devices should have the same orientation.

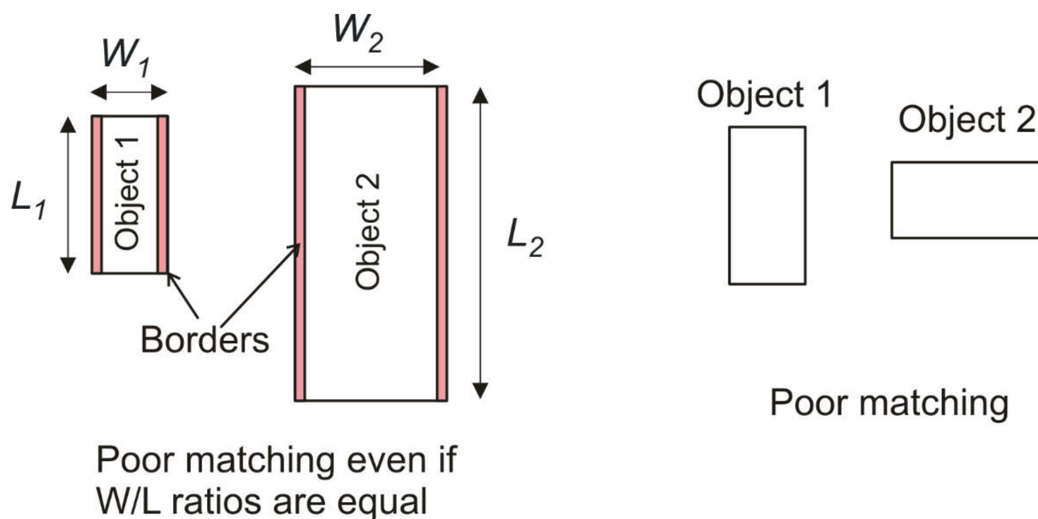

Fig. 4.12. Two common errors leading to poor matching: (left) fifferent device areas and (right) different device orientation.

Figure 4.13 illustrates two optional rules that have to be adopted when very low matching errors have to be achieved. The rule represented on the left regards the direction of current in the device. To obtain a good matching the direction of the current in the two devices should be the same. The reason is that unavoidable temperature gradients introduce an additional voltage drop whose sign depends on the

relative direction of the current with respect to the direction of the gradient. Another rule, indicated with “common surroundings” or “common environment” is illustrated in Fig. 4.13 (right).

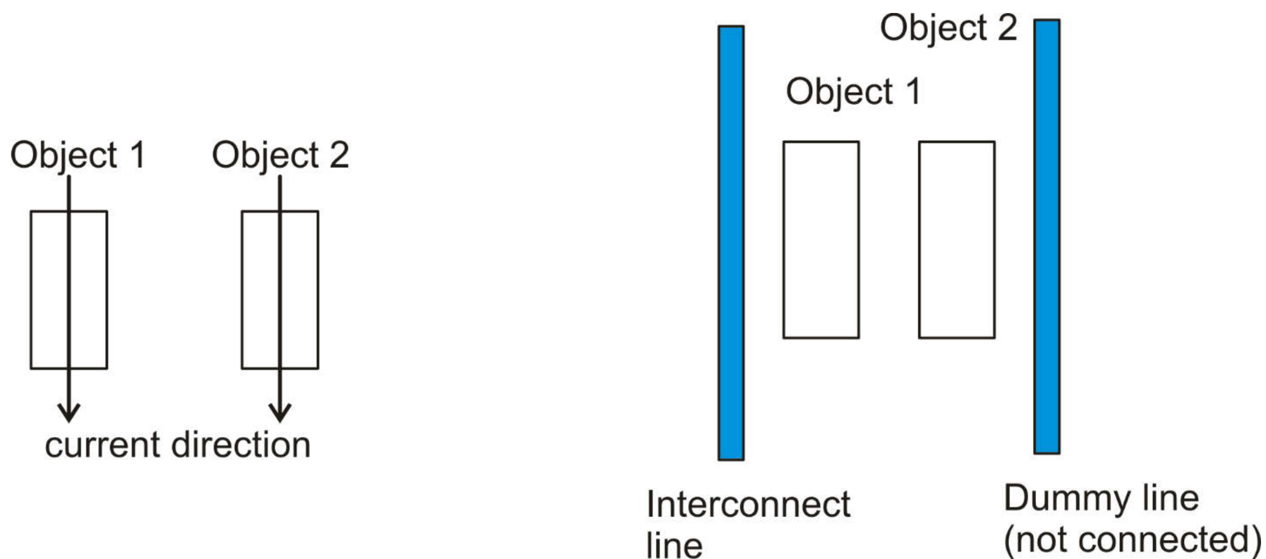

Fig. 4.13. Common direction (left) and common surroundings (right). .

If object 1 is close to a layout object (an interconnect line in the figure), then, to obtain an excellent matching, also object 2 should be close to a similar object. In other words, it is not sufficient that the objects are symmetrical, but also the environment where the object are placed must be symmetrical. If a metal is not passing close to object 2, we must place a metal (dummy line) that is not used for interconnection but only to make the environment symmetric. The dummy line can be left floating or, preferably, connected to gnd.

4.7 Rules for accurate ratios.

Frequently, important properties of electronic circuits are expressed as ratios of values of different components. A very simple example is shown in Fig. 4.14, depicting an inverting amplifier formed by an operational amplifier and two resistors.

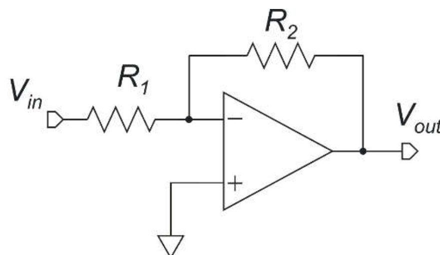

Fig. 4.14. Opamp-based inverting amplifier.

If the loop gain is large enough, the amplifier gain is simply given by:

$$A_V = -\frac{R_2}{R_1} \tag{4.10}$$

In many cases the gain magnitude (R_2/R_1) must be accurate, with differences from the ideal case of less than 1 %. If the gain magnitude is one, than R_1 should be equal to R_2 and obtaining a precise gain becomes simply a problem of good matching between the two resistors. If the gain to be obtained is different than one, than the problem is different. The more intuitive approach would be simply to introduce two resistance and set their value in order to obtain the required ratio. Unfortunately, many automated design kits assign the entered value of the resistance to the body of the resistor, leving out of the resistance computation the contact resistance,

Figure 4.15 (a) shows what happen if we simply try to obtain the required resistance ratio *r* by setting $r=L_2/L_1$. This sets the ratio of the resistor body approximately to the correct ratio. However, considering also the contribution of the contact resistances, as shown by the equivalent circuit of Fig. 4.15 (b), the actual ratio will be.

$$\frac{R_2}{R_1} = \frac{2R_C + rR}{2R_C + R} = r\frac{1 + 2R_C / rR}{1 + 2R_C / R} \tag{4.11}$$

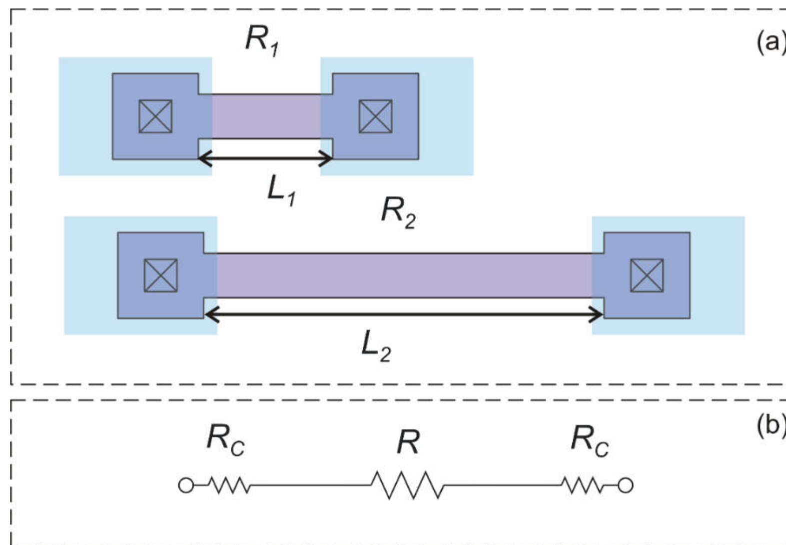

Fig. 4.15. Ratio obtained by simply scaling the resistor length (a). Equivalent circuit of a resistor with the contavt resistances (b).

The actual resistance ration is not r, unless r=1. For example, if r=3 and R_C/R=0.1, we would get: R_2/R_1=2.81, committing an error of nearly -6 %. In most cases, such an error is not acceptable. Clearly, it is possible to redesign the resistance values in order to take into account the contact resistance and obtain a more precise resistance ratio. Modern design kit of processes oriented to analog and mixed signal

design do so automatically. Unfortunately, contact resistance are not as accurate as resistor bodies, thus there would be still an important process-dependent inaccuracy. Furthermore, border effects that are not documented, makes also the resistor body close to both ends different from the central region. Again, border effects have an higher impact on the shorter resistor.

A much more accurate resistor ratio can be obtained by the so called modular approach. This is illustrated in Figure 4.16 (a) for a target ratio $r=R_2/R_1=3$. A single resistor module R_0 is used to form R_1, while R_2 is obtained by simply connecting three identical module R_0 in series. A possible layout for R_0 is shown in Fig. 4.16 (b), while the layout of the two resistors R_1 and R_2 is shown in Fig. 4.16 (c).

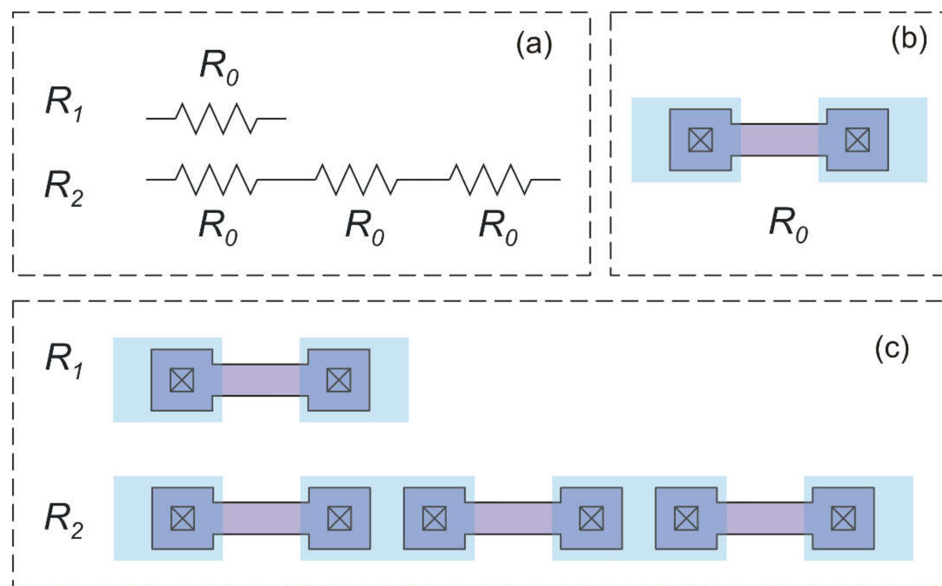

Fig. 4.16. Example of resistance ratio obtained with the modular approach.

By this arrangement, contact resistances and all other systematic non-idealities will affect in the same way all the instances of module R_0, so that the ratio R_2/R_1 will not change with respect to the ideal value (r=3 in the example). The method can be easily extended to non-integer ratios of the form M/N as shown in Fig. 4.17, where series of N and M modules (R_0) are used for R_1 and R_2, respectively.

In addition, it is possible to arrange the modules in parallel. This is advantageous when the resistances used to implement the ratio are particularly small. Combinations of series and parallel combinations of the same module R_0 can be used when a very large or a very small ratio has to be obtained. Using only pure series or parallel combinations would lead to the requirement of a large number of modules. For example, to implement a ratio r=100, the number of module involved is 101. The same occurs, obviously, if the ratio to be obtained is 1/100. Using a parallel of 10 R_0 modules for R_1 and a series of 10 R_0 modules for R_2 allows obtaining the required ratio of 100 with only 20 instances of module R_0. This approach is illustrated in Fig. 4.18.

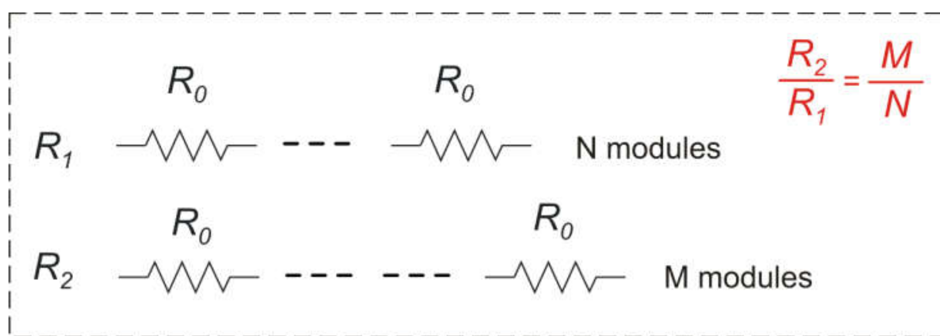

Fig. 4.17. Non-integer ratios R_2/R_1 obtained by the modular approach. .

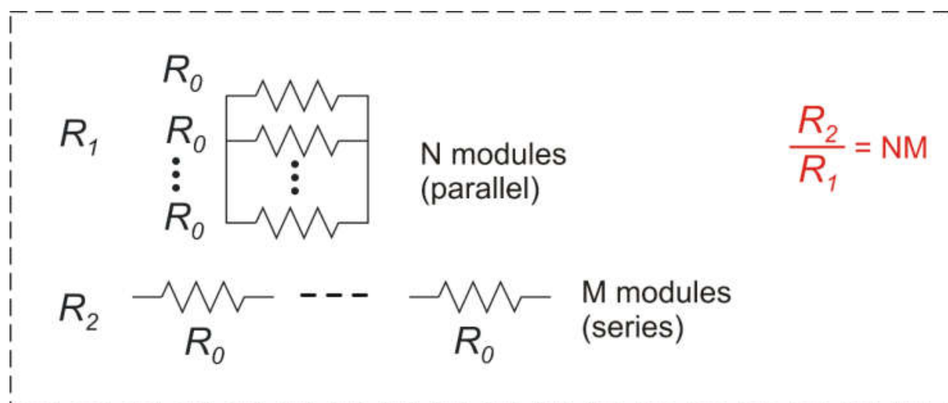

Fig. 4.18. Combination of series and parallel connections of modules to obtain very large ratios with a smaller number of components than pure series or parallel modular approaches.

Precise capacitance ratios are generally obtained with parallel connections, for the same reason mentioned for the common centroid configurations.

In the case that the requirement of precise ratios refers to the beta factor (β) of MOSFETs, high accuracy can obtained only with the parallel approach, although series or mixed connections have been proposed in the scientific literature [2]. An example of modular approach applied to a MOSFET-based current mirror is shown in Fig. 4.19. In both the input and output branch of the mirror, composite MOSFETs formed by parallel of a different number of nominally identical modules (M_0) are used. The composite MOSFET in the input branch, M1, is formed by *N* modules, while in the output branch M2 is formed by *M* modules. This corresponds to set the nominal ratio β_2/β_1 exactly equal to M/N. Doing the same using single MOSFETs for M1 and M2 and varying the aspect ratio (W/L) to obtain the required β_2/β_1 ratio results in a significant difference with respect to the target value due to the mentioned border effects (effective W and L are different from the drawn dimensions) and by unwanted effects that both *W* and *L* has on the threshold voltage. However, in all cases that a precise ratio is not required, it is preferable to act on the aspect ratio since it generally allows saving silicon area.

In the case of BJTs precise ratios of the saturation currents (I_S), are obtained only using parallel connections of a single module.

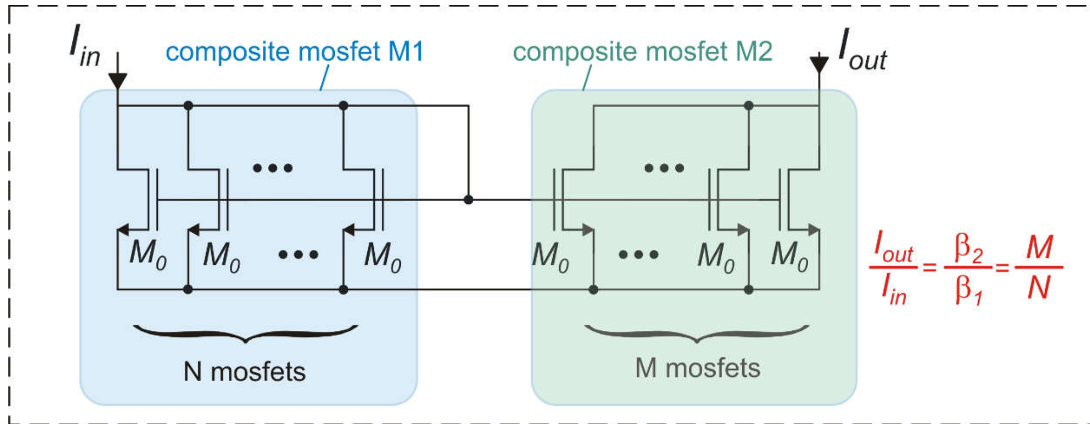

Fig. 4.19. Modular approach applied to a MOSFET-based current mirror.

4.8 Error propagation elements applied to matching errors

Generally, expressions that allow prediction of process errors (both global and local) are available for a few important parameters of the main devices of the process. An example is given in paragraph 4.4, where expressions for the standard deviations of matching errors are given for the beta factor and threshold voltage of MOSFETs. The problem that has frequently to be solved is finding how global properties of a circuit (e.g. an amplifier gain), are affected by the errors on the parameters of each component of the circuit. This is a particular case of a general problem called error propagation, which consists in finding the error on a quantity G that depends on variables A, B, C and so forth , resulting from the errors on the variables it depends on.

In this paragraph, a few remarkable cases that are easy to remember and that recur often in analog electronics. The focus will be on matching errors, but the results can be directly applied to a much wider spectrum of cases.

One-dimensional case

Let us start with a simple problem, illustrated in Fig. Quantity G depends on a single quantity A in a non-linear fashion. We are interested at finding the difference of the values assumed by G for two distinct values of A, indicated with A_1 and A_2. We introduce the following definitions:

$$\begin{aligned} &G_1 = G(A_1);\ G_2 = G(A_2) \\ &\Delta G = G_1 - G_2\,;\ \Delta A = A_1 - A_2 \end{aligned} \tag{4.12}$$

In the case of matching errors, we have two object, named object 1 and object 2, that should be as equal as possible. Thus, A_1,G_1 refer to object 1 while A_2, G_2 to object 2. However, this model can represent also other useful situations. Examples are:

- we have a single component and A_1, G_1 refer to the nominal case, while A_2, G_2 to the real case that will be affected by both global and local errors.

- A_1, G_1 and A_2,G_2 are simply two different statuses in which a given real component can be found.

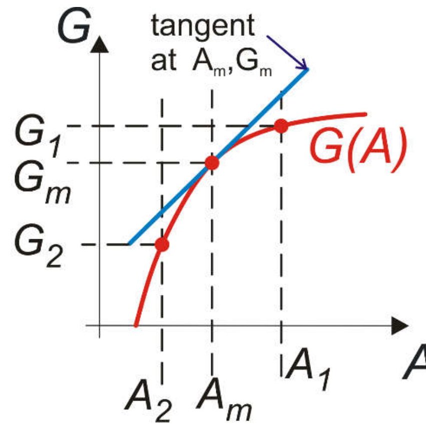

Fig. 4.20. Simple example of error propagation problem used to introduce the basic definitions.

We now introduce a third value of A, indicated with A_m. whose position with respect to A1 and A2 is completely arbitrary. Particular cases are when Am coincide with either A_1 or A_2 or is placed just in the middle of them. Values A1 and A2 can be expressed through their deviations with respect to A_m:

$$\begin{cases} A_1 = A_m + \Delta A_1 \\ A_2 = A_m + \Delta A_2 \end{cases} \tag{4.13}$$

Using a first order approximation of the $G(A)$ function around point $A=A_m$, we can express $G(A_1)$ and $G(A_2)$ as:

$$\begin{cases} G_1 = G\left(A_m + \Delta A_1\right) \cong G\left(A_m\right) + \Delta A_1 \left.\dfrac{dG}{dA}\right|_{A=A_m} \\ G_2 = G\left(A_m + \Delta A_2\right) \cong G\left(A_m\right) + \Delta A_2 \left.\dfrac{dG}{dA}\right|_{A=A_m} \end{cases} \tag{4.14}$$

$$\Delta G = G_1 - G_2 \cong \Delta A \left.\frac{dG}{dA}\right|_{A=A_m} \tag{4.15}$$

where $\Delta A = A_1 - A_2 = \Delta A_1 - \Delta A_2$.

The result represented in (4.15) is the well-known first order approximation of the relationship between the increment in the dependent variable G and the corresponding increment in the independent variable A. What we want emphasize is that the approximation shown in (4.15) is independent on the point where the derivative is calculated (point A=Am). Changing the point affects the accuracy of the approximation, but not the form of the latter.

Multi-dimensional case

In the general case, the quantity G depends on several independent variables, indicated with A, B, C and so forth. For the sake of simplicity, we will consider the case of three independent variables. We are interested at two points in the space of the independent variables and we will indicate these points as $\boldsymbol{P_1}=(A_1,B_1,C_1)$ and $\boldsymbol{P_2}=(A_2,B_2,C_2)$. Repeating the considerations made for the one-dimensional case, we can find a linear approximation of the function $G(A,B,C)$ around an arbitrary point $\boldsymbol{P_m}=(A_m,B_m,C_m)$. In the same way as in Eq. (4.13) we can express points $\boldsymbol{P_1}$ and $\boldsymbol{P_2}$. through their deviations with respect to $\boldsymbol{P_m}$:

$$\begin{cases} A_1 = A_m + \Delta A_1 \,, B_1 = B_m + \Delta B_1 \;, C_1 = C_m + \Delta C_1 \\ A_2 = A_m + \Delta A_2 \,, B_2 = B_m + \Delta B_2 \;, C_2 = C_m + \Delta C_2 \end{cases} \tag{4.16}$$

Then, the first order approximation of $G(A_1,B_1,C_1)$ and $G(A_2,B_2,C_2)$ can be written in the form:

$$\begin{cases} G_1 = G(A_m,B_m,C_m) + \Delta A_1 \left.\dfrac{\partial G}{\partial A}\right|_{P_m} + \Delta B_1 \left.\dfrac{\partial G}{\partial B}\right|_{P_m} + \Delta C_1 \left.\dfrac{\partial G}{\partial C}\right|_{P_m} \\ G_2 = G(A_m,B_m,C_m) + \Delta A_2 \left.\dfrac{\partial G}{\partial A}\right|_{P_m} + \Delta B_2 \left.\dfrac{\partial G}{\partial B}\right|_{P_m} + \Delta C_2 \left.\dfrac{\partial G}{\partial C}\right|_{P_m} \end{cases} \tag{4.17}$$

The difference $\Delta G=G_1-G_2$ can then be easily obtained from (4.17) as a function of the deviations $\Delta A=A_1-A_2$, $\Delta B=B_1,B_2$ and $\Delta C=C_1-C_2$.

$$\Delta G = G_1 - G_2 = \Delta A \left.\frac{\partial G}{\partial A}\right|_{P_m} + \Delta B \left.\frac{\partial G}{\partial B}\right|_{P_m} + \Delta C \left.\frac{\partial G}{\partial C}\right|_{P_m} \tag{4.18}$$

The advantage of using an arbitrary point for the calculation of the derivatives is that we can use the point that is more advantageous for the particular situation. Clearly, for the approximation to be accurate enough, point $\boldsymbol{P_m}$ should be close to both $\boldsymbol{P_1}$ and $\boldsymbol{P_2}$. As already stated for the one-dimensional case, Pm may be one of the two points $\boldsymbol{P_1}$ and $\boldsymbol{P_2}$, or $(\boldsymbol{P_1}+\boldsymbol{P_2})/2$. Another possible choice that can be convenient in some occasions is using the nominal value for $\boldsymbol{P_m}$, since it is the only value that we know *a priori*.

Useful examples of relationships that occurs frequently

The first relationships that will be analyzed are linear relationships. The following properties, where k is a constant, can be easily demonstrated:

$$\begin{cases} G = A + B \Rightarrow \Delta G = A_1 + B_1 - \left(A_2 + B_2\right) = \Delta A + \Delta B \\ G = kA \;\Rightarrow\; \Delta G = kA_1 - kA_2 = k\Delta A \end{cases} \tag{4.19}$$

Then, linearity can be applied to deviations. After that, it is interesting to analyze the so-called posynomial function, defined by the following formula:

$$G(A,B,C) = A^{\alpha} B^{\beta} C^{\gamma} \tag{4.20}$$

where α, β and χ are constant coefficients. Applying (4.18), we find:

$$\Delta G = \alpha A_m^{\alpha-1} B_m^{\beta} C_m^{\gamma} \cdot \Delta A + \beta A_m^{\alpha} B_m^{\beta-1} C_m^{\gamma} \cdot \Delta B + \gamma A_m^{\alpha} B_m^{\beta} C_m^{\gamma-1} \cdot \Delta C \quad (4.21)$$

This is not an expression that can be easily remembered. The formula becomes much simpler and meaningful if we calculate the relative error, ΔG/G, the particular form $\Delta G/G_m$, where G_m is given by:

$$G_m = G(A_m, B_m, C_m) = A_m^{\alpha} B_m^{\beta} C_m^{\gamma} \quad (4.22)$$

With simple calculations, we find:

$$\frac{\Delta G}{G_m} = \alpha \frac{\Delta A}{A_m} + \beta \frac{\Delta B}{B_m} + \gamma \frac{\Delta C}{C_m} \quad (4.23)$$

This expression can be summarized by saying that the relative error of a posynomial dependent variable is the sum of the relative errors of the independent variables, weighted by the respective exponents. Clearly, in the case that we need to calculate the absolute error ΔG, then we can simply obtain by multiplying expression (4.23) by Gm, given by expression (4.22).

Finally, we will consider the following remarkable case:

$$G(A, B, C) = \ln\left(A^{\alpha} B^{\beta} C^{\gamma}\right) \quad (4.24)$$

Defining a variable Z equal to the argument of the logarithm in (4.24), i.e. $Z=A^{\alpha}B^{\beta}C^{\gamma}$ we can write:

$$\Delta G = \Delta Z \left.\frac{dG}{dZ}\right|_{Z=Z_m} = \frac{\Delta Z}{Z_m} \qquad with\ Z_m = A_m^{\alpha} B_m^{\beta} C_m^{\gamma} \quad (4.25)$$

Considering that *Z* is a posynomial, its relative error $\Delta Z/Z_m$ us given by (4.23), then:

$$\Delta G = \alpha \frac{\Delta A}{A_m} + \beta \frac{\Delta B}{B_m} + \gamma \frac{\Delta C}{C_m} \quad (4.26)$$

We can summarizing expression (4.26) saying that, in the case that the posynomial expression is the argument of a natural logarithm, it is the absolute error Δ*G* to be the sum of the weighted relative errors of the independent variables.

4.9 References

[1] M. J. M. Pelgrom, A. C. J. Duinmaijer and A.P.G. Welbers, "Matching Properties of MOS Transistors", *IEEE J. Solid State Circuits*, vol. 24, No. 5, pp. 1433-1440, October 1989.

[2] C. Galup-Montoro, M. C. Schneider and I. J. Loss,. "Series-parallel association of FET's for high gain and high frequency applications" *IEEE Journal of Solid-State Circuits*, vol. 29, No 9, pp. 1094-1101, September 1994.