VLSI Design 4/M - Analogue

Edward Wasige
EEE, University of Glasgow
74 Oakfield Ave., x8662
edward.wasige@glasgow.ac.uk

1 Introduction

CMOS (complimentary metal oxide semiconductor) technology has been the dominant technology for fabricating integrated circuits (ICs or chips). The process of creating an IC by combining thousands of transistors into a single chip is *very-large-scale-integration (VLSI)*. CMOS is reliable, manufacturable, low power, low cost, and perhaps most importantly, scalable. For several decades the evolution of integrated circuits has followed *Moore's law*, according to which the number of transistors per square millimetre of silicon doubles every 18 months. At the same time transistors have become faster, making possible ever-increasing clock rates in digital circuits. This trend seems set to continue for at least another decade without slowing down. Thus, in the foreseeable future, the processing power of digital circuits will continue to increase at an accelerating pace. The gate lengths of early CMOS transistors were in the micrometer range (long-channel devices), the feature sizes of the current CMOS devices are in the nanometer range (short-channel devices).

For analog circuits the evolution of technology is not as beneficial. Thus, there is a trend to move signal prcessing functions from the analog domain to the digital one, which, besides allowing for higher levels of accuracy, provides savings in power consumption and silicon area, increases robustness, speeds up the design process, brings flexibility and programmability, and increases the possibilities for design re-use. In many applications the input and output signals of the system are inherently analog, preventing all-digital realisations; at the very least a conversion between analog and digital is needed at the interfaces. Typically, moving the analog-digital boundary closer to the outside world increases the bit rate across it.

In telecommunications systems the trend to boost bit rates is based on employing wider bandwidths and a higher signal-to-noise ratio. At the same time radio architectures in many applications are evolving towards software-defined radio, one of the main characteristics of which is the shifting of the analog-digital boundary closer to the antenna. Because of these trends, there is an urgent need for data converters with increasing conversion rates and resolution. A part of

this needed performance upgrade comes with the technology evolution, but often the demand is higher than this alone can provide. Thus, there is still room and need for innovations in circuit design.

The increasing integration level leads to systems with a smaller number of chips, the ultimate goal being a single chip solution, the system on a chip (SoC). This means that analog and digital circuits have to live on the same silicon die, which brings additional challenges in analog design, such as mixed signal issues and limitations in the choice of technology. Data converters are inherently mixed-signal circuits and face the same challenges on a smaller scale even without going as far as SoC. Furthermore, the evolution of technology has been driven by the microprocessor industry and hence does not always go in the best direction for analog. However, the recent rapid growth of the wireless telecommunications devices market has given a boost to the development of advanced mixed signal technologies, such as silicon germanium-based BiCMOS.

The main challenges in data converter design are decreasing supply voltage, short channel effects in MOS devices, mixed signal issues, the development of design and simulation tools, and testability. In analog-to-digital converters (ADCs), they need to be met at the same time as the requirements for sampling, linearity, conversion rate, resolution, and power consumption are becoming tighter. This course will address these issues, including analysis of the circuit toplogies of the various required components.

2 Basic semiconductor concepts

2.1 The pn junction

A semiconductor is a crystal lattice structure that can have free electrons and/or free holes (which are an absence of electrons and are equivalent to positive carriers). Silicon is the most common type of semiconductor material. This material has a valence of four, implying that each atom has four free electrons to share with neighbouring atoms when forming the covalent bonds of the crystal lattice. *Intrinsic* silicon, i.e. undoped silicon, is a very pure crystal structure having an equal number of electrons and holes. These free carriers are due to those electrons that have gained enough energy due to thermal agitation to break from their bonds. At room temperature there are $\sim 1.5 \times 10^{10}$ carriers of each type per cm^3. The number of carriers approx doubles for every 11°C increase in temperature.

If one dopes silicon with a pentavalent impurity, i.e. atoms of an element having a valence of five, or equivalently, five electrons in the outer shell available when bonding with neighbouring atoms, there will be an extra free electron for every impurity atom (assuming none recombines with the holes). These free electrons can be used to conduct current. A pentavalent impurity (e.g. P or As) is said to *donate* free electrons to the silicon crystal, and so the impurity is known as a *donor*. These impurities are also called n-type dopants since the free carriers resulting from their use have a negative charge. Similarly, if one dopes silicon with atoms having a valence of three, e.g. with boron (B), each of the impurity atoms *accepts* one one electron from the silicon crystal so that they may form covalent bonds in the lattice structure. Thus each impurity atom results in a vacancy or hole of positive charge, i.e. a trivalent impurity is an *acceptor* or a p-type dopant. It should emphasized that a piece of n-type or p-type silicon is electrically neutral; the majority free carriers (electrons in n-type silicon and holes in p-type silicon) are netralized by *bound* charges associated with the impurity atoms.

A pn junction or diode is a semiconductor with one part doped n-type adjacent to another part doped p-type (Fig.1). The large number of free positive carriers, holes, in the p side will tend to diffuse into the n side, where the free electrons in the n side will tend to diffuse to the p side. As the two types of carriers diffuse together, they recombine. Every electron that diffuses from the n side to the p side leaves behind a *bound* positive charge close to the transition region. Similarly, every hole that diffuses from the p side leaves behind a bound electron near the pn interface. This results in a region without mobile carriers, a *depletion or space-charge region*, at the pn interface (Fig.2a). The charges on both sides of the depletion region cause an electric field to be established across the region; hence a potential difference results across the region, with the n side at a positive voltage relative to the p side (Fig.2b). This field opposes further diffusion of holes into the n region and electrons into the p region, i.e. the voltage drop across the depletion region acts as a *barrier*. The bound charges are like two plates of a capacitor and so also give rise to a parasitic capacitance

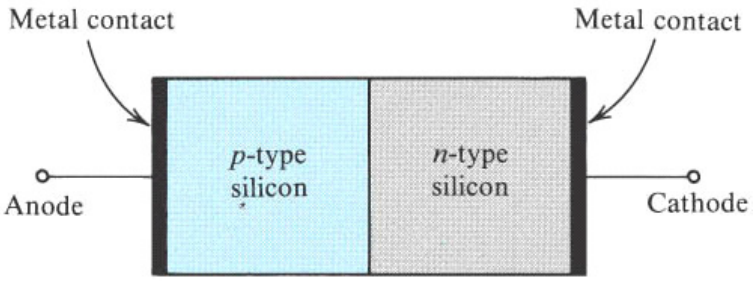

Figure 1: Simplified physical structure of the junction diode.

called the depletion or junction capacitance.

If the *pn* junction is reverse biased, i.e. the *p* side connected to the negative terminal and the *n* side to the positive terminal of a DC source (a *p* side to *n* side voltage of V_0 or less), the depletion region increases and no appreciable current flows. On the other hand, a positive voltage applied from the *p* side to the *n* side of a diode reduces the electric field opposing the diffusion of free carriers across the depletion region. The width of the depletion region is also reduced. If the forward voltage is large enough (a *p* side to *n* side voltage of V_0 or higher), the carriers will start to diffuse across the junction, resulting a current flow from the *p* to the *n* side. For Silicon, appreciable diode current starts to occur for a forward-bias voltage around 0.5V. An idealised IV characteristic of a pn junction is shown in Figure 3.

2.2 MOS Field-Effect Transistors (MOSFETs)

CMOS circuits normally use two complementary types of transistors - *n*-channel and *p*-channel. *n*-channel devices conduct with a positive gate voltage, while *p*-channel devices conduct with a negative gate voltage. Electrons are used to conduct current in *n*-channel transistors, while holes are used in *p*-channel transistors. Figure 4 shows the physical structure of the n-channel enhancement-type MOSFET. The transistor is fabricated on a *p*-type substrate, which is a single-crystal silicon wafer that provides physical support for the device (and for the entire circuit in the case of an integrated circuit). Two heavily doped *n*-type regions, n^+ source and n^+ drain regions, are created in the substrate. A thin layer of silicon dioxide (SiO_2) of thickness t_{ox} (typically 2-50nm), which is an excellent insulator, is grown on the surface of the substrate, covering the area between the source and drain regions. Metal is deposited on top of the oxide layer to form the gate electrode of the device (in fact, modern MOSFETs are fabricated using a process known as silicon-gate technology in which polysilicon, which is heavily doped noncrystalline (or amorphous) silicon, is used to form the gate electrode). Metal contacts are also made to the source region, the drain region, and the substrate, also known as the body. The gate is electri-

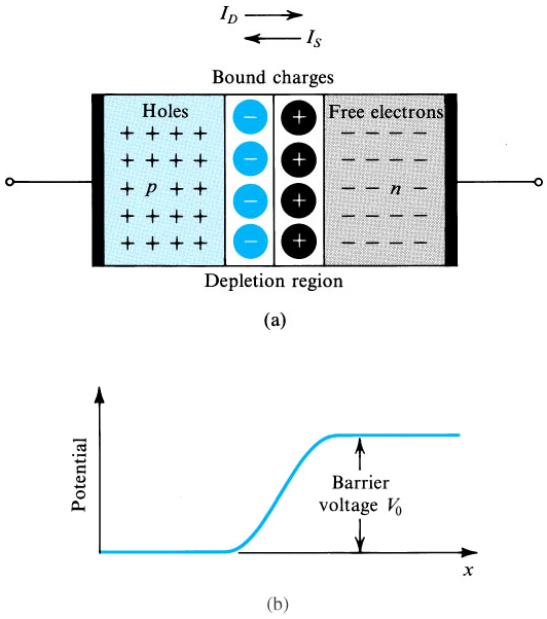

Figure 2: (a) The *pn* junction with no applied voltage (open-circuited terminals). (b) The potential distribution along an axis perpendicular to the junction.

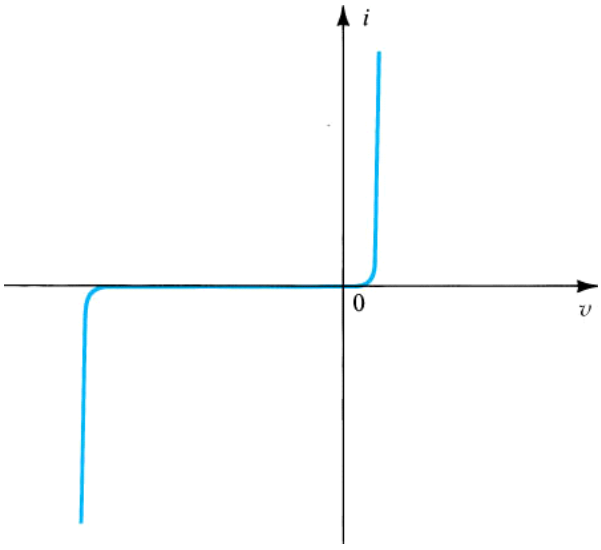

Figure 3: The IV characteristic of a silicon junction diode.

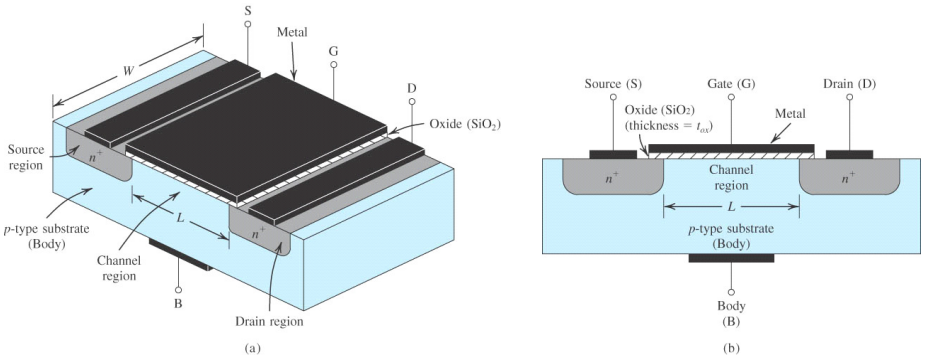

Figure 4: Physical structure of the enhancement-type NMOS transistor: (a) perspective view; (b) cross-section. Typically L = 0.1 to 3 μm, W = 0.2 to 100 μm, and the thickness of the oxide layer (t_{ox}) is in the range of 2 to 50 nm.

cally isolated from the channel by the SiO_2 and affects the channel (and hence, the transistor current) only through electrostatic coupling, similar to capacitive coupling. Therefore, the gate never conducts dc current. Normally, the p substrate is connected to the most negative voltage in the circuit which in analog circuits might be the negative power supply, but in digital circuits it is normally ground or 0 V. This connection results in all transistors placed in the substrate being surrounded by reverse-biased junctions, which electrically isolates the transistors and thereby prevent conduction through the substrate between transistors.

2.2.1 Basic operation

With no bias voltage applied to the gate, two back-to-back diodes exist in series between drain and source. One diode is formed by the pn junction between the n^+ drain region and the p-type substrate, and the other diode is formed by the pn junction between the p-type substrate and the n^+ source. These diodes prevent current conduction from drain to source when a voltage V_{DS} is applied. Now consider the case where the source, drain and substrate are all connected to ground, Fig.5. In this case, the MOS transistor operates similarly to a capacitor. The gate acts as one plate of the capacitor, and the surface of the silicon, just under the thin insultaing SiO_2, acts as the other plate. For small positive gate voltages, the positive carriers in the channel under the gate are initially repulsed and the channel changes from a p-doping level to a depletion region. As a more positive gate voltage is applied, the gate attracts negative charge from the source and drain regions, and the channel becomes an n region with mobile electrons connecting the drain and source regions. In short, a sufficiently large positive gate-source voltage changes the channel beneath the gate to an n region, and the channel is said to be *inverted*. The gate-to-source voltage, V_{GS}, for which

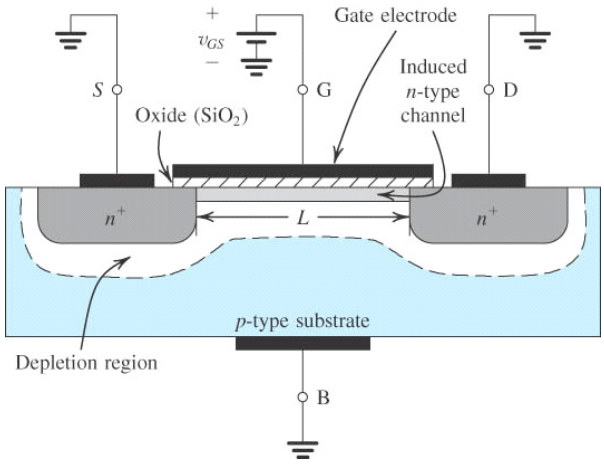

Figure 5: The enhancement-type NMOS transistor with a positive voltage applied to the gate. An n channel is induced at the top of the substrate beneath the gate.

the concentration of electrons under the gate is equal to the concentration of holes in the p substrate far from the gate is known as the *threshold voltage*, V_{tn}, and typically lies in the range 0.5 V to 1.0 V.

When $V_{GS} > V_{tn}$ the channel is present. As V_{GS} is increased, the density of electrons in the channel increases. The carrier density, and therefore the charge density, is proportional to $V_{GS} - V_{tn}$, which is often called the *effective* or *overdrive* gate-source voltage, V_{eff}.

$$V_{eff} = V_{GS} - V_{tn} \tag{1}$$

The charge density of the electrons is then given by

$$Q_n = C_{ox} \left(V_{GS} - V_{tn} \right) \tag{2}$$

Here, C_{ox}, is the gate capacitance per unit area and is given by

$$C_{ox} = \frac{K_{ox} \varepsilon_0}{t_{ox}} \tag{3}$$

where K_{ox} (the symbol ϵ_r is used in other branches of electronics) is the relative permittivity of SiO_2 (≈ 3.9) and t_{ox} is the thickness of SiO_2. The total gate capacitance, C_{gs}, is thus given by

$$C_{gs} = C_{ox} W L \tag{4}$$

where WL is the effective gate area, i.e. W is the gate width and L is the effective gate length. The gate capacitance is one of the major load capacitances that

circuits must be capable of driving. It is also important because when a MOS transistor is being turned off, the channel charge must flow from under the gate out through the terminals to other places in the circuit.

Next, with $V_{GS} > V_{tn}$, if the drain voltage V_{DS} in increased above 0 V, a drain-source potential difference exists which results in a current flowing from drain to source, I_D. The relationship between I_D and V_{DS} is the *same as for a resistor*, assuming V_{DS} is small, and is given by

$$I_D = \mu_n Q_n \frac{W}{L} V_{DS} \tag{5}$$

where $\mu_n \simeq 0.06\ m^2/Vs$ is the mobility of electrons near the silicon surface, and so

$$I_D = \mu_n C_{ox} \frac{W}{L} V_{eff} V_{DS} \tag{6}$$

Note that this relationship is only valid for V_{DS} near zero, i.e. $V_{DS} \ll V_{eff}$. In this case I_D is linearly related to V_{DS}. Note also that the relationship

$$v_d = \mu_n E \tag{7}$$

has been used, where v_d is the drift velocity of electrons, μ_n is the electron mobility and E is the applied electric field.

As the drain-source voltage increases, the channel charge concetration decreases at the drain end due to the smaller gate-to-channel voltage difference as one moves closer to the drain, i.e. since the drain sits at a higher voltage (potential) than the source, there is an increasing voltage gradient from the source to the drain, resulting in smaller gate-to-channel voltage near the drain. The charge density therefore decreases across the channel as shown in Figure 6, i.e. the channel becomes more tapered and its resistance increases correspondingly. Thus the I_D-V_{DS} curve does not continue in a straight line but bends as shown in Figure 7. As the drain voltage is increased, at some point the gate-to-channel voltage at the drain end will decrease to the threshold value V_{tn} - the minimum gate-to-channel voltage needed for n carriers in the channel to exist. Thus at the drain end the channel depth decreases to almost zero, and the channel becomes *pinched off*, with pinch-off occuring at $V_{GD} = V_{tn}$ or $V_{GS} - V_{DS} = V_{tn}$ or $V_{DS} = V_{GS} - V_{tn}$. If the gate-drain voltage falls below this critical pinch-off voltage of V_{tn}, the charge concentration in the channel remains constant (to a 1st order approximation) and the drain current no longer increases with increasing V_{DS}. The drain current thus saturates at this value and the MOSFET is said to be in the *saturation* (now commonly referred to as the *active*) region of operation. The result is the current-voltage relationship shown in Fig. 7 for a given gate-source voltage. The region where I_D changes with V_{DS} is called the *triode* region. When MOS transistors are used in analog amplifiers, they almost always are biased in the active region.

It is important to point out that the channel does not become inverted suddenly, but rather gradually. In most circuit applications, non-cutoff MOSFETs are operated in *strong inversion*, with $V_{eff} > 100\ mV$ (many designers use

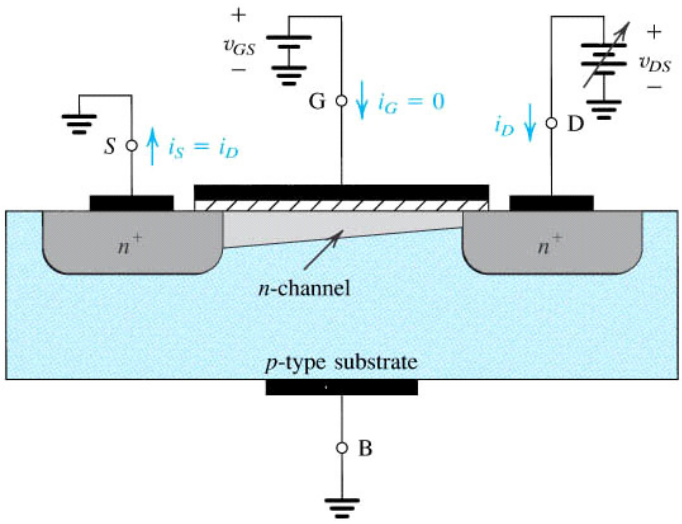

Figure 6: Operation of the enhancement NMOS transistor as V_{DS} is increased. The induced channel acquires a tapered shape, and its resistance increases as V_{DS} is increased. Here, V_{GS} is kept constant at a value $> V_{tn}$.

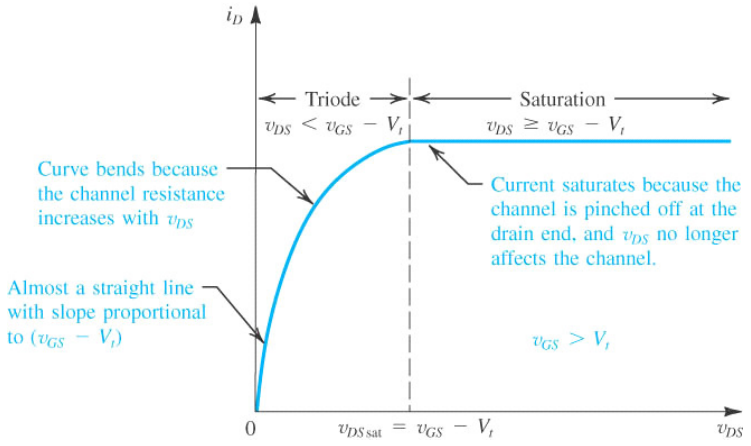

Figure 7: The drain current I_D versus the drain-to-source voltage V_{DS} for an enhancement-type NMOS transistor operated with $V_{GS} > V_{tn}$.

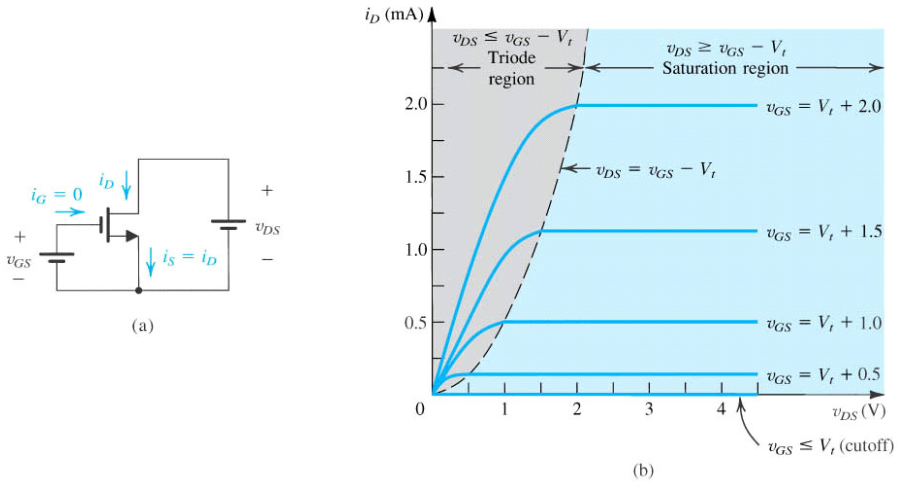

Figure 8: An n-channel enhancement-type MOSFET with V_{GS} and V_{DS} applied and with the normal directions of current flow indicated. (b) The I_D–V_{DS} characteristics for a device with k'$_n$ (W/L) = 1.0 mA/V^2.

$V_{eff} > 200\ mV$). Weak inversion occurs when V_{GS} is approx. 100 mV or less above V_{tn}.

The circuit of Figure 8a is used to determine the current-voltage characteristics of a MOSFET (Fig.8b). In the triode region, it can be shown that the drain current is given by

$$I_D = k'_n \frac{W}{L} \left[(V_{GS} - V_{tn})\, V_{DS} - \frac{V_{DS}^2}{2} \right] \tag{8}$$

where $k'_n = \mu_n C_{ox}$ is the process transconductance parameter; its value is determined by the fabrication technology. In the active region, the drain current is given by

$$I_D = \frac{k'_n}{2} \frac{W}{L} (V_{GS} - V_{tn})^2 \tag{9}$$

Exercise 1 *Draw the cross-section of a p-channel MOSFET fabricated on a p-type substrate and explain clearly the operation of this device.*

3 CMOS processing and layout

3.1 IC fabrication basics

To develop a fundamental understanding of CMOS integrated circuit design and layout, a good understanding of the fabrication processes is necessary. CMOS integrated circuits are formed by patterning different layers on and in the silicon wafer. This wafer is doped with acceptor atoms such as boron for a p-type wafer (this is the most common substrate used in CMOS processing), or donor atoms such as phosphorous for an n-type wafer. When designing CMOS ICs with a p-type wafer, n-channel MOSFETs (NMOS for short) are fabricated directly on the p-type wafer, while p-channel transistors, PMOS, are fabricated in an *n-well* (Depending on the choice of starting material (substrate), CMOS processes can be identified as *n-wel*l, *p-well* or *twin-well* processes).

The following sequence of events apply, in a fundamental way, to any layer that we need to pattern. We start out with a clean, bare wafer, as shown in Fig. 9a. The distance given by the line A to B will be used as a reference is Figs. 9b-j, which are cross-sectional views of the dashed line shown in (a) . The small box in Fig. 9a is drawn with a layout program (and used for mask generation) to indicate where to put the patterned layer.

First an oxide, SiO_2 or glass, a very good insulator is grown on the wafer. This is done using a reaction with steam, H_2O, or with O_2 alone. The oxide resulting from the reaction with steam is called a wet oxide, while the reaction with O_2 is a dry oxide. Both are called thermal oxides due to the increased temperature used during oxide growth. Next, a photosensitive resist layer is spun across the wafer (Fig. 9)[1]. After the resist is baked, the *mask* derived from the layout program, Figs. 9e and f, is used to selectively illuminate areas of the wafer, Fig. 9g. The photoresist is developed (Fig. 9h), removing the areas that were illuminated. This process illustrated here is a positive resist process because the area that was illuminated is removed. A negative resist process removes the areas of resist that were not exposed to the light.

The next step in the patterning process is to remove the exposed oxide areas (Fig. 9i). Notice that the etchant etches under the resist, causing the opening in the oxide to be larger than what was specified by the mask. Fig. 9j shows the cross-sectional view of the opening after the resist has been removed. Because creating the masks is expensive, lowering the number of masks is equated with lowering the cost of a process. Standard CMOS processes require only 10 to 12 masking levels (compared with e.g. 15 to 20 for BiCMOS).

3.2 n-Well CMOS process

The starting material for the n-well CMOS is a p-type substrate. The process begins with an n-well diffusion (Fig.10a). The n-well is required wherever p-type MOSFETs are to be placed. A thick SiO_2 is etched to expose the regions

[1] Note that the dimensions of the layers, i.e. oxide, resit and wafer are not drawn to scale. Typical wafer thickness is $\sim 500\mu m$ while thickness of grown oxide or spun resist is $\sim 1\mu m$.

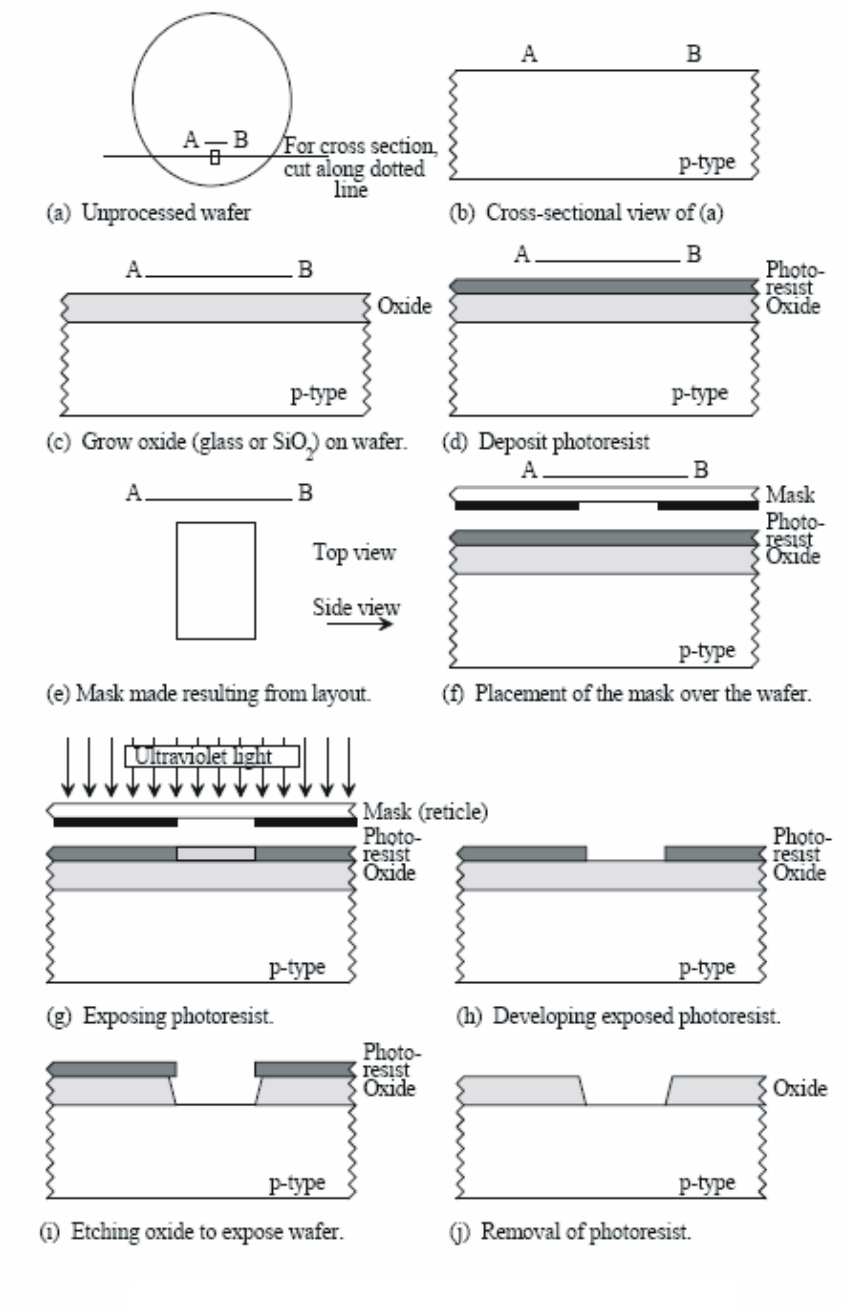

Figure 9: Generic sequence of events used in photo patterning.

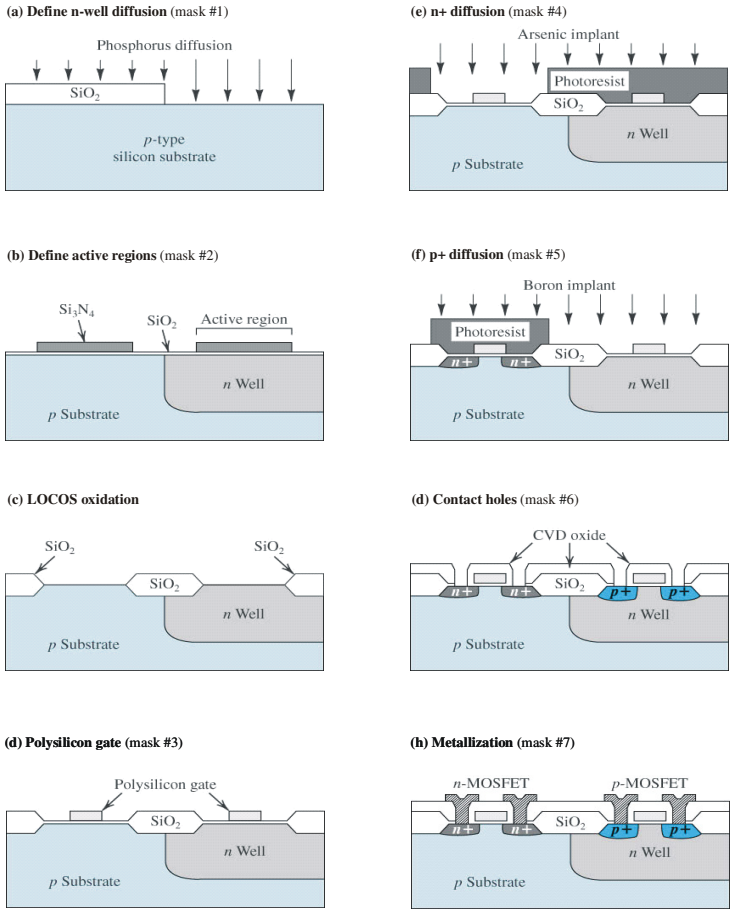

Figure 10: A typical n-well CMOS process flow.

for n-well diffusion. The unexposed regions will be protected from the n-type phosphorous impurity.

The second step is to define the active region (where transistors are to be placed) using a technique called *local oxidation* (LOCOS). A silicon nitride (Si_3N_4) layer is deposited and patterned relative to the previous n-well regions (Fig.10b). The nitride-covered regions will not be oxidized. After a long wet oxidation step, thick field-oxide will appear in regions between transistors (Fig.10c). This thick field-oxide is necessary for isolating the transistors. It also allows interconnection layers to be routed on top.

The next step is the formation of the polysilicon gate (Fig.10d). This is one of the most critical steps in the CMOS process. The thin oxide layer in the active region is first removed using wet etching followed by the growth of a high quality thin gate oxide (0.13 μm and 0.18 μm processes use oxide thicknesses

as thin as 2nm to 50nm). A polysilicon layer, usually arsenic doped (n type), is then deposited and patterned. The photolithography is most demanding in this step, since the finest resolution is required to produce the shortest possible MOS channel length.

The polysilicon gate is a *self-aligned* structure. A heavy arsenic implant can be used to form the $n+$ source and drain regions of the n-MOSFETs. The polysilicon acts as a barrier for this implant to protect the channel region. A layer of photoresit can be used to block the regions where the p-MOSFETs are to be formed (Fig.10e). The thick field-oxide stops the implant and prevents $n+$ regions forming outside the active regions. A reversed photolithography step can be used to protect the n-MOSFETs during the $p+$ boron source and drain implant for the p-MOSFETs (Fig.10f). In both cases the separation between the source and drain diffusions - the channel length - is defined by the polysilicon gate mask alone, hench the self-alignment.

A thick layer of CVD[2] oxide is next deposited over the entire wafer before contact holes are opened. A photomask is used to define the contact window opening (Fig.10g) followed by a wet or dry oxide etch. A thin aluminium layer is then evaporated or sputtered onto the wafer. A final masking and etching step is used to pattern the interconnection (Fig.10h). A final passivation step the follows prior to packaging and wire bonding in which a thick CVD oxide or pyrex glass is deposited on the wafer to serve as a protective layer. A minimum of 7 masking levels are necessary, but in practice additional levels such as n and p guards for better latch-up immunity, a second polysilicon layer for capacitors, and multilayer metals for high-density interconnections are used.

3.3 Available components: integrated devices

3.3.1 Resistors

Besides the n- and p-channel devices, other devices including pn junction diodes, MOS capacitors and resistors can be fabricated on the same wafer alongside the transistors. Resistors can be made from various diffusion regions as shown in Fig.11 (different diffusion regions have different resistivity). The $n+$ and $p+$ diffusion regions are useful for low-value resistors, while the n-well is for higher value resistors with the actual resistance defined by the length and width of diffused regions. The tolerance of the resistor value is very poor (20-50%), but matching of two similar resistor values is quite good (5%). Thus circuits should be designed to exploit resistor matching and not specific resistor values.

All diffused resistors are self-isolated by the reverse-biased pn junctions, and so exhibit a variation of resistance with bias and also include a parasitic jucntion capacitance making them unsuitable for high-frequency applications. They also exhibit a significant temperature coefficient (since carrier mobility varies with temperature).

[2] Chemical-vapour deposition (CVD) is a process by which gases or vapours are chemically reacted, leading to formation of solids on a substrate (can be used for SiO_2, Si_3N_4 and polysilicon).

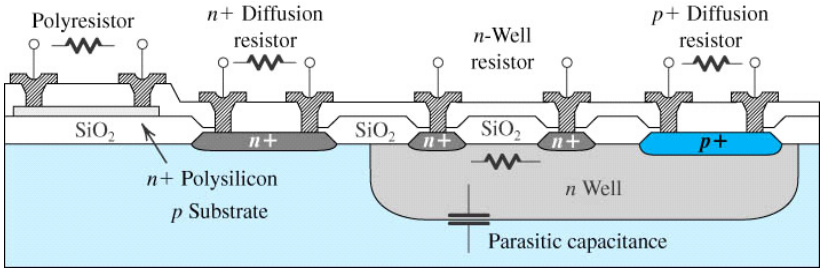

Figure 11: Cross sections of resistors of various types available from a typical n-well CMOS process.

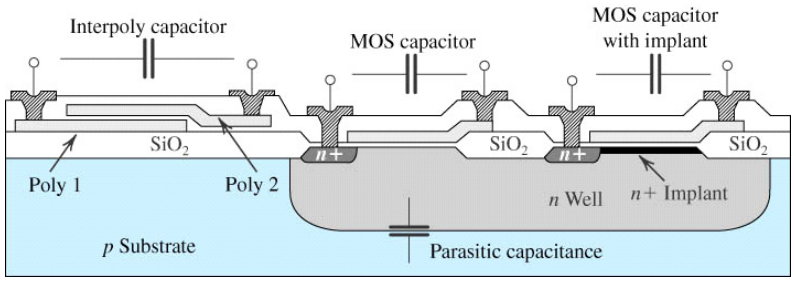

Figure 12: Interpoly and MOS capacitors in an n-well CMOS process.

Better resistors can be fabricated using the polysilicon layer that is placed on top of the thick field-oxide. The thin polysilicon provides better surface area matching and hence more accurate resistor ratios. Since this resistor (called polyresistor) is physically separated from the substrate, it much lower parasitic capacitance.

3.3.2 Capacitors

Two types of capacitors are available in CMOS processes, MOS and interpoly capacitors (see Fig. 12). The MOS gate capacitance is basically the gate-to-source capacitance. It exhibits a large voltage dependence which can be eliminated by an additional $n+$ implant as shown in the figure. MOS capacitors exhibit a large a large parasitic pn junction capacitance. The interpoly capacitor requires deposition of a second polysilicon layer but exhibits near ideal characteristics. Capacitance values can be controlled to within 1%, with values ranging from $\sim$0.5pF to a few tens of pF. Matching between similar-size capacitors can be within 0.1%.

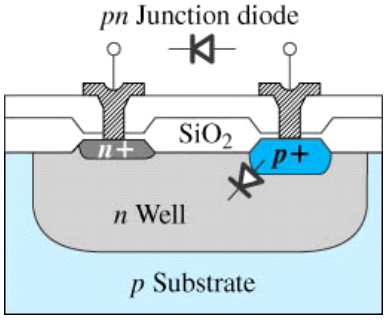

Figure 13: A pn junction diode in an n-well CMOS process.

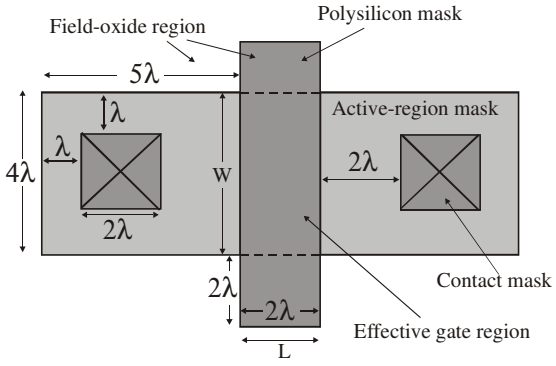

Figure 14: The layout of the active, polysilicon,and contact masks.

3.3.3 pn Junction diodes

An n-well diode is shown in Fig.13. It has high breakdown voltage, and finds use in input clamping circuits for protection against electrostatic discharge. It is also very useful as an on-chip temperature sensor (by the variation of its forward voltage drop).

3.4 CMOS Layout and design rules

Layout is the design portion of integrated-circuit manufacturing in which the geometry of circuit elements and wiring connections is defined. This process leads to the development of photographic masks used in manufacturing a microcircuit. The two most important masks are those for the active region and for the gate polysilicon. The intersection of these two masks becomes the channel region of MOS transistors (Fig.14).

The design rules for laying out transistors are often expressed in terms of a quantity, λ, where λ is half the gate length. Fig.14 shows the smallest possible

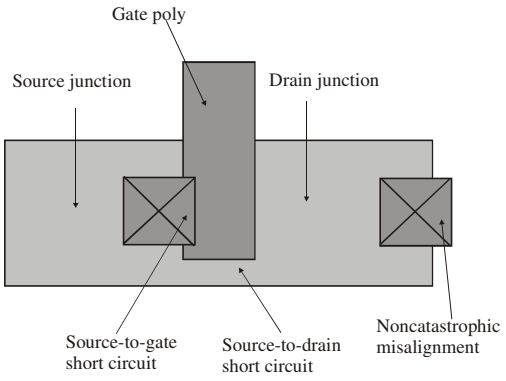

Figure 15: Mask misalignment that results in catastrophic short circuits and an example of a noncatastrophic misalignment.

transistor that can be realised when a contact is made to each junction. Also shown are many of the minimum dimensions in terms of λ. Assuming a worst-case (mis)alignment of under 0.75λ for each mask, we can guarantee that the relative misalignment between any two masks is under 1.5λ. If an overlap between any two regions of a circuit would cause a destructive short circuit, then a separation between the corresponding regions in a layout of 2λ guarantees this will never happen. For example, consider the poly mask and the contact mask. If these two regions overlap in the fabricated circuit, then the metal used to contact the source junction is also short-circuited to the gate poly, causing the transistor to be always turned off (Fig.15). To prevent this type of short, the contact openings must be kept at least 2λ away from the polysilicon gates.

Misalignment may also result in a gate that does not fully cross the active region (Fig.15). Since the active region is implanted everywhere in the active region except under the gate, this misalignment causes a short circuit between the source and drain - hence the design rule that polysilicon must always extend at least 2λ past the active region.

Another design rule is that active regions should surround contacts by at least 1λ. This guarantees an overlap between the metal contact and the $n+$ active region of at least 1.5λ (since the minimum contact wifth is 2λ). No disastrous shorts occur if overlap exists between the contacts and edge of the active region (Fig.15).

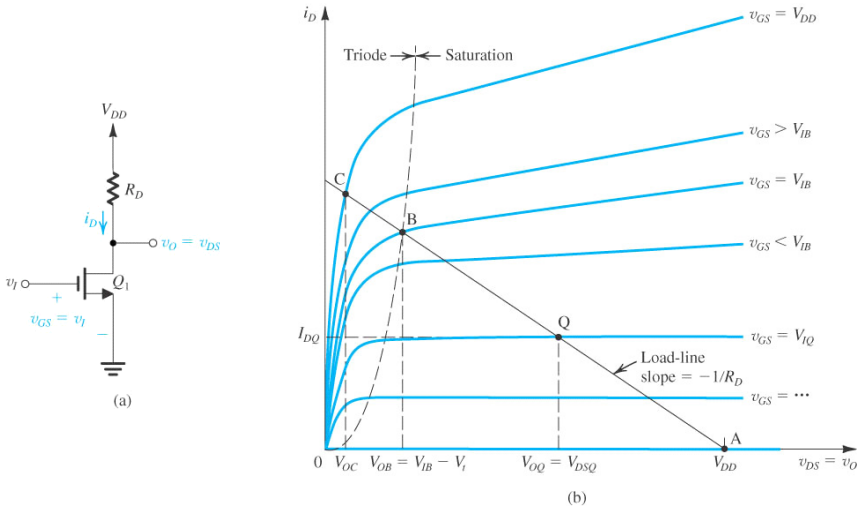

Figure 16: (a) Basic structure of the common-source amplifier. (b) Graphical construction to determine the transfer characteristic of the amplifier in (a).

4 Building blocks for analog IC design

4.1 The MOSFET as an amplifier

When operated in the active (saturation) region, MOSFETs can be used in the design of amplifier circuits. In the active region, the MOSFET acts as a voltage-controlled current source, i.e. changes in the gate-to-source voltage v_{GS} gives rise to changes in the drain current i_D. Amplification can be achieved because a small or low power signal controls a larger or higher power signal. The MOSFET can be biased to operate at a certain V_{GS} and corresponding I_D and the superimposing the voltage to be amplified, v_{gs}, on the dc bias voltage V_{GS}. By keeping "small", the resulting change in drain current, i_d, can be made proportional to v_{gs}, and so obtaining linear amplification.

Figure 16a shows the basic structure of the most commonly used MOSFET amplifier, the common-source configuration. Although the basic control action of the MOSFET is that changes v_{GS} in (here, $v_{GS} = v_I$) give rise to changes in i_D, a resistor R_D is used here to obtain an output voltage v_O,

$$v_O = v_{DS} = V_{DD} - i_D R_D \tag{10}$$

or equivalently

$$i_D = \frac{V_{DD}}{R_D} - \frac{1}{R_D} v_{DS} \tag{11}$$

i_D-v_{DS} are the transistor current and voltage, i.e. and lie on the transistor output IV characteristics. From equation (11) i_D-v_{DS} must also lie on

the straight line (of slope -1/R_D, 'y'-intercept $(0, V_{DD}/R_D)$ and 'x'-intercept $(V_{DD}, 0)$) known as the *load line.* The transistor IV curves and the load line are shown in Figure 16b.

Qualitatively, the circuit works as follows: for $v_I < V_{tn}$, the transistor will be cut off, i.e.

$$i_D = 0; \quad v_O = v_{DS} = V_{DD} \tag{12}$$

which is point A on Fig.16b. As v_I exceeds V_{tn}, the transistor turns on, i_D increases, and v_O decreases. Since v_O will initially be high, the transistor will be operating in the active region, corresponding to points along the segment of the load line from A to B. Active region operation continues until $v_I - v_O = V_{tn}$ or $v_{DS} = v_{GS} - V_{tn}$, the point at which the MOSFET enters its triode region of operation (point B on Fig.16b). At point B

$$V_{OB} = V_{IB} - V_{tn} \tag{13}$$

For $v_I > V_{IB}$, the transistor is driven deeper into the triode region. Because the characteristic curves in the triode region are bunched together, the output voltage decreases slowly towards zero. For $v_I = V_{DD}$, the operating point will be C as shown on Fig.16b, and the output voltage V_{OC} will be very small. A plot of v_I versus v_O results in a transfer characteristic shown in Fig.17. To operate the MOSFET as an amplifier the active/saturation-mode segment of the transfer curve is used. The device is located somewhere close to the middle of the curve, e.g. point Q, which is also known as the *quiescent* point. The voltage to be amplified v_i is then superimposed on the dc voltage V_{IQ}. By keeping v_i sufficiently small to restrict operation to an almost linear segment of the transfer curve, the resulting output voltage v_o will be proportional to v_i. i.e. the amplifier will be very nearly linear, and v_o will have the same waveform as v_i except that it will be larger by a factor equal to the voltage gain of the amplifier at Q.

4.2 CMOS small-signal modeling in the active region

The most commonly used small-signal model for a MOS transistor operating in the active region is shown in Fig.18. The voltage-controlled current source, $g_m v_{gs}$, is the most important component of the model, with the transistor transconductance g_m defined as

$$g_m = \frac{\partial I_D}{\partial V_{GS}} \tag{14}$$

where I_D is the drain-source current and V_{GS} the gate-source voltage. The resistor r_{ds} accounts for the finite output resistance, i.e

$$\frac{1}{r_{ds}} = g_{ds} = \frac{\partial I_D}{\partial V_{DS}} \tag{15}$$

An alternative low-frequency model, known as the T model, is shown in Fig.19.

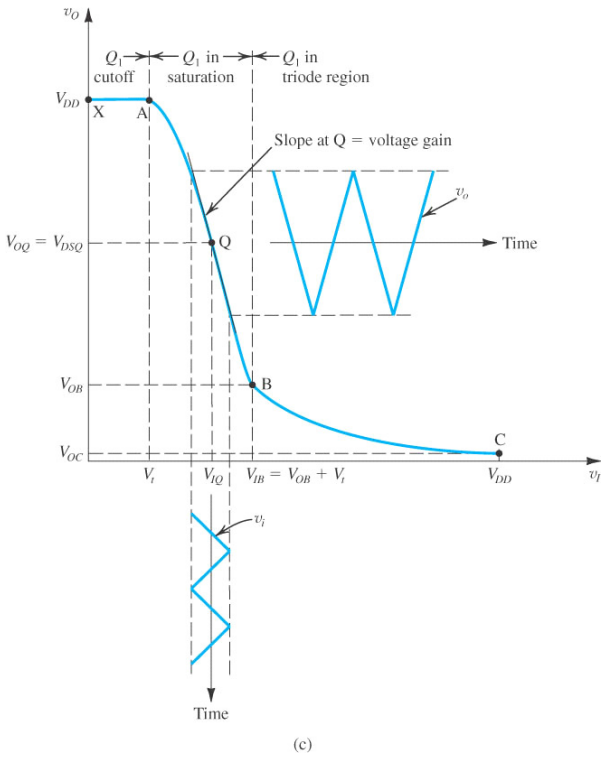

Figure 17: Transfer characteristic showing operation as an amplifier biased at point Q.

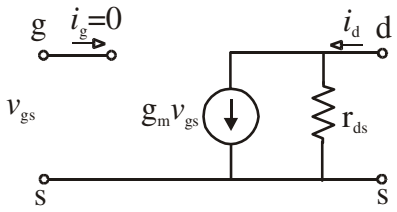

Figure 18: The low-frequency, small-signal model for a MOS transistor in the active region.

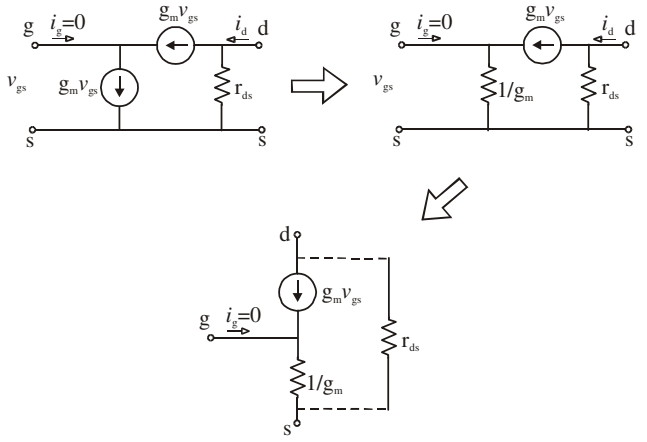

Figure 19: The small-signal, low-frequency T model for an active MOS transistor.

Its derivation from the model of Fig.18 is illustrated. Note that the drain current must always equal the source current, and, therefore, the gate current must always be zero.

4.3 Simple CMOS current mirror

A simple CMOS current mirror is shown in Fig.20, in which it is assumed that both transistors are in the active region. If both transistors are the same size, then Q1 and Q2 will have the same current since they both have the same gate voltage, i.e.

$$I_{out} = I_{in} \tag{16}$$

Note that transistor Q1 is diode connected, i.e. its drain and gate are connected. The output resistance of the current mirror, r_{out}, is given by

$$r_{out} = r_{ds2} \tag{17}$$

The current mirror active load is a way to accomplish high gain for a single stage (differential) amplifier. By using an active load, a high-impedance output load can be realised without using excessively large resistors or a large power-supply voltage. As a result, for a given power-supply voltage, a larger voltage gain can be achieved using an active load than would be possible if a resistor were used for the load.

4.4 Common-source amplifier

A common-source amplifier with an active load is shown in Fig.22. This common-source topology is the most popular gain stage, especially when high-input im-

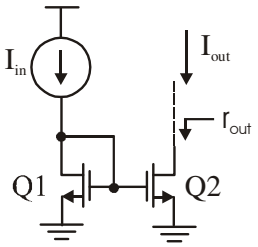

Figure 20: A simple CMOS current mirror.

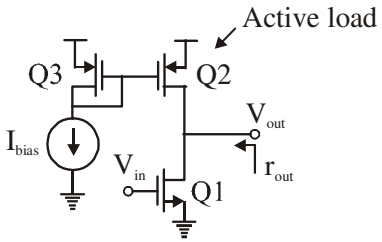

Figure 21:

pedance is desired. Here, an n-channel common-source amplifier has a p-channel current mirror used as an active load and to supply the bias current for the drive transistor. A small-signal equivalent circuit for low-frequency analysis of the common-source amplifier of Fig.22 is shown in Fig.23. V_{in} and R_{in} are the Thévenin equivalent of the input source. It is assumed that the bias voltages are such that both transistors are in the active region. Using small-signal analysis, the voltage gain as

$$A_v = \frac{V_{out}}{V_{in}} = -g_{m1}R_2 = -g_{m1}\left(r_{ds1} \parallel r_{ds2}\right) \tag{18}$$

Depending on the device sizes, currents, and the technology used, a typical gain for this circuit is in the range of -10 to -100.

4.5 Cascode gain stage

The cascode configuration consists of a common-source-connected transistor feeding into a common-gate-connected transistor (see Fig.24). Cascode stages can have quite large gain for a single stage due to the large impedances at the output. To enable this high gain, the current sources connected to the output nodes are realised using high-quality cascode current mirrors. Normally this high gain is obtained without any degradation in speed (explain? **hint:** miller effect).

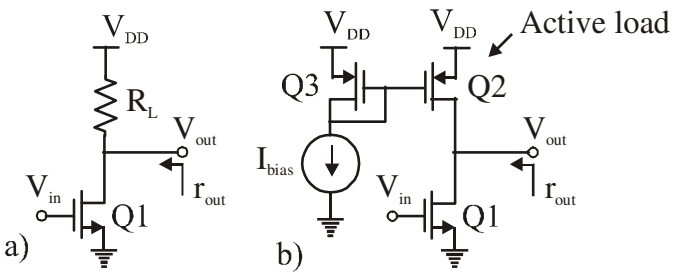

Figure 22: A common-source amplifier with a) resistive load and b) current-mirror active load.

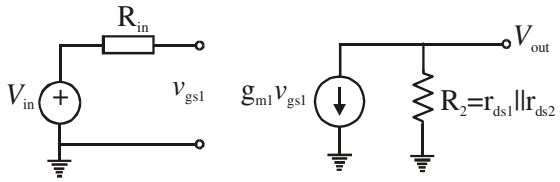

Figure 23: A small-signal equivalent circuit for the common-source amplifier (with active load)

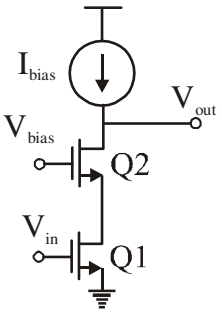

Figure 24: The cascode configuration.

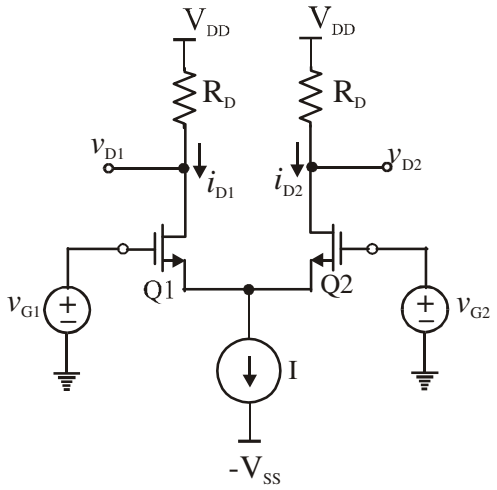

Figure 25: The basic MOS differential-pair configuration.

4.6 MOS differential pair and gain stage

The differential pair or differential amplfier configuration is the most widely used building block in analog integrated-circuit design. For example, the input stage of every op amp is a differential amplfifier. Compared to single-ended amplifiers, differential circuits are much less sensitive to noise and interference. Consider for instance two wires carrying a small differential signal as the voltage difference between the two wires. If an interference signal is coupled to the two wires (either capacitvely or inductively), the interference voltages on the two wires (i.e. between each wire and ground) will be equal. This will be so because the two wires are physically close together. Since in a differential system only the difference signal between the two wires is sensed, it will contain no interference component!

The differential configuration also enables us to bias the amplifier and to couple amplifier stages together without the need for bypass or coupling capacitors. Since large capacitors are impossible to fabricate economically, differential amplifiers are very well suited for IC fabrication. Also, the performance of the differential pair depends critically on the matching between two sides of the circuit, and IC fabriaction is capable of providing such matched devices whose parameters track over wide ranges of changes in environmental conditions. To realize this differential input, almost all amplifiers use what is commonly called a differential pair (see Fig.25). It consists of two matched transistors, Q1 and Q2, whose sources are joined together and biased by a constant-current source I. Resistive loads R_D will be used here to explain the essence of the differential pair operation, but usually active (current-source) loads are employed.

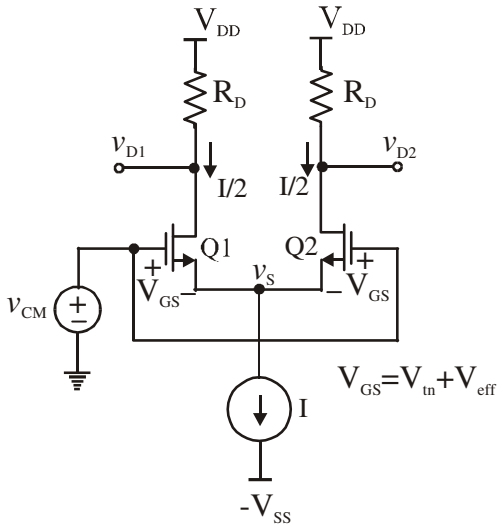

Figure 26: The MOS differential pair with a common-mode input voltage v_{CM}.

4.6.1 Operation with a common-mode input voltage

Consider first the case of the two gate terminals joined together and connected to a voltage v_{CM}, called the *common-mode voltage* (Fig.26), i.e.

$$v_{G1} = v_{G2} = v_{CM} \tag{19}$$

Since Q1 and Q2 are matched, it follows from symmetry that the current I will divide equally between the two transistors. Thus

$$i_{D1} = i_{D2} = \frac{I}{2} \tag{20}$$

and the voltage at the sources, v_S, will be

$$v_{CM} = V_{GS} + v_S \quad or \quad v_S = v_{CM} - V_{GS} \tag{21}$$

where V_{GS} is the gate-to-source voltage corresponding to a drain current $I/2$. The voltage at each drain will be

$$v_{D1} = v_{D2} = V_{DD} - \frac{I}{2}R_D \tag{22}$$

If v_{CM} is varied, the current I will divide equally between Q1 and Q2 so long as the transistors remain in the active region, and the voltages at the drains will not change. Thus the differential pair does not respond to (i.e. it *rejects*) common-mode input signals.

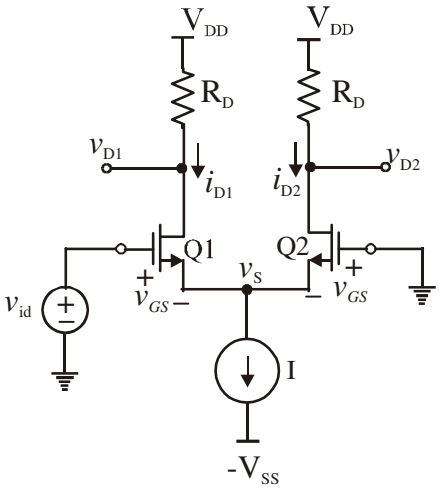

Figure 27: The MOS differential pair with a differential input signal v_{id} applied.

4.6.2 Operation with a differential input voltage

Now consider the case when a difference or differential input voltage is applied as shown in Fig.27. Since

$$v_{id} = v_{GS1} - v_{GS2} \tag{23}$$

if v_{id} is positive, $v_{GS1} > v_{GS2}$, and $v_{D1} < v_{D2}$; thus the difference output voltage $(v_{D2} - v_{D1})$ will be positive. On the other hand, when v_{id} is negative, $v_{GS1} < v_{GS2}$, and $v_{D1} > v_{D2}$; and the differential output voltage $(v_{D2} - v_{D1})$ will be negative. Thus the differential pair responds to *difference-mode* or *differential input signals* by providing a corresponding differential output signal between the two drains.

4.6.3 Operation of the MOS differential pair as a linear amplifier

Figure 28 shows the MOS differential amplifier with input voltages

$$v_{G1} = V_{CM} + \frac{1}{2}v_{id} \tag{24}$$

and

$$v_{G2} = V_{CM} - \frac{1}{2}v_{id} \tag{25}$$

where V_{CM} is a common-mode dc voltage needed to set the dc voltage of the MOSFET gates (typically $0\,V$ where two complementary supplies are utilised). Note that each of the transistors Q1 and Q2 is biased at a dc current of $I/2$ and is operating at an effective or overdrive voltage V_{eff}.

The differential input signal v_{id} is applied in a *complimentary* (or *balanced*) manner, i.e. v_{G1} is increased by $v_{id}/2$ and v_{G2} is decreased by $v_{id}/2$ (e.g. if

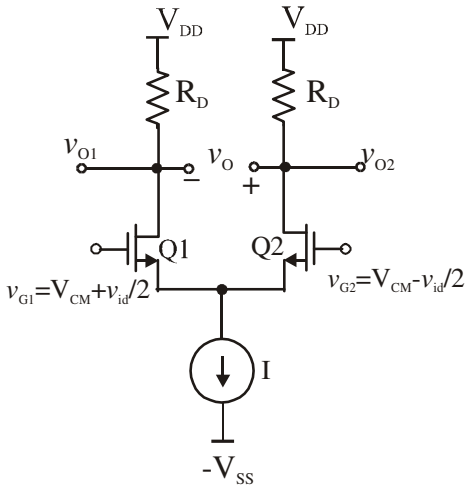

Figure 28: The MOS differential amplifier with a common-mode voltage applied to set the dc bias voltage at the gates and with v_{id} applied in a complementary (or balanced) manner.

the differential amplifier were fed from the output of another differential amplifier). Single-ended outputs v_{01} and v_{02} ride on top of dc voltages at the drains $\left(V_{DD} - \frac{I}{2}R_D\right)$, but the differential output v_0 taken between the two drains will be an entirely signal (or ac) component (having a $0\,V$ dc component).

The small-signal equivalent circuit of the differential amplifier in Fig.28 is shown in Fig.29a. From the symmetry of the circuit as well as because of the balanced manner in which v_{id} is applied, the signal voltage at the joint source connection must be zero (it acts as a virtual ground). Thus

$$v_{gs1} = v_{id}/2 \quad and \quad v_{gs2} = -v_{id}/2 \tag{26}$$

and so Q1 will have a drain current increment $g_m\,(v_{id}/2)$ and Q2 will have a drain current decrement $g_m\,(v_{id}/2)$. Thus complimentary current signals are available at the drains. These may be converted into voltages by passing them through R_D, in which case

$$v_{01} = -g_m\frac{v_{id}}{2}R_D \quad and \quad v_{02} = +g_m\frac{v_{id}}{2}R_D \tag{27}$$

and the resulting gain becomes

$$\frac{v_{01}}{v_{id}} = -\frac{1}{2}g_mR_D \quad and \quad \frac{v_{02}}{v_{id}} = \frac{1}{2}g_mR_D \tag{28}$$

if the output is taken in a single-ended fashion. Alternatively, if the output is taken differentially, the gain becomes

$$A_d = \frac{v_{02} - v_{01}}{v_{id}} = g_mR_D \tag{29}$$

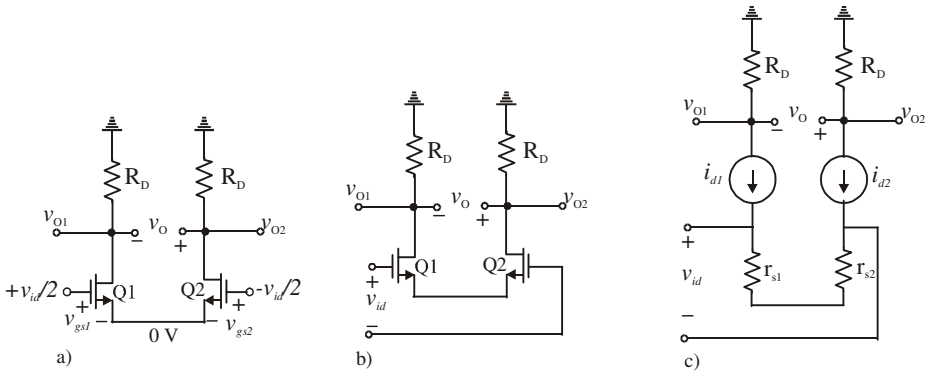

Figure 29: Small-signal equivalent circuit of the MOS differential amplifier. (a) Circuit as seen by an ac/RFor small-signal. (b) An alternative way of looking at the small-signal operation of the circuit. (c) Same circuit but with transistors replaced by their small-signal equivalent circuit T-models.

An alternative way of analysing the differential pair is shown in Fig.29b and c, in which the transistor T-model is used.

$$i_{d1} = \frac{v_{id}}{r_{s1} + r_{s2}} = \frac{v_{id}}{\left(\frac{1}{g_{m1}} + \frac{1}{g_{m2}}\right)} \tag{30}$$

Since both Q1 and Q2 have the same bias currents, $g_{m1} = g_{m2} = g_m$, and therefore

$$i_{d1} = \frac{g_m}{2} v_{id} \tag{31}$$

Also, since $i_{d2} = -i_{d1}$, then

$$i_{d2} = -\frac{g_m}{2} v_{id} \tag{32}$$

4.6.4 Differential pair with an active load (differential amplifier)

If a differential pair has a current mirror as an active load, a complete differential-input, single-ended-output gain stage can be realised as shown in Fig.30a. The differential pair is formed by transistors Q1 and Q2, loaded in a current mirror formed by transistors Q3 and Q4. If the two input terminals are connected to a dc voltage ($0\,V$), the bias current I divides equally between Q1 and Q2 (assuming perfect matching), Fig.30b. The drain current of Q1 is fed to the input transistor of the current mirror, Q3. Thus, a replica of this current is provided by Q4. Observe that at the output node, the two currents $I/2$ balance each other out, leaving a zero current to flow out to the next stage.

With a differential input v_{id} applied to the input (Fig.30c), a virtual ground develops at the common-source terminal of Q1 and Q2. Q1 will conduct a drain

signal current

$$i = g_{m1} v_{id}/2 \tag{33}$$

and Q2 will conduct an equal and opposite current i. The drain signal current i of Q1 is fed to the input of the Q3-Q4 current mirror, which responds by providing a replica in the drain of Q4. Therefore, at the output node we have two currents, each equal to i, which sum together to provide an output current $2i$. Thus

$$i_o = g_{m1} v_{id} \tag{34}$$

Since output resistance is given by

$$r_{out} = r_{ds2} || r_{ds4} \tag{35}$$

the differential gain becomes

$$A_d = g_{m1} r_{out} \tag{36}$$

A figure of merit for this amplifier is the so called *common mode rejection ratio* (CMRR)

$$CMRR \triangleq \frac{differential\, gain}{common\, mode\, gain} \tag{37}$$

$$\simeq \infty \tag{38}$$

4.7 Basic Opamp Circuit Topology

The two-stage circuit architecture has historically been the most popular approach for CMOS opamps. A block diagram of a typical two-stage CMOS opamp is shown in Fig.31. "Two-stage" refers to the number of gain stages in the opamp. Fig.31 actually shows three stages - two gain stages and a unity-gain stage. The output buffer is normally present only when resistive loads need to be driven. In a CMOS IC, opamp loads are often, but not always, purely capacitive, so the output buffer is seldom included. The first gain stage is a differential-input single ended output, often very similar to that shown in Fig.30a. The second gain stage is normally a common-source gain stage that has an active load, often very similar to that shown previously in Fig.22. Capacitor C_{cmp} is included to ensure stability when the opamp is used with feedback.

The CMOS realisation of a basic opamp is shown in Fig.32. Note that the first stage has a p-channel differential input pair with an n-channel current mirror active load. The gain of the differential stage is given by

$$A_{v1} = g_{m1} \left(r_{ds2} || r_{ds4} \right) \tag{39}$$

for the second inverter stage by

$$A_{v2} = -g_{m7} \left(r_{ds6} || r_{ds7} \right) \tag{40}$$

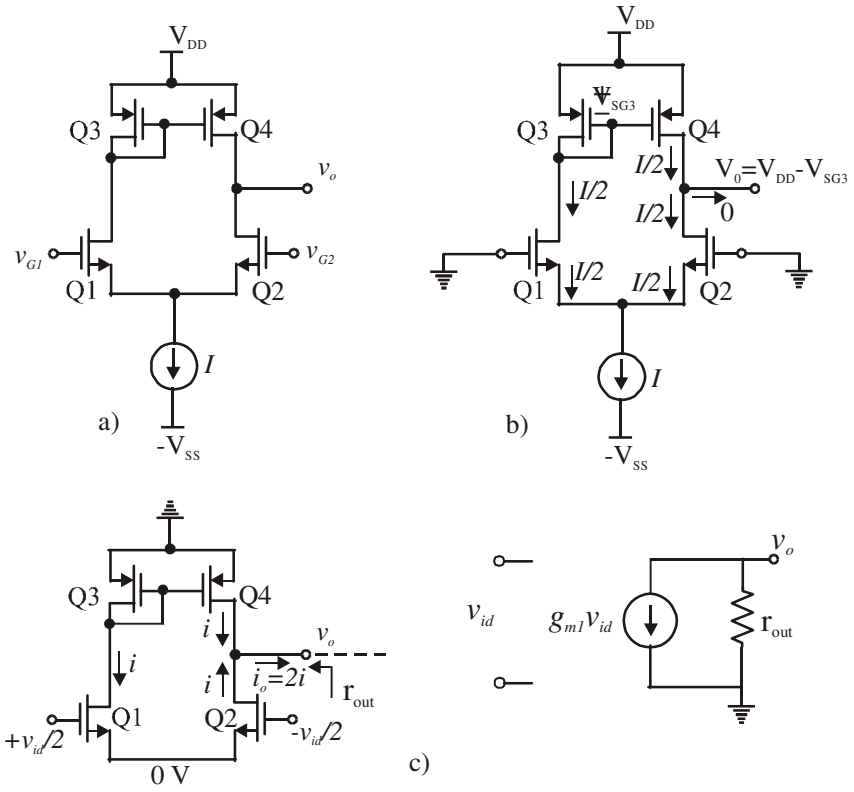

Figure 30: a) The active-loaded MOS differential pair. b) The circuit at equilibrium assuming perfect matching. c) The circuit with a differential input signal applied and its small-signal model.

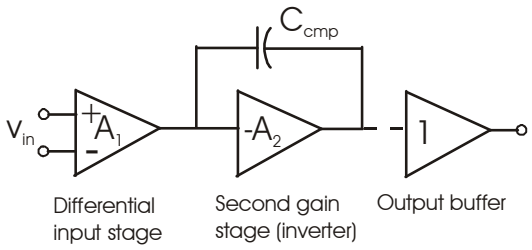

Figure 31: A block diagram of a two-stage opamp

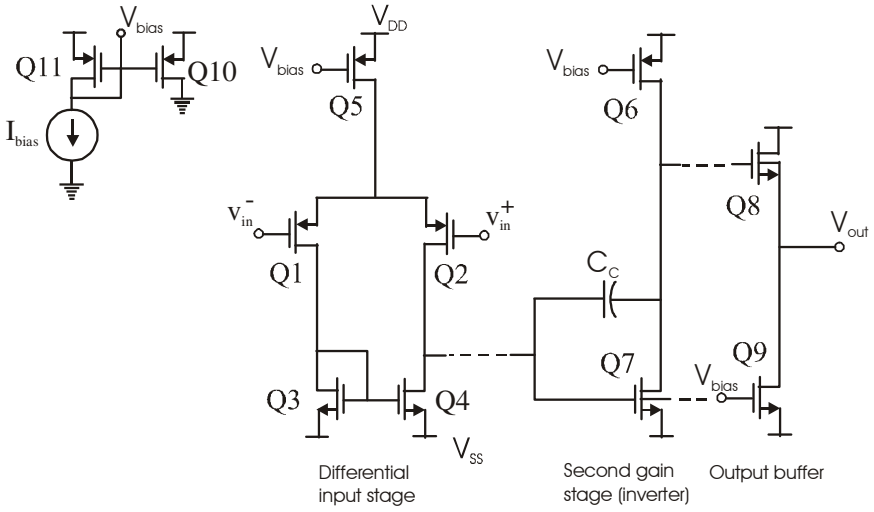

Figure 32: A CMOS realisation of a two-stage amplifier. The bias circuitry is usually more complex than shown.

The third stage is a source follower.

The two-stage opamp in Fig.32 has p-channel input transistors. It is also possible to realise a complimentary opamp where the first stage has an n-channel differential pair and the second stage is a common source amplifier having a p-channel input drive transistor. However, having a p-channel input first stage implies that the second stage has an n-channel drive transistor. This arrangement maximises the transconductance of the drive transistor of the second stage, which is critical when high-frequency operation is important. Also a n-channel source follower is preferable because this will have less voltage drop. Also, for a given power dissipation, and therefore bias current, having a p-channel input-pair maximises the slew rate. It also minimises the 1/f noise.

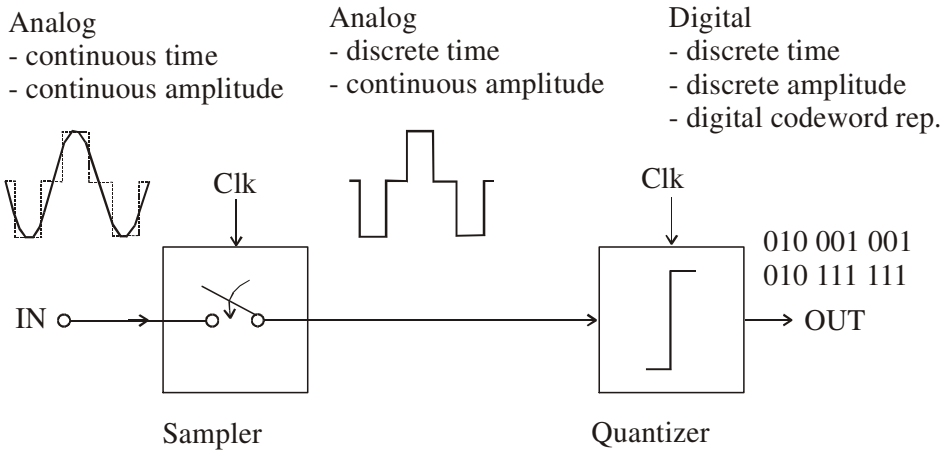

Figure 33: Principle of A/D conversion

5 Analog-to-Digital Conversion

Analog-to-digital (A/D) conversion can be separated into two distinct operations: sampling and quantization. Sampling transforms a continuous time signal into a corresponding discrete time signal, while quantisation converts continuous amplitude distribution into a set of discrete levels, which can be expressed with digital codewords. Figure 33 shows the principle of A/D conversion.

5.1 Sampling

A sample-and-hold (S/H) circuit is used to take samples of its input signal and hold these samples in its output for some period of time. Typically, the samples are taken at uniform time intervals; thus, the sampling rate (or clock rate) of the circuit can be determined. This is illustrated in Figure 34 for ideal sampling (Fig.34a) and the sample-and-hold case (Fig.34b).

Figure 35 shows an analog signal of bandwidth B (i.e. highest freqeuncy component is B Hz). An ideal S/H circuit takes samples of this signal at uniform intervals T. In the time domain this corresponds to multiplying the signal by an impluse train

$$y(t) = x(t) \cdot \sum_{n=-\infty}^{\infty} \delta(t - nT) \tag{41}$$

where $\delta(t)$ represents Dirac's delta function. The result is a train of impulses whose values correspond to the instantaneous values of the input signal. The spectrum of the sampled signal is a convolution of the input spectrum and the

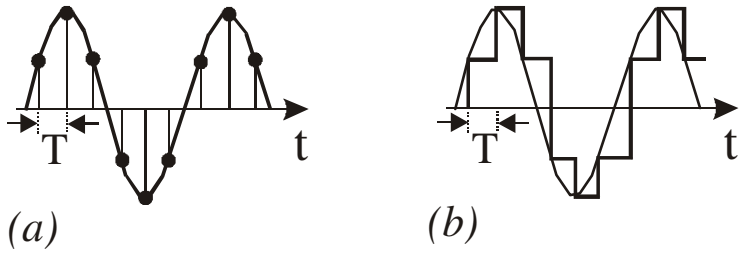

Figure 34: Sampling in time domain. Figure (a) shows the sampling instants. Figure (b) is a sample-and-held signal.

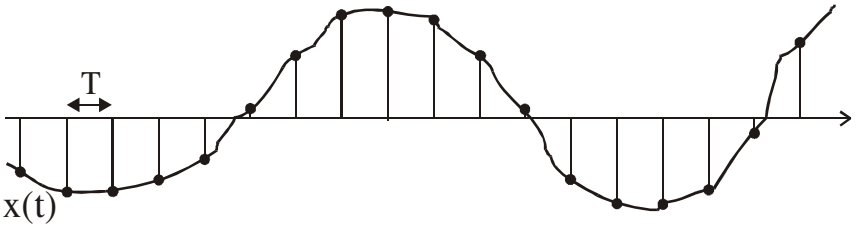

Figure 35: Sampling in time domain of signal of bandwidth B.

spectrum of the impulse train, which is also an impulse train (see Sec 5.1.1).

$$Y(f) = X(f) * \sum_{n=-\infty}^{\infty} \frac{1}{T}\delta\left(f - \frac{n}{T}\right) \tag{42}$$

This is illustrated in Figure 36, where f_s is the sampling freqency and B the signal bandwidth. The resulting spectrum is the original spectrum plus an infinite number of images of the original spectrum centered at multiples of the sampling frequency. The figure also clearly shows that as long as the bandwidth of the input signal is less than half the sampling frequency the images do not overlap and thus the original signal can be restored by filtering - this is the Nyquist criterion. If this condition is not satisfied, a part of the image is aliased (mixed with) into the desired signal, causing irreversible distortion. Because of this, the input signal usually has to be bandlimited before sampling in order to avoid the aliasing of noise and other unwanted signals present outside the desired signal band.

In practice, the output waveform of a sampling circuit cannot be a train of infinitely narrow impulses. In most practical implementations the sample is held in the output of the circuit until the next sample is taken (Fig.34). The spectrum of a sample-and-held signal is also shown in Fig.36; which in time domain is a convolution of the sampled signal with a square pulse.

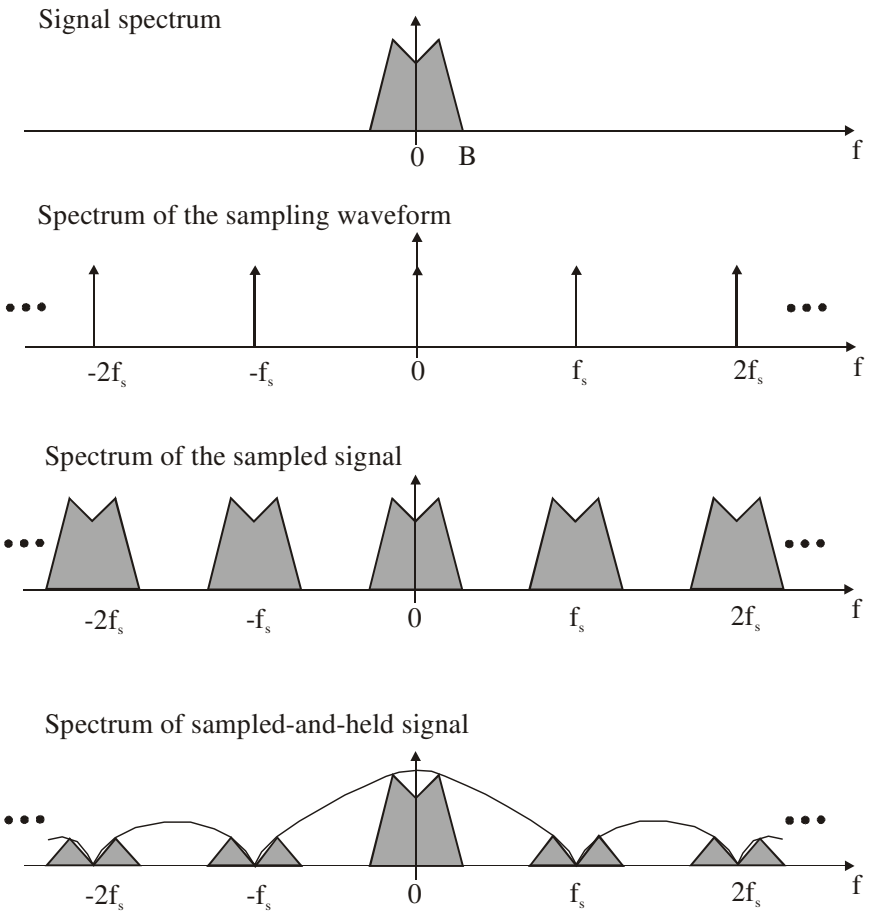

Figure 36: Spectrum of a sampled signal

5.1.1 More on impulse train

We have two expressions for a periodic impulse train,

$$s(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT) = \frac{1}{T} \sum_{n=-\infty}^{\infty} e^{j2\pi nt/T} \tag{43}$$

The Fourier transform of each expression is

$$S(f) = \sum_{n=-\infty}^{\infty} e^{-j2\pi nTf} = \frac{1}{T} \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T}\right) \tag{44}$$

Therefore, the Fourier transform[3] of a periodic impulse train in time is a periodic impulse train in frequency.

5.2 Quantization: an ideal A/D converter

The block diagram representation for an A/D converter (ADC) is shown in Figure 37a, where B_{out} is the digital output word while V_{in} and V_{ref} are the analog input and reference signals, respectively. A transfer curve for an A/D converter can be sketched as shown in Figure 37b for a 2-bit and 3-bit converter. We use the 2-bit ADC as an example. Since there are 4 digital output levels, we can only also divide the input amplitude into 4 levels. We therefore need a reference voltage V_{ref} such that if:

$$V_{in} = 0 \quad then \quad B_{out} \equiv 00$$

$$V_{in} = \frac{1}{4}V_{ref} \quad then \quad B_{out} \equiv 01$$

$$V_{in} = \frac{2}{4}V_{ref} \quad then \quad B_{out} \equiv 10$$

$$V_{in} = \frac{3}{4}V_{ref} \quad then \quad B_{out} \equiv 11$$

Since V_{in} is continuous in amplitude, we split it up as shown on the transfer curve in Fig.37b. Now defining

$$V_{LSB} \equiv \frac{V_{ref}}{2^N} \tag{45}$$

which is the change in input voltage required for the output to change the least significant bit (for an N-bit ADC). It becomes clear that there is a range of valid input values that produce the same digital output word. This signal ambiquity

[3] The Fourier transform of a function $x(t)$ is $X(f) = \int_{-\infty}^{\infty} x(t)\, e^{-j2\pi ft} dt$ and the inverse transform is given by $x(t) = \int_{-\infty}^{\infty} X(f)\, e^{j2\pi ft} df$

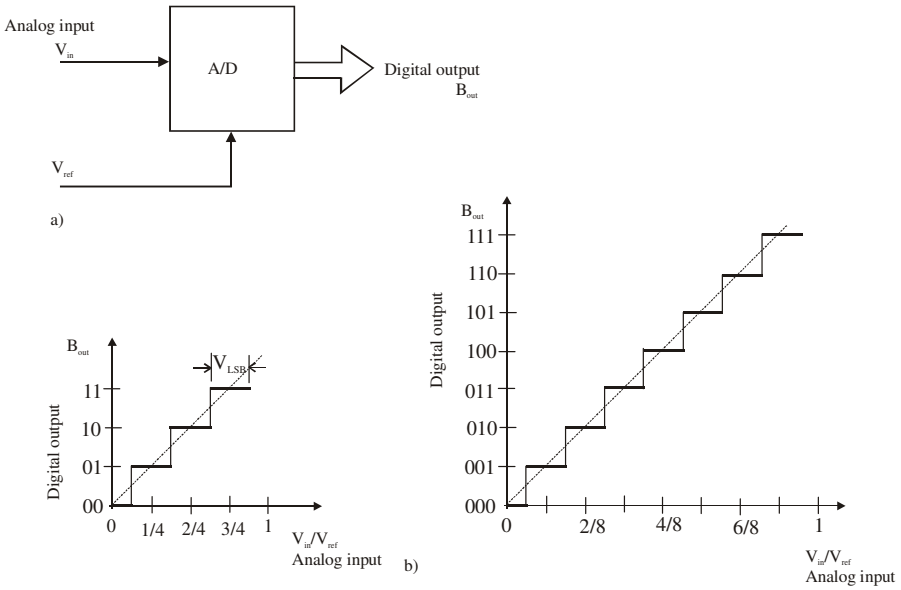

Figure 37: a) A block diagram representing an A/D converter, b) Input-output transfer curve for a 2-bit and 3-bit A/D converter.

V_x is known as *quantization error*. Note also that this signal ambiguity has a value of

$$\frac{V_x}{V_{ref}} \leq \left| \pm \frac{1}{2} \cdot \frac{1}{2^2} \right| \quad or \quad V_x \leq \left| \pm \frac{1}{2} V_{LSB} \right| \tag{46}$$

for the 2-bit converter and

$$\frac{V_x}{V_{ref}} \leq \left| \pm \frac{1}{2} \cdot \frac{1}{2^3} \right| \quad or \quad V_x \leq \left| \pm \frac{1}{2} V_{LSB} \right| \tag{47}$$

for the 3-bit converter.Therefore, for an A/D converter, the following equation relates the input and output signals

$$V_{in} \pm V_x = kV_{ref} \tag{48}$$

where k is a scaling factor determined from the digital output word (k $<$ 1), and

$$-\frac{1}{2} V_{LSB} \leq V_x \leq \frac{1}{2} V_{LSB} \tag{49}$$

with

$$V_{LSB} \equiv \frac{V_{ref}}{2^N} \tag{50}$$

and

$$V_{in,\max} = \left(1 - \frac{1}{2^N}\right) V_{ref} \tag{51}$$

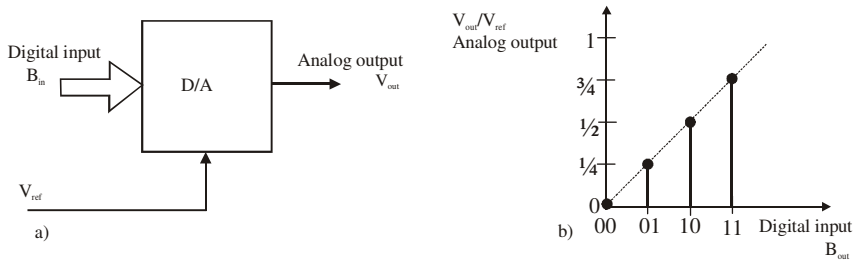

Figure 38: a) A block diagram representing a D/A converter, b) Input-output transfer curve for an ideal 2-bit DAC.

Since k < 1, conversion of an N-bit digital output word to decimal is via

$$k = \frac{b_1}{2^1} + \frac{b_2}{2^2} + \cdots + \frac{b_N}{2^N} \tag{52}$$

Exercise 2 *An 8-bit A/D converter has $V_{ref} = 5V$. What is the input voltage when $B_{out} = 10110100$?.*

5.2.1 Ideal D/A converter

Consider the block diagram of an n-bit D/A converter (DAC) shown in Fig.38a. It convertes/decodes a digital (coded) signal B_{in} into an analogue signal V_{out} according to the expression

$$V_{out} = B_{in} V_{ref} \tag{53}$$

where V_{ref} is an analogue reference voltage. For an n-bit binary DAC

$$B_{in} = \frac{b_1}{2} + \frac{b_2}{2^2} + \frac{b_3}{2^3} + \cdots + \frac{b_n}{2^n} \tag{54}$$

and so

$$V_{out} = \left(\frac{b_1}{2} + \frac{b_2}{2^2} + \frac{b_3}{2^3} + \cdots + \frac{b_n}{2^n} \right) V_{ref} \tag{55}$$

(B_{in} is a fractional binary value since $V_{out} < V_{ref}$). B_{in} can assume 2^n equally spaced values from 0 (when all bits are '0') to $1 - \frac{1}{2^n}$ (when all bits are '1'). Spacing between adjacent values is $\frac{1}{2^n}$. The DAC output is the result of multiplying the analogue input signal V_{ref} by the digital variable B_{in}. A transfer curve for a 2-bit DAC is shown in Fig.38b.

5.3 Basic Digital-to-Analog Converters

ADCs and DACs are available in a variety of architectures and technologies. In the section, we will examine the most common examples starting with DACs (since many ADCs utilize DACs in their configurations).

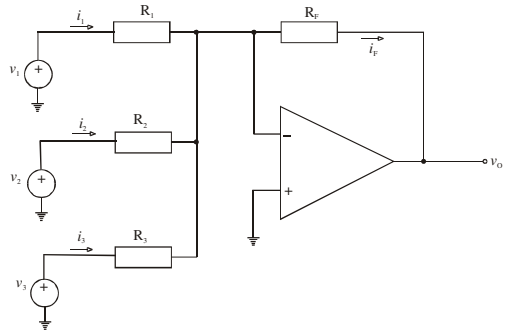

Figure 39: Summing amplifier

5.3.1 Weighted-Resistor DACs

This configuration is based on the use of operational amplifier (OPAMP) as a summing circuit. Consider an example of such a circuit as shown in Fig.39. To obtain a relationship between the output and inputs, we impose that the total current entering the virtual ground node equals that exiting it (standard assumptions on OPAMP characteristics apply, i.e. infinite input impedance and gain), or

$$i_1 + i_2 + i_3 = i_F \tag{56}$$

Using Ohm's law, we obtain

$$\frac{v_1}{R_1} + \frac{v_2}{R_2} + \frac{v_3}{R_3} = -\frac{v_O}{R_F} \tag{57}$$

and therefore

$$v_O = -R_F \left(\frac{v_1}{R_1} + \frac{v_2}{R_2} + \frac{v_3}{R_3} \right) \tag{58}$$

Though in this example three inputs have been used, this analysis can readily be generalized to an arbitrary number of them.

The weighted-resistor DAC is the simplest and most straight forward of all DACs and is based on the summing amplifier concept described above. The DAC of Fig.40 uses an OPAMP to sum n binary-weighted currents derived from V_{ref} via the scaling resistors 2R, 4R, 8R, ..., 2^nR. Whether the current $i_k = \frac{V_{ref}}{2^k R}$ appears in the sum depends on whether the corresponding switch is closed ($b_k = 1$) or open ($b_k = 0$) . The output voltage is given by (show?)

$$v_O = -\frac{V_{ref} R_F}{R} \left(\frac{b_1}{2} + \frac{b_2}{2^2} + \frac{b_3}{2^3} + \cdots + \frac{b_n}{2^n} \right) \tag{59}$$

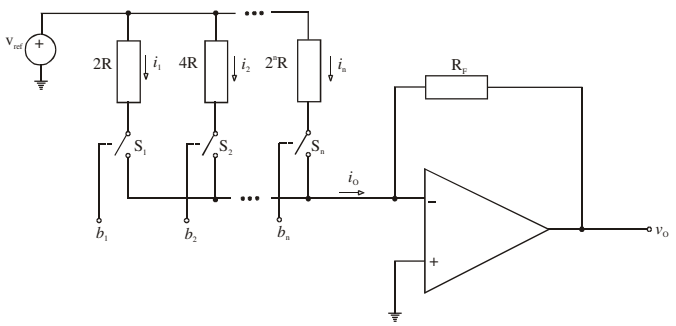

Figure 40: Weighted-resistor DAC

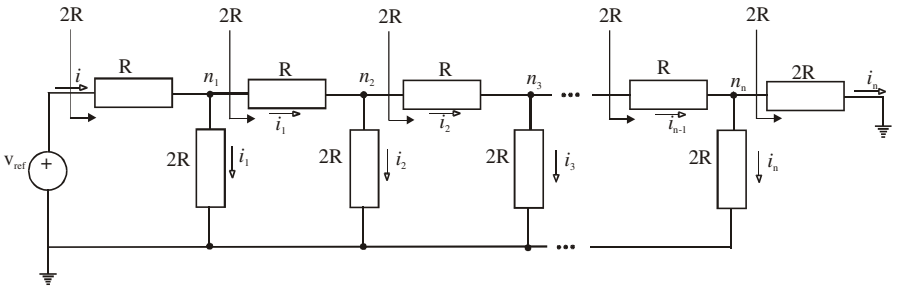

Figure 41: Resistor-ladder (R-2R) network.

5.3.2 R-2R Ladder DACs

Most DACs architectures are based on the popular R-2R ladder depicted in Fig.41. Starting from the right and working toward the left, the equivalent resistance to the right of each labelled node (n_1, n_2, ... n_n) equals 2R. Thus, the current at each node divides equally into two components, i.e. the current i sourced by V_{ref} divides along the branches as follows

$$i_1 = \frac{i}{2}; \quad i_2 = \frac{i_1}{2}; \quad i_3 = \frac{i_2}{2}; \quad \cdots \quad i_n = \frac{i_{n-1}}{2} \tag{60}$$

or

$$i_1 = \frac{i}{2}; \quad i_2 = \frac{i}{2^2}; \quad i_3 = \frac{i}{2^3}; \quad \cdots \quad i_n = \frac{i}{2^n} \tag{61}$$

where

$$i = \frac{V_{ref}}{2R} \tag{62}$$

Therefore, at each node the current decreases in a binary weighted sequence. Note that the rightmost 2R resistance serves a purely terminating function. By summing up the appropriate branch currents, this R-2R network is utilized to realise DACs. With a resistance spread of only 2-to-1 (*compare this with the weighted-resistor DACs and comment*), R-2R ladders can be fabricated monolithically to a high degree of accuracy and stability. DACs with $n \geq 12$ can be realised.

The architecture of Fig.42 utilizes the R-2R ladder network to realise a DAC. Here, the branch currents are

$$i_1 = \frac{V_{ref}}{2R} = \frac{V_{ref}/R}{2}; \tag{63}$$

$$i_2 = \frac{i_1}{2} = \frac{V_{ref}/R}{2^2}; \cdots \tag{64}$$

$$i_n = \frac{i_{n-1}}{2} = \frac{V_{ref}/R}{2^n} \tag{65}$$

and they are diverted either to the ground bus or to the virtual ground bus. Using bit b_i to identify the status of switch S_i and since

$$v_O = -R_F i_O \tag{66}$$

gives

$$v_O = -V_{ref}\frac{R_F}{R}\left(\frac{b_1}{2} + \frac{b_2}{2^2} + \frac{b_3}{2^3} + \cdots + \frac{b_n}{2^n}\right) \tag{67}$$

5.3.3 Weighted-Capacitor DACs

Complex MOS ICs such as microcomputers require on-chip data conversion capabilities using only MOSFETs and capacitors, which are the natural components of this technology. The DAC of Figure 43 can be viewed as the switched-capacitor counterpart of the weighted-resistor DAC discussed in Sec. 3.1. Its heart is an array of binary-weighted capacitances plus a terminating capacitance equal in value to the LSB capacitance.

Circuit operation alternates between two cycles called the *reset* and *sample* cycles. During the *reset* cycle, shown in the figure, all switches are connected to ground to completely discharge all capacitors. During the *sample* cycle, switch S_0 is opened while each of the remaining switches is either left at ground or connected to V_{ref}. The capacitors thus form a potential divider with some of them connected to V_{ref} and the others to ground. The voltage across the capacitors to ground corresponds to the binary input and is connected the positive input of the OPAMP (which is connected in a voltage follower configuration). As a reminder, consider the simple capacitor circuits shown in Fig. 44. Here,

$$V_{out} = V_{ref} \frac{C_1}{C_1 + C_2} \tag{68}$$

and

$$C_{eq} = C_1 + C_2 \tag{69}$$

Thus the output voltage of the DAC in Fig 43 is given by

$$v_O = V_{ref} \frac{C_r}{C_t} \tag{70}$$

where C_r represents the sum of all capacitances connected to V_{ref}, and C_t the total capacitance of the array. Therefore, we can write

$$C_r = b_1 C + b_2 \frac{C}{2} + \cdots + b_n \frac{C}{2^{n-1}} \tag{71}$$

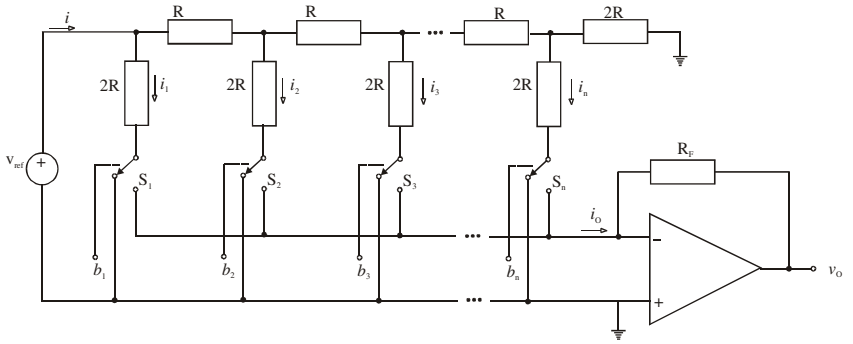

Figure 42: R-2R ladder DAC.

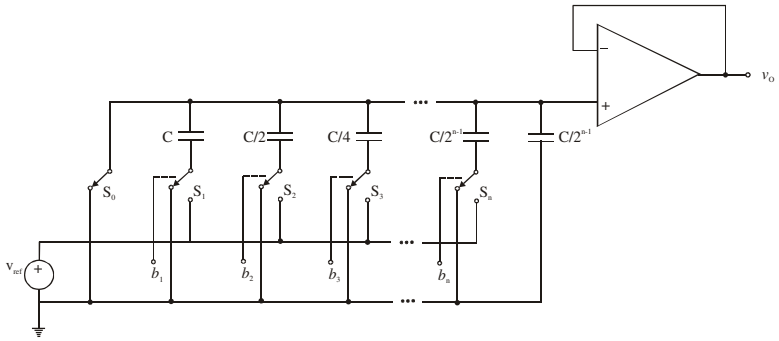

Figure 43: Weighted-capacitor DAC.

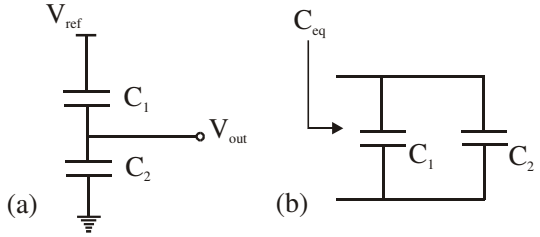

Figure 44: a) Potential division with capacitors, b) capacitors connected in parallel - equivalent capacitance equals sum of the individual capacitors.

and

$$C_t = C + \frac{C}{2} + \cdots + \frac{C}{2^{n-1}} + \frac{C}{2^{n-1}} = 2C \tag{72}$$

and so

$$v_O = V_{ref}\left(\frac{b_1}{2} + \frac{b_2}{2^2} + \frac{b_3}{2^3} + \cdots + \frac{b_n}{2^n}\right) \tag{73}$$

indicating that the sample cycle provides an n-bit D-A conversion. (Comment on the drawbacks of the DAC architecture).

5.3.4 Potentiometric DACs

One of the first integrated MOS 8-bit DACs was based on selecting one tap of a segmented resistor string by a switch network. The switch network was connected in a tree-like decoder, as in the 3-bit DAC shown in Fig.45. Notice that there will be one, and only one, low-impedance path between the resistor string and the input of the amplifier, and the path is determined by the digital input word, B_{in}:

$$B_{in} = \frac{b_1}{2} + \frac{b_2}{2^2} + \frac{b_3}{2^3} \tag{74}$$

A string of 2^n resistors partition Vr_{ef} into 2^n identical intervals. With this approach, the DAC has guaranteed *monotonicity* since any tap on the resistor string must have a lower voltage than its upper, neighbour tap. No matter how mismatched the resistors, v_0 will always increase as the amplifier is switched from one tap to the next, up the ladder, hence the inherent monotonicity. This is to be contrasted with the impact of component mismatches in the most significant bit positions of the previously discussed DACs may have on differential nonlinearity and monotoncity. Another advantage of the potentiometric DAC is that if the top and bottom nodes of the resistive string are biased at some arbitrary voltages V_H and V_L, the DAC will interpolate between V_L and V_H with a resolution of 2^n steps. However, the large number of resistors (2^n) and switches ($2^{n+1} - 2$) required limits practical potentiometric DACs to $n \leq 8$, even though the switches can be fabricated very efficiently in MOS technology.

5.3.5 Segmentation

The matching and tracking capabilities of IC components limit the resolution of the DAC structures considered so far to $n \leq 12$. However, the areas of precision instrumentation and test equipment, process control, industrial weighing systems, and digital audio playback often require resolutions and linearity performance well in excess of 12 bits. One of the most important performance requirements is monotonicity. For instance, to ensure a high signal-to-noise ratio, digital audio playback systems use 16 bits or more of differential linearity.

In conventional binary-weighted DACs, monotonicity is hardest to realize at the point of major carry due to the difficulty in realizing the required degree of match between the MSB and the combined sum of all remaining bits. To ensure monotonicity, this match must be better than one part in 2^{n-1}, indicating

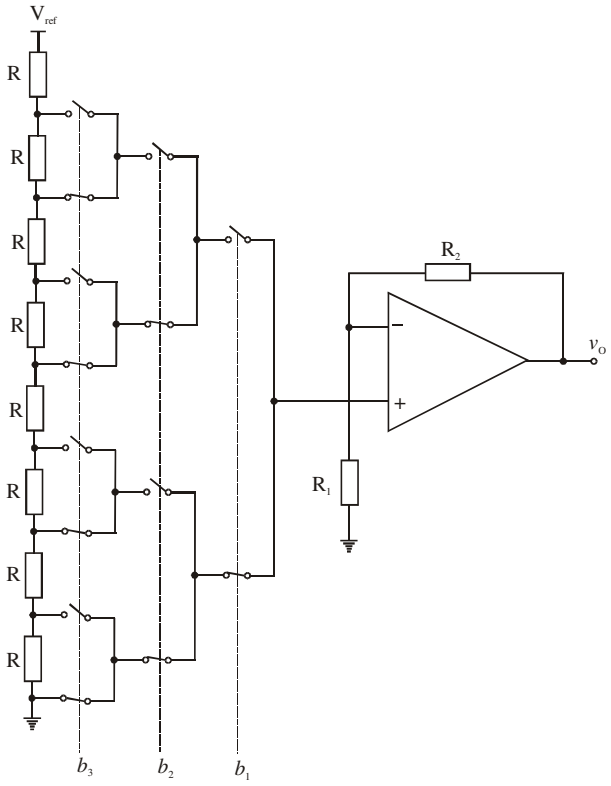

Figure 45: Potentiometric (resistor-string) DAC. An n-bit DAC uses 2^n resistors.

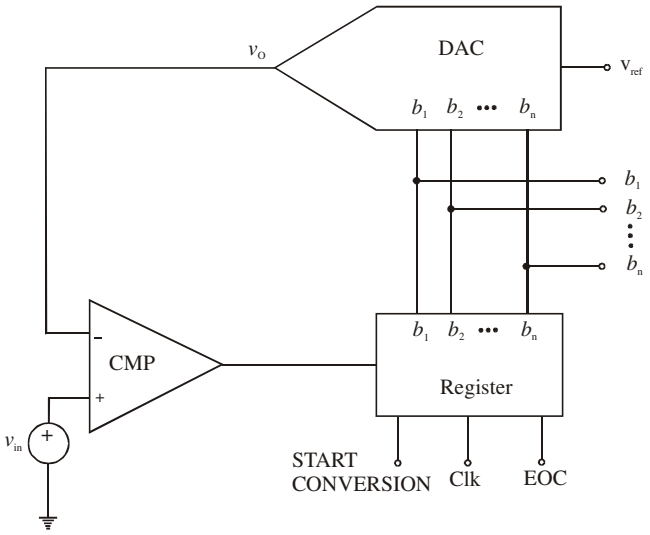

Figure 46: Functional diagram of a DAC-based ADC.

that the difficulty increases expontially with n. High-resolution DACs achieve monotonicity by a technique know as segmentation. Here the reference range is partitioned into a sufficiently large number of contigous segments, and a DAC of lesser resolution is then used to interpolate between the extremes of the selected segment (the interested student can refer to the course texts for more information).

5.4 DAC-Based A-D Conversion

A-D conversion can be accomplished by using a DAC and a suitable register to adjust the DAC's input code until the DAC's output comes within $\pm\frac{1}{2}V_{LSB}$ of the analog input. The code that achieves this is the desired ADC output b_1, b_2, ...b_n. As shown in Fig.46, this technique requires suitable logic circuitry to direct the register to perform the code search on the arrival of the START command, and a voltage comparator to announce when v_O has come within $\pm\frac{1}{2}V_{LSB}$ of v_{in} and thus issue an end-of-conversion (EOC) command. Moreover, to centre the analog range properly, the DAC output must be offset by $+\frac{1}{2}V_{LSB}$, per Fig.37.

The simplest code search is the *sequential search*, obtained by operating the register as a binary counter. As the counter steps through consecutive codes starting from 00 ...0, the DAC produces an increasing staircase, which the comparator then compares against v_{in}. As soon as this staircase reaches v_{in}, CMP fires and stops the counter. This also serves as an EOC command to notify that the desired code is sitting in the counter. The counter must be stepped at a low enough frequency to allow the DAC to settle within each clock

cycle. Considering that a conversion can take as many as $2^n - 1$ clock periods, this techniques is limited to low-speed applications.

A better approach is to allow the counter to start counting from the most recent code rather than restarting from zero. If v_{in} has not changed drastically since the last conversion, fewer counts will be needed for v_O to catch up with v_{in}. Also referred to as a *tracking* or a *servo converter*, this scheme uses the register as an up/down counter with the count direction controlled by the comparator: counting will be up when $v_O < v_{in}$, and down when $v_O > v_{in}$. Whenever v_O crosses v_{in}, the comparator changes state and this is taken as an EOC command. Clearly, conversions will be relatively fast only as long as v_{in} does not change too rapidly between consecutive conversions. For a full-scale change, the conversion will still take 2^n-1 clock periods.

The fastest code-search strategy uses binary search algorithm to complete an n-bit conversion in just n clock periods, regardless of v_{in}. Following is a description of two implementations: the *successive-approximation* and the *charge-redistribution* ADCs.

5.4.1 Successive-Approximation Converters

This technique uses the register as a *successive-approximation register* (SAR) to find each bit by trial and error. Starting from the MSB, the SAR inserts a trial 1 and the interrogates the comparator to find whether this causes v_O to rise above v_{in}. If it does, the trial bit is changed back to 0; otherwise it is left at 1. The procedure is repeated for all subsequent bits, one bit at a time.

To illustrate this consider the conversion of a 10.8V input to a 4-bit code with a $V_{ref} = 16V$. To ensure correct results, the DAC output must be offset by $+\frac{1}{2}V_{LSB}$ or 0.5V. Following the arrival of the START command, the SAR sets b_1 to 1 with all remaining bits at 0 so that the trial code is 1000. This causes the DAC output $v_O = 16\left(\frac{1}{2}\right) + 0.5 = 8.5V$. At the end of clock period T_1, v_O is compared against v_{in} and since $8.5 < 10.8$, b_1 is left at 1.

At the beginning of T_2, b_2 is set to 1, so the trial code is now 1100 and $v_O = 16\left(\frac{1}{2} + \frac{1}{2^2}\right) + 0.5 = 12.5V$. Since $12.5 > 10.8$, b_2 is changed back to zero at the end of T2.

At the beginning of T_3, b_3 is set to 1, so the trial code is now 1010 and $v_O = 10.5V$. Since $10.5 < 10.8$, b_3 is left at 1.

At the beginning of T_4, b_4 is set to 1, so the trial code is now 1011 and $v_O = 11.5V$. Since $11.5 > 10.8$, b_4 is changed back to zero. Thus when leaving T4, the SAR has generated the code 1010, which corresponds to 10.5V. Since the entire conversion takes a total of n clock cycles, a SA ADC offers a major speed improvement over a sequential search ADC.

5.4.2 Charge-Redistribution ADCs

The circuit of Fig.47 performs a successive-approximation conversion using a weighted-capacitor DAC of the type of Fig.43. Its operation involves three cycles called the *sample*, *hold*, and *redistribution* cycles. During the sample

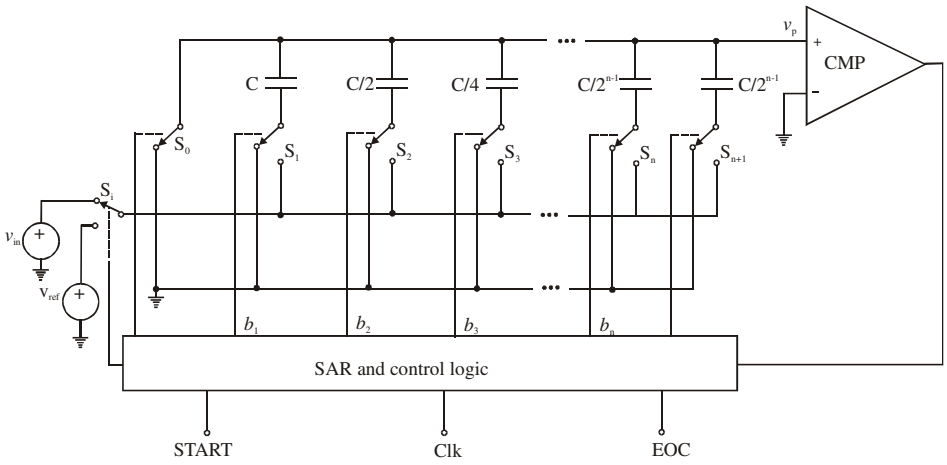

Figure 47: Charge-redistribution ADC.

cycle, S_0 grounds the top-plate bus while S_i and S_1 through S_{n+1} connect the bottom plates to v_{in}, thus precharging the entire capacitor array to v_{in}. During the hold cycle, S_0 is opened and the bottom plates are switched to ground, thus causing the top-plate voltage to swing to -v_{in}. The voltage presented to the comparator at the end of this cycle is thus $v_p = -v_{in}$. During the redistribution cycle, S_0 is still open, S_i is connected to V_{ref}, and the remaining switches are sequentially flipped from ground to V_{ref}, and possibly back to ground, to perform a successive-approximation for the desired code.

Flipping switch S_1 from ground to V_{ref} causes v_p to increase by the amount

$$v_p = -v_{in} + V_{ref}\left(\frac{C}{C + \frac{C}{2} + \frac{C}{4} + \cdots + \frac{C}{2^{n-1}} + \frac{C}{2^{n-1}}}\right) \quad (75)$$

$$= -v_{in} + \frac{V_{ref}}{2} \quad (76)$$

Similarly, flipping switch S_2 from ground to V_{ref} gives (with all other switches grounded)

$$v_p = -v_{in} + \frac{V_{ref}}{2^2} \quad (77)$$

and so, in general, flipping switch S_k from ground to V_{ref} gives

$$v_p = -v_{in} + \frac{V_{ref}}{2^k} \quad (78)$$

If it is found that this increase causes the comparator to change state, then S_k is returned to ground; otherwise it is left at V_{ref} and the next switch is tried. This procedure is repeated at each bit position, starting from the MSB and progressing down to the LSB (excluding the terminating capacitor switch,

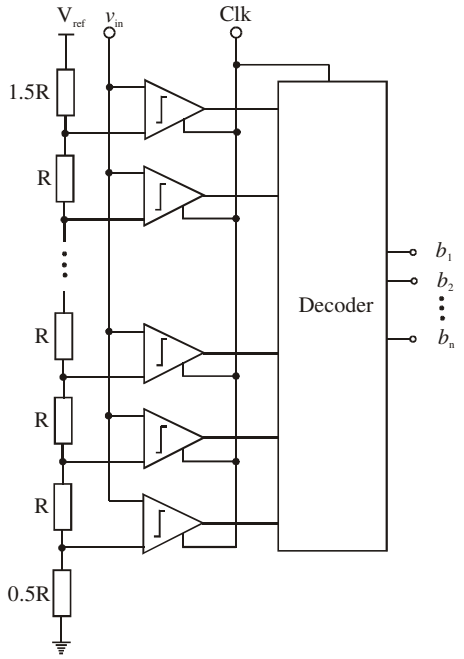

Figure 48: n-bit flash converter

which is left permanently grounded). Therefore at the end of the search the voltage presented to the comparator is

$$v_p = -v_{in} + V_{ref}\left(\frac{b_1}{2} + \frac{b_2}{2^2} + \cdots + \frac{b_n}{2^n}\right) \tag{79}$$

and that v_p is within $\pm\frac{1}{2}V_{LSB}$ of 0V. Thus the final switch pattern provides the desired output code.

Exercise 3 *Consider a charge-redistribution ADC of the type of Fig.47 with $n = 4$, $V_{ref} = 3.0V$, and $C = 8pF$. Assuming node v_p has a parasitic capacitance of 4pF toward ground, find the intermediate values of v_p during the conversion of $v_{in} = 1.00V$.*

Exercise 4 *Find the intermediate node voltages at the inverting input of the comparator during the operation of a 6-bit charge redistribution ADC. Vin=1.23V and Vref=5V. What is the final output digital word? If a parastic capacitance of C/9 exists at the comparator input, what would be the final output digital word?*

5.4.3 Flash Converters

The circuit of Fig.48 uses a resistor string to create $2^n - 1$ reference levels separated from each other by 1 V_{LSB}, and a bank of $2^n - 1$ high speed latched

comparators to simultaneously compare v_{in} against each level. Note that to position the analog signal properly, the top and bottom resistors must be 1.5R and 0.5R, as shown. As the comparators are strobed by the clock, the ones whose reference levels are below v_{in} will output a logic 1, and the remaining ones a logic 0. The result, referred to as a *bar graph*, or also as a *thermometer code*, is then converted to the desired output code b_1, b_2, ... b_n by a suitable decoder, such as a priority encoder[4].

Since the input sampling and latching take place during the first phase of the clock period, and decoding during the second phase, the entire conversion takes only one clock cycle, so this ADC is the fastest possible. The *flash converter* is therefore used in high-speed applications, such as video and radar signal processing, where conversion rates on the order of millions of samples per second are required, and successive approximation ADCs are generally not fast enough. Most high-speed oscilloscopes and some RF test instruments use flash ADCs because of their fast digitizing rate, which now reaches 5 Gsamples/s for off-the-shelf devices and 20 Gsamples/s for proprietary designs. The typical flash converter resolves analog voltages to 8 bits, although some flash converters can resolve 10 bits.The main disadvantage of this converter is the number of comparators required. *How many does an 8-bit converter require*? The large number of comparators required, associated high power dissipation, and stray input capacitance makes flash converters impractical for $n > 10$.

Exercise 5 *Consider a 3-bit flash ADC with $V_{ref} = 3\,V$. i) how many comparators does this ADC use and what are their voltage reference levels? ii) find the digital output for $V_{in} = 1\,V$.*

5.4.4 Two-step (Subranging) Converters

Two-step (or subranging) converters trade speed for circuit complexity by splitting conversion into two subtasks, each requiring less complex circuitry. Compared to flash converters, two-step converters require less silcon area, dissipate less power, have less capacitive loading, and the voltages the comparators need to resolve are less stringent than for flash equivalents. However, two-step converters do have a larger latency delay, although their throughput approaches that of flash converters.

The block diagram for a two-step ADC is shown in Fig.49. The 4-bit MSB ADC determines the first four MSBs. To determine the remaining LSBs, the quantisation error (also called the *residue*) is found by reconverting the 4-bit digital signal to an analog value using the DAC and subtracting that value from the input signal. To ease requirements in the circuitry for finding the remaining LSBs, the quantisation error is first multiplied by 2^4 using the gain amplifier, and the LSBs are determined using the 4-bit LSB ADC. Besides the sample-and-hold (SH), the DAC and the gain amplifier, the circuit uses $2\left(2^4 - 1\right) = 30$ comparators, indicating a substantial saving compared to the 255 comparators

[4] A priority encoder circuit basically converts an n-bit input into a binary representation. If the input m is active, all lower inputs (m-1 .. 0) are ignored.

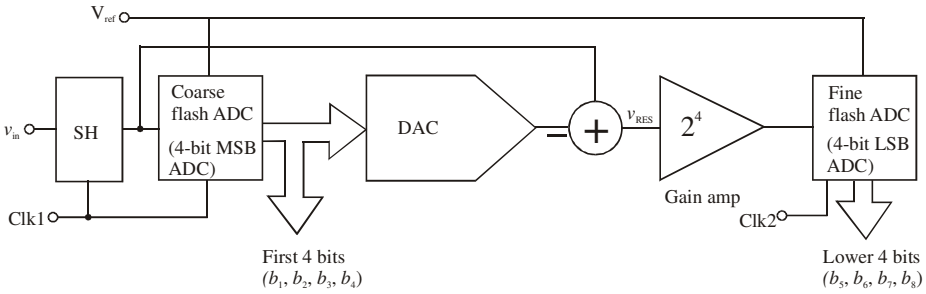

Figure 49: 8-bit subranging ADC.

required by full-flash. (This saving is even more dramatic for $n \geq 10$). The main price for this saving is a longer conversion time. Also, this approach requires that the DAC be n-bit accurate which may be a heavy requirement.

Although not as fast as a parallel ADC, subranging (also called pipelined) ADCs can digitize at speeds greater than 100 Msamples/s at 8-bit resolution. They can resolve signals to 16 bits at slower speeds. Subranging ADCs often find use in RF test equipment, lower-speed digitizing oscilloscopes, and high-end PC plug-in digitizer cards and PC-external data-acquisition systems.

5.4.5 Pipelined Converters

The two-stage architecture described in the preceeding section can be generalized to multiple stages, where each stage finds a single bit. Specifically, the first stage finds the most significant bit, b_1, the second stage finds the next bit, b_2, and so on. To avoid having to wait until residual errors ripple through the entire converter for a given sample, the approach incorporates *pipelining* such that once the first stage completes its work, it does not sit idle while the remaining lower bits are found, but immediately starts to work on the next input sample.

A block diagram of a pipelined ADC is shown in Fig.50. Each subtask stage consists of an SH, and ADC, a DAC, a subtractor, and a gain amplifier. The SH in each stage stores the input signal. This SH allows the preceeding stage to be immediately used to process its next input before the succeeding stage has finished, as long as the preceeding stage's digital output is also stored. Pipeline architectures are used in a variety of formats ($k > 1$), but with $k = 1$ this architecture results in the simplest per-stage circuitry, though n such stages are needed.

Exercise 6 *Assume the 8-bit subranging ADC of Fig.49 has* $V_{ref} = 2.560V$. *(a) Find the total number of comparators, their voltage reference levels, and the maximum level tolerances allowed for a* $\pm\frac{1}{2}V_{LSB}$ *accuracy. (b) Find* b_1 *...* b_8, v_{RES}, *and the quantization error for* $v_{in} = 0.5V$, $1.054V$, *and* $2.543V$.

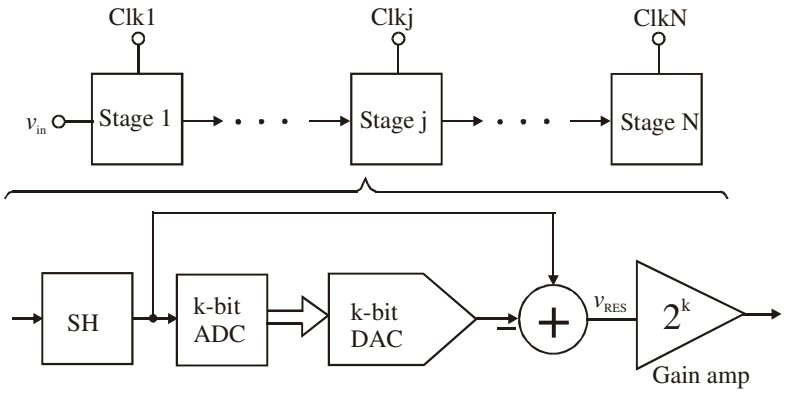

Figure 50: Pipeline ADC architecture

6 Switched-Capacitor Circuits

Switched capacitor (SC) filters simulate resistors by periodically operating MOS capacitors with MOSFET switches, and produce time constants that depend on capacitance ratios rather than RC products. Note that conventional filters produce time constants that depend on RC time constants. IC processes do not lend themselves to the fabrication of resistances and capacitances with the magnitudes (10^3 to $10^6\,\Omega$ and 10^{-9} to 10^{-6} F) and accuracies (1% or better) typically required in audio and instrumentation applications. The switched capacitor technique is useful in simulating a large value resistor, generally $> 1\,M\Omega$.

Consider the basic MOSFET-capacitor arrangement of Fig.51. The transistors are characterised by a low channel resistance (typically $<10^3\,\Omega$) when the gate voltage is high, and a high resistance (typically $>10^{12}\,\Omega$) when the gate voltage is low. With an off/on ratio this high, a MOSFET can be regarded for all practical purposes as a switch. If the gates are driven with nonoverlapping out-of-phase clock signals, the transistors will conduct on alternate half-cycles, thus providing a single-pole double throw (SPDT) switch function with make-before-break characteristics.

Assuming $V_1 > V_2$, flipping the switch to the left charges C to V_1, and flipping it to the right discharges C to V_2. The net charge transfer from V_1 to V_2 is

$$\Delta Q = C\,(V_1 - V_2) \tag{80}$$

If the switch is flipped back and forth at rate of f_{CK} cycles per second, the charge transferred in 1 second from V_1 to V_2 defines an average current

$$\begin{aligned} I_{avg} &= f_{CK}\Delta Q & (81)\\ &= C f_{CK}\,(V_1 - V_2) & (82) \end{aligned}$$

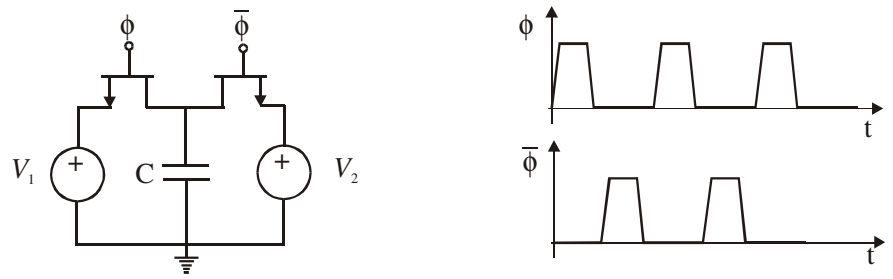

Figure 51: Switched capacitor using a MOSFET SPDT switch, and clock drive for the MOSFETs.

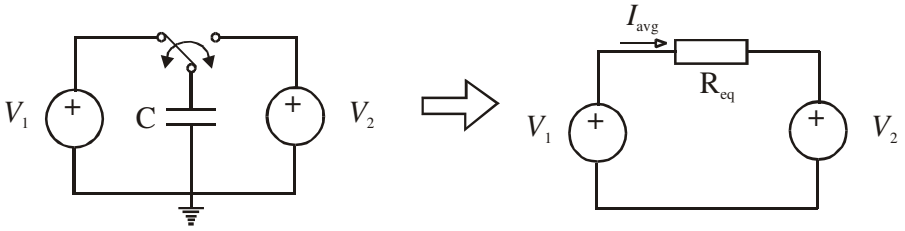

Figure 52: Resistance simulation using a switched capacitor.

Note that the charge is flowing in packets rather than continously. However, if f_{CK} is made sufficiently higher than the highest frequency components of V_1 and V_2, the process can be regarded as continuous and the switch capacitor combination can be modeled with an equivalent resistance (see Fig.52):

$$R_{eq} = \frac{V_1 - V_2}{I_{avg}} = \frac{1}{C f_{CK}} \tag{83}$$

6.1 SC integrators

We will now investigate how a 'simulated' switched capacitor resistance can be used to implement the integrator (the workhorse of active filters). The RC integrator of Fig.53a yields

$$\frac{V_{in}}{R_1} = -\frac{V_0}{1/j\omega C_2} \tag{84}$$

or

$$H(j\omega) = \frac{V_0}{V_{in}} = -\frac{1}{j\omega/\omega_0} \tag{85}$$

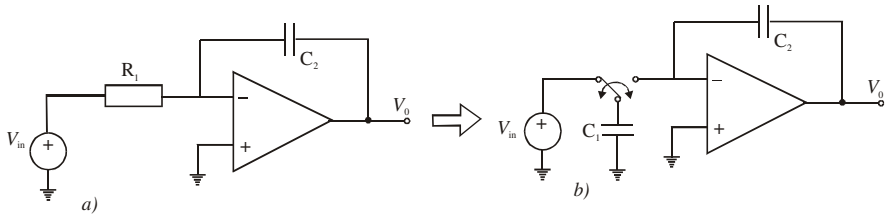

Figure 53: Converting an RC integrator to an SC integrator.

where the unity gain frequency is given by

$$\omega_0 = \frac{1}{R_1 C_2} \tag{86}$$

Replacing R_1 by an SC resistance gives the SC integrator of Fig.53b. If the input frequency is such that $\omega \ll \omega_{CK}$ then the current flow from V_{in} to the summing node can be regarded as continuous, and R_1 would be given by

$$R_1 = \frac{1}{C_1 f_{CK}} \tag{87}$$

or

$$\omega_0 = \frac{C_1}{C_2} f_{CK} \tag{88}$$

This expression reveals three important features that hold for SC filters in general:

1. There are no resistors. This is highly desirable from the viewpoint of IC fabrication, since monolithic resistors are plagued by large tolerances and thermal drift, and also take up precious chip area. Swiches, on the other hand, are implemented with MOSFETs, which are the basic ingredients of VLSI technology and occupy very little chip area.

2. The characteristic frequency ω_0 depends on capacitance ratios, which are much easier to control and maintain with temperature and time than RC products.

3. ω_0 is proportional to f_{CK} indicating that SC filters are inherently of programmable type.

It is desirable to keep C_1 larger than the associated parasitics present in the circuit (e.g. depletion capacitances of the source/drain implants and the stray capacitances to substrate). Therefore, practical SC integrators are implemented with SPDT switch pairs to minimise the effect of parasitic capacitances and also increase circuit versatility (Fig.54). The switched-capacitor integrator of Fig.54 is not sensitive to parasitic or stray capacitances. The parasitic capacitances are

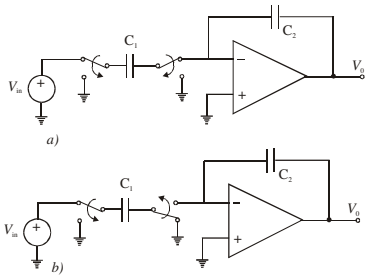

Figure 54: a) Inverting and b) noninverting SC integrators.

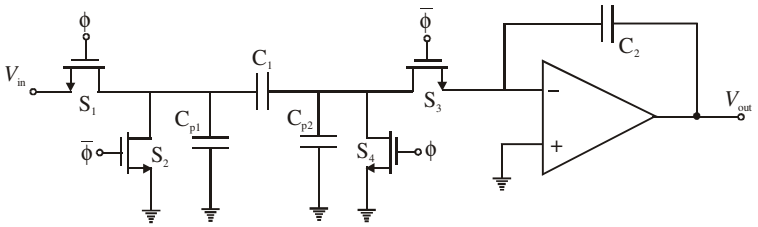

Figure 55: Parasitic capacitances associated with a switched-capacitor resistor.

shown explicitly in Fig.55: consider first the parasitic capacitance C_{p2}, this is always grounded either through S_4 or through the virtual ground of the inverting input of the op-amp. Thus C_{p2} does not have a change in the charge stored on it. Next, the parasitic capacitance C_{p1} is charged to V_{in} when S_1 is closed and then discharged when S_2 closes. Since none of the charge stored on C_{p1} when S_1 is closed is transferred to C_1, it does not affect the integrating function.

6.1.1 Basic filters

Some special cases of the very basic filter types with passive (RC) and active (op amp-RC) realizations are shown in Fig.56 and Fig.57. For the passive low pass filter (Fig.56a), the DC gain and RC time constant are given by

$$DC\,gain = 1 \tag{89}$$

$$RC = \frac{1}{\omega_0} \tag{90}$$

while for the op amp realization (Fig.56b)

$$DC\,gain = -\frac{R_2}{R_1} \tag{91}$$

$$R_2C = \frac{1}{\omega_0} \tag{92}$$

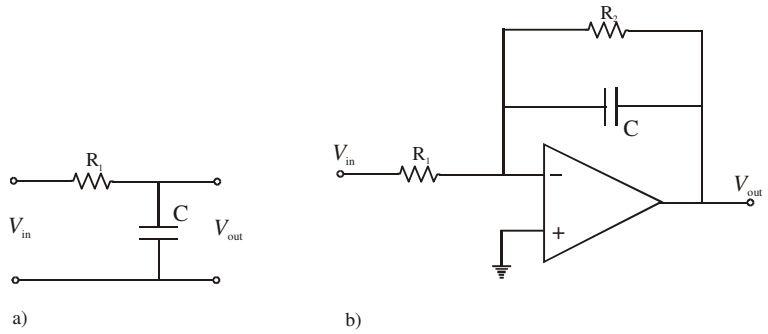

Figure 56: Low pass first-order filter a) passive realization, b) Op amp realization.

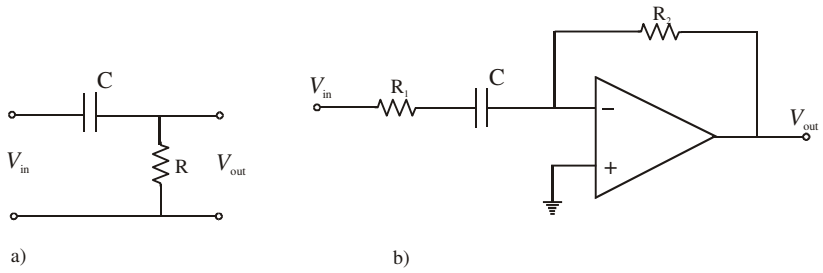

Figure 57: High pass first-order filter a) passive realization, b) Op amp realization.

And for the passive high pass filter (Fig.57a), the high frequency gain and RC time constant are given by

$$High\,Frequency\,gain = 1 \tag{93}$$

$$RC = \frac{1}{\omega_0} \tag{94}$$

while for the op amp realization (Fig.57b)

$$High\,Frequency\,gain = -\frac{R_2}{R_1} \tag{95}$$

$$R_1 C = \frac{1}{\omega_0} \tag{96}$$

Exercise 7 *Obtain the switched-capacitor counterparts of the op amp realizations of the filters in Fig.56b and Fig.57b.*

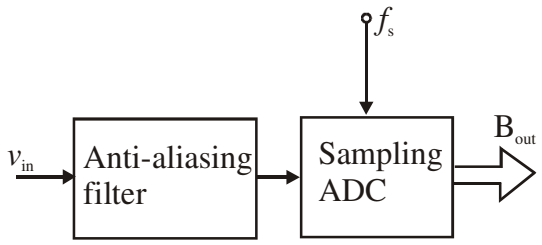

Figure 58: Nyquist sampling with analog filtering.

7 Oversampling ADCs

Oversampling A/D and D/A converters have become popular for very high-resolution ($\geq$ 16 bits) medium-to-low speed applications such as high-quality digital audio. The major reasons for this include: *i)* oversampling converters relax the requirements placed on the analog circuitry at the expense of more complicated digital circuitry. This trade-off becomes more desirable for modern submicron technologies with 3.3V power supplies where complicated high-speed digital circuitry is more easily realized in less area, but realisation of high-resolution analog circuitry is complicated by low power-supply voltage. With oversampling data converters, the analog components have reduced requirements on matching tolerances and amplifier gains. Oversampling converters simplify the requirements placed on the analog anti-aliasing filters for ADCs and smoothing filters for DACs. Also, a sample-and-hold is usually not required at the input of an oversampling ADC; and *ii)* Oversampling followed by digital filtering also results in reduced quantization-noise.

7.1 Nyquist-Rate Sampling

The digitisation process, depicted in Fig.58, has a profound impact on the frequency spectrum of the input signal. As earlier discussed in Sec. 2.1, digitisation, viewed as *discretisation in time*, creates additional spectral components, called *images*, at locations symmetric about the midpoint $f_S/2$. *Nyquist's criterion* states that if we want to recover or reconstruct a signal of a given bandwidth B (i.e. highest frequency component B Hz) from its digitised version, the sampling rate f_S must be such that

$$f_S > 2B \tag{97}$$

where $2B$ is called the *Nyquist rate*. This requirement can be met either by bandlimiting v_{in} below $f_S/2$, or by raising f_S above the Nyquist rate. In order to prevent any noise or spurious input spectral components above $f_S/2$ from folding into the baseband (DC to $f_S/2$), an anti-aliasing filter is required.

Digitisation, viewed as d*iscretisation in amplitude*, introduces quantisation noise. Consider the ideal characteristic of a 3-bit ADC shown in Fig.59. The con-

version process partitions the analog input into 2^n intervals called *code ranges*, and all values if v_{in} within a given code range are represented by the same code, namely, that corresponding to the midrange value. For example, code 011, corresponding to the midrange value $\frac{V_{in}}{V_{ref}}=\frac{3}{8}$, actually represents all inputs within the range$\frac{3}{8} \pm \frac{1}{16}$. Due to the inability by the ADC to distinquish among different values within this range, the output code can be in error by as much as $\pm\frac{1}{2}V_{LSB}$. This uncertainity, called *quantisation error*, or also called *quantisation noise* v_Q, is an inherent limitation of any digitisation process. An obvious way to improve it is by increasing n. This error is modelled as being equivalent to an additive noise source, see Fig.60. From Fig.60 , where both n-bit converters are considered ideal, we have

$$v_1 = v_{in} + v_Q \tag{98}$$

where the quantised signal v_1 has been modelled as the input signal v_{in} plus some additive quantisation noise signal v_Q. As shown in Fig.59 (and Fig. 61 for a ramp input), v_Q is a sawtooth-like variable with a peak value of $\frac{1}{2}V_{LSB}$ (this is for a ramp input but is typical). Its rms value is given by

$$V_{Q(rms)} = \sqrt{\left[\frac{1}{T}\int_{-T/2}^{-T/2} V_Q^2 dt\right]} = \sqrt{\left[\frac{1}{T}\int_{-T/2}^{-T/2} V_{LSB}^2 \left(\frac{-t}{T}\right)^2 dt\right]} \tag{99}$$

$$= \frac{V_{LSB}}{\sqrt{12}} \tag{100}$$

Note from equation (100) that the noise power is independent of the sampling rate. If v_{in} is a sinusoidal signal, the signal-to-noise ratio is maximised when v_{in} has a peak value of $V_{ref}/2$, or an rms value of $(V_{ref}/2)/\sqrt{2}$. Thus,

$$SNR_{max} = 20\log 10\left[\frac{(V_{ref}/2)/\sqrt{2}}{V_{LSB}/\sqrt{12}}\right] \tag{101}$$

$$= 6.02n + 1.76 \quad dB \tag{102}$$

7.2 Quantisation noise spectrum & oversampling

For Nyquist rate sampling, the spectrum of the quantisation noise signal will be limited by the highest frequency component, which in this case is the rate of sampling, see Fig. 61, or

$$f_{\max}(v_Q) = \frac{f_S}{2} \tag{103}$$

If the input signal is a relatively active or busy signal, its quantisation noise can be treated as white noise, i.e of constant amplitude across the bandwidth. Since the noise power is given by

$$P_Q = V_Q^2 = \left(\frac{V_{LSB}}{\sqrt{12}}\right)^2 \tag{104}$$

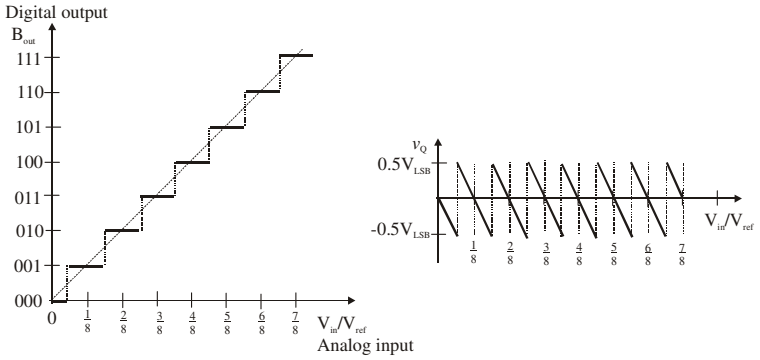

Figure 59: ADC ideal transfer characteristic and quantisation noise for $n = 3$.

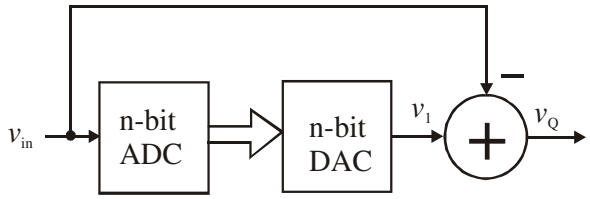

Figure 60: A circuit to investigate quantisation noise behaviour.

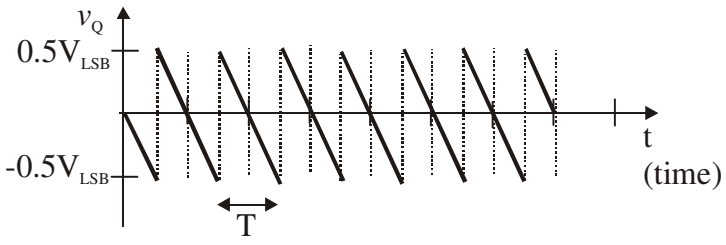

Figure 61: Quantisation noise signal for a ramp input as a function of time.

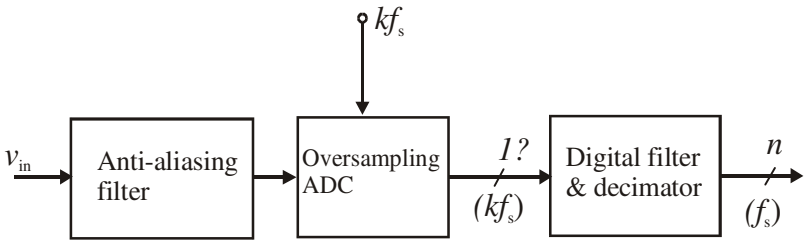

Figure 62: Oversampling with analog and digital filtering.

the noise (power) spectral density[5] k_x in $V/\sqrt{Hz}$ can be derived from

$$\int_{-f_s/2}^{f_s/2} k_x^2 df = k_x^2 f_s = \left(\frac{V_{LSB}}{\sqrt{12}}\right)^2 \tag{105}$$

or

$$k_x = \frac{V_{LSB}}{\sqrt{12}\sqrt{f_s}} \tag{106}$$

Now consider the effect of speeding up the sampling rate by a factor of k, $k \ggg 1$ (see Fig.62). The ensuing benefits are twofold:

- The frequency band between the signal and the first image about kf_s is now much wider. This is the transition band of the analog filter preceding the digitizer, and thus the circuit complexity of this filter can be reduced. In fact, in oversampling converters of the Σ-Δ type, this filter can be as simple as a mere RC stage.

- The quantisation noise is now spread over a wider band, or

$$k_x = \frac{V_{LSB}}{\sqrt{12}\sqrt{kf_s}} \tag{107}$$

  indicating a spectral density reduction by $\sqrt{k}$.

The price for the preceeding benefits is the need for a *digital filter* at the output of the digitizer to (a) suppress any spectral components and noise above $f_s/2$, and (b) reduce the data rate from kf_s back to f_s, a process known as *decimation*.

We observe that the rms noise at the output of the digitizer is still $V_{ref}/2^n\sqrt{12}$; however, only the spectra up to $f_s/2$ will make it past the filter/decimator, so

[5] Power spectral density (PSD) describes how the power of a signal is distributed with frequency. It can be defined as the squared value of the signal, that is, as the actual power if the signal was a voltage applied to a 1-ohm load. For voltage signals, it is customary to use units of V^2Hz^{-1} for PSD.

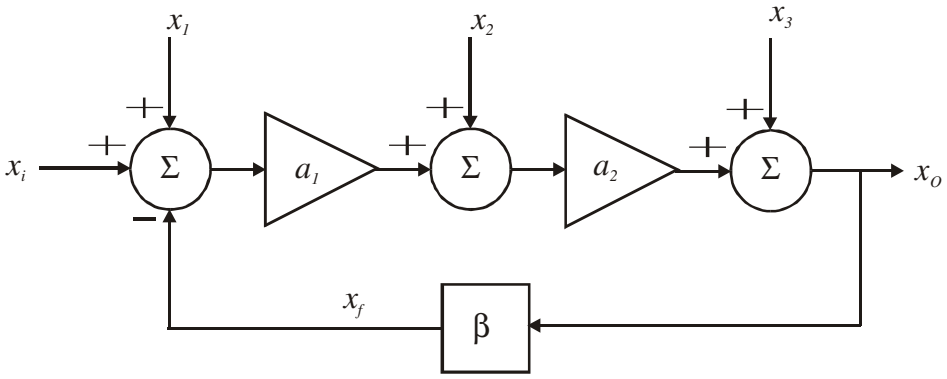

Figure 63: Investigating the effect of negative feedback on disturbances and noise.

the rms noise at the output is (calculated from the noise power)

$$V_{Q1(rms)} = \sqrt{\int_{-f_s/2}^{f_s/2} k_x^2 df} = \frac{V_{LSB}}{\sqrt{12}\sqrt{k}} \tag{108}$$

Expressing k in the form $k = 2^m$, we now have

$$SNR_{max} = 6.02\,(n + 0.5m) + 1.76 \quad dB \tag{109}$$

indicating a $\frac{1}{2}$-bit improvement for every octave of oversampling. Its actual resolution (often known as its *Effective Number of Bits* or ENOB) is defined as

$$ENOB = \frac{(SNR - 1.76)\ dB}{6.02\,dB} \tag{110}$$

Example 8 *An audio signal is oversampled with a 12-bit ADC. Find the oversampling frequency needed to achieve a 16-bit resolution.* $f_s = 44.1kHz$. *What is the corresponding* SNR_{max}*?* ***Ans****:11.29MHz, 98.09 dB*

7.3 Effect of feedback on noise

Negative feedback provides a means also for reducing circuit sensitivity to certain types of disturbances. Figure 63 illustrates three types of disturbances x_1, x_2, and x_3 entering the circuit at different stages as shown. The output is found as

$$x_0 = x_3 + a_2\,[x_2 + a_1\,(x_i - \beta x_0 + x_1)] \tag{111}$$

or

$$x_0 = \frac{a_1 a_2}{1 + a_1 a_2 \beta}\left(x_i + x_1 + \frac{x_2}{a_1} + \frac{x_3}{a_1 a_2}\right) \tag{112}$$

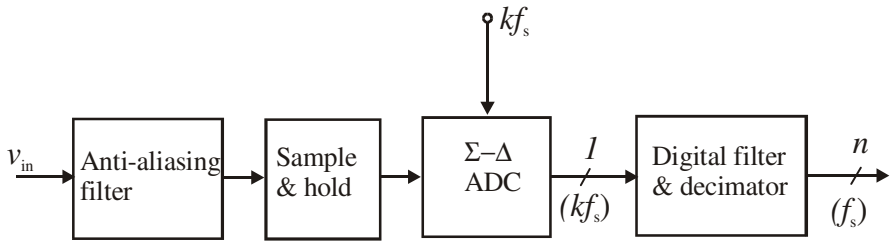

Figure 64: Block diagram of an oversampling ADC.

Note that x_1 undergoes no attenuation relative to x_i. However, x_2 and x_3 are attenuated by the forward gains from the input to the points of entry of the disturbances themselves. This feature is exploited in oversampling converters to reduce quantisation noise (aka noise shaping).

7.4 Oversampling, noise shaping and $\Sigma - \Delta$ converters

In this section, the advantage of noise shaping the quantisation noise through the use of feedback is discussed. Note that equation (109) indicates that we need four samples to increase the resolution by 1 bit, sixteen samples to increase by 2 bits, sixty-four samples to increase by 3 bits, and so forth. The system architecture of a Σ-Δ oversampling ADC is shown in Fig.64 (see also Fig.62).Σ-Δ ADCs use feedback for the double purpose of *a)* generating dither to keep the input busy - ensures quantisation noise of uniform spectral density, and *b)* reshaping the noise spectrum to reduce the amount of oversampling required. In its simplest form depicted in Fig.65, a Σ-Δ ADC consists of a 1-bit digitizer or modulator to convert v_{in} to a high-frequency serial data stream v_O, followed by a digital filter/decimator to convert this stream to a sequence of n-bit words of fractional binary value B_{out} at a lower rate of f_S words per second. The modulator is made up of a latched comparator acting as a 1-bit ADC, a 1-bit DAC, and an integrator to integrate (Σ) the difference (Δ) between v_{in} and the DAC output; hence the name Σ-Δ ADC. The comparator is strobed at a rate of kf_S samples per second, where k, usually a power of 2, is called the *oversampling ratio (OSR).*

To understand how noise shaping comes about, refer to Fig.66, where the quantisation error is modeled additively. By inspection, the various Fourier transforms are related as

$$V_0 = V_Q + H\left(V_{in} - V_0\right) \tag{113}$$

or

$$V_0 = \frac{H\left(f\right)}{1 + H\left(f\right)} V_{in} + \frac{1}{1 + H\left(f\right)} V_Q \tag{114}$$

Choosing H(f) such that its magnitude is sufficiently large over the frequency band of interest will provide simultaneous benefits of *a)* making V_0 closely track

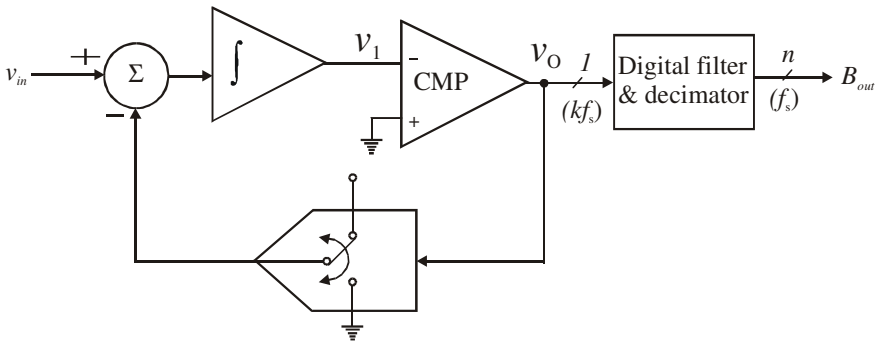

Figure 65: First-order ΣΔ ADC.

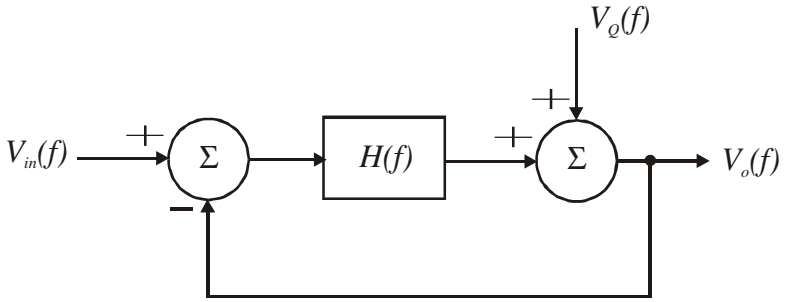

Figure 66: Linear system model of a Σ-Δ ADC.

V_{in} over the given band, and *b)* drastically reducing quantisation noise over the same band.

For frequency bands extending down to dc, H(f) is usually implemented with integrators. In mixed-mode IC processes, H(f) is implemented using switched capacitor techniques.

Example 9 *Given that a 1-bit ADC has a 6-dB SNR, what sample rate is required to obtain a 96-dB SNR (or 16 bits) if f_S = 25kHz for straight oversampling as well as first-order noise shaping?* ***Ans:*** *27000 GHz, 819 MHz. Note: $SNR_{\max} = 6.02(n + 1.5m) - 3.41dB$ for a first order Σ-Δ ADC.*

7.5 Oversampling with noise shaping using an SC integrator

Referring to the noninverting SC integrator circuit of Fig.54, we observe that ϕ pulses charge C_1 to V_{in}, while $\bar{\phi}$ pulses pull the charge accumulated in C_1 out of C_2, causing a step increase in V_0. Letting n denote an arbitrary clock period,

we have

$$v_0\left[nT_{CK}\right] = v_0\left[(n-1)\,T_{CK}\right] + \frac{\Delta Q\left[(n-1)\,T_{CK}\right]}{C_2} \tag{115}$$

or

$$v_0\left[nT_{CK}\right] = v_0\left[(n-1)\,T_{CK}\right] + \frac{C_1}{C_2} v_{in}\left[(n-1)\,T_{CK}\right] \tag{116}$$

where $\Delta Q\left[(n-1)\,T_{CK}\right] = C_1 v_{in}\left[(n-1)\,T_{CK}\right]$ denotes the charge accumulated by C_1 during the previous ϕ pulse. Equation (116) represents a discrete time sequence relating input and output values. In the frequency domain[6], equation (116) becomes

$$V_0\left(j\omega\right) = V_0\left(j\omega\right) e^{-j\omega T_{CK}} + \frac{C_1}{C_2} V_{in}\left(j\omega\right) e^{-j\omega T_{CK}} \tag{117}$$

Letting $C_1 = C_2$ and $T_{CK} = \frac{1}{kf_S}$, we can express the SC integrator transfer function as

$$H\left(f\right) = \frac{V_0}{V_{in}} = \frac{1}{e^{j\omega T_{CK}} - 1} \tag{118}$$

$$= \frac{1}{e^{j2\pi f/kf_S} - 1} \tag{119}$$

Here, k is the oversampling ratio while f_S is the Nyquist rate. Now referring to Fig.65, Fig.66 and equation (114), the noise transfer function $N\left(f\right)$ becomes

$$N\left(f\right) = \frac{1}{1 + H\left(f\right)} = 1 - e^{-j2\pi f/kf_S} \tag{120}$$

or

$$\left|N\left(f\right)\right| = 2\sin\left(\frac{\pi f}{kf_S}\right) \tag{121}$$

where the identity $\sin\theta = \left(e^{\theta} - e^{-\theta}\right)/2j$ has been used. Thus

$$\left|V_{Qout}\right| = 2\sin\left(\frac{\pi f}{kf_S}\right)\left|V_{Qin}\right| \tag{122}$$

Since the spectral density of V_{Qin} was found to be (see equation (107))

$$k_x = \frac{V_{LSB}}{\sqrt{12}\sqrt{kf_s}} \tag{123}$$

the spectral density of V_{Qout}, i.e. after feedback, is given by

$$k'_x = 2\sin\left(\frac{\pi f}{kf_S}\right)\frac{V_{LSB}}{\sqrt{12}\sqrt{kf_s}} \tag{124}$$

$$\simeq 2\frac{\pi f}{kf_S}\frac{V_{LSB}}{\sqrt{12}\sqrt{kf_s}} \quad for \quad k \gg \pi \quad and/or \quad f < \frac{f_S}{2} \tag{125}$$

$$= \frac{2\pi V_{LSB}}{\sqrt{12}\sqrt{k^3}\sqrt{f_S^3}} f \tag{126}$$

[6] $\int_{-\infty}^{\infty} x\left(t-T\right) e^{-j\omega t} dt = X(j\omega) e^{-j\omega T}$ i.e delaying a signal by one period T is equivalent to multiplying its Fourier transform by $e^{-j\omega T}$.

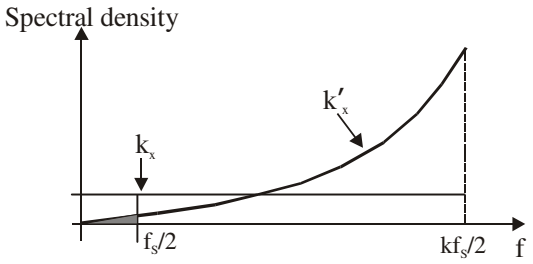

Figure 67: First-order noise shaping ($k = 16$).

The approx. $\sin\theta \simeq \theta$ for $\theta \ll 1$ has been used. Fig.67 shows a sketch of k_x and k'_x which reveals that the modulator shifts most of the noise energy toward higher frequencies. Only the shaded portion will make it past the filter/decimator. Thus the output noise power is

$$\int_{-f_S/2}^{f_S/2} (k'_x)^2 \, df \tag{127}$$

from which the output rms noise signal is found to be

$$V_{Qout}\,(rms) = \sqrt{\int_{-f_S/2}^{f_S/2} (k'_x)^2 \, df} \tag{128}$$

$$= \frac{2\pi V_{LSB}}{\sqrt{12}\sqrt{k^3}\sqrt{f_S^3}} \sqrt{\int_{-f_S/2}^{f_S/2} f^2 df} \tag{129}$$

$$= \frac{\pi V_{LSB}}{6\sqrt{k^3}} \tag{130}$$

Expressing k in the form $k = 2^m$ gives, for a first-order Σ-Δ ADC,

$$SNR_{\max} = 6.02(n + 1.5m) - 3.41\,dB \tag{131}$$

indicating a 1.5 bit improvement for every octave of oversampling; this is better than the 0.5-bit improvement without noise shaping. Here,

$$ENOB = \frac{(SNR + 3.41)\ dB}{6.02\,dB} \equiv n + 1.5m \tag{132}$$

The benefits of noise shaping can be enhanced further by using higher-order modulators.

Besides offering the aforementioned advantages of undemanding and mixed-mode-compatible analog circuitry, 1-bit quantizers are inherently linear; since only two output levels are provided, a straight characteristic results. Practical upper limits on sampling rates currently restrict Σ-Δ ADCs to moderate speed but high-resolution applications, such as digital audio, digital telephony, and

low-frequency measurement instrumentation, with resolutions ranging from 16 to 24 bits. Note that since the digital filter/decimator computes each high-resolution sample using many previous low-resolution samples, there is *latency* as information progress from input to output through the various stages of the filter. This delay may be intolerable in certain real-time applications, such as control.

References

[1] R Jacob Baker, *CMOS: Circuit Design,Layout & Simulation*, Wiley-IEEE, 2010.

[2] A S Sedra & K C Smith, *Microelectronic Circuits*, 5th Ed., Oxford Univ. Press, 2004.