1 Operational amplifiers.

1.1 Definition and specifications.

The operational amplifier (op-amp) is an amplifier with the following characteristics:

- Differential input
- Very large DC gain
- High input impedance on both inputs
- Stability over a wide range of negative feedback conditions.

The op-amp is a very versatile cell that can be used to build feedback loops, which, thanks to the large gain and high input impedance of the amplifier, perform operations showing low sensitivity to the amplifier characteristics.

The operational amplifier shares many of its specifications with generic amplifiers (instrumentation amplifiers, low noise pre-amplifiers, power amplifiers etc.) but there are a few that are specific to this kind of amplifier and are related to the fact that operational amplifiers are designed to be used in closed loop configurations.

In order to understand these specifications, we can focus on the typical closed loop configuration shown in Fig. 1.1 (a). The input signal is indicated with v_s, and the output signal is v_{out}. The open loop gain (with open output port) of the amplifier, is indicated with A_{OL}, while the amplifier output impedance is indicated with Z_{out}.; Z_L is the load impedance. The figure refers to the small signal equivalent circuit. With such a network, we generally intend to synthesize a transfer function v_{out}/v_s that depends only on the transfer functions of the feedback network, defined by the following formulas referring to the configuration of Fig. 1.1 (b):

$$\alpha_N \equiv \left(\frac{v_e}{v_s}\right)_{v_o=0} \quad ; \quad \beta_N \equiv \left(\frac{v_e}{v_o}\right)_{v_s=0} \tag{1.1}$$

If the amplifier is ideal, i.e. its output impedance is zero and its input impedance is infinite, the transfer function turns out to be:

$$\frac{v_{OUT}}{v_S} = -\frac{\alpha_N}{\beta_N}\left(\frac{\beta_N A_{OL}}{\beta_N A_{OL} - 1}\right) \tag{1.2}$$

If the gain A_{OL} of the amplifier is high enough to make $|\beta_N A_{OL}|>>1$, then the transfer function can be approximated by

$$\frac{v_{OUT}}{v_S} \cong -\frac{\alpha_N}{\beta_N} \tag{1.3}$$

In these conditions, the transfer function is set only by the feedback network while the amplifiers provides only the required gain to make this result occurs and the required power to drive the load and the feedback network itself. The latter can be designed with only passive (no power gain is required) components, that generally provide transfer functions that can be designed to be precise and stable with respect to temperature, process variations and time.

Unfortunately the amplifier is not ideal. While the condition of high input impedance can be generally fulfilled, at least in a restricted frequency range, the output impedance is not negligible.

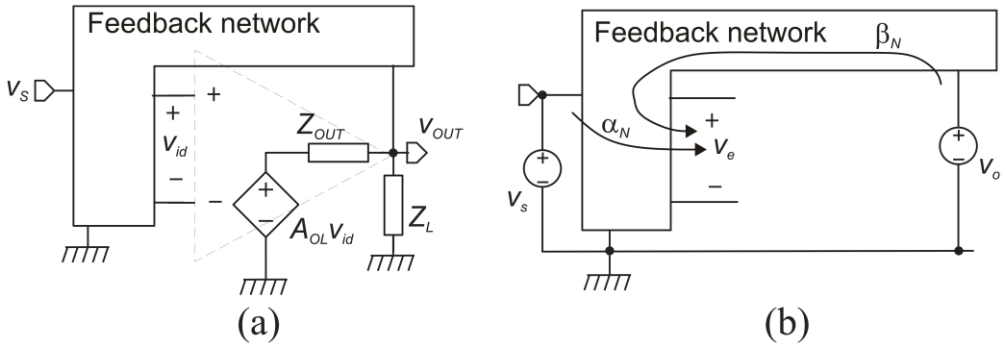

Fig. 1.1. An op-amp based closed loop network (a); definitions of the feedback network transfer functions (b)

The circuit can be rigorously analyzed using Pellegrini's cut-insertion theorem [1], cutting the network at the amplifier input as in Fig. 1.2 (a). Since the amplifier can generally be assumed to be unidirectional, then $Z_p=Z_i$, where Z_i is the input impedance of the amplifier. Considering appendix 3.2, we can write the overall transfer function as:

$$\frac{v_{out}}{v_s} = -\frac{\alpha^*}{\beta^*}\frac{\beta^* A}{\beta^* A - 1} + \frac{\gamma}{1-\beta^* A} \tag{1.4}$$

where the transfer function α^* and β^* are calculated considering the network of Fig.1.2 (b), which differs from Fig. 1.1 (b) only by the presence of the impedance $Z_p=Z_i$ in parallel to the error port of the feedback network:

$$\alpha^* \equiv \left(\frac{v_r}{v_s}\right)_{v_o=0} \quad ; \quad \beta^* \equiv \left(\frac{v_r}{v_o}\right)_{v_s=0} \tag{1.5}$$

Note that the higher $|Z_{i,}|$, the closer α^* and β^* are to α_N and β_N, respectively.

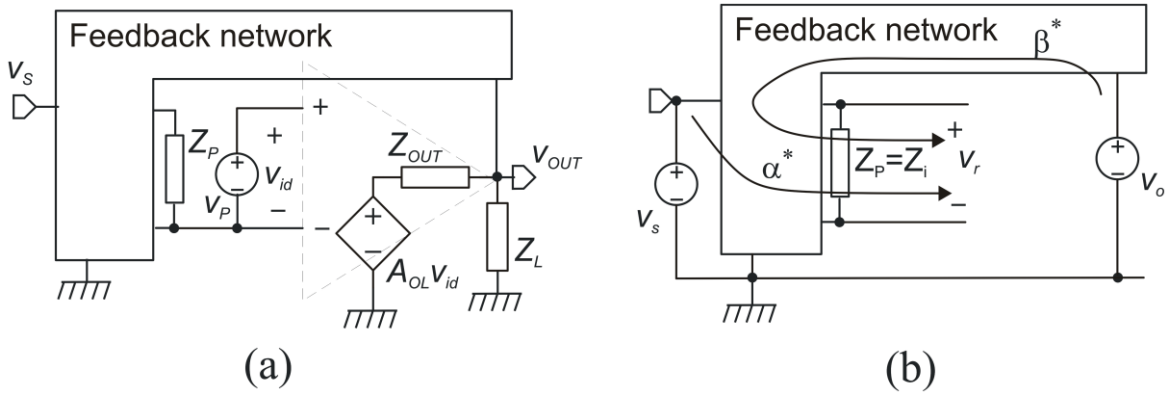

Fig. 1.2: Application of the cut-insertion theorem to the network of Fig. 1.1 (a); network used to define α^* and β^*.

Parameters A and γ are calculated on the cut network of Fig. 1.2(a) and are given by:

$$\begin{aligned} A &= \left(\frac{v_{out}}{v_p}\right)_{v_s=0} = A_{OL}\,\frac{Z_\beta \,//\, Z_L}{Z_{out} + Z_\beta \,//\, Z_L} \\ \gamma &= \left(\frac{v_{out}}{v_s}\right)_{v_p=0} \end{aligned} : \tag{1.6}$$

where Z_β is the impedance seen by voltage source v_o in Fig. 1.2 (b) and represents the loading effect of the feedback network on the amplifier output port. Note that, due to the output impedance Z_{out} and the combined loading effect of Z_L and Z_β , $|A|$ is often significantly smaller than $|A_{OL}|$. If we consider (1.4), we can compute the relative error of the transfer function with respect to the ideal one ($-\alpha^*/\beta^*$):

$$\varepsilon_R = \left|\left[\frac{v_{out}}{v_s} - \left(-\frac{\alpha^*}{\beta^*}\right)\right] \Big/ \left(-\frac{\alpha^*}{\beta^*}\right)\right| = \left|\frac{1}{\beta^* A - 1}\right| \left|1 + \frac{\beta^*\gamma}{\alpha^*}\right| \cong \frac{1}{|\beta^* A|}\left|1 + \frac{\beta^*\gamma}{\alpha^*}\right| \le \frac{1}{|\beta^* A|}\left(1 + \left|\frac{\gamma}{A_L}\right|\right) \tag{1.7}$$

Expression (1.7) shows that the relative error is the of the order of $1/|\beta^* A|$. For more details on this topics and for a demonstration of (1.4), see appendix 3.3.

Therefore, the gain A, resulting from loading the amplifier with the specified load and feedback network, should be still high enough to make the relative error, given by (1.7), smaller than the maximum value allowed by the target application. This means that a low output impedance is desirable, but not mandatory for an op-amp. Many integrated operational amplifiers are marked by output resistances (R_{out}) of the order of several kΩ, or even tens of kΩ. In many applications, the total load resistance is much smaller than R_{out}, so that $|A|<<|A_{OL}|$. This does not represent a problem as far as the residual $|\beta^* A|$ value is still >>1.

Maximum output current

In the above discussion, we have stated that a low output impedance is not strictly required, provided that the resulting loop gain is still high enough. Dealing with the output impedance, we have implicitly assumed that we are working with small signal circuits. The scheme of Fig. 1.1 (a) implement the ideal transfer function $-\alpha^*/\beta^*$ regardless of the value of the load impedance Z_L, obviously provided that $|Z_L|$ does not get so low that $|\beta^* A|$ drops below the minimum value for keeping the error ε_R negligible. This is equivalent to say that the closed loop circuit has a very low output impedance independently of the op-amp open loop output impedance. This is a well-known benefit of negative feedback.

The situation can be different if we consider large signals. Due to saturation of one or more stages in the op-amp, the output stage can be unable to feed the current needed to produce the required voltage level in the output load (comprehensive of the feedback network load Z_β). Then, the real specification that we will be obliged to consider is the maximum current that the output stage can feed to the load. Generally, it is important to take into account both the maximum positive and negative output currents.

The maximum output current affects the maximum output swing (at a certain total load R_L) according to the following conditions:

$$-R_L I_{ON} \le V_{out} \le R_L I_{OP} \tag{1.8}$$

where I_{ON} and I_{OP} are the maximum currents (absolute values) that the output port can sink (negative current) or source (positive current), respectively.

A maximum output current can be required also if a pure capacitive load has to be driven. In that case, the currents I_{ON} and I_{OP} determine the limits imposed by the output stage on the falling and rising slopes of the output voltage, respectively.

Stability

The stability of an op-amp clearly refers to cases where it is used in closed loop configuration to implement blocks different from oscillators and latches. The important quantity to be taken into consideration is the loop gain βA. A general-purpose op-amp should be designed to be stable in a wide variety of negative feedback configurations. A typical requirement is that the amplifier is stable when $\beta = -1$ (conventionally indicated as "unity gain condition"). Figure 1.3 shows a sketch of a possible Bode plot of the βA magnitude and phase. In order to have d.c. stability, since $|\beta A|>>1$ for the reasons exposed above, βA should be negative at zero frequency, that is the phase diagram asymptotically tends to 180° when the frequency tends to zero. Due to the reactive elements present in the circuit (capacitances), the phase will decrease as the frequency is progressively increased. At the same time,

the magnitude will also decrease. In order to assure stability, it is important to guarantee that for no frequency the following conditions simultaneously hold:

$$\begin{cases} |\beta A| > 1 \\ \angle \beta A = 0 \end{cases} \tag{1.9}$$

In practice, we have to avoid that the phase becomes zero at a frequency where the magnitude has not yet fallen below the zero dB line. The frequency, at which |βA|=1 (0 dB), is called the unity gain frequency and is indicated with f_0. The residual phase at $f=f_0$ is called phase margin (ϕ_m). A high phase margin is required to guarantee stability even in the case of large device parameter variations due to temperature and process spread. Large phase margins also reduce unwanted features of the closed loop step response, such as overshoot and ringing. Generally, particular techniques are required to shape the open loop frequency response of an operational amplifier to obtain a frequency response such that of Fig. 1.3, with a sufficient phase margin. These operations are indicated as frequency compensation of the amplifier.

The frequency response of general-purpose op-amps is generally of dominant pole type, as that shown in Fig. 1.3, where the dominant pole frequency is indicated with f_p. It should be desirable, when possible, to place all non-dominant singularities beyond f_0. In the case of unity gain configuration, shown in Fig. 1.3, |β| reaches the maximum value achievable with a resistive feedback network, that is |β|=1. The corresponding magnitude response of βA is shown in Fig. 1.3. If we have to design an amplifier with a gain > 1, the required |β| will be < 1. In this case, the phase diagram will be unchanged while the magnitude diagram will shift down, as shown in the figure. As a result, f_0 will be lower than in the unity gain case and the phase margin increases, improving stability. For this reason, the unity gain condition is generally considered as a worst case in terms of stability requirements.

In the case that only a single non-dominant pole (indicated with f_2 in the figure) affects the phase at f_0, the phase margin (in radians) will be given by:

$$\phi_m = \pi - \frac{\pi}{2} - \arctan\left(\frac{f_0}{f_2}\right) = \frac{\pi}{2} - \arctan\left(\frac{f_0}{f_2}\right) = \arctan\left(\frac{f_2}{f_0}\right) \tag{1.10}$$

In two stage op-amps, it is often possible to assume that only one non-dominant pole exists. By design, the frequency f_2 is set at a value equal to σf_0, where σ is a confidence factor greater than one. A typical value is σ =3, which, according to Eq. (1.10), provides a phase margin of nearly 72°. Phase margins for various values of σ are given in Table 1.1.

σ	0.5	1	2	3	5
ϕ_m	26.6°	45°	63.4°	71.6°	78.7°

Table 1.1. Phase margin for various $\sigma=\omega_2/\omega_0$ values.

The effect of singularities at frequencies higher than f_2 will usually further reduce the actual phase margin; therefore, this procedure should be used as an approximate method to set the phase margin.

More precise estimation and refinement of the phase margin has to be performed by means of an electrical simulator.

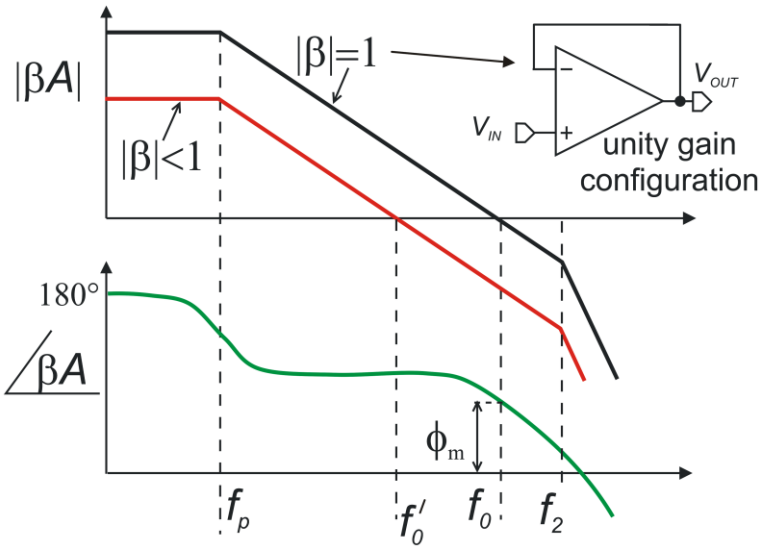

Fig. 1.3. Magnitude and phase diagram of the loop gain for only two singularities.

As stated earlier, use of the op-amp in closed loop configurations with $|\beta|<1$ facilitates the achievement of stability. For this consideration, we have assumed that β is a real, negative number. Examples of feedback networks that satisfy this property are either all resistor networks or all capacitor networks such those of Fig. 1.4. If an op-amp is designed to be unity gain stable, it will remain stable if used with $|\beta|<1$. Op-amps for special applications may be designed to be stable for $|\beta| \le \beta_{MAX} < 1$. In this case, it is not guaranteed that the amplifier is stable in unity gain configuration. If such an op-amp is used in non-inverting configurations, where the closed loop gain is $|A_V|=|1/\beta|$, there will be a minimum gain $A_{VMIN} = 1/\beta_{MAX} > 1$, below which the amplifier may be unstable. For examples, commercial amplifiers such as the Analog Devices OP37 are guaranteed stable for $A_V > 5$.

Speed Specifications.

For an operational amplifier, speed is expressed by two specifications: the Gain-Band-Width product (GBW) and the slew rate. The GBW is a small signal parameter. It is particularly useful if we can consider that the amplifier open loop frequency response is characterized by a single pole, i.e. the open loop gain A_{OL} is given by:

$$A_{OL}(s) = \frac{A_0}{1 + s/\omega_p} \tag{1.11}$$

where $\omega_p = 2\pi f_p$ and f_p is the pole frequency. In this case, if $A_0 >> 1$, (which is always true for an op-amp), the unity gain frequency f_0 is practically equal to $A_0 f_p$, so that f_0 coincides with the GBW. An approximate very useful equation that is strictly applicable if f_p is the only singularity, but works fine also in the case that f_p is the dominant pole, is the expression that gives the upper band limit (-3 db) of an amplifier built connecting an op-amp in closed loop with an all-resistor network:

$$f_H = \frac{GBW}{A_{V0}} \tag{1.12}$$

where f_H and A_{V0} are is the upper band limit and d.c. gain of the closed loop amplifier. Expression (1.12) is valid for the non-inverting amplifier of Fig. 1.1 (b) and can be applied also to the typical inverting amplifiers of Fig. 1.4 (a), substituting A_{V0} in (1.12) with $|A_{V0}|+1$. It can be shown that the validity of (1.12) can be extended to all those cases of feedback network characterized by a pure real transfer function (β) from d.c. to frequencies > f_0. This is the case of all-capacitors network, such that of Fig. 1.4 (b), which are commonly used in switched capacitors architectures.

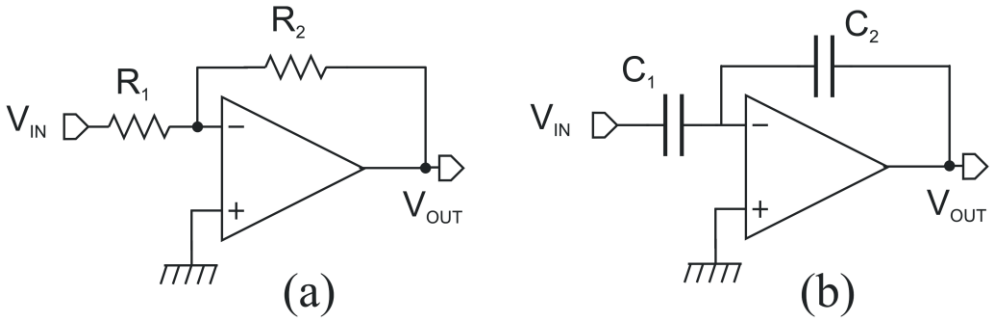

Fig. 1.4. Feedback configuration resulting in a pure real β transfer function.

The singularities that can be calculated using the *GBW* are suitable to describe the response to small signals. For example, the time constant, which characterizes an inverting or non-inverting amplifier like those depicted above, is $\tau=(2\pi f_H)^{-1}$. If a step of small amplitude is applied to those amplifiers, the output voltage settles to the finale value after several time constants (e.g. 4.6 τ for 1 % residual error). However, f_H (and thus GBW, from which f_H is derived) is not sufficient to calculate the response to signals as large as to saturate the amplifier input stages. In these cases, the amplifier output voltage varies at a fixed rate, independent of the input signals.

The maximum rising and falling rate of the output voltage is indicated as slew rate (s_R). If we apply a step of large amplitude to an amplifier built using an op-amp in closed loop configuration, the response is initially a ramp with a slope equal to the s_R. This slope is smaller (even much smaller in the case of very large steps) than in the case of pure linear response. When the output voltage gets close to the final value, the input stages exit from saturation and the response is linear again, so that an exponential approach begins. In the case of very large steps, most of the settling time is spent with the amplifier in slew rate, so that the settling time itself can be approximated by $\Delta Vout / s_R$. where $\Delta Vout$ is the amplitude of the output step.

In the particular case of sinusoidal input, the slew rate produces a distortion of the output voltage that begins to show at the points of maximum derivative, i.e at the instants where the sinusoid crosses the zero value. The maximum undistorted output amplitude allowed to a sinusoid of frequency *f* is given by:

$$\max(V_M) = \frac{S_R}{2\pi f} \tag{1.13}$$

If the amplitude is much higher than the limit given by (1.13), the output voltage approximates a triangular waveform.

To summarize, a reasonably complete set of specifications for a MOSFET operational amplifier is the following:

- DC gain A_0
- Speed: Gain BandWidth product (GBW) and Slew rate (Sr)
- Closed loop stability: e.g. phase margin in unity gain configuration and particular load conditions (typically a maximum load capacitance C_L is specified)
- Input referred voltage noise: Thermal noise density: $S_{vT,}$ Flicker noise: $k_F=fS_{vF}(f)$
- Offset (Input offset voltage: σ_{io})
- Static power consumption ($I_{supply,}$ minimum V_{dd})
- Maximum output current (positive and negative)
- Ranges: Input common mode range (CMR), output swing.
- Common mode rejection ratio (CMRR) and power supply rejection ratio (PSRR).

1.2 Operational amplifier design: general considerations.

As for other analog circuits, the design of an operational amplifier can be divided into two well distinct steps:

1. choice of the topology;
2. device sizing.

For the first step, it can be useful to consider a small set of different topologies, which can efficiently satisfy the most common requirements. It is a good idea to start from the simplest circuit and then increase its complexity only if the specifications cannot be met. The most important parameter that discriminates operational amplifier topologies is the number of stages. For an operational amplifier, the stages to be counted are only the gain stages. The following considerations about the number of stages can be drawn:

- Single stage amplifiers are characterized by only one high resistance node. The dominant pole will be associated to that node. Frequency compensation can be achieved by simply connecting a capacitor between the high impedance node and ground, lowering the pole frequency and, consequently, the unity-gain frequency f_0. If the high impedance node is the output node, the amplifiers are indicated also as OTAs (Operational Transconductance Amplifiers). Relatively high voltage gains (up to 80 dB) can be obtained with single-stage cascode OTAs. However, their application is limited only to op-amp intended to drive capacitive loads, since the high gain relies on the high output resistance that vanishes when resistive loads are applied. Application of source follower stages between the high impedance node and the output port may eliminate this problem, but its use is discouraged due to the output swing penalty introduced by these stages.
- Two-stage op-amps are characterized by two high impedance nodes. For the same reasons of OTAs, source-follower output stages are seldom used, so that the high impedance node of the second stage coincides with the output port. With non-cascode two stage amplifiers, it is possible to obtain the same gains of cascode OTAs but without the necessity of an extremely high output resistance. Much higher gains can be obtained by using a cascode first stage. In this way, it is only the internal high resistance node to be boosted, with no effect on the output resistance. Open loop gains up to 120 dB can be reached with two-stage op-amps. Frequency compensation of two-stage op-amps is generally accomplished through Miller compensation.
- Three or more stage op-amps are necessary when the required gain cannot be achieved with two-stage architectures. This occurs when MOSFETs with sub-micron lengths are used, for example to achieve very high GBWs. Short lengths increase the MOSFET λ, reducing the device output resistances. Low supply voltages, preventing the use of cascode stages, contribute to reduce the maximum gain available with two-stage op-amps, increasing the demand for multiple stage amplifiers. Compensation is more critical in multiple stage amplifiers. Nested Miller schemes are often adopted.

In the next part of this chapter, we will focus on two stage amplifiers, which still represent the mostly used op-amp category.

1.3 Operational amplifier design: two stage operational amplifiers

We will analyze the basic two-stage CMOS op-amp, for which we will describe a relatively simple sizing strategy, aimed at satisfying important specifications.

The simplest two-stage CMOS operational amplifier is shown in Fig. 1.5, where the two high resistance nodes are indicated with H and O. The circuit is biased by M8, which will be omitted in next figures. The bias voltage V_{BIAS}, produced by M8, drives M7 and M6 current sources, providing the bias currents of the first and second stage, respectively. The group R_C-C_C is introduced to perform frequency compensation. We have considered that the amplifier is powered by a single supply voltage V_{dd}. In the case of dual power supply, *gnd* should be replaced with V_{SS} <0.

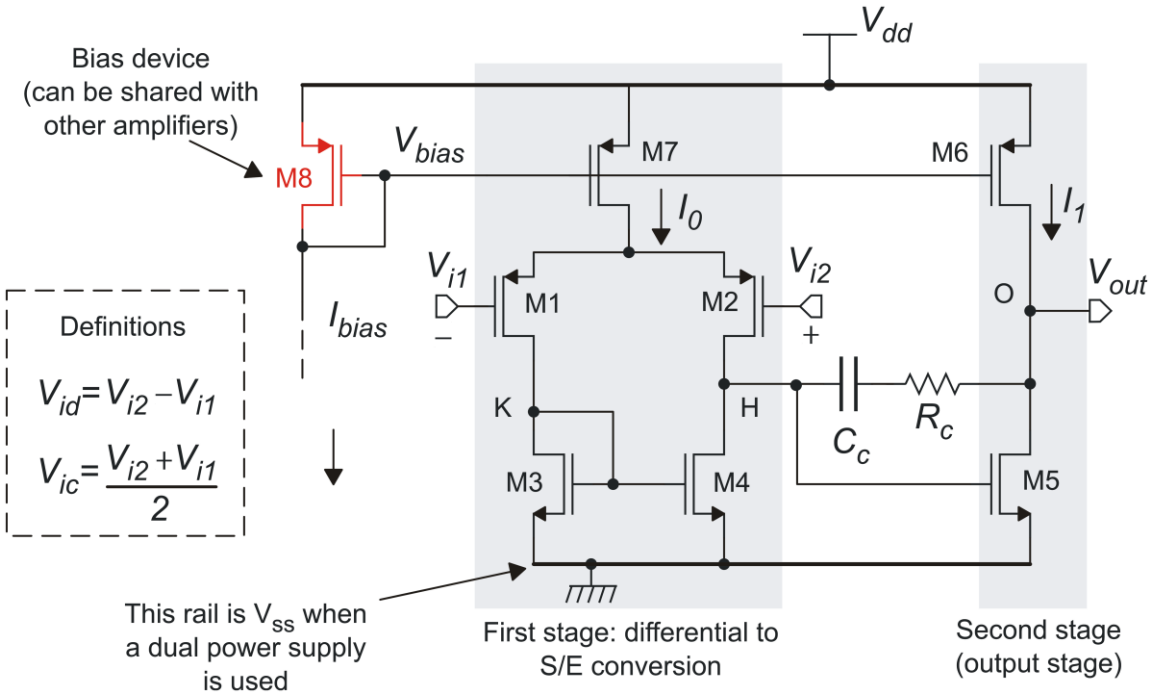

Fig. 1.5. The simplest two-stage CMOS operational amplifier

Degrees Of Freedom

Sizing is the operation by which the device parameters (e.g. width / length for MOSFETs) and bias are chosen in order to obtain the required performances. Not all parameters are independent, so that it is possible to find a restricted set of values, from which all the others can be derived. Parameters in this set are called Degrees Of Freedom (DOFs). Clearly, there is not a single set of possible DOFs for a circuit but the choice of a proper DOF-set facilitates the sizing operation. If we exclude M8, which can

be considered an external element of the cell (a single device can provide V_{BIAS} to several op-amps), the circuit has five independent MOSFETs (M1=M2 and M3=M4) and two passive elements (R_C, C_C). Considering that each MOSFET introduces two parameters (W, L), we have an initial set of 12 DOFS (including C_C and R_C), to which I_0 has to be added. Note that $I_1 = I_0\beta_6/\beta_7$, thus it is not a free parameter. The total DOFS are then 13. We can start considering only the DOFs that affect the operating point, neglecting for now the compensation network. The resulting set of 11 DOFs can be further reduced introducing a few relationships that will be explained in the following paragraph.

Static equations

-) Null systematic offset. In the nominal design, we usually require that for null input differential voltage (V_{id}=0), the output voltage is zero. Since no random mismatch errors are present in the nominal design, deviation from this condition means that a systematic offset is present. In a single power supply circuit, the conventional zero voltage for signals cannot coincide with the *gnd* potential, otherwise negative signals could not be represented, being the output node prevented from getting lower than the lowest power supply rail (i.e. *gnd*). We generally specify an intermediate point between V_{dd} and *gnd* as the signal zero level. A typical choice fro the zero level is V_{dd}/2, but also other solutions are possible. In same cases, the zero level is chosen to suit the input range of the stage that follows the op-amp. Note that it is not possible to precisely determine the output voltage of the op-amp of Fig. 1.5. The reason is that M5 and M6 behave as two opposed current sources and, if both are in saturation region, the resulting voltage strongly depends on the effect of V_{DS} on the drain current, effect that is not precisely predictable. We can impose a much simpler relationship regarding the output short-circuit current instead of the output voltage. If we short-circuit the output port by connecting it to a voltage source equal to the required zero level (e.g. *Vdd*/2), which should keep both M5 and M6 in saturation, the short circuit current will be:

$$I_{out-SC} = I_{D5} - I_{D6} \tag{1.14}$$

Considering all MOSFETs in saturation region, the current I_{D6} is given by:

$$I_{D6} = \frac{\beta_6}{\beta_7} I_0 \tag{1.15}$$

For zero input differential voltage, currents and voltages in the input stage are symmetrical. Current I_0 is then split into two equal parts flowing through the M1 and M2 branch. Furthermore, in these conditions, points K and H in Fig. 1.5 are at the same potential. As a result, M5 and M3 has the same V_{GS} (remember that this condition occurs only for null differential voltage) and the current I_{D5} can be written as:

$$I_{D5} = \frac{\beta_5}{\beta_3}\frac{I_0}{2} \tag{1.16}$$

Substituting I_{D6} and I_{D5} from (1.15) and (1.16) into (1.14) and imposing $I_{out\text{-}SC}$=0, we get a condition for the betas:

$$\frac{\beta_6}{\beta_7} = \frac{1}{2}\frac{\beta_5}{\beta_3} \tag{1.17}$$

Note that a small $I_{out\text{-}SC}$ is able to produce large output voltage variations, due to the relatively high output resistance (equal to the parallel of r_{d5} and r_{d6}). Even $I_{out\text{-}SC}$ values of a few percent of the MOSFET quiescent drain currents can produce output voltage variations as large as to push either M5 or M6 into triode region (saturation of the output voltage).

-) Symmetrical output swing. The output voltage can approach the *gnd* and V_{dd} rails, but should maintain a distance from them equal to M5 and M6 saturation voltage, in order to prevent them from getting into triode region. The output swing will then be given by:

$$\left(V_{GS} - V_t\right)_5 \le V_{OUT} \le V_{dd} - \left|V_{GS} - V_t\right|_6 \tag{1.18}$$

The output swing is symmetrical if the minimum distances from the *gnd* and V_{dd} rails are identical. This condition can be useful for general-purpose op-amps and is obtained imposing:

$$\left(V_{GS} - V_t\right)_5 = \left|V_{GS} - V_t\right|_6 \tag{1.19}$$

-) Precise current mirroring. Condition (1.17) is based on the assumption that I_{D6}/I_{D7} and I_{D5}/I_{D3} current ratios coincides with the respective beta ratios. This clearly requires that the threshold voltages of the MOSFETs involved in the ratios are equal. Considerable differences in the threshold voltages can be caused by using device with different lengths. Note that errors in the current ratios in (1.17) may introduce a non-negligible output short circuit current, which would introduce a systematic offset. This problem can be mitigated by imposing:

$$L_6 = L_7 \tag{1.20}$$

$$L_3 = L_5 \tag{1.21}$$

Choice of the static DOF set

Reducing the number of DOFs may significantly simplify amplifier sizing. Every additional equation reduces the DOFs by one unit. Equations regarding the DOFs are called “equality constraints”. This distinguishes equations from “inequality constraints”, consisting in inequalities generally related to amplifier specifications (e.g. *GBW* > 10 MHz). Starting from the initial set of 11 static DOFs, if we consider equations (1.17), (1.19), (1.20), and (1.21) we reduce the DOFs to only 7 parameters. We will follow this approach in the next part of this chapter. However, it should be observed that only (1.17) is strictly mandatory. Adopting also the other three conditions, might results in design limitations that completely counterbalance the advantages deriving from precise mirroring and symmetrical output swing. In these cases, the designer can selectively remove one or more arbitrary constraint and estimate the possible negative effect using proper simulation tools.

As we have stated earlier, there are several equivalent way to choose the residual 7 DOFs. A possible criterion, followed in the rest of this chapter, is that of assigning as much as possible DOFs to the device that perform the main operations. In the circuit of Fig. 1.5 these devices are the active

MOSFETs of the first and second stage, i.e. M1 (together with M2) and M5. To define M1 and M5 completely, we have decided to include their dimensions (W and L) and the overdrive voltage (V_{GS}-V_t) into the DOF set. In this way also M1 and M5 operating point is fixed. To reach the total number of 7 (static) DOFs, L_6 has also been included into the DOF set. The complete set of static DOFs is then:

$$DOFs = \left\{W_1,\ L_1,\ \ \left|V_{GS}-V_t\right|_1,\ W_5,\ L_5,\ \ (V_{GS}-V_t)_5,\ L_6\right\} \tag{1.22}$$

From these DOFs, it is possible to derive all the other static circuit parameters. We will refer to the case of all devices in strong inversion, but extension to the case of moderate/weak inversion is possible by using the respective equations of the current in place of the square-law approximation used here. First, note that M1 and M5 currents are known, being given by:

$$I_{D1} = \frac{\mu_p C_{OX}}{2}\frac{W_1}{L_1}\left(V_{GS}-V_t\right)_1^2 \qquad I_{D5} = \frac{\mu_n C_{OX}}{2}\frac{W_5}{L_5}\left(V_{GS}-V_t\right)_5^2 \tag{1.23}$$

Device M3 is then completely determined by:

$$(V_{GS}-V_t)_3 = (V_{GS}-V_t)_5,\ \ L_3 = L_5,\ \ I_{D3} = I_{D1} \Rightarrow\ W_3 = \frac{2I_{D3}}{\left(V_{GS}-V_t\right)_3^2}\frac{L_3}{\mu_n C_{OX}} \tag{1.24}$$

As far as M6 is concerned its length L_6 belongs to the DOFs. The other parameters can be obtained adding condition (1.19) and considering that I_{D6}=I_{D5}:

$$\left\{\left|V_{GS}-V_t\right|_6 = (V_{GS}-V_t)_5,\quad I_{D6} = I_{D5}\right\} \Rightarrow \beta_5 = \beta_6 \tag{1.25}$$

Finally, M7 is completely determined considering that, using (1.20):

$$(V_{GS}-V_t)_7 = (V_{GS}-V_t)_6,\ \ L_7 = L_6,\ \ I_{D7} = 2I_{D1} \tag{1.26}$$

Small signal equivalent circuit

Figure 1.6 shows the small signal equivalent circuit of a two-stage amplifier. This circuit is representative of a large class of two stage topologies. In this circuit, with the uppercase letter G_m we have indicated the transconductance of a whole stage. G_{m1} and G_{m2} are then the transconductances of the first and second stage, respectively. The transconductance of single devices will be indicated with the lowercase symbol g_m, as customary. Parasitic capacitances related to the input terminals (M1 and M2 gates) are neglected in this analysis.

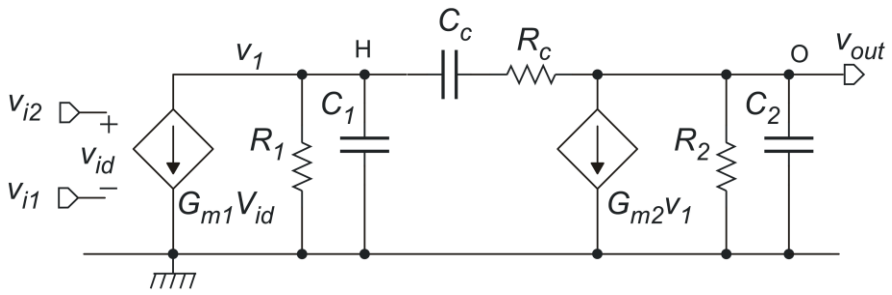

Fig. 1.6. Simplified small signal equivalent circuit of a generic two stage op-amp.

In the case of the amplifier of Fig. 1.5, the parameters of the small signal circuit are related to the device parameters according to the following relationships:

$$\begin{aligned} G_{m1} &= g_{m1} \\ G_{m2} &= g_{m5} \end{aligned} \tag{1.27}$$

$$\begin{aligned} R_1 &= r_{d4} \| r_{d2} = r_{d1} \| r_{d3} \\ R_2 &= r_{d5} \| r_{d6} \end{aligned} \tag{1.28}$$

$$\begin{aligned} C_1 &= C_{DB2} + C_{DB4} + C_{GS5} \\ C_2 &= C_2' + C_L \end{aligned} \tag{1.29}$$

where C_L is the load capacitance while C_2' is the part of C_2 due to the parasitic capacitances, i.e.:

$$C_2' = C_{DB5} + C_{DB6} \tag{1.30}$$

D.C gain.

The circuit in Fig. 1.6 can be easily solved to calculate the d.c. gain $A_0 = v_{out}/v_{id}$:

$$A_0 = G_{m1} R_1 G_{m2} R_2 \tag{1.31}$$

Using the general expressions: $g_m = I_D/V_{TE}$ and $r_d = 1/(\lambda I_D)$ into (1.27) and (1.28) to calculate $G_{m1}.G_{m2}$, R_1,R_2 and considering that $I_{D1}=I_{D3}$ and $I_{D5}=I_{D6}$, we get:

$$A_0 = \frac{1}{V_{TE1}} \frac{1}{V_{TE5}} \frac{1}{(\lambda_1 + \lambda_3)} \frac{1}{(\lambda_5 + \lambda_6)} \tag{1.32}$$

This expression indicates that, in order to obtain a large d.c. gain, it is necessary to:

- Set the $V_{GS}-V_t$ of the active transistors (M5 and M1) to the minimum level in order to get a small V_{TE} value. Note that V_{TE} asymptotically tends to ζV_T when $V_{GS}-V_t$ gets so low that the devices enter weak inversion. Below that point, there is no significant advantage to reduce $V_{GS}-V_t$ any further.

- Use device lengths considerably greater than the minimum value, since the lambda parameters increase as the channel lengths are reduced.

Frequency response

Solution of the small signal circuit of Fig. 1.6 in the *s* domain gives three poles at angular frequencies ω_p, ω_2, ω_3, and one zero s_z. Approximate expressions of these singularities are the following [2]:

$$\begin{aligned}
\omega_p &\cong \frac{1}{R_1 G_{m2} R_2 C_c} \\
\omega_2 &\cong \frac{G_{m2}}{C_1 + C_2}\left(1 + \frac{C_S}{C_c}\right)^{-1} \qquad \text{where } C_S = \left(\frac{1}{C_1} + \frac{1}{C_2}\right)^{-1} \\
\omega_3 &\cong \frac{1}{R_c C_{S3}} \quad \text{where } C_{S3} = \left(\frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_c}\right)^{-1} \\
s_z &= \frac{1}{C_c\left(\frac{1}{G_{m2}} - R_c\right)}
\end{aligned} \tag{1.33}$$

Note that C_S is the series of C_1 and C_2, while C_{S3} is the series of all three capacitors present in Fig. 1.6. The approximations used to obtain expressions (1.33) are valid if $\omega_p << \omega_2$. i.e when ω_p is really a dominant pole. This condition is automatically verified, provided that C_c is of the same order of the other capacitors (C_1 and C_2), R_1 and R_2 are of the same order of magnitude and $G_{m2}R_2 >> 1$. Note that all these conditions are generally valid in all practical two-stage op-amps. The zero s_z may be positive, and this has unfavorable consequences on the phase margin, since it adds phase delay to the unavoidable contribution of the poles. The zero is certainly positive if R_C is zero. Choosing a proper value for R_C it is possible to eliminate the zero or make it negative.

Considering that the frequency response is dominated by ω_p up to the unity gain frequency (ω_0), the latter is then given by:

$$\omega_0 \cong A_0 \omega_p = \frac{G_{m1}}{C_c} \tag{1.34}$$

Therefore, the gain bandwidth product will be given by

$$GBW \cong f_0 = \frac{\omega_0}{2\pi} \tag{1.35}$$

Design for GBW and stability

Stability in closed loop configuration is the main aspect of op-amp design since it has to be guaranteed independently from other performances. As we have seen in previous paragraphs, stability is closely related to the *GBW*. Nevertheless, even if we are designing an op-amp to be used only for d.c. signals,

it has to be stable in closed loop configuration, thus the arguments described in this chapter have to be mandatorily addressed. Figure 1.7 shows an idealized magnitude and phase diagrams of the open loop amplification *A*. Differently from Fig. 1.3, we have not considered here the value of β. The plots can be used to investigate the case β= –1 (unity gain configuration) that we have seen to be the worst case for stability. The magnitude diagram of *A* coincides with that of β*A* in unity gain configuration while, due to the negative sign of β, the βA phase diagram can be obtained by shifting the phase diagram of *A* by 180°. As a result, the phase margin in Fig. 1.7 is the difference with respect to the line –180°. Furthermore, here we have indicated the angular frequencies instead of frequencies to simplify the discussion that follows.

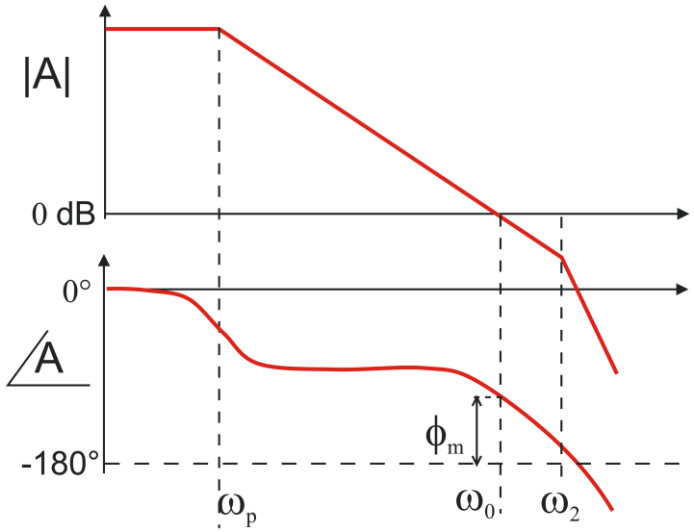

Figure 1.7. Open loop magnitude and phase response with wide phase margin: excellent design.

Figure 1.7 represents an ideal situation, where a relatively large phase margin is present. Note that we have indicated only the dominant pole and the first non-dominant pole. Equations (1.33) indicate that also a zero is present. Let us indicate the angular frequency of the zero with $\omega_z = |s_z|$. If R_C=0, the zero is positive and its angular frequency becomes equal to G_{m2}/C_c. This value is comparable to ω_0, since G_{m1} and G_{m2} are of the same order of magnitude. In this condition, the presence of the zero will modify the frequency response as shown in Fig. 1.8. With such a response, the op-amp will result unstable in unity gain configuration.

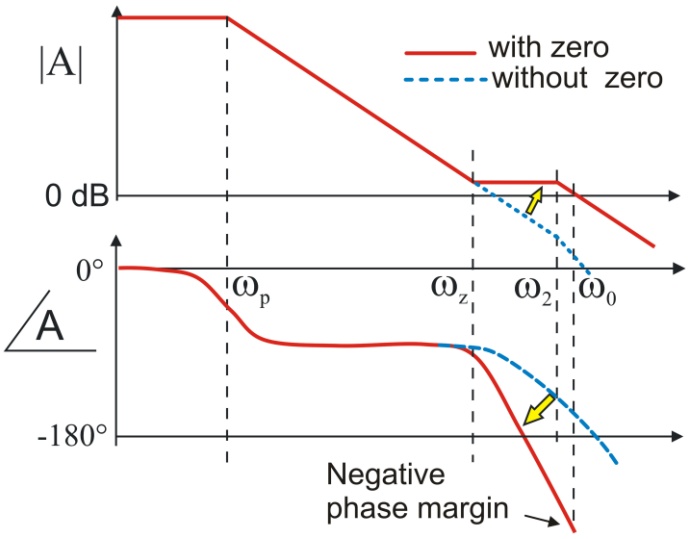

Fig. 1.8. Adverse effect of the right-half-plane (RHP) zero.

Clearly, Fig. 1.8 shows one of the possible cases. If ω_z is significantly larger than ω_0, the response can still be stable. However, even in this case, the zero will reduce the phase margin. Resistor R_C is then added to modify ω_z. Various strategies are possible. The first (and simpler) strategy is pushing ω_z to infinity, practically cancelling the zero. This is obtained by choosing $R_C=1/G_{m2}$. An alternative strategy is choosing a particular value of R_C that makes the zero negative (with $R_C>1/G_{m2}$) and equal to ω_2. In this way, the zero and first non-dominant pole cancels out each other [3]. This attracting strategy is less robust than the former, due to the necessity of precise matching between ω_2 and ω_z. Furthermore, note that ω_2 is dependent on the load capacitance, therefore cancellation occurs only for a given load condition. It can be objected that also making $R_C=1/G_{m2}$ requires an excellent matching, but in this case, if a mismatch is present, ω_z is no more infinite but still remains at very high frequencies and can be neglected as well. Resistor R_C can also be chosen to be slightly larger than $1/G_{m2}$ to assure that s_z remains negative even in the presence of mismatch but still at much higher frequencies than in the case $R_C=0$. In the next part of this discussion, we will assume:

1. The choice $R_C=1/G_{m2}$ is made, so that the zero can be neglected;

2. The phase margin is determined only by ω_2 and ω_0, according to Eqn. (1.10). This is equivalent to considering that ω_3 and other possible singularities are located at frequencies much higher than f_2.

In order to obtain the required phase margin ϕ_m we will impose:

$$\omega_2 \cong \sigma\omega_0 \tag{1.36}$$

where σ is chosen using (1.10) to provide the desired phase margin. As we will see, σ is paid in terms of performance, so that a tradeoff should be sought. A typical choice is $\sigma=3$, resulting in a theoretical phase margin of around 70 degrees. The real phase margin of the amplifier will generally be worse,

owing to the contribution of other singularities that we are neglecting in this simplified analysis. For this reason, it is necessary to aim at phase margin significantly higher than the real target value.

By substituting the expressions of ω_2 and ω_0, given in (1.33) and (1.34), into (1.36), one could calculate the value of the compensation capacitor [4]:

$$C_c = \frac{\sigma}{2}\frac{G_{m1}}{G_{m2}}(C_1 + C_2)\left[1 + \sqrt{1 + \frac{4}{\sigma}\frac{G_{m2}}{G_{m1}}\frac{C_S}{C_1 + C_2}}\right] \tag{1.37}$$

In this way, C_c is no more a DOF, since it can be expressed in terms of the 7 static DOFs. Indeed, G_{m1}, G_{m2} and C_1, C_2 are actually functions of the DOFs indicated in (1.22). Since also R_C has been eliminated from the DOFs with the choice $R_C=1/G_{m2}$, the only remaining DOFs are those specified in (1.22). Furthermore, this expression could be put into (1.34) and (1.35) in order to calculate the *GBW* as a function of the DOFs. In practice, this does not lead to equations that can be used to perform manual design of the amplifier, since too many DOFs still appear in (1.37).

It is then necessary to introduce approximations aimed at simplifying the analysis and provide clear indications to the designer. We will then introduce two hypotheses that have to be always checked at the end of the design work:

Hyp. 1 $$C_1 << C_2, C_C \tag{1.38}$$

Hyp. 2 $$C_2' << C_L \tag{1.39}$$

In practice, hypotheses 1 and 2 means that the parasitic capacitances C_1 and C_2' are negligible with respect to the load and compensation capacitances. Hypothesis 1 can be used to simplify the expression of ω_2 given in Eqns.(1.33). If $C_1 << C_2$, then the series capacitance C_S is nearly equal to C_1. Using hypothesis 1 again, $C_1 << C_C$, thus we can write:

$$\omega_2 \cong \frac{G_{m2}}{(C_1 + C_2)} \cong \frac{G_{m2}}{C_2} \tag{1.40}$$

Using Hyp.2, we can finally write:

$$\omega_2 \cong \frac{G_{m2}}{C_L} \tag{1.41}$$

Now, using (1.35) and stability condition (1.36), we can write a useful expression of the *GBW*:

$$GBW \cong \frac{1}{2\pi}\frac{\omega_2}{\sigma} \cong \frac{1}{2\pi}\frac{G_{m2}}{\sigma C_L} \tag{1.42}$$

Equation (1.42) states that, to obtain a given *GBW*, it is necessary to start designing the output stage. This seems in contrast with (1.34) and (1.35), which relate the *GBW* to parameters of the first stage (G_{m1}) and the compensation capacitor. Actually, (1.34) and (1.35) are correct for the analysis of the amplifier, but not immediately useful during the synthesis. In fact, ω_0 is not a free parameter but, once

the phase margin has been chosen, it cannot be larger than ω_2/σ. Thus, in order to obtain a certain value of ω_0, we have to "make room to it" by setting ω_2. If Hyp.1 and Hyp.2 are valid, ω_2 depends only on parameters of the output stage and this explains Eq. (1.42).

We can further transform (1.42) to highlight the role of the various DOFs involved. Considering that, for the amplifier of Fig. 1.5, $G_{m2}=g_{m5}$, we have:

$$GBW = \frac{1}{2\pi\sigma C_L}\frac{I_{D5}}{V_{TE5}} \tag{1.43}$$

Equation (1.43) directly relates the *GBW* to the current consumption of the output stage. A high *GBW* is paid mainly in terms of current consumption. Once the target *GWB* has been fixed, it is possible to choose a small $(V_{GS}\text{-}V_t)_5$ value in order to reduce V_{TE5} and then obtain the GBW with an as small as possible current consumption. Unfortunately, this condition can be in contrast with other constraints, as it will be shown in the section regarding offset and noise. In next part of this paragraph we will show also that a small $(V_{GS}\text{-}V_t)_5$ may lead to violate Hyp. 1 and 2.

Equation (1.43) seems to indicate that there is no upper limit to the GBW, the problem being only power consumption. Furthermore, this equation is independent of process parameters. This is true until Hyp.1 and 2 are valid. To understand what happens when Hyp. 1 and 2 are not applicable, let us consider the expression of I_{D5}, referred to strong inversion operation, for simplicity:

$$I_{D5} = \mu_n C_{OX}\frac{W_5}{L_5}\frac{(V_{GS}-V_t)_5^2}{2} \tag{1.44}$$

In order to increase I_{D5} with fixed $V_{GS}–V_t$ and L, it is necessary to increase W_5. This produces an increase of the parasitic capacitances of M5 and, in particular, C_{DB5} and C_{GS5}. According to Eqns.(1.29) , these two capacitances appears in the expression of C_1 and C_2', respectively. Therefore, there will be a value of I_{D5}, over which the hypotheses will not hold any more and Eqn. (1.42) cannot be applied any more. It can be useful to rewrite (1.40) without the approximations given by Hyp.1, and 2. The only exception will be the approximation $C_S/C_C<<1$, which will be considered to be still valid, since C_C can be properly chosen high enough to make this happen. Using the strong inversion approximation for $G_{m2}=g_{m5}$, we get:

$$GBW = \frac{1}{2\pi}\frac{G_{m2}}{\sigma(C_1+C'_2+C_L)} = \frac{1}{2\pi\sigma}\frac{\mu_n C_{OX}\frac{W_5}{L_5}(V_{GS}-V_t)_5}{(2/3C_{ox}W_5L_5+C_JL_C(W_5+W_6)+C_L)} \tag{1.45}$$

Parameters C_J and L_C are the junction capacitance per unit area and the minimum length of the drain/source diffusion, respectively. Both L_C and C_J have been assumed identical for *p*-MOS and *n*-MOS devices, for simplicity. Dividing both the numerator and denominator in (1.45) by W_5, we get the following expression:

$$GBW = \frac{1}{2\pi\sigma} \cdot \frac{\mu_n C_{OX} \frac{(V_{GS} - V_t)_5}{L_5}}{2/3 C_{OX} L_5 + C_J L_C \left(1 + \frac{W_6}{W_5}\right) + \frac{C_L}{W_5}} \tag{1.46}$$

The ratio W_6/W_5 in (1.46) can be regarded as a constant term, since an increase of W_5 should be followed by an identical increase of W_6 to maintain $\beta_6=\beta_5$, imposed by Eqn. (1.25).

If I_{D5} is increased to obtain a large *GBW*, according to Eqn. (1.43), W_5 should be increased as well. Eqn. (1.46) shows that the *GBW* does not increase indefinitely, but tends to an asymptotic value given by:

$$GBW^* = \frac{1}{2\pi\sigma} \cdot \frac{\mu_n C_{OX} \frac{(V_{GS} - V_t)_5}{L_5}}{\frac{2}{3} C_{ox} L_5 + C_J L_C \left(1 + \frac{W_6}{W_5}\right)} \tag{1.47}$$

This value is deeply process dependent. The behavior of *GBW* as a function of W_5 (i.e. I_{D5}) is sketched in Fig. 1.9.

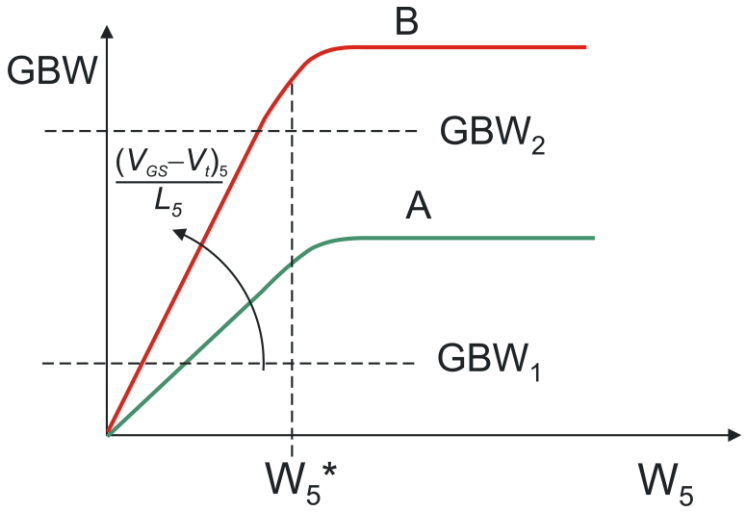

Fig. 1.9. GBW dependence of W_5 for fixed $(V_{GS}\text{-}V_t)_5$ and L_5.

First, consider curve A, obtained varying W_5 while keeping $(V_{GS}-V_t)_5$ and L_5 fixed. Starting from low W_5 values, corresponding to low M5 drain currents, we find a linear region. Here the *GBW* is proportional to I_{D5}, thus (1.43) is applicable. We have indicated with W_5* the upper limit of the linear region. For $W_5 > W_5$* the GBW begins to saturate to the final value given by (1.47). If we choose different $(V_{GS}-V_t)_5$ and L_5 values, so that the ratio $(V_{GS}\text{-}V_t)_5/L_5$ is larger, then we get a similar *GBW* curve that saturates to a higher *GBW* value. This is represented in Fig. 1.9 by curve B. The transition between the linear to the saturated region occurs when the load and parasitic capacitances becomes comparable. This occurs at the same W_5* value for both curves A and B. Nevertheless, considering the

drain current expression given by (1.44), the value of I_{D5} for which the linear relationship holds is higher for curve B.

It is interesting to consider two different design cases where we will assume that the length L_5 is fixed while $(V_{GS}\text{-}V_t)_5$ is a free parameter (i.e. it is not fixed by other specifications), Referring to Fig. 1.9, in the first case the required *GBW* is indicated by the value GBW_1. We note that both curves A and B satisfy the specification with W_5 in the linear zone, then (1.43) is valid. In this case, it is convenient to choose curve A, since the $(V_{GS}-V_t)_5$ is smaller and, from (1.43) the power consumption (due to I_{D5}) is lower. In the second case, it is required to reach GBW_2. This value cannot be achieved with curve A and the designer should use a larger $(V_{GS}-V_t)_5$ and, from (1.43), a larger power consumption. Clearly, for a given process and power supply voltage, there will be a maximum *GBW* that cannot be exceeded. The continuous improvement of CMOS processes, manly aimed at scaling down the minimum channel length, has produced a corresponding increase in the maximum achievable *GBW*. In the following part of this discussion, we will assume that Hyp.1 and 2 are valid, so Eqn. (1.41) holds.

Once the value of G_{m2} has been determined, it is necessary to calculate the value of the compensation capacitor C_C and of the first stage transconductance G_{m1}.

Going back to the expressions of ω_0 and ω_2, given by Eqns. (1.34) and (1.41), respectively, and combining them with the stability condition (1.36), we find:

$$\frac{G_{m2}}{C_L} = \sigma \frac{G_{m1}}{C_C}, \tag{1.48}$$

from which we obtain C_C:

$$C_C = \sigma \frac{G_{m1}}{G_{m2}} C_L \tag{1.49}$$

In order to calculate C_C the designer has to decide the value to assign to G_{m1}/G_{m2}. The parameter to be considered is the value of C_L. It is possible to start from a commonly used rule of thumb that assigns:

Rule of thumb:
$$\frac{G_{m1}}{G_{m2}} = \frac{1}{\sigma} \qquad \Rightarrow C_C = C_L \tag{1.50}$$

Since generally $\sigma \sim 3$, this means $G_{m1}=G_{m2}/3$. The rule of thumb has the advantage of being simple and making easier to satisfy Hyp. 1, because if $C_C=C_L$ then:

$$\text{if} \quad C_L >> C_1 \quad \text{then} \quad \text{also} \; C_C >> C_1 \tag{1.51}$$

The rule of thumb is no more convenient if C_L is so large that C_C occupies too much silicon area. Note that the amplifier can be designed to drive loads that are external to the chip, so that C_L may be relatively large, even up to a few nF. In these cases, it can be convenient to choose smaller G_{m1}/G_{m2} ratios in order to reduce C_C. For example, we can choose:

$$\frac{G_{m1}}{G_{m2}} = \frac{1}{5\sigma}$$
$$\Rightarrow C_C = \frac{1}{5} C_L \tag{1.52}$$

It should be observed that there is no risk that C_C becomes too small to satisfy Hyp.1, since we have assumed that C_L is particularly large.

Finally, it is important to consider what happens if we have designed an op-amp for a certain C_L value and the same op-amp is used with a smaller C_L. Note that C_L is generally a maximum load capacitance specification, thus the amplifier has to remain stable in closed loop conditions even if C_L is completely removed. Let us consider the consequences of reducing C_L (or completely removing it) with respect to the value used to design the amplifier:

- The unity gain angular frequency ω_0 does not change, since it is given by (1.34) where C_L is not present.
- The first non-dominant pole ω_2 is affected by C_L since it is included into C_2. Reducing C_L will probably lead to violate Hyp.1 and Hyp.2. Therefore, the expression of ω_2 given in (1.33) should be used. Reducing C_2, reduces also C_S. Thus, ω_2 shifts to higher frequencies for two concurrent reasons: (i) denominator (C_1+C_2) decreases and (ii) also the ratio C_S/C_C decreases.

For these reasons, if we reduce C_L with respect to the specified value, the ratio ω_2/ω_0 increases and so does the phase margin. On the contrary, if we increase the value of C_L over the value used to design the amplifier, the phase margin will progressively decrease and will eventually get negative, resulting in unstable closed loop configurations. This situation is common to most two-stage amplifiers.

Relationship between GBW and slew rate

The slew rate is the maximum rate by which the amplifier output voltage can increase and decrease. The effect occurs when one of the amplifier stage saturates and its current reaches the maximum value. In a two-stage op-amp, slew rate is generally due to saturation of the input stage. As Fig. 1.10 shows, the combination of the amplifier second stage (inverting) and compensation capacitor forms a Miller integrator having the output current of the first stage as its input. It can be demonstrated that this schematization is valid for frequencies much higher than the dominant pole f_p.

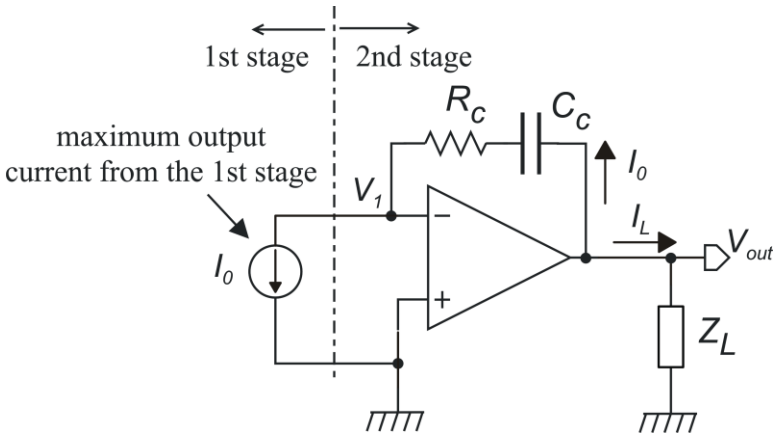

Fig.1.10. Simplified model used to determine the slew rate.

Due to its small value, the resistance R_C can be neglected while considering a first order approximation of the integrator step response. The output load, represented in Fig. 1.10 by Z_L, can be considered equal to capacitance C_2, which, as we have seen, typically coincides with the load capacitance. If the first stage saturates, it feeds the integrator with a constant current. For the op-amp that we have considered so far (Fig. 1.5), the maximum current is $\pm I_0$. If the output stage is capable of providing a current equal to $I_0 + I_L$, where I_L is the current delivered to the load, then the maximum time derivative of the output voltage, corresponding to the slew rate s_R is given by:

$$s_R = \max\left|\frac{dv_{out}}{dt}\right| = \frac{I_0}{C_C} \tag{1.53}$$

Expression (1.53) is useful to relate the slew rate to the gain bandwidth product. Considering (1.34), the unity gain angular frequency ω_0 is given by G_{m1}/C_C. Therefore:

. $$s_R = \omega_0 \frac{I_0}{G_{m1}} \tag{1.54}$$

For the considered amplifier topology, $G_{m1}=g_{m1}$ and $I_0=2I_{D1}$ and the slew rate is given by:

$$s_R = \omega_0 \frac{2I_{D1}}{g_{m1}} = 2\omega_0 V_{TE1} \tag{1.55}$$

Note that $\omega_0 = 2\pi \cdot GBW$. Thus, if the *GWB* is fixed, a higher slew rate can be obtained with a larger input-equivalent thermal voltage (V_{TE}), which, in turn, means an high overdrive voltage. In particular, for a MOSFET in strong inversion, Eq. (1.55) becomes: $s_R=\omega_0(V_{GS}-V_t)$. On the other hand, if the same op-amp of Fig. 1.5 is implemented with BJTs, V_{TE} is simply given by $V_T=kT/q$. In this respect, for the same GBW, the BJT version will have a smaller slew rate than all MOSFET versions.

Finally, we note that for particularly large load capacitances, the output stage could not be able to produce a total current equal to I_0+I_L, where $I_L=C_L dV_{out}/dt$. In this case, it is the output stage to limit the slew rate and the expressions found in this section do not apply.

Design for input referred voltage noise.

In order to calculate the noise of a given amplifier, we have to add noise sources to each device of the circuit. For MOSFETs up to frequencies of several hundred MHz it simply possible to model noise by adding a noise current source across the drain and source terminals. After that, the output voltage (output noise voltage) caused by the simultaneous action of all device noise sources is calculated. From the output noise, the input noise can be obtained by simply dividing the output noise by the amplification. In a multistage circuit, like that the amplifier that we are analyzing, it is convenient to study each stage separately obtaining simplified equivalent circuits, which can then be used to build the complete amplifier noise model. When the amplifier has a relatively large output resistance, the input noise estimation can be simplified if we calculate the effect of the noise sources on the output short circuit current instead of the output voltage. In the absence of noise, we can express the relationship between the output short circuit current ($i_{o\text{-}sc}$) as:

$$i_{o-sc} = Y_m v_{in} \tag{1.56}$$

where Y_m is an admittance that is generally a function of frequency and we have considered that an output current that enters the output node is positive. The amplifier voltage gain, A_V, is then simply given by:

$$A_V = -Y_m Z_{out} \tag{1.57}$$

where Z_{out} is the output impedance of the amplifier. When we consider the effect of noise sources, we can express the output noise voltage as:

$$v_{on} = -Z_{out} i_{on-sc} \tag{1.58}$$

where $i_{on\text{-}sc}$ is the output short circuit current produced by the noise sources. Then, the input referred noise, $v_n = -v_{on}/A_V$ is simply equal to:

$$v_n = \frac{-i_{on-sc}}{Y_m} \tag{1.59}$$

The advantage of this approach is that calculation of the output short circuit current is generally simpler in amplifiers that has a large output resistance.

We can start by applying this method to first stage of the operation amplifier of Fig. 1.5. The first stage, depicted in Fig. 1.11 (a), is replaced by the small-signal equivalent circuit shown in Fig. 1.11 (b), where the noise sources of all circuit have been introduced and the output port is short-circuited. The noise output short circuit current of this stage is indicated with $i_{1n\text{-}sc}$.

This circuit presents only a high impedance node, which coincides with the output port. The output impedance does not affect the output short circuit current, thus the singularities (poles and zeroes) that we have to take into account for the calculation of both Y_m and $i_{1n\text{-}sc}$ comes from low resistances nodes (such as node K, for example) and then they are located at frequencies of the same order of f_0. In this noise analysis, we will consider only frequencies that are significantly smaller than f_0 (at least one

decade lower), and thus we will neglect these singularities and we will perform the calculation of both Y_m and $i_{occ\text{-}n}$ in the low frequency limit. It can be easily found that:

$$Y_m = G_{m1} = g_{m1} \tag{1.60}$$

As far as $i_{1n\text{-}sc}$ is concerned, we can introduce current gains A_{Ik}, which represent the transfer function from current source i_{nk} and the output short circuit current. With this definition:

$$i_{1n-sc} = \sum_{k=1}^{M} A_{Ik} i_{nk} \tag{1.61}$$

where, for the amplifier in Fig. 1.11, M=5. Coefficients A_{Ik} for i_{n3} and i_{n4} can be easily found: i_{n4} flows directly into the output short circuit, and then A_{I4}=1, while i_{n3} reaches node H through the M3-M4 mirror that introduces a sign inversion. Then, A_{I3}= −1.

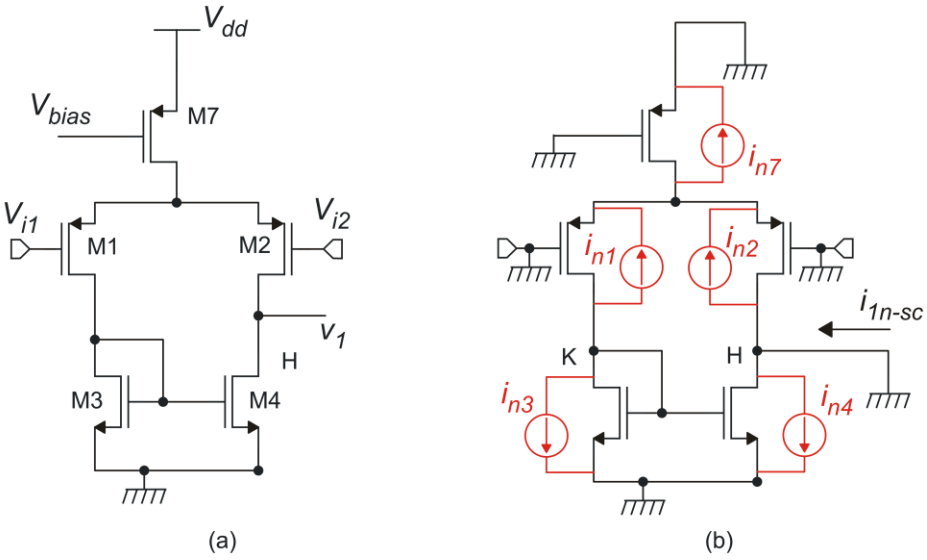

Fig. 1.11. Fisrt stage of the op-amp (a) and its equivalent small-signal circuit for the calculation of the noise output short circuit current.

Source i_{n7} produces negligible effects on the output current since it is split into two equal components that flows through M1 and M2, respectively. The component through M2 reaches node H directly, while the component through M1 encounters M3-M4 mirror, and then it is reversed. As a result, the two components give opposite contributions to the output current and then the net effect if i_{n7} is zero. Note that this occurs only in quiescent conditions, i.e. when V_{id} = 0. If a non-zero input differential signal is present, in7 is split into two components that are not equal any more, and then they do not cancel each other on the output node H. In the following part of this section, we will continue to perform noise analysis in the quiescent point of the amplifier, so that we will consider $A_{I7} \cong 0$. However, it is important to know that if a relatively large input signal is present, also i_{n7} gives a non-negligible contribution to the noise of the first stage.

The situation is slightly more complicated for i_{n1} and i_{n2}, since they do not have a terminal connected to gnd (they are "floating sources"). To simplify the analysis, they can be split into two current sources as shown in Fig. 1.12. For the equivalence to hold true, currents $i_{n1\text{-}a}$ and $i_{n1\text{-}b}$ should be equal to i_{n1}.

Component $i_{n1\text{-}a}$ reaches the output node H in the same way as i_{n3}, then it is simply multiplied by –1. Component $i_{n1\text{-}b}$ follows the same paths as i_{n7} and then do not give a significant contribution. As a result, i_{n1} gives a contribution to the output short circuit current equal to $-i_{n1}$, so that $A_{I1}= -1$. With the same procedure, we can split i_{n2} into two components that have one terminal connected to gnd. The only component that gives a significant contribution is connected directly to node H and then is gives a contribution such that $A_{I2}=1$.

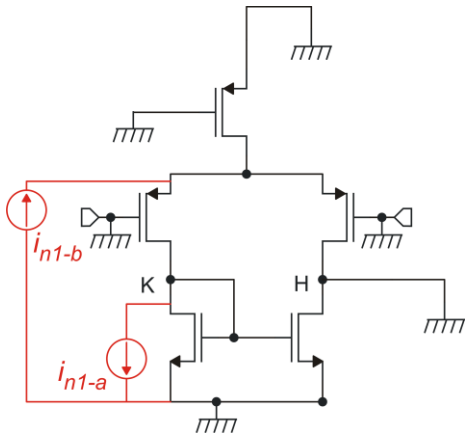

Fig. 1.12. Substitution of floating source i_{n1} with two sources $i_{n1\text{-}a}$ and $i_{n1\text{-}b}$ with one terminal to gnd.

Then, we are ready to write an approximate expression of the output short circuit current of the first stage:

$$i_{1n-sc} = i_{n2} - i_{n1} + i_{n4} - i_{n3} \tag{1.62}$$

The input referred noise of the first stage can be simply found by dividing $i_{1n\text{-}sc}$ by the admittance Y_m of the first stage, given by (1.60). Before doing that, we need to consider also the effect of the second stage. Repeating the same procedure to the second stage, depicted in Fig. 1.13, it can be easily found that:

$$i_{2n-sc} = i_{n5} + i_{n6} \tag{1.63}$$

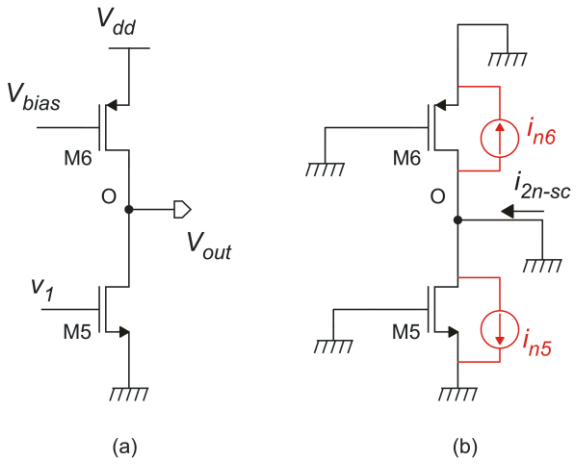

Fig. 1.13. (a) Second stage of the op-amp and (b) its equivalent circuit for the output noise short circuit current.

Now we can model the op-amp noise using a simplified equivalent small signal circuit similar to that of Fig. 1.6, where the noise of the first and second stage is simply represented by their output noise short circuit currents. This equivalent circuit is shown in Fig. 1.14 (a).

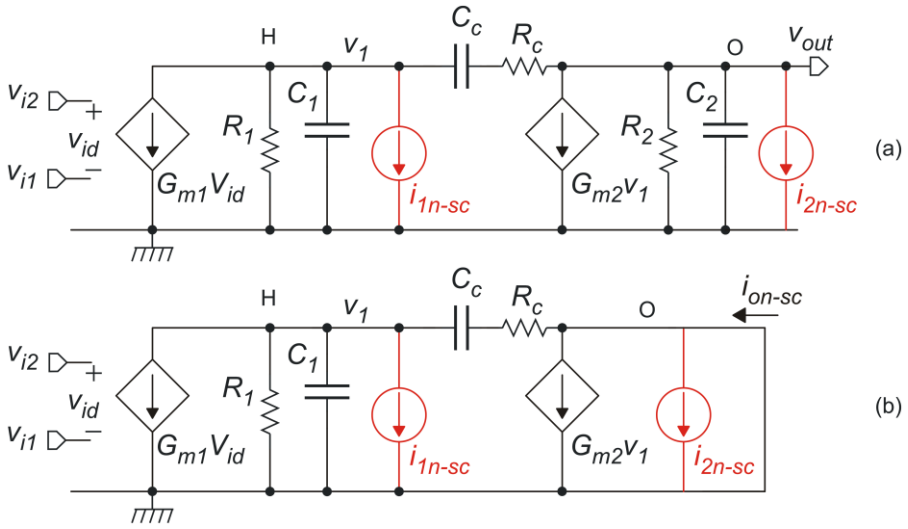

Fig. 1.14. Low frequency small signal equivalent model with the first stage and second stage noise sources.

We can apply the procedure used for the two individual stages to the overall amplifier and calculate the output short circuit current, $i_{on\text{-}sc}$ and the admittance $Y_m = i_{o\text{-}sc}/v_{id}$. The circuit to be considered is shown in Fig. 1.14 (b), where the output port has been short-circuited and superfluous components have been removed. Let us start by considering $i_{on\text{-}sc}$:

$$i_{on-sc} = i_{2n-sc} + A_{Ih} i_{1n-sc} \tag{1.64}$$

where A_{Ih} is the current transfer function from current source $i_{1n\text{-}sc}$ and the output short circuit current. By simple considerations, we can write:

$$A_{Ih} = -Z_A\left(G_{m2} - Y_C\right) \tag{1.65}$$

where Z_A is the impedance seen by source $i_{1n\text{-}sc}$ and Y_C is the admittance of the series R_C,C_C. Then:

$$Y_C = \frac{1}{R_C + \dfrac{1}{sC_C}} = \frac{sC_C}{1 + sR_CC_C} \tag{1.66}$$

$$Z_A = \frac{1}{\dfrac{sC_C}{1 + sC_CR_C} + \dfrac{1}{R_1} + sC_1} = \frac{R_1\left(1 + sC_CR_C\right)}{1 + s\left(C_CR_1 + C_CR_C + C_1R_1\right) + s^2R_CC_CR_1C_1} \tag{1.67}$$

Substituting the expressions of Y_C and Z_A into (1.65), we obtain:

$$A_{Ih} = -\frac{R_1}{1 + s\left(C_CR_1 + C_CR_C + C_1R_1\right) + s^2R_CC_CR_1C_1}\left[G_{m2} - sC_C\left(1 - G_{m2}R_C\right)\right] \tag{1.68}$$

With the choice R_C=1/G_{m2} that we have done to eliminate the RHP zero, we finally obtain:

$$A_{Ih} = -\frac{G_{m2}R_1}{1 + s\left(C_CR_1 + C_CR_C + C_1R_1\right) + s^2R_CC_CR_1C_1} \tag{1.69}$$

Considering that A_{Ih} is also the transfer function from source $G_{m1}v_{id}$ and the output short circuit current, we can also express the admittance Y_m as:

$$Y_m = G_{m1}A_{Ih} \tag{1.70}$$

Considering (1.64) and (1.59) we can now calculate the input referred noise of the op-amp:

$$v_n = -\frac{i_{2n-sc} + A_{Ih}i_{1n-sc}}{G_{m1}A_{Ih}} = -\frac{1}{G_{m1}}\left(i_{1n-sc} + \frac{i_{2n-sc}}{A_{Ih}}\right) \tag{1.71}$$

Notice that current $i_{2n\text{-}sc}$ is divided by A_{Ih} before being summed to $i_{1n\text{-}sc}$. At low frequencies, A_{Ih} is equal to the $G_{m2}R_1$ product, which is much greater than one. This can be easily understood because G_{m2} is generally larger than G_{m1} (see the section on stability and *GBW*), thus $G_{m2}R_1$>$G_{m1}R_1$ and $G_{m1}R_1$>>1 because $G_{m1}R_1$ is the gain of the first stage. Therefore, at low frequencies the contribution of $i_{2n\text{-}sc}$ can be neglected. The magnitude of A_{Ih} begins to reduce at high frequencies due to the presence of s and s^2 terms in the denominator. Then, at high frequencies $i_{n2\text{-}sc}$ cannot be neglected any longer and becomes the dominating terms at frequencies where A_{Ih} becomes <<1. This cause the well-known increase of the input referred noise at frequencies around and beyond the amplifier *GBW*. These effects are generally not important, since the increase of the input referred noise, which is a mathematical effect well described by (1.71) and (1.69), is counteracted by the filtering effects produced by the amplifier in all possible closed loop configurations. For these considerations, we will use (1.71), taking into account $i_{1n\text{-}sc}$ contribution. Taking expression (1.62) for $i_{1n\text{-}sc}$, we obtain:

$$v_n = \frac{i_{n1} - i_{n2} + i_{n3} - i_{n4}}{G_{m1}} \tag{1.72}$$

The power spectral density (PSD) of the input referred noise, S_{vn}, will be given by:

$$S_{vn} = 2\frac{S_{i1} + S_{i3}}{G_{m1}^2} \tag{1.73}$$

where S_{i1} and S_{i3} are the current PSDs of i_1 and i_3. Due to the symmetry of the circuit we have considered $S_{i2}=S_{i1}$ and $S_{i4}=S_{i3}$. All the noise currents have been considered to be independent stochastic processes. A more useful expression can be obtained by transforming the MOSFET current PSDs into the corresponding gate-referred voltage PSDs, according to:

$$S_I = g_m^2 S_V \tag{1.74}$$

The input referred noise PSD of the amplifier becomes:

$$S_{vn} = 2\frac{g_{m1}^2 S_{v1} + g_{m3}^2 S_{v3}}{G_{m1}^2} \tag{1.75}$$

In the considered amplifier topology, $G_{m1}=g_{m1}$, therefore:

$$S_{vn} = 2\left(S_{v1} + \frac{g_{m3}^2}{g_{m1}^2} S_{v3} \right) \tag{1.76}$$

Indicating with F the ratio g_{m3}/g_{m1} we can write the following compact formula:

$$S_{vn} = 2\left(S_{v1} + F^2 S_{v3}\right) \tag{1.77}$$

The parameter F can be used to reduce the effect of the mirror MOSFETs on the input referred noise. Writing g_m as I_D/V_{TE}, we get:

$$F = \frac{I_{D3}}{I_{D1}}\frac{V_{TE1}}{V_{TE3}} \qquad \text{since } I_{D1} = I_{D3} \Rightarrow \quad F = \frac{V_{TE1}}{V_{TE3}} \tag{1.78}$$

In order to reduce the noise contribution of the mirror MOSFETs, we have to make the equivalent thermal voltage V_{TE} of the mirror devices (M_3,M_4) much larger than that of the input devices. This is usually accomplished by biasing M1 and M2 in weak inversion, or, at least at the lower end of the strong inversion, with $|V_{GS}-V_t|_1$ nearly equal to 100 mV. The mirror devices should be biased in strong inversion, preferably with $(V_{GS}-V_t)_3$ of the order of several hundreds mV. This is paid in terms of input common mode range and, considering that $(V_{GS}-V_t)_3=(V_{GS}-V_t)_5$, also in terms of output swing.

Case 1: Thermal Noise.

Thermal noise can be calculated using (1.76) with the expression $S_v=(8/3)kT/g_m$ for S_{v1} and S_{v3}:

$$S_{vn} = 2\left(\frac{8}{3}kT\frac{1}{g_{m1}} + \frac{g_{m3}^2}{g_{m1}^2}\frac{8}{3}kT\frac{1}{g_{m3}}\right) \tag{1.79}$$

Applying elementary simplifications, we obtain:

$$S_{vn} = 2\left(\frac{8}{3}kT\frac{1}{g_{m1}}\right)(1+F) \tag{1.80}$$

The input thermal noise can is then given by the voltage thermal noise of the input devices (M1 and M2) multiplied by a factor (1+*F*) > 1, which takes into account the additional contribution of the mirror devices M3 and M4. Again, in order to minimize the effect of he latter and be enabled to consider that the noise comes from only the input devices, $(V_{GS}-V_t)_3$ should be much larger than $|V_{GS}-V_t|_1$, penalizing the input and output ranges.

Now let us express g_{m1} in (1.80) as I_{D1}/V_{TE1}. After obvious simplifications, we obtain:

$$S_{vn} = \frac{16}{3}kT\frac{V_{TE1}}{I_{D1}}(1+F) \tag{1.81}$$

The following considerations can be drawn:

- A small thermal noise PSDs is mainly paid in terms of current, i.e power consumption. Remember that the bias current of the first stage, I_0, is equal to $2I_{D1}$. The higher I_0, the smaller S_{v1}.
- Keeping small the input device overdrive voltage, $(V_{GS}-V_t)_1$, helps reducing the input noise.

Case 2: Flicker Noise.

For M1 and M3 gate referred noise the expression $fS_v(f)=N_f/(WL)$ will be used. Equation (1.77) becomes:

$$f \cdot S_{vn}(f) = 2\left(\frac{N_{fp}}{W_1L_1} + F^2\frac{N_{fn}}{W_3L_3}\right) \tag{1.82}$$

where N_{fn} and N_{fp} are the flicker noise process parameters for *n*-MOS and *p* MOS, respectively. Note that:

- a low flicker noise is paid in terms of silicon area
- it is possible to save area by reducing the effect of the mirror devices choosing small *F* factors.

Design for input offset voltage.

In previous chapter, we have seen that MOSFET parameter variations can be modeled as d.c. current sources placed across the drain and source of the nominal (ideal) device. As in the case of noise, the circuit of Fig.1.11 and the analysis that follows can still be used. Then, expression (1.72) can be used also for the offset voltage:

$$v_{io} = \frac{i_{1p} - i_{2p} + i_{3p} - i_{4p}}{G_{m1}} \tag{1.83}$$

where currents i_{ip} is the currents modeling the process variations of i-th MOSFET. In the following analysis, we will consider that all devices are in strong inversion. Since M1 and M2 as well as M3 and M4 form matched pairs, we can write:

$$\begin{aligned} i_{1p} - i_{2p} &\equiv \Delta I_{D1,2} = I_{D1}\left(\frac{\Delta\beta_{1,2}}{\beta_{1,2}} - \frac{2\Delta V_{t1,2}}{\left|V_{GS} - V_t\right|_1}\right) \\ i_{3p} - i_{4p} &\equiv \Delta I_{D3,4} = I_{D3}\left(\frac{\Delta\beta_{3,4}}{\beta_{3,4}} - \frac{2\Delta V_{t3,4}}{\left(V_{GS} - V_t\right)_3}\right) \end{aligned} \tag{1.84}$$

Substituting the expressions in (1.84) into (1.83) and considering that $I_{D1}=I_{D3}$, and $G_{m1}=g_{m1}$, we get:

$$v_{io} = \frac{I_{D1}}{g_{m1}}\left(\frac{\Delta\beta_{1,2}}{\beta_{1,2}} + \frac{\Delta\beta_{3,4}}{\beta_{3,4}} - \frac{2\Delta V_{t1,2}}{\left|V_{GS} - V_t\right|_1} - \frac{2\Delta V_{t3,4}}{\left(V_{GS} - V_t\right)_3}\right) \tag{1.85}$$

Finally, in strong inversion $I_{D1}/g_{m1}=(V_{GS}-V_t)_1/2$, so that the expression of the input offset voltage becomes:

$$v_{io} = \frac{\left(V_{GS} - V_t\right)_1}{2}\left(\frac{\Delta\beta_{1,2}}{\beta_{1,2}} + \frac{\Delta\beta_{3,4}}{\beta_{3,4}}\right) - \Delta V_{t1,2} - F\Delta V_{t3,4} \tag{1.86}$$

where F is given by equation (1.78).

Equation (1.86) can be used to calculate the standard deviation of the offset voltage. Considering that all random variables are independent, we can write:

$$\sigma^2_{vio} = \frac{\left(V_{GS} - V_t\right)_1^2}{4}\left(\sigma^2_{\frac{\Delta\beta_{1,2}}{\beta_{1,2}}} + \sigma^2_{\frac{\Delta\beta_{3,4}}{\beta_{3,4}}}\right) + \sigma^2_{\Delta V_{t1,2}} + F^2\sigma^2_{\Delta V_{t3,4}} \tag{1.87}$$

The standard deviations can be expressed in terms of the p-MOS and n-MOS matching parameters $C_{\beta p}$, C_{Vtp}, $C_{\beta n}$ and C_{Vtn} according to:

$$\sigma_{\frac{\Delta\beta_{1,2}}{\beta_{1,2}}} = \frac{C_{\beta p}}{\sqrt{W_1 L_1}};\ \sigma_{\frac{\Delta\beta_{3,4}}{\beta_{3,4}}} = \frac{C_{\beta n}}{\sqrt{W_3 L_3}};\ \sigma_{\Delta V_{t1,2}} = \frac{C_{Vtp}}{\sqrt{W_1 L_1}};\quad \sigma_{\Delta V_{t3,4}} = \frac{C_{Vtn}}{\sqrt{W_3 L_3}} \tag{1.88}$$

Using these expressions equations (1.87) becomes:

$$\sigma_{vio}^2 = \frac{A}{W_1 L_1} + \frac{B}{W_3 L_3} \tag{1.89}$$

where *A* and *B* are given by:

$$A = \frac{(V_{GS} - V_t)_1^2}{4} C_{\beta p}^2 + C_{Vtp}^2 \qquad B = \frac{(V_{GS} - V_t)_1^2}{4} C_{\beta n}^2 + F^2 C_{Vtn}^2 \tag{1.90}$$

Equation (1.89) indicates that a low offset voltage is paid mainly in terms of area. Once the target $_{vio}$ is given, constants *A* and *B* have to be minimized as much as possible in order to save area. As shown by (1.90), the input overdrive voltage $(V_{GS}\text{-}V_t)_1$ has to be reduced to the minimum value (around 0.1 V), while *F* should be made small by making $(V_{GS}\text{-}V_t)_3$ several times larger than $(V_{GS}\text{-}V_t)_1$.

Once these operations have been performed, M1 and M3 gate areas have to be chosen to obtain the required offset voltage ($_{vio}$). Clearly, there are infinite solutions to Eqn. (1.89), because we have two unknowns (W_1L_1, W_3L_3). If there are no other specifications, it is convenient to find the solution that minimizes the total gate area of M1 and M3, defined as $S = W_1L_1 + W_3L_3$. If we introduce the following unknown:

$$a = \frac{W_3 L_3}{W_1 L_1} \tag{1.91}$$

we can rewrite (1.89) as:

$$\sigma_{vio}^2 = \frac{1}{W_1 L_1}\left(A + \frac{B}{a}\right) \tag{1.92}$$

Calculating W_1L_1 from (1.92) and substituting it into the expression of the area, we find:

$$S = W_1 L_1 (1 + a) = \frac{1}{\sigma_{vio}^2}\left(A + \frac{B}{a}\right)(1 + a) \tag{1.93}$$

It can be easily shown that the previous expression tends to infinity when *a* tends either to zero or to infinity. Therefore, a minimum should exist. Calculating the derivative of (1.93) with respect to *a* and equating it to zero, we find the optimum value of *a* that minimizes M1 and M3 total area:

$$a_{OPT} = \sqrt{\frac{B}{A}} \tag{1.94}$$

Substituting a_{OPT} into (1.92) we finally find W_1L_1:

$$W_1 L_1 = \frac{1}{\sigma_{vio}^2}\left(A + \frac{B}{a_{OPT}}\right) = \frac{1}{\sigma_{vio}^2}\left(A + \sqrt{AB}\right) \tag{1.95}$$

Note: the same optimization procedure can be applied to the flicker Noise, since Eqn. (1.82) is formally identical to Eqn. (1.89).

Power consumption.

The power consumption is given by:

$$P = V_{DD} I_{Supply} \tag{1.96}$$

where I_{supply} is the total supply current, equal to:

$$I_{supply} = I_0 + I_1 = 2I_{D1} + I_{D5} \tag{1.97}$$

Considering that we can write the drain current $I_D = g_m V_{TE}$, the supply current becomes:

$$I_{supply} = 2g_{m1}V_{TE1} + g_{m5}V_{TE5} \tag{1.98}$$

This equation can be specialized to emphasize the role of either g_{m1} or g_{m5}. In the case that the dominant specification is the input thermal noise, it is important to show the dependence of g_{m1}, since the thermal noise input spectral density is marked by an inverse proportionality to g_{m1}. Then:

$$I_{\text{supply}} = g_{m1}V_{TE1}\left(2 + \frac{1}{r_{gm}F}\right) \tag{1.99}$$

where:

$$r_{gm} = \frac{g_{m1}}{g_{m5}}; \quad F = \frac{V_{TE1}}{V_{TE3}} = \frac{V_{TE1}}{V_{TE5}} \tag{1.100}$$

In the case that the dominant role is the *GBW*, it is important to design g_{m5} in such a way that (1.43) holds. Therefore, we can transform (1.98) to highlight the role of g_{m5}:

$$I_{\text{supply}} = g_{m5} V_{TE5} \left(1 + 2 r_{gm} F\right) \tag{1.101}$$

Considering, (1.43) we can directly relate the current consumption to the *GBW*:

$$I_{\text{supply}} = 2\pi\sigma V_{TE5} C_L GBW \left(1 + 2 r_{gm} F\right) \tag{1.102}$$

1.4 References

[1] B. Pellegrini, "Considerations on the feedback theory," Alta Frequenza, vol. 41, pp. 825-820, Nov. 1972. Available at: http:// brahms.iet.unipi.it/elan/scompos.pdf.

[2] G. Gregorian, G.C. Temes, "Analog MOS Integrated Circuits for Signal Processing", 1st edition, John Wiley & Sons, New York, 1986, pp.172-176.

[3] G. Palmisano, G. Palumbo, S. Pennisi "Design Procedure for Two-Stage CMOS Transconductance Operational Amplifiers: A Tutorial" Analog Integrated Circuits and Signal Processing, vol. 27, pp. 179-189, 2001.

[4] P. Bruschi, D. Navarrini, G. Tarroboiro, G. Raffa, "A Computationally Efficient Technique for the Optimization of Two Stage CMOS Operational Amplifiers", proceedings of ECCTD'03 - vol. III, pp. 305-308, Cracovia, September 1-4 2003