EE 348: LECTURE AID #08

Small Signal Models of MOS Technology Transistors

Dr. John Choma
Professor of Electrical Engineering

University of Southern California
Ming Hsieh Department of Electrical Engineering
University Park: Mail Code: 0271
Los Angeles, California 90089–0271

213–740–4692 [USC Office]
213–740–7581 [USC Fax]
johnc@usc.edu

PRELUDE:

In this supplement to formal class lectures, we develop the small signal models for metal-oxide-semiconductor field-effect transistor (MOSFET) technologies. Despite the exclusive focus on MOS technologies, the reader will surmise that the fundamental concepts and strategies of small signal modeling presented herewith are generally applicable to bipolar, III-V compound, and other semiconductor devices that offer an input port for signal application and an output port for signal responses. To this end, we shall assert that while the specific branch parameter values of the models we develop assuredly change as a function of the addressed device technology, the circuit topologies of virtually all small signal device models remain invariant with technology.

In addition to formulating and parametrically defining the aforementioned models, we seek to establish a general design strategy that optimizes the performance of MOSFET analog networks. In this set of notes, performance optimization is largely construed as ensuring low power dissipation while achieving nominally linear input/output (I/O) signal processing. However, a cursory investigation of performance optimization at high signal frequencies is also embraced.

MARCH 2012

8.1.0. INTRODUCTION

The MOSFET, like all semiconductor transistors, is inherently nonlinear. In the case of the n-channel (or NMOS) MOSFET abstracted in Figure (8.1a), this nonlinearity is witnessed at low frequencies as a drain current, I_{dn}, which nonlinearly relates to gate-source voltage V_{gs}, drain-source voltage V_{ds}, and bulk-source voltage V_{bs}. For the p-channel (or PMOS) transistor in Figure (8.1b), similar nonlinearities prevail in that drain current I_{dp} is a nonlinear function of source-gate voltage V_{sg}, source-drain voltage V_{sd}, and source-bulk voltage V_{sb}. We may generalize these assertions by writing for NMOS transistors,

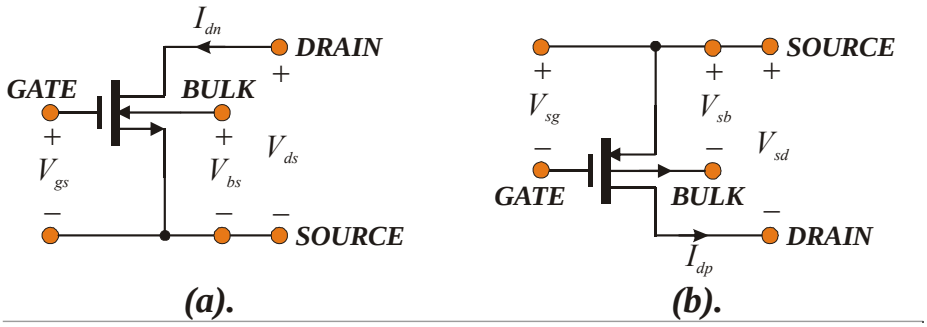

Figure (8.1). **(a).** Schematic portrayal of an n-channel (NMOS) MOSFET. **(b).** Schematic portrayal of a p-channel (PMOS) MOSFET.

$$I_{dn} = f_n(V_{gs}, V_{ds}, V_{bs}), \tag{8-1}$$

while for PMOS,

$$I_{dp} = f_p(V_{sg}, V_{sd}, V_{sb}). \tag{8-2}$$

The static nonlinearities that the foregoing algebraic relationships imply are exacerbated at high signal frequencies by key device capacitances whose values vary with certain device voltages. We delineate these capacitances for the NMOS transistor exemplified in Figure (8.2). In this diagram, the bulk-drain capacitance, C_{bd}, is a depletion capacitance that is dependent on an inverse square root or cube root function of bulk-drain voltage V_{bd}. While the bulk terminal of a MOSFET is generally incident with signal ground, the drain voltage varies in response to the applied signal, thereby incurring a frequency dependent distortion and possibly, even a circuit bandwidth that fluctuates with signal frequencies. An analogous assertion applies to the bulk-source capacitance, C_{bs}, which is nonlinearly related to the bulk-source voltage, V_{bs}. To a lesser extent, the gate-source and gate-drain capacitances, C_{gs} and C_{gd}, respectively, also enter into the high frequency distortion fray for analog systems.

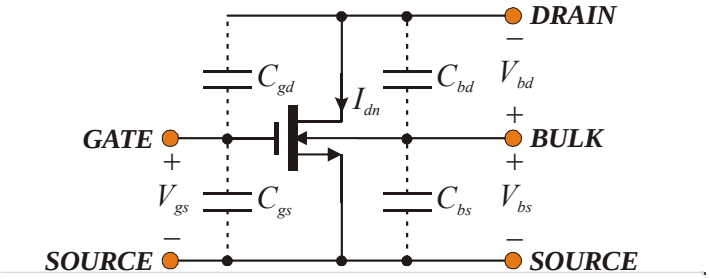

Figure (8.2). Schematic diagram of NMOS transistor showing key device capacitances that may degrade I/O linearity at high signal frequencies.

The fundamental dilemma posed for the analog circuit designer by transistor

nonlinearities is that the design objective for a majority of analog systems is input -to- output (I/O) behavior that closely emulates linearity. For example, in a stereo system, I/O linearity commendably guarantees that the signal frequencies witnessed at the system output port are only those that reign at the input port. In other words, the spectral response of the ideal stereo system is comprised of only those frequencies that are present in the compact disc (CD), audio tape, or digitized iPod signal. Such output is therefore free of the harmonic distortion precipitated by I/O nonlinearities. Similar issues prevail in a modern communications network in that nonlinearity, which produces harmonically rich versions of processed input signals, contaminates or even obscures, the input signals that are representative of the audio, video, or other data forms intended for I/O transmission. For example, if we are using a wireless means to monitor temperature, nonlinearities in the communications network utilized to transmit the electronically sensed temperature may promulgate erroneous or inaccurate thermal measurements.

In an attempt to ensure reasonable linearity in an electronic system, a three-prong design strategy is typically adopted. The first of these strategic procedural steps entails biasing, which ensures that for all values of the applied input signal, the transistors operate in desirable regions of their static volt-ampere characteristic curves where linearity can be at least viably approximated. The implication herewith is that certain portions of the MOSFET static characteristics reflect or promote nominal I/O linearity better than do other regions. Biasing, which derives from constant voltage and/or current sources that are applied to the analog network, constrains the embedded transistors to operate in these "more linear" regimes. The second strategic step involves applying signals that are sufficiently small to ensure that under dynamic operating conditions, wherein these signals modulate in accordance with the input information earmarked for electronic processing, all transistors remain confined to the nominally linear regimes of their static curves. This strategy may entail dividing large input signals by resistive or capacitive dividers or it may involve some sort of preprocessing that limits the amplitude excursions of the applied signals. To the latter end, a common preprocessor is a logarithmic amplifier in which the large amplitudes of given input signals are compressed logarithmically[1]. A third and very common design strategy applies suitable feedback around the offending amplifier. Such a tack capitalizes on the ability of stable feedback to reduce, often dramatically, the dependence of I/O network responses on the transistor parameters or characteristics that engender nonlinearities. Feedback also mitigates other circuit performance problems caused by parametric vulnerability to manufacturing tolerances and routine production uncertainties[2],[3].

8.2.0. COMMON SOURCE AMPLIFIER

The methods of small signal circuit analyses and the philosophical concepts that underpin these methods are best developed in conjunction with a design-oriented consideration of a practical MOSFET amplifier. To this end, consider the two common source amplifiers shown in Figure (8.3). We shall largely concentrate on the NMOS realization depicted in Figure (8.3a) but in the interest of completeness, the PMOS counterpart to the NMOS common source cell is offered in Figure (8.3b). Although much of the literature associates the word, "common" in the phrase, "common source amplifier," to the term "grounded," we note in Figure (8.3) that the source terminal is ungrounded in both configurations. In particular, a resistance, R_{ss}, which is known as a ***source degeneration resistance***, returns the source of the NMOS transistor to circuit ground, while in the PMOS version of the amplifier, R_{ss} serves as the branch vehicle that returns the transistor source terminal to the positive supply line, $+V_{ss}$. As we shall ultimately reveal, R_{ss} establishes the feedback to which the preceding section alludes as a means of promoting en-

hanced I/O network performance.

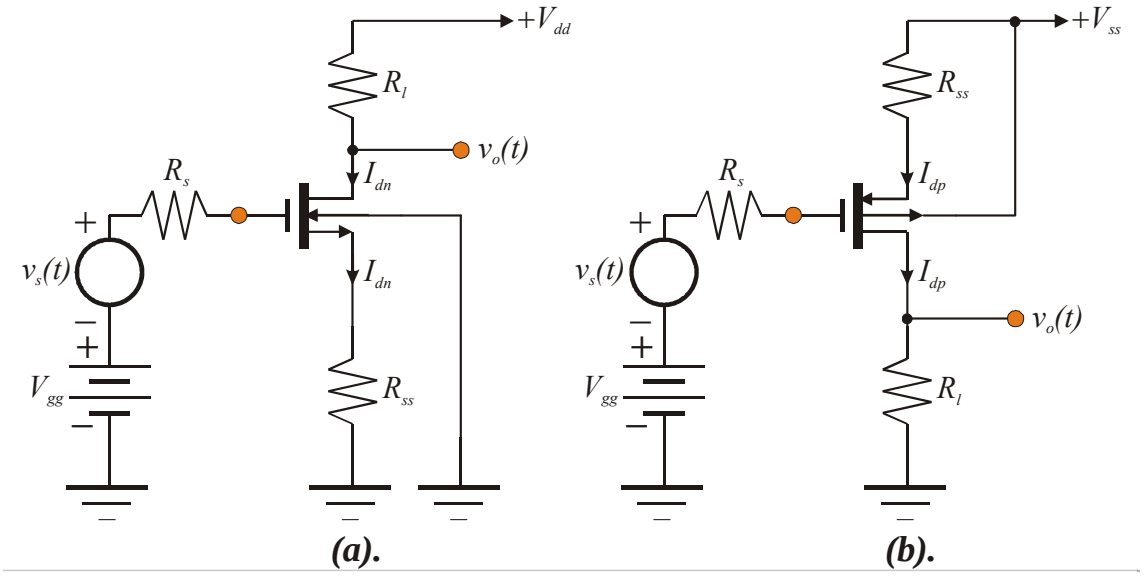

Figure (8.3). **(a).** Basic schematic diagram of a common source amplifier realized with an NMOS transistor. **(b).** The PMOS version of the NMOS common source amplifier in (a).

In contrast to much of the aforementioned literature, we shall conflate the word, "common," in the phrase, "common source amplifier," with the terminal that is used neither to receive the applied input signal nor to establish the resultant response to this signal. In Figure (8.3), we see that the signal voltage, V_s, which carries the information earmarked for amplifier processing, is applied to the gate terminal. On the other hand, the network response, $v_o(t)$, is extracted as a voltage with respect to ground at the drain terminal. Thus, the source terminal is used neither for input signal application nor for the delivery of the output signal response and therefore, we refer to the stage at hand as a common source amplifier. By way of illustrative reinforcement, the gate of a common gate amplifier need not be grounded. But the input signal in a common gate unit is always applied to the source terminal, while the response to this applied signal always derives from the drain terminal. Thus, the gate terminal neither receives the input signal nor delivers the response to the applied input excitation in a common gate amplifier. Analogously, the drain of a common drain amplifier does not receive the applied input, which is applied to the transistor gate terminal, and it does not deliver the output response, which is established at the source terminal.

The source signal, $v_s(t)$, in Figure (8.3) is the time varying signal that is to be amplified by the network. It can be a pulse train, an amplitude modulated signal, a frequency modulated signal, or any other form of time domain function that is reflective of the information or data to be processed. In our test deliberations, we shall take V_s to be the simple sinusoid,

$$v_s(t) = V_s \cos(\omega t + \theta), \tag{8-3}$$

where ω represents the radial frequency of the sinusoid, and θ is an arbitrary phase angle.

8.2.1. TRANSISTOR TURN ON

The fundamental purpose of voltage V_{gg} in Figure (8.3a) is to ensure turn of the transistor by establishing, for all values of applied signal, a gate -to- source voltage, V_{gs}, that is at least as large as the threshold potential, say V_{hn}, of the NMOS device. Since the transistor gate conducts no current at low signal frequencies, Figure (8.3a) confirms that

$$V_{gs} = \left(V_s + V_{gg}\right) - R_{ss} I_{dn} > V_{hn} . \tag{8-4}$$

It is important to understand that this inequality must be satisfied for all time; that is, for all values of the input signal, $v_s(t)$. Thus, when $v_s(t)$, which serves to perturb the value of the constant voltage, V_{gg}, assumes its most positive value of V_s, (8-4) delivers

$$\left(V_s + V_{gg}\right) - R_{ss} I_{dn+} > V_{hn} , \tag{8-5}$$

where I_{dn+} is the value of drain current corresponding to $v_s(t) = V_s$. On the other hand, if the drain current, I_{dn}, reduces to I_{dn-} when $v_s(t) = -V_s$,

$$\left(-V_s + V_{gg}\right) - R_{ss} I_{dn-} > V_{hn} . \tag{8-6}$$

The simultaneous satisfaction of (8-5) and (8-6) comprise the necessary and sufficient conditions for ensuring that the transistor in the network of Figure (8.3a) remains turned on, or conductive, for all anticipated values of the applied input signal. However, care must be exercised when applying (8-5) and (8-6) because we have presumed that the device threshold voltage, V_{hn}, is a constant. In truth, the presence of the source degeneration resistance establishes a nonzero bulk-source voltage for the transistor in the amplifier of Figure (8.3a) of $V_{bs} = -R_{ss} I_{dn}$. This nonzero voltage modulates threshold voltage V_{hn} in accordance with

$$V_{hn} = V_{ho} + 2\sqrt{V_\theta V_F}\left(\sqrt{1 - \frac{V_{bs}}{2V_F}} - 1\right) = V_{ho} + 2\sqrt{V_\theta V_F}\left(\sqrt{1 + \frac{R_{ss} I_{dn}}{2V_F}} - 1\right), \tag{8-7}$$

where V_{ho} is the zero bias (meaning $V_{bs} = 0$) value of the threshold voltage. Additionally,

$$V_\theta = \frac{qN_{sub}\varepsilon_s}{C_{ox}^2} = qN_{sub}\varepsilon_s\left(\frac{T_{ox}}{\varepsilon_{ox}}\right)^2 \tag{8-8}$$

is known as the ***bulk effect,*** or ***body effect, voltage***. The ***Fermi potential***, V_F in (8-7), which is one-half the potential established at the interface of the gate oxide layer and the strongly inverted channel that is indigenous to transistor turn on, is given by

$$V_F = V_T \ln\left(\frac{N_{sub}}{n_i}\right), \tag{8-9}$$

where V_T is the familiar Boltzmann voltage,

$$V_T = \frac{kT}{q} , \tag{8-10}$$

with k denoting Boltzmann's constant[1], q representing the magnitude of electron charge[2], and T symbolizing the absolute temperature of the inverted channel. In (8-9), n_i is the ***intrinsic carrier concentration*** of silicon. At $27\ ^\circ C$, $n_i \approx 10^{10}\ atoms/cm^3$, but n_i doubles -to- as much as quadruples for every $10\ ^\circ C$ rise in operating temperature. In (8-8) and (8-9), N_{sub} denotes the average impurity concentration of the bulk substrate, while ε_s is the dielectric constant of silicon[3].

Our tacit assumption of constant threshold potential becomes increasingly more tenable

[1] Boltzmann's constant, k, is $k = (1.38)(10^{-23})$ joules/°K.
[2] The magnitude, q, of electron charge is $q = (1.6)(10^{-19})$ coulombs.
[3] The dielectric constant, ε_s, of silicon is $1.05\ pF/cm$.

with decreased oxide thicknesses, T_{ox}. In particular, the voltage metric V_θ is proportional to the square of the oxide thickness, whence V_{hn} in (8-7) derives as the superposition of voltage V_{ho} and a corrective voltage term that is proportional to T_{ox}. An example of the amount by which the threshold potential is altered by the voltage developed across the source degeneration resistance in the amplifier of Figure (8.3a) is displayed graphically in Figure (8.4). In this graph, the substrate concentration of the utilized transistor is set to the reasonable value of 10^{15} *atoms/cm*^{3}. Note that the change in threshold potential, with respect to its $V_{bs} = 0$ value, is only about *6.8 mV* for a *40 Å* oxide thickness and $R_{ss}I_{dn}/2V_F = 1$ and double that amount *(≈ 13.6 mV)* for an *80 Å* gate oxide thickness.

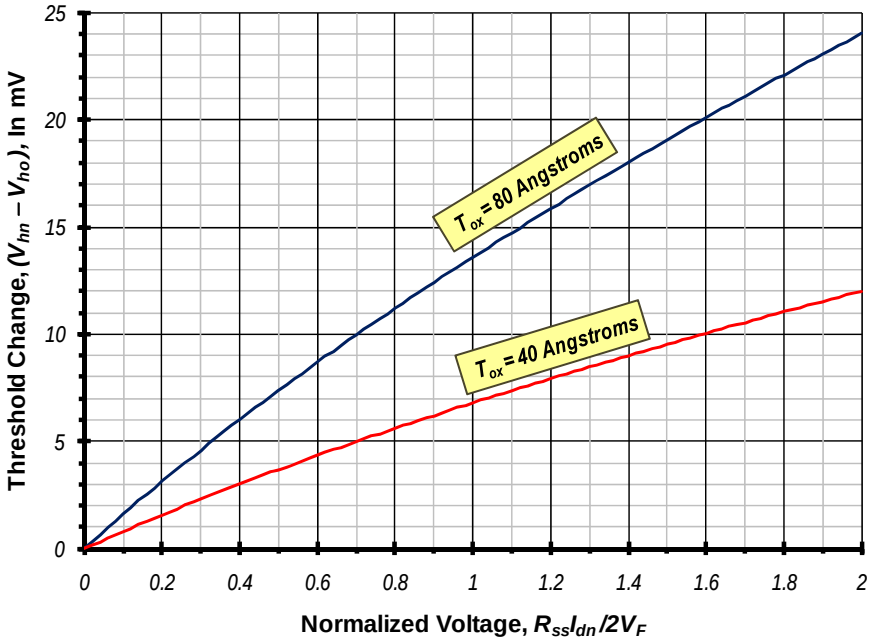

Figure (8.4). The threshold voltage sensitivity as a function of the voltage dropped across the source degeneration resistance in the amplifier of Figure (8.2a). The NMOS transistor is assumed to operate at room temperature *(27 °C)* and has a substrate concentration of $N_{sub} = 10^{15}$ *atoms/cm3*.

It is worthwhile interjecting that a commonly made design error is to ensure device turn on only for the zero signal condition implicit to $v_s(t)$ (or V_s) equal to zero. This zero signal circumstance defines the ***quiescent state*** of the amplifier. The quiescent operating state is also commonly referred to as the ***standby condition*** in that we can mentally paint the picture of a circuit "standing by" or just "hanging out" with biasing in the form of voltage V_{gg} (and voltage V_{dd}) applied, waiting for the signal to visit the amplifier input port. Under the no signal, or standby operating constraint, the drain current, I_{dn}, assumes its quiescent value, I_{dnQ}, and thus, (8-5) or (8-6) becomes

$$V_{gg} - R_{ss} I_{dnQ} > V_{hn} \,. \tag{8-11}$$

Clearly, (8-11) is a special case of (8-5) or (8-6) and as such, it forms only a necessary, but not sufficient, condition for device turn on. Instead, necessity and sufficiency of sustained turn on requires the simultaneous satisfaction of (8-5) and (8-6); that is, engineering design prudence dictates our consideration of the relevant effects of both positive and negative amplitude extremes in the input signal voltage.

8.2.2. OHMIC AND SATURATION REGIMES

When an NMOS transistor is turned on by ensuring that its gate-source voltage, V_{gs}, satisfies the constraint, $V_{gs} > V_{hn}$, it enters either its ohmic or saturation regimes of operation. In the ohmic domain, where the drain-source voltage, V_{ds}, of an NMOS is small enough to satisfy the inequality, $V_{ds} < (V_{gs} - V_{hn})$, the drain current, I_{dn}, abides by

$$I_{dn} \approx 2\beta_n V_{ds}\left(V_{gs} - V_{hn} - \frac{V_{ds}}{2}\right), \tag{8-12}$$

where the approximation sign is invoked to reflect the presumption that threshold voltage V_{hn} is nominally constant. Recall that constant, voltage invariant, threshold potential requires that the bulk-source voltage, V_{bs}, be maintained at zero or the gate oxide thickness, T_{ox}, be small. In (8-12), β_n, the so-called ***transconductance coefficient***, is given by

$$\beta_n = \frac{\mu_n C_{ox}}{2}\left(\frac{W}{L}\right) = \frac{\mu_n}{2}\left(\frac{\varepsilon_{ox}}{T_{ox}}\right)\left(\frac{W}{L}\right), \tag{8-13}$$

where (W/L), which symbolizes the ratio of gate width (W) -to- channel length (L) is known as the ***gate aspect ratio***. The gate aspect ratio is a designable variable in that the integrated circuit design engineer can select any gate width W and any channel length L for a particular transistor, as long as neither W nor L are smaller than the minimum geometric feature size allowed by the process foundry. Equations (8-12) and (8-13) confirm that through the gate aspect ratio, device drain currents can be scaled for given gate-source and drain-source voltage biases. Of course, design reasonableness must prevail when selecting the appropriate value of W. In particular, large W for fixed channel length fosters high drain currents, and thus high circuit power dissipation. Large W also manifests large device capacitances, which may incur degraded circuit bandwidth and circuit response speeds. In (8-13), μ_n denotes the mobility of the free electrons in the strongly inverted channel lying immediately beneath the gate oxide layer. For a bulk substrate whose impurity concentration does not exceed 10^{15} *atoms/cm*^{3}, mobility μ_n is about *420 cm*^{2}*/volt-sec*.

If V_{gs} remains larger than V_{hn} but now, $V_{ds} \geq (V_{gs} - V_{hn})$, the transistor leaves its ohmic regime of operation and enters its saturation domain, where drain current I_{dn} satisfies

$$I_{dn} \approx \beta_n\left(V_{gs} - V_{hn}\right)^2\left(1 + \frac{V_{ds} - V_{dsat}}{V_\lambda}\right). \tag{8-14}$$

In this expression,

$$V_{dsat} = V_{gs} - V_{hn} \tag{8-15}$$

is the ***drain saturation voltage***. The parameter, V_λ, accounts for ***channel length modulation***. This device metric is determined from the semi-empirical expression,

$$V_\lambda = L\left(\frac{V_j}{V_F}\right)^2\sqrt{\frac{32\,qN_{sub}}{\varepsilon_s}\left(V_{ds} - V_{dsat} + V_j\right)}, \tag{8-16}$$

where V_F is the Fermi potential introduced in (8-9), and, with N_d signifying the average donor impurity concentration in the implanted source and drain volumes of an NMOS device,

$$V_j = V_T \ln\left(\frac{N_d N_{sub}}{n_i^2}\right) \tag{8-17}$$

stipulates the ***built-in potential*** of the source-bulk and drain-bulk junctions. While (8-14) accounts for channel length modulation, it tacitly neglects other higher order phenomena, such as carrier mobility degradation, which arises from the high channel electric fields that commonly prevail in short channel and/or thin gate oxide devices, and drain-induced barrier lowering[4],[5]. We should note that when $V_{ds} = V_{dsat}$, I_{dn} takes on the value, I_{dsat}, which derives from (8-14) as

$$I_{dsat} \triangleq I_{dn}\Big|_{V_{ds}=V_{dsat}} = \beta_n\left(V_{gs} - V_{hn}\right)^2 = \beta_n V_{dsat}^2 \ . \tag{8-18}$$

Then, the immediate effect of channel length modulation is seen to increase the drain current above I_{dsat} as a function of the excess drain voltage, $(V_{ds} - V_{dsat})$. Specifically,

$$I_{dn} \approx I_{dsat}\left(1 + \frac{V_{ds} - V_{dsat}}{V_\lambda}\right). \tag{8-19}$$

This increased drain current is a mildly nonlinear function of both V_{ds} and V_{gs} owing to the dependence of parameter V_λ on a square root function of $(V_{ds} - V_{dsat})$. Figure (8.5) displays a representative set of drain current -versus- drain-source voltage characteristics for a transistor that features, $L = 0.25\ \mu m$, $T_{ox} = 80\ \text{Å}$, $N_d = 10^{21}\ atoms/cm^3$, $N_{sub} = 10^{15}\ atoms/cm^3$, and $W/L = 10$. Superimposed on these characteristic curves is a dashed curve that separates the ohmic and saturation regimes. This curve derives from (8-18) by allowing current I_{dsat} to assume the general drain current value of I_{dn} when drain-source voltage V_{ds} is set to its drain saturation value, $(V_{gs} - V_{hn})$.

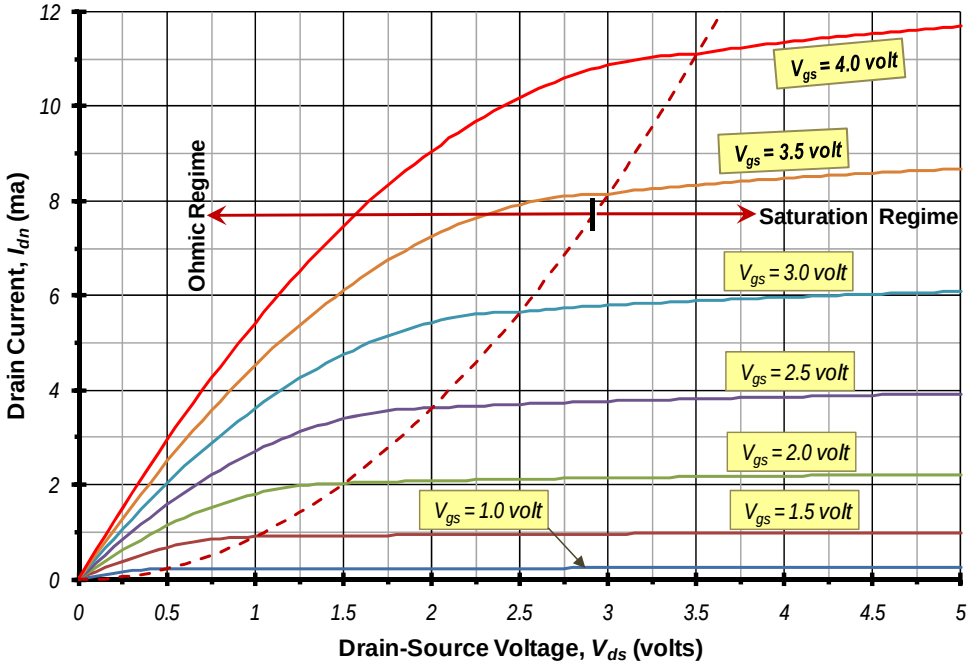

Figure (8.5). Static characteristic curves of an NMOS device having $L = 0.25\ \mu m$. The transistor boasts a gate aspect ratio of $W/L = 10$, a zero bias threshold voltage, V_{ho}, of 0.5 *volts*, a source/drain implant impurity concentration, N_d, of $10^{21}\ atoms/cm^3$, and a substrate impurity concentration, N_{sub}, of $10^{15}\ atoms/cm^3$. The bulk-source voltage, V_{bs}, is zero, and the operating temperature of the transistor channel is $27\ °C$.

Mere inspection of the characteristic curves in Figure (8.5) supports the apparent prudence of operating MOSFETs in their saturation domains when linear analog signal processing is a key design consideration. In particular, we note that the curves in saturation appear to reflect linearity better than do the curves that are indigenous to the ohmic domain. Accordingly, the nonlinear function in (8-1) that is most relevant to analog MOSFET networks is, by (8-14),

$$I_{dn} = f_n(V_{gs}, V_{ds}, V_{bs}) \approx \beta_n \left(V_{gs} - V_{hn}\right)^2 \left(1 + \frac{V_{ds} - V_{dsat}}{V_\lambda}\right). \tag{8-20}$$

8.3.0. COMMON SOURCE AMPLIFIER MODELING

The operation of the common source amplifier in Figure (8.3a), and in particular, the relevant effects exerted by both the biasing supplies, V_{dd} and V_{gg}, and the signal source, $v_s(t)$, is best digested by an amalgam of graphical and analytical considerations. To this end, we begin by introducing the relatively simple concept of the amplifier load line. Ultimately, we shall rely, albeit qualitatively, on the engineering concepts underscored by this load line to develop a mathematical model that is useful for establishing reasonable estimates of I/O network performance.

8.3.1. AMPLIFIER LOAD LINE

In the amplifier of Figure (8.3a), KVL applied to the drain-source-ground loop provides

$$V_{dd} = R_l I_{dn} + V_{ds} + R_{ss} I_{dn} = \left(R_l + R_{ss}\right) I_{dn} + V_{ds}\ , \tag{8-21}$$

where we have made use of the fact that zero gate current at low frequencies implies identical drain lead and source lead currents. The foregoing relationship is equivalent to

$$I_{dn} = -\frac{V_{ds}}{R_l + R_{ss}} + \frac{V_{dd}}{R_l + R_{ss}}. \tag{8-22}$$

It is crucial to underscore the basic fact that this simple linear relationship between the drain current, I_{dn}, and the drain-source voltage, V_{ds}, must be satisfied for all values of V_{dd}, V_{gg}, and indeed the signal, $v_s(t)$, in Figure (8.3a). At low frequencies of course, drain current I_{dn} and drain-source voltage V_{ds} must also interrelate in accordance with the static drain characteristics curves for the transistor embedded in the subject amplifier. A sampling of such characteristics is displayed in Figure (8.6). Since these curves and (8-22) must be simultaneously satisfied under all operating conditions, it makes sense to superimpose the plot of (8-22) onto the static characteristics. Because (8-22) is the equation of a straight line, its plot, which is appended to Figure (8.6), is often termed the ***amplifier load line***. We note, in particular, that the current axis intercept of the load line is $V_{dd}/(R_l + R_{ss})$, while the voltage axis intercept is simply V_{dd}. It follows that the load line slope is $-1/(R_l + R_{ss})$. Thus, adjustments to the power line voltage, V_{dd}, merely displace the subject amplifier load line parallel unto itself, without altering the aforementioned (negative) slope.

In the quiescent operating state, which reflects $v_s(t) = 0$ and correlates with a drain current of I_{dnQ}, a drain-source voltage of V_{dsQ}, and a gate-source voltage of V_{gsQ}, KVL invoked on the gate circuit in Figure (8.3a) yields

$$V_{gg} = V_{gsQ} + R_{ss} I_{dnQ}\ . \tag{8-23}$$

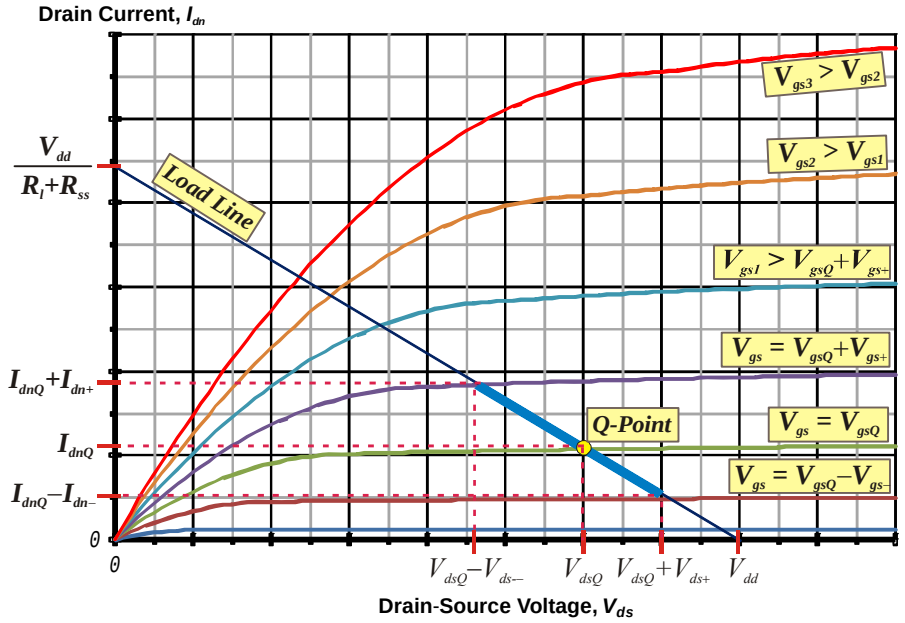

Figure (8.6). Representative, common source static characteristic curves for the NMOS transistor deployed in the amplifier of Figure (8.3a). Superimposed on these curves is the **load line** for the common source amplifier, which depicts the static operating point, or Q-point, and the signal excursion (heavy **BLUE** line segment) about the Q-point arising from the application of input signal.

Thus, the intersection of the load line defined by (8-22) with the $V_{gs} = V_{gsQ}$ static characteristic curve defines the ***Q-point*** (otherwise known as ***quiescent state*** or ***standby state***), as we depict in Figure (8.6). In this figure, the current axis projection of the Q-point is noted as I_{dnQ}, while the voltage axis projection of the same point is highlighted as V_{dsQ}. Several points are worthy of itemized advancement herewith.

(1). Since the static characteristic curves in the saturation domain of transistor operation are the graphical implications of (8-20), (8-23) satisfies, within the context of the modeling approximations propounded earlier, the relationship,

$$V_{gg} = V_{gsQ} + R_{ss} I_{dnQ}$$

$$\approx V_{gsQ} + R_{ss}\beta_n \left(V_{gsQ} - V_{hn}\right)^2 \left[1 + \frac{V_{dsQ} - \left(V_{gsQ} - V_{hn}\right)}{V_\lambda}\right], \tag{8-24}$$

where the quiescent value, V_{dsQ}, of drain-source voltage V_{ds} is, from (8-21),

$$V_{dsQ} = V_{dd} - \left(R_l + R_{ss}\right) I_{dnQ} \,. \tag{8-25}$$

The solution for the quiescent gate-source voltage, V_{gsQ}, and the corresponding quiescent drain-source voltage, V_{dsQ}, entails the combination of (8-24) with (8-25), followed by a messy mathematical iteration of a high order polynomial in V_{gsQ} and V_{dsQ}. Clearly, the load line method for discerning the quiescent state comprises a preferred computational strategy.

(2). As signal voltage $v_s(t)$ rises to its positive amplitude, V_s, we can surmise from Figure (8.3a) that the gate-source voltage, V_{gs}, similarly rises, reaching a maximum value of $(V_{gsQ} + V_{gs+})$, where V_{gs+} is the peak positive perturbation of gate-source voltage (over and above the quiescent value, V_{gsQ}, of gate-source voltage) corresponding to the applied signal amplitude,

V_s. Since all solutions for gate-source and drain-source voltages must lie on the load line superimposed on the characteristic curves of Figure (8.6), the Q-point effectively moves along the heavy **blue** segment of said line to the intersection of the line with the $V_{gs} = (V_{gsQ} + V_{gs+})$ static characteristic curve, as diagrammed in the figure. This intersection defines the corresponding maximum drain current, which we have indicated in the figure as $(I_{dnQ} + I_{dn+})$.

(3). As $v_s(t)$ falls to its negative amplitude, $-V_s$, V_{gs}, diminishes to its minimum value of $(V_{gsQ} - V_{gs-})$. Now, the Q-point clearly moves downward along the heavy **blue** segment of the load line to the intersection of the line with the $V_{gs} = (V_{gsQ} - V_{gs-})$ static characteristic curve, as shown in the figure. This intersection defines the corresponding minimum drain current, which we have indicated in the figure as $(I_{dnQ} - I_{dn-})$.

(4). We now see that an increasing signal voltage applied to the gate port of a common source amplifier incurs a rise in drain current and a corresponding decrease in drain-source voltage. Because the drain-source voltage is a component of the output voltage, $v_o(t)$, of the amplifier, we submit that ***phase inversion*** takes place between the input signal and the output voltage response. In other words, increasing $v_s(t)$ spawns decreasing $v_o(t)$ and vice versa. The net change, or ***swing*** induced by the applied signal in the drain-source voltage is $(V_{dsQ} + V_{ds+}) - (V_{dsQ} - V_{ds-}) = (V_{ds+} + V_{ds-})$. Analogously, the companion swing in drain current is $(I_{dn+} + I_{dn-})$, and the swing in gate-source voltage is $(V_{gs+} + V_{gs-})$. Even though the applied input sinusoidal signal is symmetrical about zero signal strength, the nonlinearities implicit to the static characteristics preclude our mindlessly presuming that $V_{ds+} = V_{ds-}$, $I_{dn+} = I_{dn-}$, and $V_{gs+} = V_{gs-}$. In the idealized circumstance of a linear response, these three equalities are indeed satisfied.

(5). The amplitude of the applied signal logically influences the proper placement of the operating point and even the slope of the load line. To wit, suppose the power supply voltage, V_{dd}, is reduced, perhaps to reduce circuit power dissipation and/or to facilitate circuit operation from a reduced battery voltage. Then, the load line indicated in Figure (8.6) is displaced parallel unto itself toward the origin. But if V_{dd} is reduced sharply, while maintaining the presumably desirable Q-point gate-source voltage, V_{gsQ}, the Q-point may fall on or near the transitional cusp between ohmic and saturated operated regimes. If the perturbed Q-point is on or near this cusp, positive signal swing almost assuredly displaces the Q-point into the ohmic regime, where the linearity of the curves is not as transparent as is the linearity of the curves in the saturated regime. If the quiescent operating point lies within the ohmic regime, most, if not all, of the signal swing occurs in the ohmic domain.

(6). If the magnitude of the slope of the load line is reduced through increased R_l and/or increased R_{ss}, an operational intrusion into the ohmic domain is likely. On the other hand, if the magnitude of the load line slope is increased through reduced R_l and/or reduced R_{ss}, operation in saturation is indeed likely to be sustained. While sustained operation in saturation is laudable, the drain-source signal swing is necessarily constrained more than is indicated in the figure.

(7). The output voltage response, $v_o(t)$, of the common source amplifier depicted in Figure (8.3a) is comprised of a static (or "quiescent" or "standby" or simply, "DC") component, say V_{oQ}, and a time varying component, $v_{os}(t)$, which is manifested by the applied input signal $v_s(t)$. In other words,

$$v_o(t) = V_{oQ} + v_{os}(t) . \tag{8-26}$$

Appealing to Figure (8.3a), the quiescent component, V_{oQ}, of the net output response is

$$V_{oQ} = V_{dd} - R_l I_{dnQ} = V_{dsQ} + R_{ss} I_{dnQ} . \quad (8\text{-}27)$$

It should be understood that V_{oQ} in (8-27) is the voltage observed at the amplifier output port when zero signal conditions prevail. In order to reduce power dissipation and thus extend the life of the battery used to energize the circuit, it is obviously advantageous to minimize the quiescent drain current, I_{dnQ}, drained from the battery. But from Figure (8.6), I_{dnQ} too small, which requires a correspondingly small V_{gsQ}, pushes the amplifier Q-point progressively closer to the horizontal, or zero current, axis. As a result, care must be assiduously exercised to ensure that the Q-point is not driven by the applied input signal into a device cutoff condition. It should be noted that the Q-point voltage defined by (8-27) can be used to bias the input port of a succeeding stage, which we do not depict in Figure (8.3). In other words, a separate biasing subcircuit for the next stage may not be necessary.

(8). Returning to (8-26), if amplifier linearity is closely approximated, the sinusoidal input at radial frequency ω that is proscribed by (8-3) ideally delivers a sinusoidal output at frequency ω; that is,

$$v_{os}(t) = V_{os} \cos(\omega t + \varphi) . \quad (8\text{-}28)$$

In this declarative, the output sinusoid, like the input sinusoid, is symmetrical about its zero signal value. The output signal amplitude is V_{os}, and at low signal frequencies, the previously highlighted I/O phase inversion implies that

$$\varphi = \theta - 180° , \quad (8\text{-}29)$$

where θ is recalled as the phase angle associated with the applied input signal defined by (8-3). The voltage gain, say A_{vcs}, of the common source configuration follows as

$$A_{vcs} = \frac{V_{os}}{V_s} . \quad (8\text{-}30)$$

It is appropriate to emphasize that the voltage gain or transfer function, in general, of an active network is not the ratio of the output response -to- the input signal. Rather, the transfer function is the ratio of the amplitude of the signal component of the output response -to- the amplitude of the signal component of the net input excitation. To the extent that the considered active network behaves linearly, we may also assert that its I/O transfer metric is identical to the ratio of the peak -to- peak amplitude of the signal component of the output response -to- the peak -to- peak amplitude of the net input voltage or current. Because the output signal voltage in a common source amplifier is intimately related to the transistor drain-source voltage signal swing, any action taken to render increased load line slope, which has been shown to limit the drain-source voltage swing, limits the voltage gain that can be generated by the amplifier.

8.3.2. SMALL SIGNAL MODEL

When the transistor deployed in the common source amplifier of Figure (8.3a) is properly biased to establish a reasonable approximation of network I/O linearity, the transistor currents and voltages are confined for all time to a restricted, and presumably reasonably linear, region of the static characteristic curves. Figure (8.6) clearly projects this point in that the confined region is bracketed by the indicated heavy **blue** segment of the amplifier load line. Observe that the vector length of this **blue** segment is progressively smaller, and hence the constrained region is more restricted, if the gate-source voltage swing, $(V_{gs+} - V_{gs-})$, becomes

progressively smaller. In turn, of course, smaller gate-source voltage swings are incurred by correspondingly smaller input signals, as might be monitored by the amplitude, V_s, of the test sinusoid in (8-3).

The logical conclusion to be drawn from the foregoing observations is the arguable imprudence of exploiting (8-20), which pertains to the entire saturation domain, as a low frequency transistor model. Instead, we might find it profitable (in a circuit design sense) and more expedient (in a computational sense) merely to concentrate our analytical attention to the saturation domain in the immediate neighborhood of the quiescent operating point. With pun intended, the modeling representation problem before us is tailor made for a judicious application of Taylor's series[6]. In particular, let us represent the saturation domain defined by (8-20) only in the immediate neighborhood of the Q-point through a suitably truncated Taylor series. Prior to such truncation, we have the infinite power series,

$$\begin{aligned} I_{dn} \;=\; I_{dnQ} &+ \left.\frac{\partial I_{dn}}{\partial V_{gs}}\right|_{Q}\left(V_{gs}-V_{gsQ}\right)+\left.\frac{\partial I_{dn}}{\partial V_{bs}}\right|_{Q}\left(V_{bs}-V_{bsQ}\right) \\ &+\left.\frac{\partial I_{dn}}{\partial V_{ds}}\right|_{Q}\left(V_{ds}-V_{dsQ}\right)+\left(H.\,O.\,T.\right)+\cdots, \end{aligned} \tag{8-31}$$

where each of the indicated partial derivatives is numerically evaluated at the quiescent operating point, $(V_{gs},\, V_{bs},\, V_{ds}) = (V_{gsQ},\, V_{bsQ},\, V_{dsQ})$, of the utilized transistor. The parenthesized term, *(H. O. T.)*, on the right hand side of (8-31) symbolizes all ***H**igher **O**rder **T**erms* in the power series expansion. These latter terms are proportional to the second and higher order derivatives of the saturation regime drain current with respect to the three transistor voltages. They are also proportional to second and higher order terms involving the voltage differences, $(V_{gs}-V_{gsQ})$, $(V_{bs}-V_{bsQ})$, and $(V_{ds}-V_{dsQ})$.

When signal is applied to the common source amplifier, we know from our examination of Figure (8.6) that the drain current, I_{dn}, is perturbed, positively or negatively, from its Q-point value, I_{dnQ}. This awareness implies that $(I_{dn}-I_{dnQ})$ denotes the ***signal-induced change***, say I_{ds}, in the transistor drain current; that is,

$$I_{ds} \triangleq I_{dn} - I_{dnQ}\,. \tag{8-32}$$

In effect, I_{ds} is the signal component of the net drain current, $I_{dn} = (I_{dnQ} + I_{ds})$, manifested by the combined effects of biasing and applied signal. It is crucial to understand that I_{ds} is null if no signal activates the input port of the subject amplifier. And depending on the phase angle and polarity of any applied (nonzero) signal, I_{ds} can be a positive or a negative drain current component. Similarly,

$$\left.\begin{aligned} V_{1s} &\triangleq V_{gs}-V_{gsQ} \\ V_{2s} &\triangleq V_{bs}-V_{bsQ} \\ V_{3s} &\triangleq V_{ds}-V_{dsQ} \end{aligned}\right\rangle \tag{8-33}$$

define the signal-induced components of the gate-source voltage, bulk-source voltage, and drain-source voltage, respectively. As in the case of I_{ds}, each of these signal induced voltage components are zero when no input signal is applied. Moreover, each voltage constituent can be positive or negative, depending on the nature of, and ultimate responses to, the input signal.

It should be conceptually clear that fundamentally, the drain current of a transistor can be perturbed from its Q-point, or standby, value, I_{dnQ}, for three different, voltage-related reasons. First, a gate-source voltage change, delineated as V_{1s} in (8-33), alters the transistor drain current because of the square law relationship that prevails between saturation domain drain current and gate-source voltage. Second, a bulk-source, signal-induced voltage perturbation, V_{2s}, shifts the threshold voltage, V_{hn}, of a transistor. Since the drain current is functionally dependent on the excess gate-source voltage, $(V_{gs}-V_{hn})$, a drain current ramification of the bulk-source voltage change can be expected as a result of threshold voltage sensitivity to bulk-source voltage. However, we do expect this drain current change to be minimal in transistors that extol thin gate oxide layers for, as we witnessed in (8-7) and (8-8), thin gate oxides breed essentially constant threshold potentials that are independent of bulk-source biasing and signals. Finally, because of channel length modulation, which gives rise to the non-zero slope in the saturation regime I-V characteristic curves, we expect V_{3s}, the signal induced change in drain-source voltage, to alter the observable transistor drain current. As in the circumstance of thin gate oxides incurring a minimal impact of bulk-source voltage on drain current, long channel lengths, to which the channel length modulation voltage, V_{λ}, is directly proportional, minimize the effect that drain-source voltage perturbations have on drain currents observed in the saturation regime. It bears repeating that all of these voltage changes are zero when no input signal is applied, whence no drain current change is incurred in (8-32). Armed with (8-32) and (8-33), we may write (8-31) as

$$I_{ds} = \left.\frac{\partial I_{dn}}{\partial V_{gs}}\right|_{Q}\left(V_{1s}\right)+\left.\frac{\partial I_{dn}}{\partial V_{bs}}\right|_{Q}\left(V_{2s}\right)+\left.\frac{\partial I_{dn}}{\partial V_{ds}}\right|_{Q}\left(V_{3s}\right)+\left(H.O.T.\right)+\cdots . \tag{8-34}$$

In (8-34), each of the partial derivative coefficients of the indicated signal-induced voltage components must have units of conductance since all terms on the right hand side must sum to a drain current value. We shall call the first derivative term on the right hand side of (8-34) the ***forward transconductance***, g_m, of the transistor. Specifically, if the channel length modulation voltage, V_{λ}, is big in comparison to $(V_{ds}-V_{dsat})$ in (8-20), (8-13) and (8-20) combine to deliver

$$g_m \triangleq \left.\frac{\partial I_{dn}}{\partial V_{gs}}\right|_{Q} \approx 2\sqrt{\beta_n I_{dnQ}} = \sqrt{2\mu_n C_{ox}\left(\frac{W}{L}\right)I_{dnQ}} . \tag{8-35}$$

Since parameter g_m quantifies a differential change in the drain current about its Q-point value, which is an output signal quantity, in reference to a corresponding differential change in gate source voltage, which is an input port signal variable, g_m is a measure of the forward gain that can be provided by a transistor. In Figure (8.6), it quantifies the separation between proximate V_{gs}-curves for $V_{ds} = V_{dsQ}$ and $V_{bs} = V_{bsQ}$. In a word, the larger g_m is, the larger is the magnitude of gain that can be realized in the common source amplifier. To this end, it is notable that

$$g_m \ \alpha \ \sqrt{\left(\frac{W}{L}\right)I_{dnQ}} , \tag{8-36}$$

This result asserts that g_m can be increased by increasing the gate aspect ratio and/or increasing the quiescent drain current of the transistor used in the common source amplifier. In both of these scenarios, engineering prudence must be exercised. For example, increasing the gate aspect ratio by the normal route of increasing the gate width, W, increases transistor capacitances, which may spell a degradation in circuit response speed and *3-dB* bandwidth. On the other hand, an increasing quiescent drain current begets increased circuit power dissipation, whence reduced

battery life in portable applications. The best approach for augmenting the forward transconductance generally entails a combination of expanded gate width and increased transistor current. Note, however, that even a four-fold increase in the product of gate aspect ratio and quiescent drain current gives rise to only a factor of two enhancement in the forward transconductance.

Returning to (8-31), the ***bulk transconductance***, which is also referred to as the ***body effect transconductance***, g_{mb}, is

$$g_{mb} = \left.\frac{\partial I_{dn}}{\partial V_{bs}}\right|_{Q} = \lambda_b g_m \,, \tag{8-37}$$

where it can be shown that

$$\lambda_b = \sqrt{\frac{V_\theta/2}{2\left(V_F - V_T\right) - V_{bsQ}}} \,. \tag{8-38}$$

In (8-38), V_θ is the bulk effect voltage in (8-8), V_F is the Fermi potential introduced in (9-9), and, of course, V_T is the Boltzmann voltage, per (8-10). We note that for a small gate oxide thickness, T_{ox}, V_θ is proportionately small, which renders the bulk transconductance, g_{mb}, correspondingly small. In turn, and as we hypothesized earlier, thin gate oxides render minute the impact of bulk-source signal voltage on the transistor drain current.

In (8-31), we conclude our model parameterization by introducing the ***channel resistance***, r_o, of a MOSFET in accordance with the stipulation,

$$\frac{1}{r_o} \triangleq \left.\frac{\partial I_{dn}}{\partial V_{ds}}\right|_{Q} = \frac{I_{dnQ}}{V_\lambda + V_{dsQ} - V_{dsatQ}} \,; \tag{8-39}$$

equivalently,

$$r_o = \frac{1}{\left.\frac{\partial I_{dn}}{\partial V_{ds}}\right|_{Q}} = \frac{V_\lambda + V_{dsQ} - V_{dsatQ}}{I_{dnQ}} \,. \tag{8-40}$$

The channel conductance, $1/r_o$, is the slope, evaluated at the transistor Q-point, of the saturation regime static characteristic curves. Virtually zero slope in this characteristic curve implies that the drain current is essentially independent of drain-source voltage, V_{ds}, which makes the signal component of the drain current invulnerable to signal-induced drain-source voltage changes. For nearly zero characteristic curve slope, (8-39) shows that r_o is necessarily very large. In (8-40), we see that large r_o requires either a small quiescent drain current, I_{dnQ}, and/or a large channel length modulation voltage, V_λ. In turn, (8-16) affirms that a geometrically long channel length, L, nurtures large V_λ.

Equation (8-34) can now be written as

$$I_{ds} = g_m V_{1s} + g_{mb} V_{2s} + \frac{V_{3s}}{r_o} + \left(H.O.T.\right) + \cdots \,. \tag{8-41}$$

with the understanding that parameters g_m, g_{mb}, and r_o, which derive from (8-35), (8-37)/(8-38), and (8-40), respectively, rely on *a priori* awareness of the transistor operating point. Several reasons encourage investigating the plausibility neglecting the high order terms in (8-41). The first, and most obvious, reason is a profoundly simplified analytical relationship. In particular, if

(H.O.T.) ≈ 0, (8-41) collapses from an infinite power series representation of the Q-point neighborhood of the static characteristic curves to the truncated, albeit approximate, expression,

$$I_{ds} \approx g_m V_{1s} + g_{mb} V_{2s} + \frac{V_{3s}}{r_o}. \tag{8-42}$$

The second reason derives directly from the fruit of the first reason. Specifically, (8-42) is a linear volt-ampere relationship and as such, it necessarily mirrors the equilibrium node current relationship a linear circuit. This linear circuit, which we reference as the ***small signal, low frequency model*** of a MOSFET is displayed for an NMOS transistor in Figure (8.7).

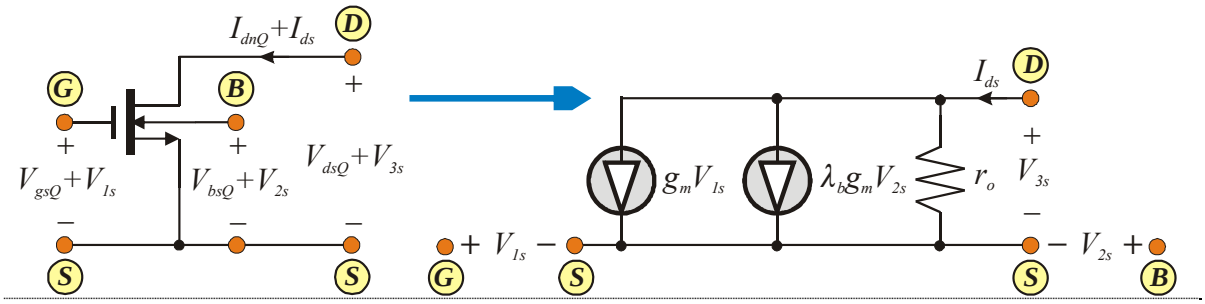

Figure (8.7). Small signal, low frequency model of an NMOS transistor whose saturation regime Q-point is defined by $(I_{dn}, V_{gs}, V_{bs}, V_{ds}) = (I_{dnQ}, V_{gsQ}, V_{bsQ}, V_{dsQ})$.

Before addressing the salient features of the model in Figure (8.7), it is necessary to point out that either of two operating circumstances encourage the tacit neglect of the high order terms *(H.O.T.)* in (8-41). First and arguably foremost, if the signal components, V_{1s}, V_{2s}, and V_{3s}, of the net gate-source voltage, net bulk-source voltage, and net drain-source voltage, respectively, are small, the high order terms collapse into negligible obscurity. This collapse stems from the previously noted fact that each of the high order terms in the Taylor series expansion of the MOSFET drain current is proportional to second order and higher order powers of the voltage differences, V_{1s}, V_{2s}, and V_{3s}. Thus, for example, if V_{1s} is a very small signal component of the gate-source voltage, its square, its cube, its fourth order value, etc., are each progressively smaller, thereby rendering these higher order powers of voltage ultimately negligible. The self-fulfilling prophecy herewith is that the required smallness of signal voltages V_{1s}, V_{2s}, and V_{3s}, is precisely the reason that the model offered in Figure (8.7) is called a "small signal" model, or a "small signal" equivalent circuit. Indeed, the safest answer to the somewhat rhetorical question, "How small must a signal be to render the small signal model valid?", is "Small enough to justify the neglect of all high order terms in the Taylor series expansion of the MOSFET drain current."

While the preferred operating domain for nominally linear analog signal processing is the saturation domain, it is possible to operate transistors in their ohmic regimes while still claiming nominal I/O linearity. In particular, if the signal swings remain strictly confined to a small interval about a specified operating point, the resultant very small perturbations about the Q-point promote a linearly truncated Taylor series approximation. While the ohmic regime topology and the parametric definitions for the MOSFET model remain identical to the saturation architecture displayed in Figure (8.7), the actual expressions and resultant numerical values of the model parameters differ from those presented earlier. In particular, the ohmic regime forward transconductance is, depending on quiescent operating point and gate aspect geometry, generally smaller than its saturation regime counterpart is. Accordingly, smaller I/O gains can be forecast if a transistor is biased in its ohmic regime. Moreover, the channel resistance in the ohmic regime is generally smaller than is the channel resistance in saturation.

However, it is true that a "small signal" can be cautiously permitted to assume progressively larger amplitudes if the immediate region about the quiescent operating point closely projects linear volt-ampere characteristics. For example, if the region in question exudes perfect volt-ampere linearity (which it never will), the first order voltage derivatives on the right hand sides of (8-31) and (8-34) are constant, independent of the voltage with respect to which the derivative is undertaken. This means that the second and higher order derivatives are clamped to zero. Since the high order terms are individually proportional to the second and higher order derivatives of the saturation regime drain current with respect to the three transistor voltages, the high order term component of the Taylor expansion is rendered null. Of course, in actual practice, perfect linearity cannot prevail and as a result, the high order terms never collapse perfectly to zero. But excellent linearity in the neighborhood of the Q-point does make the second and higher order voltage derivatives monotonically smaller with derivative order, thereby ensuring small *(H.O.T.)*. Obviously, a combination of reasonably approximate linearity and small signals serve to encourage further the negligible viewpoint of *(H.O.T.)*.

In the interest of completeness, the PMOS small signal model is given in Figure (8.8). It is vitally important to recognize herewith that the PMOS and NMOS small signal models are topologically identical. In particular, note in the PMOS topology that the $g_m V_{1s}$ controlled current remains directed from drain -to- source, and the controlling voltage, V_{1s}, is still viewed as the gate -to- source signal voltage. Similarly, the $g_{mb} V_{2s}$ current flows from drain -to- source, with V_{2s} understood to be the bulk -to- source voltage. And in the interest of clarity, it should be understood that neither the NMOS nor the PMOS small signal models give any information about the quiescent state of the transistor represented by the model. Not only do the models not elucidate the engineering stature of the transistor Q-point, the numerical values of the parameters embedded in the models rely on an *a priori* awareness of the transistor quiescent currents and voltages.

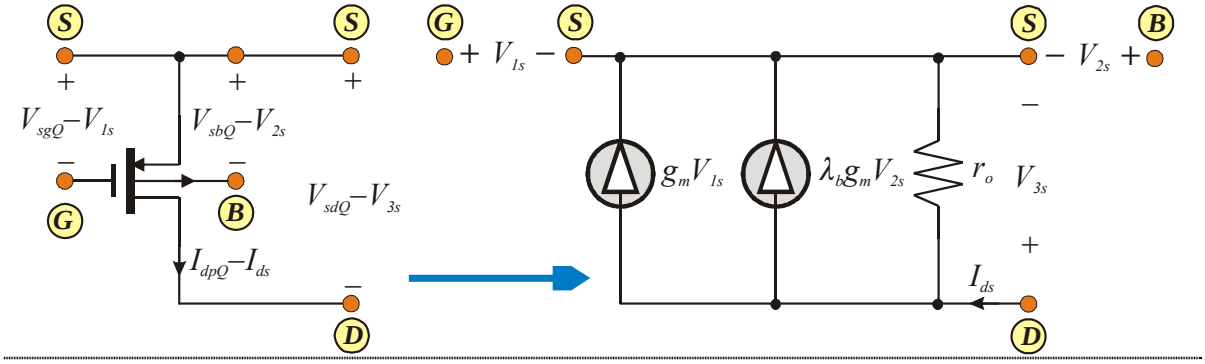

Figure (8.8). Small signal, low frequency model of a PMOS transistor whose saturation regime Q-point is defined by $(I_{dn}, V_{sg}, V_{sb}, V_{sd}) = (I_{dnQ}, V_{sgQ}, V_{sbQ}, V_{sdQ})$.

8.4.0. REFERENCES

[1]. A. B. Grebene, *Bipolar and MOS Analog Integrated Circuit Design.* New York: Wiley-Interscience, 1984, pp. 314-315.

[2]. P. R. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits.* New York: John Wiley & Sons, Inc., 2001, chap. 8.

[3]. J. Choma and W-K. Chen, *Feedback Networks: Theory and Circuit Applications*. Singapore: World Scientific Press, 2007, chap. 4.

[4]. T. H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits.* United Kingdom: Cambridge University Press, 2004, chap. 5.

[5]. C. C. Enz and E. A. Vittoz, *Charge-Based MOS Transistor Modeling.* New York: John Wiley & Sons, Ltd., 2006, pp. 189-198.
[6]. G. B. Thomas, Jr., *Calculus and Analytic Geometry.* Reading, Massachusetts: Addison-Wesley Publishing Company, 1962, pp. 787-814.