Linearity Issues In Differential Pairs
And Gilbert Multipliers

Onder Oz
Doctoral Student, USC Electrical Engineering
&
Dr. John Choma
Professor of Electrical & Systems Architecture Engineering
Fellow, Scintera Networks (San Jose, CA)
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]
626–915–0944 [Home Fax]
jchoma@usc.edu
jchoma@scintera.com

ABSTRACT:
This report undertakes a low frequency analysis of a Gilbert multiplier realized in MOS technology. A prelude to the multiplier study is a linearity analysis of a simple differential amplifier. The cases of saturation and triode region operation of the transistors in the balanced pair are addressed. In the multiplier study, two cases are likewise examined. The first is the case in which all transistors in the multiplier operate in saturation, while the second case retains saturation domain operation in the driver transistors but allows the transistors utilized in the quad array to operate in their triode regimes.

December 2006
1.0. INTRODUCTION

The balanced differential pair is a cornerstone circuit for high performance analog integrated circuits. It finds widespread use as the core amplification cell of linear voltage and transconductance amplifiers, active filters, oscillators, comparators, multipliers, and numerous other analog signal processors\cite{1}-\cite{4}. Although biasing and both the small signal steady state and transient responses of differential pairs are treated ubiquitously in the literature, comparatively little attention is devoted to dynamic signal range limitations incurred by the inherently nonlinear characteristics of the active devices embedded in these pairs\cite{5}-\cite{6}. Because dynamic range is a crucial performance index in modern communication, data processing, and information transmission systems, a formulation of mathematically sound engineering design criteria aimed toward maximizing input/output (I/O) linearity in differential stages is a prudent undertaking.

To the foregoing end, a simple balanced differential amplifier realized in MOS device technology is carefully examined for the cases in which the utilized transistors are biased in saturation and in triode, or ohmic, operating regimes. These results are carefully scrutinized and assessed prior to applying them to the problem of deducing the low frequency I/O describing function of a Gilbert multiplier, which uses three balanced differential circuits\cite{7}.

Itemized below are several simplifying assumptions invoked throughout this report to deduce mathematically tractable and enlightening results.

1. The nonlinearities precipitated by applied signal overstress are weak in the sense that the quiescent voltages and currents within considered circuits remain independent of applied signal amplitudes and their resultant voltage and current responses.

2. Only low signal frequencies are addressed so that the capacitances implicit to the transistors used in the circuit and all energy storage parasitics associated with the layout of considered circuit cells can be ignored. It should be noted that the bulk-drain, gate-source, gate-drain, and most other capacitances that prevail in MOSFETs are voltage dependent. Accordingly, the nonlinearity observations disclosed herein are likely to be degraded by progressively increased signal frequencies.

3. Balanced operation of all differential cells is presumed, which is to say that the passive and active elements in the respective cells of these half circuits are perfectly matched. For well-designed circuits whose devices are biased at reasonable standby voltages and currents, generally excellent matching is afforded through judiciously executed circuit layout.

4. Threshold voltage modulation incurred by perturbations in bulk-source potential is presumed to be zero. This presumption is straightforwardly satisfied if the process used to fabricate transistors allows the bulk terminal to be short circuited to the source terminal.

5. The MOSFETs embedded in the differential cells identified for analytical consideration are represented by simple and familiar volt-ampere characteristics. For a MOSFET operated in saturation, where its drain-source voltage, $V_{ds}$, is at least as large as its drain saturation voltage, $V_{dsat}$, the drain current, $I_d$, is given by

$$I_d = \frac{\mu C_{ox}}{2T_{ox}} \left( \frac{W}{L} \right) (V_{gs} - V_h)^2,$$

where $V_{gs}$ is the voltage applied between the gate and source terminals, and $V_h$ represents the threshold voltage of the transistor. By virtue of the preceding declaration, $V_h$ is constant, independent of signal fluctuations observed at the source terminal. In (1), $\mu$ is the
low-field mobility of mobile charge in the inverted channel between the source and drain implants, \( e_{ox} \) is the dielectric constant of the gate oxide, \( T_{ox} \) is the thickness of the gate oxide, and \( W/L \) is the gate aspect ratio. Given that the drain saturation voltage, \( V_{dsat} \), is

\[
V_{dsat} = V_{gs} - V_h, \tag{2}
\]

(1) reflects the requirement that the drain-source voltage, \( V_{ds} \), satisfies

\[
V_{ds} \geq V_{dsat} = V_{gs} - V_h. \tag{3}
\]

In the triode regime, \( V_{ds} < V_{dsat} \), and the drain current, \( I_d \), is given by

\[
I_d = \frac{\mu e_{ox}}{T_{ox}} \left( \frac{W}{L} \right) V_{ds} \left( V_{gs} - V_h - \frac{V_{ds}}{2} \right). \tag{4}
\]

It is convenient to introduce the substitution,

\[
\beta = \frac{\mu e_{ox}}{2T_{ox}} \left( \frac{W}{L} \right), \tag{5}
\]

so that (1) and (4) are respectively expressible as

\[
I_d = \beta \left( V_{gs} - V_h \right)^2 \tag{6}
\]

for \( V_{ds} \geq V_{dsat} \) and

\[
I_d = 2 \beta V_{ds} \left( V_{gs} - V_h - \frac{V_{ds}}{2} \right) \tag{7}
\]

for \( V_{ds} < V_{dsat} \). Note that the dimensional units of parameter \( \beta \) are mhos/volt. Observe further that (6) and (7) yield identical drain currents for \( V_{ds} = V_{dsat} = V_{gs} - V_h \).

### 2.0. MODEL APPROXIMATIONS

It is important to understand the implications of two fundamental approximations encumbered by (6) and (7). In deep submicron technologies, arguably the most significant of these approximations is the tacit neglect of mobility degradation arising from lateral electric fields produced in the channel by drain-source voltages. It can be shown that a first order account of this carrier mobility degradation modifies the triode region characteristic of (7) to the volt-ampere form

\[
I_d = 2 \beta \left[ \frac{V_{ds} \left( V_{gs} - V_h - \frac{V_{ds}}{2} \right)}{1 + \frac{V_{ds}}{V_{le}}} \right]. \tag{8}
\]

In (8), \( V_{le} \), which is termed a field modulation voltage, is given by

\[
V_{le} = \left( \frac{v_{max}}{\mu} \right) L, \tag{9}
\]
where $L$ is the channel length and $v_{\text{max}}$ is the saturated limited velocity of channel charge carriers (of the order of $0.15 \, \mu\text{m/pSEC}$ for electrons). Typically, the value of voltage $V_{\text{le}}$ is in the range of $1 \, V/\mu\text{m}$ to $3 \, V/\mu\text{m}$.

A complication spawned by (8) is that it no longer delivers the simple relationship witnessed in (2) for the drain saturation voltage. By definition, the drain saturation voltage, $V_{ds}^{\text{sat}}$, is the value of the drain-source voltage, $V_{ds}$, for which the slope of the ohmic regime $I_d$ versus $V_{ds}$ characteristic is zero at any fixed value of gate-source voltage $V_{gs}$. An application of this definition to (8) leads to the revised drain saturation voltage,

$$V_{ds}^{\text{sat}} = M_{\text{sat}} \left( V_{gs} - V_{h} \right),$$

where, with

$$\alpha = \frac{V_{gs} - V_{h}}{V_{\text{le}}},$$

denoting the effective gate-source voltage normalized to the field modulation voltage,

$$M_{\text{sat}} = \frac{\sqrt{1 + 2\alpha} - 1}{\alpha}.$$  

![Voltage & Current Reduction Factor](image)

Fig. (1). Functional dependencies of the saturation voltage degradation factor, $M_{\text{sat}}$, and the degradation factor, $M_{\text{sat}}^2$, of the saturated drain current on the normalized excess gate voltage, $\alpha$.  


It is easily demonstrated that $M_{\text{sat}} \leq 1$ for $\alpha \geq 0$ and thus, an impact of the carrier mobility degradation incurred by lateral fields in the inverted channel is a decrease in the drain saturation voltage evidenced at low field values. The amount of this decrease is depicted graphically in Figure (1), which depicts a plot of parameter $M_{\text{sat}}$ versus the normalized effective gate voltage, $\alpha$. While mobility degradation is generally an undesirable phenomenon from the perspective of attainable switching speed and unity gain frequency, the drain saturation voltage decrease is actually good news in low voltage analog circuit applications that require MOSFETs to function in their saturated regimes.

The drain current in saturation can now be determined by substituting (10) into (8) to obtain the aesthetically pleasing result,

$$I_d = \beta M_{\text{sat}}^2 (V_{gs} - V_h)^2. \quad (13)$$

In the limit of large channel lengths, $\alpha$ in (11) tends toward zero, whence $M_{\text{sat}}$ in (12) approaches one. With $M_{\text{sat}} \approx 1$ and $V_{le}$ relatively large, the drain saturation voltage in (10) becomes the simple voltage expression given by (2). Moreover, both the triode regime and saturation regime drain currents collapse to the simple relationships of (4) and (6), respectively. But in deep submicron technologies, $V_{le}$ is small, which means that parameter $\alpha$ can become large for even modest values of the effective gate-source voltage, $(V_{gs} - V_h)$. For large $\alpha$, $M_{\text{sat}}$ in (12) becomes

$$M_{\text{sat}} \approx 1 \quad \text{for} \quad \alpha \geq 0.$$

It follows that in the limit of very short channel lengths and at least modest values of $(V_{gs} - V_h)$, the saturated drain current in (13) becomes

$$I_d \approx \varepsilon_{ox} \nu_{\text{max}} \left( \frac{W}{T_{ox}} \right) (V_{gs} - V_h). \quad (15)$$

where (5), (11), and (13) have been applied. Figure (1) shows the extent by which lateral electric fields modify the drain current in (13) by displaying a plot of the square of $M_{\text{sat}}$ versus the normalized effective gate voltage, $\alpha$.

Note that for the sufficiently large values of $(V_{gs} - V_h)$ that render the approximation in (14) valid, the saturated drain current is linearly proportional to $(V_{gs} - V_h)$, as opposed to a proportionality to the square of this excess gate voltage. Under this state of affairs, the volt-ampere characteristics of the transistor are limited by velocity saturation, principally because the drain current in (15) is directly proportional to the saturated limited velocity of channel charge carriers and is independent of the channel length, $L$. An interesting feature of device operation in velocity saturation is that the forward transconductance, $g_m$, becomes a constant, independent of drain current and device voltages; that is

$$g_m \approx \frac{\partial I_d}{\partial V_{gs}} \Bigg|_{\text{small } L, \ V_{ds} \geq V_{dsat}} = \varepsilon_{ox} \nu_{\text{max}} \left( \frac{W}{T_{ox}} \right). \quad (16)$$

In contrast, the forward transconductance when $M_{\text{sat}}$ tends toward unity is, by (6),
which highlights a transconductance that increases as the square root of drain current.

A second approximation implicit to the simple volt-ampere functions in (6) and (7) is negligible channel length modulation. The problem at hand is that the saturation regime drain current given by (6) or (13) is premised on the presumption of a channel inversion layer length that is identical to the channel length, \( L \), which geometrically separates the source and drain regions. Since \( V_{ds} = V_{dsat} \) incurs inversion layer pinch off at the drain site, and hence an inversion layer length equal to the channel length, \( L \), \( V_{ds} > V_{dsat} \) necessarily incurs pinch off within the source-drain spacing. Accordingly, the effective channel length over which mobile charge carriers are poised for drain current conduction is reduced from \( L \) by an amount that is functionally dependent on the excess drain voltage, \((V_{ds} - V_{dsat})\). A first order account of the volt-ampere implications of channel length reduction entails replacing (6) by

\[
I_d = \beta \left( V_{gs} - V_h \right)^2 \left( 1 + \frac{V_{ds} - V_{dsat}}{V_{\lambda}} \right) .
\]

Alternatively, if an account of mobility degradation is made, (13) is supplanted by

\[
I_d = \beta M_{sat}^2 \left( V_{gs} - V_h \right)^2 \left( 1 + \frac{V_{ds} - V_{dsat}}{V_{\lambda}} \right) .
\]

with the understanding that \( V_{dsat} \) in this relationship is given by (10). In (18) and (19), the channel length modulation voltage, \( V_{\lambda} \), is stipulated by the semi-empirical equation,

\[
V_{\lambda} = \left( \frac{L}{D_b} \right) \left( \frac{V_j}{V_F} \right)^2 \sqrt{32 V_T \left( V_{ds} - V_{dsat} + V_j \right)} ,
\]

where \( V_j \) is the built in potential of the drain-substrate junction, \( V_T \) is the Boltzmann voltage, \( V_F \) is the Fermi potential of the channel, and \( D_b \) represents the carrier Debye length, which is dependent on the background impurity concentration of the substrate in regions proximate to the semiconductor surface. Note that for long channel lengths, \( V_{\lambda} \) is proportionately large and drain current corrections for channel length modulation phenomena are resultantly unnecessary. On the other hand, channel length modulation effects can be mitigated for short channel lengths if the transistor in question operates at drain-source voltages that are only slightly larger than the drain saturation voltage of the device undergoing scrutiny.

2.0. DIFFERENTIAL AMPLIFIER

Figure (2) depicts the basic schematic diagram of a differential pair utilizing two identical MOSFETs, \( M1 \) and \( M2 \). The drains of each device are terminated to the positive rail voltage, \( V_{dd} \), in identical resistances, \( R_l \). Common mode biasing, say \( V_{ci} \), is applied to each transistor gate, while the input signal is applied differentially as the indicated voltage, \( V_{di} \). To this end, it is convenient to write the gate to ground voltages, \( V_{i1} \) and \( V_{i2} \), as
Because of the voltages applied to both transistor gate terminals, drain currents, $I_{d1}$ and $I_{d2}$, which individually are comprised of quiescent and signal components, flow in the transistor drain circuits. The sum of these currents flow to ground through a constant tail current, $I_k$, which is generally realized as a high impedance, NMOS current sink. This constant current, $I_k$, is nominally independent of the voltage, $V_k$, established across the sink, as long as the transistors implicit to the active current sink operate in their saturation domains and have large drain-source channel resistances and transistors $M1$ and $M2$ are matched. The flow of currents $I_{d1}$ and $I_{d2}$ give rise to single ended output voltages, $V_{o1}$ and $V_{o2}$, as well as to a differential voltage response, $V_{do}$.

As in the case of the applied input voltages, it is expedient to represent these output voltages as a common mode component, $V_{co}$, and the aforementioned differential component, $V_{do}$, such that

$$V_{o1} = V_{co} + \frac{V_{do}}{2}$$
$$V_{o2} = V_{co} - \frac{V_{do}}{2}$$

(22)

If current $I_k$ derives from a current sink boasting high impedance, $V_{co}$ contains only a quiescent voltage component. In contrast, and assuming ideally balanced operation, $V_{do}$ embraces only a signal component that is presumably proportional to the input differential signal, $V_{di}$.

2.1. TRANSISTORS IN SATURATION REGION

If $M1$ and $M2$ operate in their saturation regimes, (6) produces

$$I_{d1} = \beta\left(V_{gs1} - V_h\right)^2$$

(23)
and

\[ I_{d2} = \beta \left( V_{gs2} - V_h \right)^2, \]  

(24)

where the respective gate-source voltages, \( V_{gs1} \) and \( V_{gs2} \), satisfy the Kirchhoff constraint,

\[ V_{di} = V_{gs1} - V_{gs2}. \]  

(25)

The differential output current, \( I_{do} \), is therefore given by

\[ I_{do} = I_{d1} - I_{d2} = \beta \left[ \left( V_{gs1} - V_h \right)^2 - \left( V_{gs2} - V_h \right)^2 \right], \]  

(26)

which is expressible as

\[ I_{do} = \beta \left[ (V_{gs1} + V_{gs2} - 2V_h)(V_{gs1} - V_{gs2}) \right] = \beta \left( V_{gs1} + V_{gs2} - 2V_h \right) V_{di}. \]  

(27)

From (23) and (24), the parenthesized term on the far right of this expression is

\[ V_{gs1} + V_{gs2} - 2V_h = \frac{1}{\sqrt{\beta}} \left( \sqrt{I_{d1}} + \sqrt{I_{d2}} \right). \]  

(28)

Since the tail current, \( I_k \), is identically the sum of the two drain currents, \( I_{d1} \) and \( I_{d2} \),

\[ \left( V_{gs1} + V_{gs2} - 2V_h \right)^2 = \frac{I_k + 2\sqrt{I_{d1}I_{d2}}}{\beta}. \]  

(29)

Digressing for a moment,

\[ V_{di}^2 = \left[ (V_{gs1} - V_h) - (V_{gs2} - V_h) \right]^2 = \frac{I_k - 2\sqrt{I_{d1}I_{d2}}}{\beta}. \]  

(30)

Combining (30) into (29),

\[ V_{gs1} + V_{gs2} - 2V_h = \sqrt{\frac{2I_k - \beta V_{di}^2}{\beta}}. \]  

(31)

The insertion of (31) into (27) leads to the desired result,

\[ I_{do} = V_{di} \sqrt{\frac{2}{\beta I_k}} \sqrt{1 - \frac{\beta V_{di}^2}{2I_k}}. \]  

(32)

The corresponding differential output voltage, \( V_{do} \), is given in Figure (2) by

\[ V_{do} = R_l (I_{d2} - I_{d1}) = -R_l I_{do} = -V_{di} \sqrt{\frac{2}{\beta I_k}} R_l \sqrt{1 - \frac{\beta V_{di}^2}{2I_k}}. \]  

(33)

The algebraic endurance required to progress to the last result pays dividends because of the engineering insights the expression fosters. To begin, appeal to the small signal transconductance relationship postured by (17). Under zero signal, or quiescent, operating
conditions, each transistor in the circuit of Figure (1) conducts one-half of the tail current, \( I_k \); that is, \( I_{d1Q} = I_{d2Q} = I_k/2 \). It follows that the small signal device transconductance at the quiescent operating point of the circuit is

\[
g_{mQ} = 2\sqrt{\beta \left( I_k/2 \right)} = \sqrt{2\beta I_k} . \tag{34}\]

Accordingly, the I/O voltage describing function, \( A_v(V_{di}) \), implied by (33) is

\[
A_v(V_{di}) \triangleq \frac{V_{do}}{V_{di}} = -g_{mQ}R_l \sqrt{\frac{2V_{di}^2 - \beta V_{di}^2}{2I_k}} , \tag{35}\]

which clearly quantifies the radical factor on the right hand side as a departure from ideal I/O voltage gain linearity. As expected, the effective gain, \( A_v(V_{di}) \), converges to its small signal value, \((-g_{mQ}R_l)\), for very small differential input signals, \( V_{di} \).

A second noteworthy point derives from the fact that the drain saturation voltage, \( V_{dsatQ} \), at a quiescent current of \( I_k/2 \) is, from (2) and either (23) or (24),

\[
V_{dsatQ} = \sqrt{\frac{I_k/2}{\beta}} . \tag{36}\]

Equation (35) can therefore be recast in the form

\[
A_v(V_{di}) \triangleq \frac{V_{do}}{V_{di}} = -g_{mQ}R_l \sqrt{1 - \left( \frac{V_{di}}{2V_{dsatQ}} \right)^2} , \tag{37}\]

which suggests heuristically that “small signal” conditions apply if the square of one-half of the differential input signal is significantly smaller than the square of the quiescent operating point value of the drain saturation voltage for the transistors utilized in a balanced differential pair.

In an attempt to quantify the degree of nonlinearity implicit to the describing function of (37), it is beneficial to write the square root function on the right hand side of this relationship as

\[
\sqrt{1 - \left( \frac{V_{di}}{2V_{dsatQ}} \right)^2} = 1 - \varepsilon(V_{di}) \tag{38}\]

so that \( \varepsilon(V_{di}) \) becomes a per-unit deviation from the idealized linear voltage gain of \((-g_{mQ}R_l)\). Figure (3) depicts this deviation in percentage form as a function of the differential input voltage, normalized to the quiescent value of the drain saturation voltage. The plot indicates a substantial linearity degradation with increasing differential voltage. As an example of this degradation, consider the case in which \( V_{di} = 250 \text{ mV} \) is to produce an output response in the considered differential amplifier that deviates from linearity by no more than \( 1\% \). Then, an assiduous investigation of the data underlying the plot at hand indicates that \( V_{di}/V_{dsat} \) must be no larger than \( 0.283 \), which implies \( V_{dsat} = V_{di}/0.283 = 250 \text{ mV}/0.283 = 883.4 \text{ mV} \). Assume that the transistor has a gate aspect ratio, \( W/L \), of 5, an oxide thickness of \( 20 \text{ Å} \), and is characterized by a carrier mobility in the inversion channel of \( 900 \text{ cm}^2/\text{V-sec} \). With an oxide dielectric constant, \( \varepsilon_{ox} \), of \( 345 \text{ fF/cm} \), (5) gives \( \beta = 7.76 \text{ \mu mho/volt} \). Then by (36), the requisite tail current is \( I_k = 12.12 \text{ mA} \), which indicates that under standby operating conditions, each transistor in the pair must conduct
6.06 mA of current. It is to be understood that this current flows through each load resistance, \( R_l \), whose value is determined by the desired linear, or small signal, voltage gain. Accordingly, the power supply voltage, \( V_{dd} \), must be sufficiently large to sustain the sum of the resultant voltage drop across \( R_l \), a drain-source voltage that exceeds –albeit by a small amount– \( V_{dsat} \), and the voltage, \( V_k \), established across the tail current sink.

![Graphical quantification of the nonlinearity associated with the simple differential pair on Figure (2). The deviation from linearity is \( \varepsilon(V_{di}) \), as introduced in (38), and the differential input voltage is normalized to the quiescent value of the drain saturation voltage.](image)

### 2.2. TRANSISTORS IN TRIODE REGION

If \( M1 \) and \( M2 \) operate in their triode, or ohmic, regimes, (7) establishes a differential output current, \( I_{do} \), deriving from

\[
\frac{I_{do}}{2\beta} \equiv \frac{I_{d1} - I_{d2}}{2\beta} = V_{ds1} \left( V_{gs1} - V_h - \frac{V_{ds1}}{2} \right) - V_{ds2} \left( V_{gs2} - V_h - \frac{V_{ds2}}{2} \right).
\]

(39)

An inspection of the schematic diagram of Figure (2) confirms that the drain-source voltage, \( V_{ds1} \), of transistor \( M1 \) is simply the difference between the indicated single ended voltages, \( V_{o1} \) and \( V_k \). On the other hand, the drain-source voltage of \( M2 \) is the difference between \( V_{o2} \) and \( V_k \). Recalling (22), these drain-source voltages can be written as
\[ V_{ds1} = V_{o1} - V_k = V_{co} + \frac{V_{do}}{2} - V_k \]
\[ V_{ds2} = V_{o2} - V_k = V_{co} - \frac{V_{do}}{2} - V_k \]  

Similarly, Figure (2) and (21) confirm for the transistor gate-source voltages, \( V_{gs1} \) and \( V_{gs2} \), that
\[ V_{gs1} = V_{i1} - V_k = V_{ci} + \frac{V_{di}}{2} - V_k \]
\[ V_{gs2} = V_{i2} - V_k = V_{ci} - \frac{V_{di}}{2} - V_k \]

The insertion of (41) and (40) into (39), followed by obligatory algebra, produces
\[ I_{do} = 2\beta(V_{co} - V_k)V_{di} + 2\beta(V_{ci} - V_{co} - V_h)V_{do} \]  

In view of the fact that
\[ V_{do} = R_l(I_{d2} - I_{d1}) = -R_lI_{do} \]
(42) implies a voltage transfer function of
\[ A_v(V_{di}) = \frac{V_{do}}{V_{di}} = \frac{-2\beta R_l(V_{co} - V_k)}{1 + 2\beta R_l(V_{ci} - V_{co} - V_h)} \]  

The first of the observations to surface from (42) and (44) is that if the tail current, \( I_k \), is realized by a current sink presenting very high impedance to the sources of \( M1 \) and \( M2 \) and if the differential input signal, \( V_{di} \), fosters only weak nonlinearities in the output response, voltage \( V_{co} \) is exclusively a quiescent voltage. If voltage \( V_{ci} \) is likewise a static voltage, the gain in (44) is a constant, independent of \( V_{di} \). In other words, the differential pair of Figure (2) functions as a linear circuit when both transistors sustain triode region operating points. This state of affairs contrasts sharply with the saturation result postured in (35), which projects an I/O voltage describing function collapsing to a constant gain value if and only if the differential input signal is sufficiently small. Of course, the gain in (44) remains constant only if transistors \( M1 \) and \( M2 \) continue to operate in their triode regimes for all values of the differential signal excitation, \( V_{di} \). Since the drain-source voltage of each of the transistors in the differential pair at hand must be smaller than a threshold potential below the respective gate-source voltage, it is easily confirmed that triode domain operation mandates
\[ \frac{V_{di}}{2} < \frac{V_{ci} - V_{co} - V_h}{l + |A_v(V_{di})|} \]  

A further confirmation of triode domain I/O linearity derives from observing that (42) and (43) give rise to the simple model shown in Figure (4). The model in question advances an effective triode regime transconductance, say \( g_{mt} \), of \( 2\beta(V_{co} - V_h) \) and an output resistance, say \( r_{ot} \), of \( 1/2\beta(V_{ci} - V_{co} - V_h) \). Observe that the voltage difference, \( (V_{co} - V_h) \), is simply the quiescent value, \( V_{dsQ} \), of the transistor drain source voltages, while \( (V_{ci} - V_{co} - V_h) \) is little more than \( (V_{gsQ} - V_{dsQ} - V_h) \), where \( V_{gsQ} \) represents quiescent gate-source voltage. These observations synergize with (7), which delivers the traditional small signal transconductance and resistance parameters,
Fig. (4). Equivalent circuit of the differential pair in Figure (2) when transistors M1 and M2 operate in their triode regimes.

\[
\begin{align*}
g_{mt} & \equiv \frac{\partial I_d}{\partial V_{gs}} = 2\beta V_{dsQ} \\
\frac{I}{r_{ot}} & \equiv \frac{\partial I_d}{\partial V_{ds}} = 2\beta(V_{gsQ} - V_{dsQ} - V_h)
\end{align*}
\]

The model also delivers a Thévenin, and thus maximum, differential voltage gain of

\[
\begin{align*}
\frac{V_{do}}{V_{di}} \bigg|_{R_l=\infty} &= -\frac{V_{co} - V_k}{V_{ci} - V_{co} - V_h} = -\frac{V_{dsQ}}{V_{gsQ} - V_{dsQ} - V_h}.
\end{align*}
\]  

This gain must exceed unity if greater than unity gain is to be achieved under terminated load conditions. Thus, the quiescent drain-source voltage, \(V_{dsQ}\), must be larger than \((V_{gsQ} - V_h)/2\), or \(V_{dsQ} > V_{dsatQ}/2\), where \(V_{dsatQ}\) obviously symbolizes the quiescent value of the drain saturation voltage. But since a necessary condition supportive of triode regime operation is \(V_{dsQ} < V_{dsatQ}\), the drain to source signal swing required to sustain triode region operation is necessary limited to \(V_{dsatQ}/2\) since the allowable drain-source quiescent operating voltage excursion is limited by

\[
\frac{V_{dsatQ}}{2} < V_{dsQ} < V_{dsatQ}.
\]

3.0. GILBERT MULTIPLIER

The basic schematic diagram of the Gilbert multiplier offered in Figure (5) consists of a driver differential pair terminated at its drain ports in a quad array of two interconnected differential pairs. The driver pair is a balanced differential amplifier formed of transistors M1 and M2, whose effective transconductance coefficient, \(\beta\), is taken herewith as \(\beta_x\). A differential voltage signal, \(V_x\), which is simply the difference of the indicated single ended voltages, \(V_{x1}\) and \(V_{x2}\), is applied to the driver pair to produce the drain currents, \(I_{d1}\) and \(I_{d2}\), conducted by M1 and M2, respectively. These currents flow through an active load comprised of the dual differential pairs, M3-M4 and M5-M6. Although M3, M4, M5, and M6 are matched transistors, their gate aspect ratios need not be identical to those of the transistors utilized in the driver pair. Accordingly, the effective transconductance coefficient of each of the active devices in the interconnected quad cell is taken to be \(\beta_y\). Note that a differential signal, \(V_y\), which is the difference between the single ended voltages, \(V_{y1}\) and \(V_{y2}\), is applied from the gate of transistor M3 to the gate of transistor M4, as well as from the gate of M6 to the gate of M5. The immediate effect of this quad cell differential input signal is to establish the drain currents, \(I_{d3}\), \(I_{d4}\), \(I_{d5}\), and \(I_{d6}\), which in turn produce the currents, \(I_a\) and \(I_b\), conducted by the resistive load terminations, \(R_l\). It should be noted that while the tail current, \(I_t\), of the M1-M2 pair is a constant, the tail currents of each of
the pairs in the quad cell are not constant. In particular, the tail current of the M3-M4 pair is the drain current, $I_{s1}$, of transistor M1, and the tail current of the M5-M6 pair is the drain current, $I_{s2}$, flowing in M2. Multiplicative action between the applied signals, $V_x$ and $V_y$, results from two topological constraints imposed by the Gilbert topology. First, the differential currents, $(I_{d3} - I_{d4})$ and $(I_{d6} - I_{d5})$, are functionally related to the signal, $V_y$, applied differentially to the quad cell. Second, currents $I_{d3}$ and $I_{d4}$ sum to the M1 drain current, $I_{s1}$, while currents $I_{d6}$ and $I_{d5}$ superimpose to form the drain current, $I_{s2}$, flowing in transistor M2. But the differential current, $(I_{s1} - I_{s2})$, which is dependent on the differential voltage excitation, $V_x$, necessarily modulates the differential voltage, $V_{sd}$, established as indicated between the drain terminals of transistors M1 and M2. In turn, voltage response $V_{sd}$ influences the gate-source voltages, and hence the drain currents, of the four quad transistors. It follows that the resultant differential output current, $(I_o - I_b)$, which is clearly related to the quad cell drain currents, is influenced by both $V_x$ and $V_y$. The analyses that follow confirm that this effect is manifested by a differential output current that is ideally proportional to the product of the voltage signals, $V_x$ and $V_y$.

Since the Gilbert multiplier is an interconnection of only balanced differential pairs, its nonlinear analysis for the case of nominally weak nonlinearities, can be pursued by exploiting the analytical depositions of the preceding section of technical material. Two cases are consid-
tered. The first is the case in which all six transistors operate in their saturation domains. In the second case, transistors $M1$ and $M2$ remain in saturation, but transistors $M3$ through $M6$ are presumed to operate in their triode regimes.

### 3.1. ALL TRANSISTORS IN SATURATION

Recalling that transistors $M1$ and $M2$ have an effective transconductance coefficient of $\beta_x$, (32) gives

$$I_{s1} - I_{s2} = V_x \sqrt{2\beta_x I_k} \sqrt{1 - \frac{\beta_x V_x^2}{2I_k}} \approx V_x \sqrt{2\beta_x I_k} \left(1 - \frac{\beta_x V_x^2}{4I_k}\right),$$

(48)

where the approximation exploits the weak nonlinearity presumption; that is, differential voltage $V_x$ is small enough to satisfy the inequality, $(\beta_x V_x^2) << 2I_k$. Similarly, and with $\beta_y$ recalled as the transconductance coefficient of the four transistors in the quad interconnection,

$$I_{d3} - I_{d4} = V_y \sqrt{2\beta_y I_{s1}} \sqrt{1 - \frac{\beta_y V_y^2}{2I_{s1}}} \approx V_y \sqrt{2\beta_y I_{s1}} \left(1 - \frac{\beta_y V_y^2}{4I_{s1}}\right)$$

(49)

and

$$I_{d6} - I_{d5} = V_y \sqrt{2\beta_y I_{s2}} \sqrt{1 - \frac{\beta_y V_y^2}{2I_{s2}}} \approx V_y \sqrt{2\beta_y I_{s2}} \left(1 - \frac{\beta_y V_y^2}{4I_{s2}}\right).$$

(50)

The differential output current, say $I_{do}$, is simply $(I_a - I_b)$, which an inspection of Figure (5) reveals to be $(I_{d3} - I_{d4}) - (I_{d6} - I_{d5})$, or simply the difference of the two preceding results. Using (23) and (24) for currents $I_{s1}$ and $I_{s2}$ in (49) and (50), it can be demonstrated that

$$I_{do} = I_a - I_b = V_x V_y \sqrt{2\beta_x \beta_y} \left[1 + \frac{\beta_y V_y^2}{4\beta_x (V_{gs1} - V_h)(V_{gs2} - V_h)}\right].$$

(51)

But

$$\beta_x (V_{gs1} - V_h)(V_{gs2} - V_h) = \sqrt{I_{s1} I_{s2}},$$

(52)

and appealing to (30)

$$\beta_x (V_{gs1} - V_h)(V_{gs2} - V_h) = \sqrt{I_{s1} I_{s2}} = \frac{I_k - \beta_x V_x^2}{2}.$$  

(52)

It follows that the differential output voltage, $V_{do}$, is

$$V_{do} = -R_l I_{do} = -V_x V_y R_l \sqrt{2\beta_x \beta_y} \left[1 + \left(\frac{\beta_y V_y^2}{V_x}\right) \left(\frac{V_y}{V_x}\right) \left(\frac{I_k}{\beta_x V_x^2} - 1\right)\right].$$

(53)
It is instructive to recast (53) into the compact form,

\[ V_{do} = -\frac{V_x V_y}{V_{GM}} \left[ 1 + \varepsilon(V_x, V_y) \right], \]  

(54)

where

\[ \frac{1}{V_{GM}} = R_l \sqrt{2\beta_x \beta_y} \]  

(55)

\((1/V_{GM})\), in units of inverse volts, might be thought of as the nominal transducer gain of the Gilbert multiplier, and

\[ \varepsilon(V_x, V_y) = \left( \frac{\beta_y}{2\beta_x} \right) \left[ \frac{I_k}{\beta_x V_x^2} - 1 \right] \]

\(\varepsilon(V_x, V_y)\) is a per unit deviation of the output response with respect to the nominal output value of \(-V_x V_y/V_{GM}\). If \(\varepsilon(V_x, V_y)\) is very small for values of the differential input signals, \(V_x\) and \(V_y\), an idealized multiplier response is achieved in the sense that the differential output voltage response is proportional to the product of the two input signals. Equation (56) highlights the fact that an acceptably small deviation requires a suitably large tail current and/or a constrained differential signal applied to the quad array. This contention is reinforced by exploiting the commonly encountered situation in which \(I_k \gg \beta_x V_x^2\). In this event, (56) reduces to

\[ \varepsilon(V_x, V_y) \approx \frac{\beta_y V_y^2}{2I_k}, \]  

(57)

which is independent of the transistor parameters in the driver pair and the differential signal applied to the pair.

The foregoing independence of the linearity deviation factor on the driver differential signal, \(V_x\), is a curiosity worthy of further exploration. Of course, (56) does convey a weak deviation factor dependence on \(V_x\), but it also suggests a potential stability problem. In particular, observe that \(\varepsilon(V_x, V_y)\) theoretically tends toward infinity as \(V_x\) approaches a signal strength rendering \(I_k = \beta_x V_x^2\). Accordingly, if the peak value, say \(V_{xp}\), of the driver differential signal is known, it is necessary to design the Gilbert circuit to ensure that

\[ I_k > \beta_x V_{xp}^2. \]  

(58)

An inspection of the schematic diagram in Figure (2) shows that this requirement is tantamount to the stipulation that the maximum positive or negative value of voltage \(V_x\) must never be so large as to force all of the available tail current through either one of the two driver transistors, \(M1\) and \(M2\).
3.2. QUAD ARRAY TRANSISTORS IN TRIODE REGIME

Before commencing with the multiplier analysis for the case in which the transistors in the quad array operate in their triode domains, it is expedient to formulate bookkeeping with respect to defining the various voltages in the schematic diagram of Figure (5) in terms of their common mode and differential mode constituents. To this end, let the voltages, \( V_{y1} \) and \( V_{y2} \), applied to the quad array be represented as

\[
V_{y1} = V_{yc} + \frac{V_y}{2} \\
V_{y2} = V_{yc} - \frac{V_y}{2}
\]

where the common mode component, \( V_{yc} \), is presumed to be a static voltage and, of course, \( V_y \) is the differentially applied signal voltage delineated in the schematic diagram. If \( V_{ac} \) symbolizes the common mode output port voltage,

\[
V_{a1} = V_{ac} + \frac{V_{do}}{2} \\
V_{a2} = V_{ac} - \frac{V_{do}}{2}
\]

In the paragraphs that follow, \( V_{ac} \) is presumed to be a static voltage. The propriety of this presumption relies on the absence of excessive signal strength at both the driver and quad array input ports, matched pairs in the quad cell, and high impedances presented to the quad subcircuit by the drains of transistors \( M1 \) and \( M2 \) in the differential driver. The latter requirement is likely to require appropriately up-scaled channel lengths in \( M1 \) and \( M2 \) to minimize the effects of their channel length modulations if the indicated differential voltage,

\[
V_{sd} = V_{s1} - V_{s2},
\]

across the drain terminals of \( M1 \) and \( M2 \) is forced to swing to relatively large magnitudes. Finally, let

\[
I_{sd} = I_{s1} - I_{s2}  \\
I_{do} = I_a - I_b
\]

respectively designate the difference between the drain currents, \( I_{s1} \) and \( I_{s2} \), conducted by \( M1 \) and \( M2 \), and the differential current of the multiplier output port.

Using (7) as a volt-ampere characteristic template for the triode regime, the drain currents, \( I_{d3} \) and \( I_{d4} \), in \( M3 \) and \( M4 \) derive from

\[
\frac{I_{d3}}{2\beta_y} = (V_{a1} - V_{s1})(V_{y1} - V_{s1} - V_h) - \frac{(V_{a1} - V_{s1})^2}{2} \\
\frac{I_{d4}}{2\beta_y} = (V_{a2} - V_{s1})(V_{y2} - V_{s1} - V_h) - \frac{(V_{a2} - V_{s1})^2}{2}
\]
while for transistors $M5$ and $M6$,

$$
\frac{I_{d5}}{2\beta_y} = \left( V_{a1} - V_{s2} \right) \left( V_{y2} - V_{s2} - V_h \right) - \frac{\left( V_{a1} - V_{s2} \right)^2}{2}, \quad (64)
$$

$$
\frac{I_{d6}}{2\beta_y} = \left( V_{a2} - V_{s2} \right) \left( V_{y2} - V_{s2} - V_h \right) - \frac{\left( V_{a2} - V_{s2} \right)^2}{2}. \quad (64)
$$

Since the differential output current is

$$
I_{do} = (I_{d3} - I_{d4}) - (I_{d6} - I_{d5}), \quad (65)
$$

(63) and (64) combine with (59), (60), and (61) to produce

$$
\frac{I_{do}}{2\beta_y} = -V_{sd}V_y + 2(V_{yc} - V_{ac} - V_h)V_{do}. \quad (66)
$$

In (66), the difference voltage, $(V_{yc} - V_{ac})$, is the common mode gate-drain voltage for each of the four transistors in the quad cell of Figure (5). In the triode operational regime, this difference potential must exceed the transistor threshold potential, $V_h$. Thus, design care must be exercised to ensure that the coefficient of voltage $V_{do}$ on the right hand side of (66) is a positive voltage. On the other hand, the coefficient, $V_{sd}$, of voltage $V_y$ can be a positive or a negative voltage, depending on the algebraic sense of the differential input voltage, $V_x$, applied to the driver cell.

Continuing with the analysis, the differential tail current, $I_{sd}$, established by the quad array is

$$
I_{sd} = I_{s1} - I_{s2} = \left( I_{d3} + I_{d4} \right) - \left( I_{d5} + I_{d6} \right). \quad (67)
$$

Upon exploiting (59), (60), and (61) once again, (67) leads to

$$
\frac{I_{sd}}{2\beta_y} = V_{do}V_y - 2(V_{yc} - V_h)V_{sd}. \quad (68)
$$

But current $I_{sd}$ is also constrained by voltage $V_x$ in accordance with the approximate relationship of (48). It follows that (68) and (48) combine to deliver

$$
V_{sd} = \frac{1}{2(V_{yc} - V_h)} \left[ V_{do}V_y - \sqrt{2\beta_xI_k} \left( 1 - \frac{\beta_y V_y^2}{4I_k} \right) V_x \right]. \quad (69)
$$

Recalling that the differential output voltage, $V_{do}$, is related to its counterpart differential output current, $I_{do}$, by $V_{do} = -R_I I_{do}$, the substitution of (69) into (66) produces the desired result,

$$
V_{do} = -V_x V_y \left\{ \frac{R_I \sqrt{2\beta_x I_k}}{2V_M} \left[ 1 - \frac{\beta_y V_y^2}{4I_k} \right] \right\}, \quad (70)
$$

where voltage parameter $V_M$ is defined as
\[ V_M \triangleq (V_{yc} - V_h) \left[ 1 + 4 \beta_y R_l (V_{yc} - V_{ac} - V_h) \right]. \]  

(71)

Because the quad transistors operate in their triode domains, \((V_{yc} - V_{ac})\), which is the common mode voltage applied from gate to drain in the quad transistors, is likely to be only slightly larger than threshold voltage \(V_h\). For reasonably small values of the load resistance, \(R_l\), \(V_M\) is thus seen as approximating \((V_{yc} - V_h)\); that is, parameter \(V_M\) is a voltage that is roughly one threshold drop below the common mode biasing applied to the gates of the transistors in the quad configuration.

A study of (70) reveals that the first parenthesized term on the right hand side is a constant, independent of the signal levels, \(V_x\) and \(V_y\). On the other hand, the second parenthesized term on the right hand side of (70) is dependent on \(V_x\) and \(V_y\) and collapses to unity when \(V_x\) and \(V_y\) are very small voltages. It is therefore meaningful to express (70) as

\[ V_{do} = -V_x V_y \left( \frac{R_l \sqrt{2 \beta_x I_k}}{2 V_M} \right) \left[ 1 + \varepsilon_t(V_x, V_y) \right], \]

(72)

where

\[ \varepsilon_t(V_x, V_y) = \frac{\beta_y R_l V_y^2}{V_M} - \frac{\beta_x V_x^2}{4 I_k} \]

(73)

where \(\varepsilon_t(V_x, V_y)\) is the deviation from an idealized voltage response that is linearly proportional to the product, \((V_x V_y)\), of signals applied to the multiplier. Unlike the linearity error of (56) for the saturated case, which is directly related to the square of the voltage ratio, \((V_y/V_x)\), the error at hand is contrasted by a dependence on the scaled differences of the squares of voltages \(V_x\) and \(V_y\). This state of affairs begets the plausibility of reduced linearity error when the devices in the quad cell are operated in triode domains. Indeed, zero error is theoretically possible if

\[ \left( \frac{V_y}{V_x} \right)^2 = \frac{\beta_x V_M}{4 \beta_y R_l I_k}. \]

(74)

This condition for zero error is hardly sacrosanct because of the numerous approximations invoked in the interest of realizing mathematically tractable conclusions. Nonetheless, it does suggest that a Gilbert multiplier operated with the quad array of transistors in triode domains is likely to deliver more acceptable error deviations from idealized responses than are possible when all transistors operate in saturation.

As in the globally saturated case, stability issues persist in the sense that it is possible for the denominator on the right hand side of (73) to vanish for a particular value of signal level, \(V_y\). To obviate such problems, the common mode biasing of the quad cell must be such that for a maximum value, say \(V_{yp}\), of signal \(V_y\),

\[ V_M / R_l > \beta_y V_{yp}^2. \]

(75)
In other words, the current that any of the quad transistors conducts when \( V_y = V_{yp} \) must be less than the current conducted by the load resistance when the common mode output voltage is a \( V_{M} \) level below the power bus voltage, \( V_{dd} \).

4.0. CONCLUSIONS

The analyses undertaken in this investigation confirm that in balanced differential amplifiers, improved linearity accrues when the transistors embedded in the signal flow path are biased for operation in their triode domains, as opposed to the more traditionally imposed saturation regime biasing. While this fact is hardly a seminal discovery, the same statement is shown to apply to the ubiquitous Gilbert multiplier when the transistors interconnected in the quad cell are operated in their triode regimes. The fundamental price paid for the improved linearity in both amplifier and multiplier applications is reduced signal swing. In the multiplier case, this reduced signal swing mandates a signal differentially applied to the quad cell that is substantively smaller than the differential signal activating the driver pair. In both the triode and saturation cases, a suitably large tail current is required to achieve exceptionally good linearity. In addition, the triode case requires that the effective impedances seen looking into the drain terminals of the driver pair be very large, thereby ensuring that the common mode voltages established at the multiplier output ports are exclusively static in nature.

A meaningful continuing study might entail relating the linearity deviation errors derived in this report to such performance indices as total harmonic distortion and intermodulation products. A high frequency linearity study is also a worthy endeavor, but this effort can likely be conducted pragmatically only through behavioral analyses.

5.0. REFERENCES