Appl. Phys. A 28, 151–160 (1982)

Applied Physics A Solids and Surfaces

© Springer-Verlag 1982

Majority Carrier Transistor Based on Voltage-Controlled Thermionic Emission

R.F. Kazarinov and Serge Luryi

Bell Laboratories, Murray Hill, NJ 07974, USA

Received 24 March 1982/Accepted 15 April 1982

Abstract. A new type of transistor is proposed based on gate-controlled charge injection in unipolar semiconductor structures. Its design has some similarity with the recently fabricated triangular barrier diodes but contains an additional input circuit which allows an independent control of the barrier height for thermionic emission. This circuit is provided by a MOS gate on the semiconductor surface. In the proposed device the current flows perpendicular to the semiconductor surface over a planar potential barrier controlled by the gate. The static transconductance characteristics and dynamical response are analyzed. The characteristic response time is limited by the time of flight of electrons across the structure and can be in the picosecond range. The gate voltage required to switch the output current at room temperature is of order 0.2 V.

PACS: 85.30. Hi, 73.40. Lq, 85.30. Tv

In the present paper a new type of a majority-carrier transistor is proposed which is based on the phenomenon of charge injection in unipolar semiconductor structures. By charge injection we mean the thermionic emission of carriers over a potential barrier when the barrier height is efficiently controlled by an applied voltage.

The physical principle involved can be illustrated by analogy with a vacuum-tube diode. In these diodes, for sufficiently large anode voltage, the current saturates at a value determined by the thermionic emission from the cathode. The value of the saturation current exponentially depends on the barrier height for thermal emission, i.e. the work function of the cathode material. Suppose for a moment that we could control the work function with the help of some ingenious input circuit (gate). Moreover, suppose that the barrier height depends linearly on the gate voltage and rapidly adjusts to its variation. The output current would then depend on the gate voltage exponentially and hence the transconductance would be proportional to the current. For a sufficiently large current we would have high values of the transconductance and therefore fast response time with low power-delay product and noise. Needless to say, we do not possess the means of controlling the work function of a metal in the indicated way.

However, in semiconductors the above physical idea can be realized. An example of the charge injection device is provided by the IGFET in its subthreshold regime [1]. Indeed the subthreshold drain current is due to the thermionic emission from the source which in this regime plays the role of a cathode. The potential barrier between the source and the channel linearly decreases with the band bending which is controlled by the gate voltage. In the subthreshold regime IGFET may be called a potential-effect rather than a field-effect device. By a reasonable definition the field effect consists in the screening of the electric field under the gate by an accumulation or depletion of the mobile charge in the channel. In IGFET this occurs only in the strong inversion regime where the surface carrier density is proportional to the field. Below threshold the electrons in the channel give no significant contribution to the screening and their concentration is determined by the surface potential rather than the field.

In the potential-effect (charge injection) mode of FET the surface charge density σ of electrons in the channel depends exponentially on the gate voltage V_G, i.e.,

0721-7250/82/0028/0151/$02.00

$\sigma \propto \exp(\beta V_G/n)$ where $\beta = q/kT$ and n is some ideality factor of order unity. For large σ the exponential dependence is lost because of the screening effect and σ becomes a linear function of V_G which can be represented by making n an increasing function of V_G. This transition between the charge-injection and the field-effect modes is not sharply defined. We shall consider the charge injection mode to terminate when n deteriorates by a factor of two. As can be shown (Appendix A) this occurs at a value of $\sigma = \sigma_c$ given by

$$\sigma_c = kT\varepsilon_{ox}/qt_{ox}, \tag{1}$$

where t_{ox} and ε_{ox} are, respectively, the oxide thickness and permittivity. For $\sigma > \sigma_c$ because of velocity saturation, the output current also becomes a linear function of V_G, i.e., the transconductance saturates[1]. This means that by going beyond the subthreshold regime one gains no advantage in the intrinsic speed of operation. On the contrary, working in the strong inversion regime one loses considerably in terms of the power-delay product.

It is clear that the charge injection mode of operation of FET would be more attractive than the field-effect mode provided that high values of the output current could be achieved in the subthreshold regime. Unfortunately, this is not the case. For $t_{ox} \sim 500$ Å and the saturated velocity, $v_s \sim 10^7$ cm/s the maximum current one can obtain in the charge injection mode of IGFET is of the order of 10^{-2} A/cm of gate width. Because of the parasitic capacitances the subthreshold current is usually insufficient for fast operation of the device.

Another obstacle against using FETs in the subthreshold regime is the uncertainty in the threshold voltage due to processing variations. The difficulty in reducing the voltage swing to several kT/q lies not with the thermal noise as is sometimes incorrectly thought (for a typical transconductance of IGFET the mean-square fluctuation of the gate voltage due to the thermal noise at room temperature is of order 1–2 mV) but with reproducibility of the device parameters. Indeed, the charge injection current depends critically on the height and the shape of the potential barrier which is not controlled accurately because of the uncertainty in the state of the surface.

The device we would like to discuss in this work is designed to overcome the above limitations. The crucial feature of the proposed device is the possibility of extending the charge injection regime to currents typical for FETs in strong inversion. This becomes possible because the troublesome accumulations of carriers under the gate which limits the subthreshold current in FET is circumvented here. In this device the output current flows perpendicular to the semiconductor surface and is controlled by potential barriers which are parallel to the surface.

Structures containing such barriers can be fabricated by Molecular Beam Epitaxy (MBE) and by using ion implantation. As is well-known, these methods give much higher resolution than any lithography. For example, the state-of-art MBE technology allows one to obtain modulation-doped semiconductor layers with the resolution of few tens of Å [2]. Rectifying diodes based on such barriers were recently fabricated by ion implantation [3] and by MBE using either variable-gap [4] or modulation-doped [5] materials. In these diodes the current is due to charge injection [6]. The structures studied in [4, 5] contained built-in potential barriers of triangular shape, either symmetric (isosceles) or asymmetric. The current-voltage characteristics were nearly exponential up to the current densities of several kA/cm^2 in both directions. For asymmetric diodes the ideality factors were different in forward and reverse directions of current which corresponds to rectification. In general, the ideality factor of a potential barrier is determined by its geometry and the doping profile. Depending on the ideality factor a barrier can be either injecting or blocking.

Although the first experimental realizations of the triangular barrier (TB) structures appeared very recently [4, 5], conceptually they represent the simplest charge injectors. In a certain sense the TB concept is a generalization of the Schottky barrier. Indeed, in a forward-biased Schottky diode electrons are injected into the metal from the semiconductor. However, because of the large concentration of electrons in the metal, the injected charge produces no tangible effect on the metal conductivity near the boundary. No charge injection into the semiconductor occurs in a reverse-biased Schottky diode (neglecting a small effect of image-force barrier lowering), and current in this case is limited by the thermionic emission over a barrier of fixed height. A similar situation takes place in all-semiconductor analogs of Schottky barriers such as camel diodes [3] and N-n heterojunctions [7]. As before, injection takes place only into a quasi-neutral material. TB offers a fundamentally new feature: efficient injection of charge into a high-field region of a semiconductor. It may be worthwhile to note that this feature also opens an attractive possibility of using TB's for making low-noise transit-time devices similar to but more efficient than the baritts [8], as will be discussed elsewhere.

In the present paper we propose a way of introducing an input gate circuit in a TB structure which allows an

[1] We consider only the case of a short-channel FET. In the long channel device the subthreshold current is limited by the slow diffusion transport through the flat portion of the channel. This further reduces the maximum transconductance achievable in the charge injection regime

independent control of the barrier height. This is an example of a unipolar transistor operating entirely in the charge injection regime. We suggest that in general such devices may be called TET which stands for gate-controlled Thermionic Emission Transistor. We expect that the fundamental advantage of the TET devices lies in the exponential transconductance extended to higher values of the output current. The maximum charge injection current in TET is space charge limited. The exponential dependence allows switching this current by a gate voltage of the order of several kT/q. The characteristic time of switching corresponds to the drift of electrons across the structure and can be in the picosecond range.

A certain analogy exists between TET and the recently proposed [9] Permeable Base Transistor (PBT) in which a grid of metal electrodes is embedded in the semiconductor between the source and the drain of the device. Indeed, in the case of a low-doped base the current in PBT is of thermonic nature over the barrier formed by the controlling electrodes. The main difference is that in TET there exists a built-in triangular barrier which allows us to transfer the controlling electrodes to the surface of the semiconductor.

The proposed design of TET is introduced in Sect. 1. In the same section we describe the physical principles of the device operation, formulate the requirements on its geometry, and discuss the expected characteristics of the device and their limitations. Some of the conclusions in Sect. 1 are presented without proof, on an intuitive level. The rigorous mathematical treatment is given in Sect. 2 where we calculate the transconductance and the characteristic response time of the device. In Sect. 3 we discuss the possible fabrication of TET and summarize our conclusions.

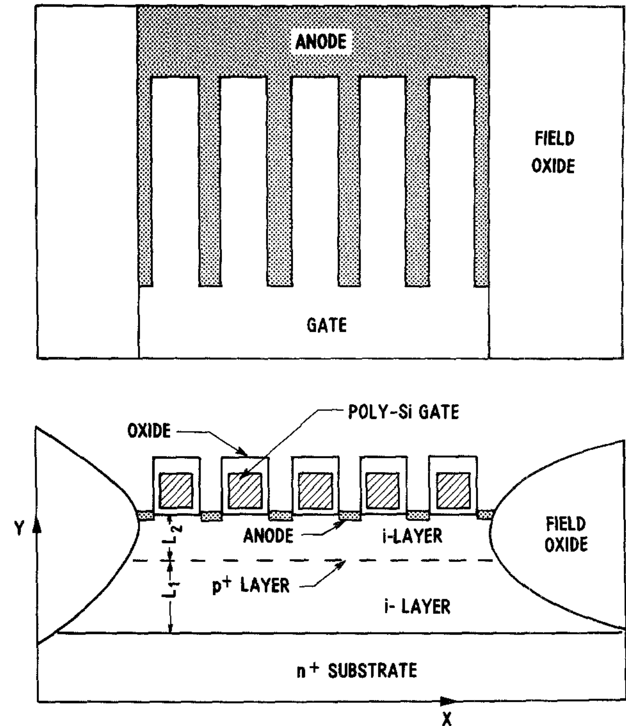

Fig. 1a and b. Schematic description of TET. (a) Top view. (b) Cross-section in the working area of the device

1. Qualitative Description of the Device

The proposed version of the charge injection or Thermionic Emission Transistor (TET) is shown in Fig. 1. The device contains an i layer grown epitaxially on an n^+ substrate. In the process of growth by MBE a p^+ layer is built in the i layer by modulation doping. The thickness δ of the p^+ layer is assumed to be infinitesimal compared to that of the i layer. In practice, δ can be as small as a few tens of Å. The acceptors in the p^+ layer are completely ionized and form a sheet of negative charge which gives rise to a triangular potential barrier (TB) similar to those studied in [5, 6]. The n^+ substrate forms one of the terminals of TET, which will be called the cathode. The other two terminals, the anode and the gate are arranged in a periodic pattern of stripes on the surface. Every other stripe represents a metallized n^+ contact or a silicide layer and these stripes are connected on one side to a metallic pad which is the anode terminal. The alternate stripes connected on the other side, form a gate terminal. From the technological point of view it appears most promising to use a poly-silicon gate structure, as shown in Fig. 1. The top view of TET is shown schematically in Fig. 1a. Figure 1b represents a vertical cross-section in the working area of the device.

As in a TB diode [5, 6], the output current in TET is due to charge injection. The cathode work function, given by the potential difference between the top of TB and the substrate, is, however, nonuniform and varies in the x direction (for the definition of the coordinate axes see Fig. 1). For any combination of the anode and gate voltages (V_A, V_G) the electric field beneath the surface can be split into a uniform part which can be considered emanating from a conducting plane at a constant average potential, and a nonuniform oscillating part. This procedure can be regarded as a multipole expansion of the appropriate symmetry. Close to the surface we have a "near zone" where the field is mainly multipolar and the oscillation of the potential in x direction is appreciable. Far from the surface the potential is uniform and is determined by the field of a parallel-plate condenser charged to the potential $\bar{V}$, viz.

$$\bar{V} = \frac{V_A S_A + V_G S_G}{S}, \qquad (2)$$

where S_A and S_G are the areas of the anode and gate electrodes, respectively, and $S = S_A + S_G$. The transition from the near zone to the far zone occurs at a characteristic distance $y = \lambda$ into the structure which approximately equals the size of the surface dipoles, i.e. the period d of the pattern of stripes on the surface. Our calculations in Sect. 2 show that in fact λ is even smaller. Beyond λ the nonuniform part of the potential drops off exponentially.

For the performance of TET it is crucial that the top of TB should fall into the far zone. In this case the cathode work function, i.e. the barrier height for thermionic emission, is determined by $\bar{V}$, (2), and therefore the thermionic current is controlled by the gate voltage V_G. Thus we can expect the transconductance to have an exponential dependence on V_G with the ideality factor increased by a factor of S/S_G compared to the ideality factor for the IV characteristic of a TB diode of same geometry and doping profile. Of course, the exponential dependence must cease at some value of the current which, therefore, limits the charge injection regime. This limitation in TET is brought about by two distinct phenomena: i) slowing down of the effective diffusion velocity on the uphill slope of TB as steepness of the latter decreases, and ii) carrier accumulation on the downhill slope which screens the influence of the gate on the barrier height. Which of these factors is more important depends on the barrier geometry and the substrate doping. As shown in the next section, in silicon, with the barrier dimensions $L_1, L_2 \sim 10^{-5}$ cm (Fig. 1) and the substrate doping $N_D \ll N_C$ where N_C is the effective density of states in the conduction band, the second of the above limitations dominates. In this case it turns out that the ideality factor increases by a factor of two at current densities J of the order of a few kA/cm^2. At higher J the exponential dependence is replaced by a power law appropriate for a space charge limited current.

Let us emphasize that the above limitations of the charge injection mode in TET apply only to the current density J per unit *area* of the device, rather than to the *linear* current density as is the case for FETs. For a TET with the surface area 10 μm × 10 μm the charge injection current is of order ten milliamps like the strong inversion current in IGFET of same linear dimensions. However, the controlling voltage required to switch such current is an order of magnitude lower, viz. several kT/q. It is important to realize that an accurate threshold control should not present a problem in TET. Indeed, the height and the shape of the TB are determined with high precision in the process of crystal growth by MBE. Moreover, inasmuch as TB is located far from the surface the interface states produce no significant variation of the barrier height. One can expect to control the TET threshold within a 25 mV margin and use as low as ~0.2 V for switching. Accordingly reduced will be the charge associated with parasitic capacitances and this allows us to expect that the speed of operation of TET in a real integrated circuit will approach its intrinsic speed. As is shown in the next section the characteristic switching time of TET is determined by the time of flight of electrons on the downhill slope of TB and can be in the picosecond range.

So far we have discussed the case when the top of the barrier is located in the far zone of the anode-gate multipole. It is instructive to consider qualitatively the opposite case when the top of TB is in the non-uniform potential region. This merely means that the period of the gate-anode stripes is larger than the thickness of the i layer. With a negative voltage applied to the gate the barrier height will vary periodically in the x direction so that the current flow will occur mainly in regions under the anode stripes. Clearly, the gate voltage will have little control over this current which shows the necessity of placing the barrier beyond the near zone, i.e. in the uniform-potential region. The exact relation between the extent λ of the near zone and the period d of the electrode pattern on the surface will be established in the next section.

2. Calculation of the Transconductance and the Characteristic Response Time

In the previous section we formulated the main requirement on the geometry of TET. It was stated that the top of TB should be located sufficiently far from the surface so that the barrier height would be uniform in x direction and be determined by the average potential $\bar{V}$ on the gate and anode electrodes. To address this question quantitatively we must consider the two-dimensional electrostatic problem corresponding to Fig. 1b.

We shall make a simplifying assumption that the oxide is negligibly thin and thus the anode and gate electrodes lie in the same plane on the semiconductor surface. The electrostatic potential $\Psi(x, y)$ is determined by the solution to the Laplace equation in the domain shown in Fig. 2a. The boundary conditions are given by $\Psi(x, 0) = 0$ on the boundary between the substrate and the i layer, and $\Psi(x, L) = V(x)$ on the surface, where $V(x)$ is the periodic function shown in Fig. 2b. We seek the solution in the form of a series,

$$\Psi(x, y) = \bar{V}y/L + \sum_{m=1} a_m \cos k_m x \sinh k_m y / \sinh k_m L$$
$$k_m = 2\pi m/d \tag{3}$$

corresponding to the multipole expansion discussed in Sect. 1. Each term in the series (3) satisfies the Laplace

equation, $\nabla^2\Psi=0$, and the boundary condition at $y=0$. In order to satisfy the boundary condition at $y=L$, the coefficients a_m must be the Fourier coefficients of the cosine-expansion of $V(x)$, viz.

$$a_m=(V_A-V_G)\frac{\sin(m\pi S_G/S)}{m\pi/2}. \qquad (4)$$

It is readily seen that at distances $L_2>\lambda=d/2\pi$ the multipole terms in the expansion (3) are exponentially small with L_2, the m-th term being of order $m^{-1}\exp(-mL_2/\lambda)$.

Our simplifying assumption (Fig. 2) consisted in neglecting the gaps between the anode and gate electrodes. Without this assumption the problem becomes mathematically more difficult and requires matching of the solutions in the upper and lower half-planes, as done in Appendix B for calculation of the input capacitance. Our present conclusions regarding the extent of the near zone are quite general and depend only on periodicity of the surface boundary condition. Away from the surface the inhomogeneous part of the potential drops off exponentially with a characteristic length λ given by

$$\lambda=d/2\pi. \qquad (5)$$

If the top of the triangular barrier is located in the far zone, $L_2\gg\lambda$, then the barrier height is constant in the x direction and the thermionic current over the barrier has a uniform density J given by [10]:

$$J=A^*T^2\,e^{\beta\Psi_1}, \qquad (6)$$

where A^* is an effective Richardson constant. It is assumed in (6) that $V_A\gg kT/q$ so that the reverse component of the current can be neglected. According to the theory [6] the barrier height Ψ_1 can be expressed in terms of the asymptotic value E_1 of the electrostatic field near the top of TB on the uphill slope as follows:

$$\Psi_1=-E_1L_1+\frac{2kT}{q}\ln\left|\frac{2E_1}{E_0}\left\{\left[1+\left(\frac{E_1}{E_0}\right)^2 e^{1+(E_1/E_0)^2}\right]^{1/2} -\left[\left(\frac{E_1}{E_0}\right)^2 e^{1+(E_1/E_0)^2}\right]^{1/2}\right\}\right|, \qquad (7)$$

where $E_0^2\equiv kTN_D/\varepsilon$ with N_D being the donor concentration in the substrate. When the conditions $L_2\gg\lambda$ and $V_A\gg kT/q$ are fulfilled the anode current in TET is described by the diode formula (6) expressed in terms of the field E_1 on the uphill slope. Expressions (6) and (7) will be used below to estimate the maximum current in the charge injection regime. However, these equations are insufficient to determine the complete IV characteristics of the transistor since they must be complemented by a relation between E_1 and the

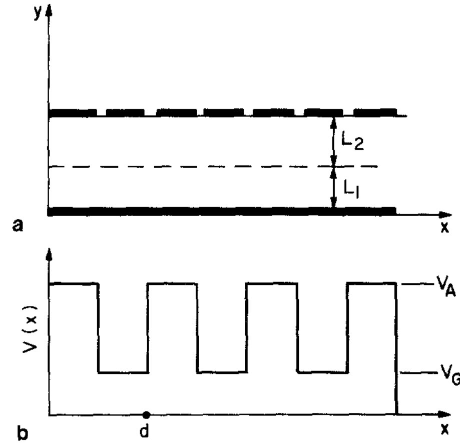

Fig. 2a and b. Simplified geometry of the electrodes (a) and potential distribution at the surface (b) used in the mathematical model of TET

applied voltages V_A and V_G. Strictly speaking, such a relation cannot be obtained from the corresponding diode relation by substituting $\bar{V}$ for V. Indeed, in the diode case the field E_1 contains a contribution associated with the variation of the electrostatic potential in the n^+ material of the anode. The corresponding contribution in TET is difficult to take into account because of the inhomogeneous field in the near zone. However, according to the diode theory [6], the drop ϕ_2 of the electrostatic potential in the n layer adjacent to the downhill slope of TB is almost independent of the bias and, moreover, ϕ_2 enters additively in the expression for the barrier height. Therefore, by replacing $V\to\bar{V}$ in the diode equation we obtain a correct expression for the TET transconductance characteristic, accurate to within a constant factor.

It was shown in [6] that the dependence of J on the applied voltage in a TB diode is well approximated by an exponential form $J\propto\exp(\beta Vl_1)$ where l_1 is an effective dimensionless length of the barrier shoulder on the cathode side, viz.

$$l_1=\frac{L_1}{L}+\frac{2\sum}{qLN_D}\left(1-2\frac{L_1}{L}\right), \qquad (8)$$

where $\sum$ is the surface density of the charge in the p^+ sheet. From the above discussion and using (2) for $\bar{V}$ we obtain the transconductance characteristic of TET in the form

$$J=J_0e^{\beta V_G/n}, \qquad (9)$$

where the ideality factor $n=S/S_Gl_1$ with l_1 given by (8) and J_0 is a function of V_A. From (9) the transconductance $g_m\equiv\partial I/\partial V_G$ (where $I=JS$ is the anode current in a device of area S) is given by

$$g_m=qI/nkT. \qquad (10)$$

The fact that g_m is proportional to the current is due to the exponential form of (9) which is the essence of the charge injection regime. The maximum value of g_m corresponds to the largest current we can obtain in this regime. As stated in Sect. 1 the limitations on the charge injection current arise in TET for two reasons.

First let us discuss the limitation due to the slowing down of the effective diffusion velocity on the uphill slope. The exponential dependence (9) with a constant prefactor is strictly valid for a sufficiently steep uphill slope, $E_1 \gtrsim E_0$. In the opposite limit, $E_1 \ll E_0$ we must use the general expressions (6) and (7) which in this limit give

$$J = A^* T^2 \left(\frac{2E_1}{E_0}\right)^2 e^{-\beta E_1 L_1}. \tag{11}$$

Equation (11) still describes an exponential IV characteristic but with a prefactor which decreases with bias. It formally predicts $J \to 0$ in the limit of a flat uphill slope, $E_1 = 0$, while it is physically clear that without a barrier we must have a space-charge limited current flow. The curve $J(E_1)$ goes through a maximum at $\beta E_1 L_1 = 2$. The reason for this unphysical behavior is that (7) and therefore (11) is valid only if

$$\beta E_1 L_1 \gg 1, \tag{12}$$

as was pointed out in [6]. Condition (12) may be interpreted as a restriction on the effective diffusion velocity μE_1 on the uphill slope which must exceed a minimum value D/L_1 (μ and D are, respectively, the mobility and the diffusion coefficient of electrons in the i layer). The above arguments show that a necessary condition for the validity of (6) and (7) is $J \ll J(E_{\min})$ where $E_{\min} = 2\,kT/qL_1$. Using (11) this condition can be written in the form

$$J \ll A^* T^2 (L_D/L_1)^2 = \frac{\varepsilon k T}{q L_1} \frac{N_C}{N_D} \frac{v_R}{L_1}, \tag{13}$$

where $L_D = (\varepsilon k T/q^2 N_D)^{1/2}$ is the Debye length in the substrate, and $v_R \equiv (kT/2\pi m^*)^{1/2}$ is the so-called effective recombination velocity [Ref. 10, p. 385], i.e. the mean thermal velocity of electrons in a given direction. Note that condition (12) automatically guarantees the fulfillment of the other assumption made in [6], namely $J \ll \varepsilon \mu \beta E_1^3$.

The obtained condition, (13), is not too restrictive when applied to silicon, where even for $N_D \sim 10^{18}$ cm^{-3} the right-hand side of (13) gives approximately 5×10^4 A/cm^2. However, it may be important in GaAs because of its smaller N_C or, equivalently, smaller Richardson constant A^*.

Let us now turn our attention to the second and more important limitation of the charge injection mode in TET, which is due to accumulation of carriers drifting with the saturation velocity v_s on the downhill slope of TB. The density ϱ of the accumulated charge is proportional to the current density, viz.

$$\varrho = \frac{J}{v_s}. \tag{14}$$

This charge increases the potential barrier Ψ_1 by the amount

$$\Delta \Psi_1 = \frac{J L_2^2}{2 \varepsilon v_s}. \tag{15}$$

This velocity-saturation effect has not been taken into account. Physically, it can be interpreted as a screening of the gate and anode field by the injected charge. To include this effect we must replace V_G by $V_G - \Delta \Psi_1$ in (9) for the current density

$$J = J_0 e^{\beta [V_G - \Delta \Psi_1(J)]/n}. \tag{16}$$

Differentiating (16) and using (15) we find an expression for the transconductance in the form

$$g_m = \frac{\beta S J}{n + \beta J L_2^2 / 2 \varepsilon v_s}. \tag{17}$$

The denominator in this formula defines an effective ideality factor $\tilde{n}$ which depends on the current density. For $J \ll J_c$ where

$$J_c = 2 n \varepsilon k T v_s / q L_2^2 \tag{18}$$

one has $\tilde{n}(J) = n$ and (17) reduces to (10). On the other hand, for $J \gg J_c$ the transconductance saturates at the value

$$g_{m\,\mathrm{sat}} = 2 \varepsilon v_s S / L_2^2 \tag{19}$$

and the exponential characteristic (9) goes over into a linear dependence of J on the applied voltages. At $J = J_c$ the ideality factor of the exponential dependence is degraded by a factor of 2. Comparing (13) and (18) assuming $L_1 \sim L_2$ one sees that at $J = J_c$ the inequality (13) is automatically satisfied, provided $N_D \ll N_C$ as is usually the case in silicon. We can regard the value J_c as a watershed separating the charge injection mode from a field-effect mode analogous to the strong inversion regime of a short-channel IGFET. However, the total current through the device $I_c = S J_c$ at threshold can be much larger than the maximum subthreshold current in FET. Indeed, taking $L_2 = 10^{-5}$ cm in (18) one has $J_c \sim 10^4$ A/cm^2. For the device area $S = 10\,\mu\text{m} \times 10\,\mu\text{m}$ the current $I_c \sim 10$ mA which equals the typical current in FET of same linear dimensions in the strong inversion regime. Such an improvement is due to the fact that the charge injection in TET is limited only with respect to the current density. The

current in this device flows across the structure through a "channel" of cross-section S equal to its total working area.

The characteristic response time $\tau = 1/2\pi f_m$ where f_m is the maximum operating frequency of a transistor is determined by the total input capacitance C_{in} and the transconductance [10]:

$$\tau = C_{in}/g_m. \tag{20}$$

At low currents the input capacitance consists of the gate-cathode and gate-anode capacitances connected in parallel. On the other hand, as shown in Appendix B, in the high-current limit C_{in} is of the form

$$C_{in} = \frac{L_2}{v_s} g_m + \frac{\varepsilon S}{\pi d}\left(1 + \frac{\varepsilon_{ox}}{\varepsilon}\right) \ln\left(\text{ctg}\frac{\pi a}{2d}\right), \tag{21}$$

where a is the length of the gap between the anode and gate electrodes. In this calculation the "worst case" of $S_A = S_G$ has been assumed for simplicity; clearly for better performance of the device one should have $S_A < S_G$. Then using (19)–(21) we find the minimum response time of TET in the form

$$\tau_{min} = \frac{L_2}{v_s}\left[1 + \frac{nL_2}{2\pi d}\left(1 + \frac{\varepsilon_{ox}}{\varepsilon}\right) \ln\left(\text{ctg}\frac{\pi a}{2d}\right)\right]. \tag{22}$$

The first term in (22) describes the delay associated with charging the gate-cathode capacitance and equals the time of flight of electrons on the downhill slope of TB. The second term in the square brackets represents a ratio of the parasitic gate-to-anode capacitance to the useful, gate-channel, capacitance. Recall that the ratio d/L_2 is limited by our requirement that the top of the barrier be located in the "far zone", i.e. beyond the non-uniform field region. Taking $L_2 \sim 2\lambda = d/\pi$ satisfies this condition. For a typical value of $n \lesssim 2.5$ and taking as an example $a/d \sim 0.1$ we find that the parasitic capacitance leads to only about 25% increase in the delay time compared to the time-of-flight L_2/v_s. We note that the above "parasitic" delay is estimated assuming no load resistance. In a logical inverter circuit, the gate-anode capacitance enters with a factor of 2 (Miller's effect) and thus the total gate delay of TET in the above example would be about 1.5 times L_2/v_s.

In conclusion of this section we present a numerical example of a possible TET structure and its expected characteristics at room temperature. We assume that the device is implemented in silicon with the following barrier parameters: $L_1 = 2000$ Å, $L_2 = 1000$ Å, $N_D \sim 2 \times 10^{17}$ cm^{-3}, and $\sum/q \sim 3 \times 10^{11}$ cm^{-2}, cf. (8). In a TB diode the same geometry and doping profile would give a barrier height $\psi_0 \sim 0.3$ V. The above choice of L_2 implies the surface electrode period $d \sim 0.3$ µm. We also

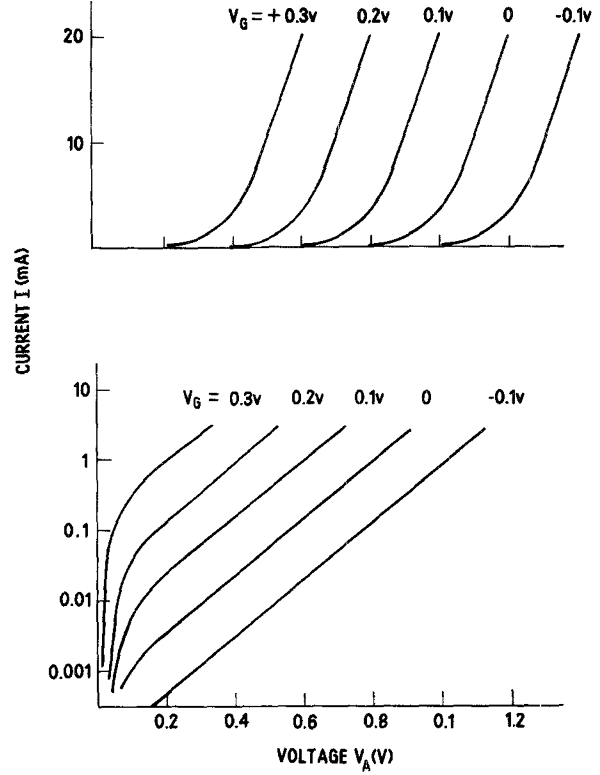

Fig. 3. A family of IV characteristics of TET for different gate voltages

assume $S_G/S_A = 2$ with the total area of the device $S = 10$ µm × 10 µm.

With these parameters we find the transconductance ideality factor (low current) $n = 2.2$, the maximum transconductance $g_{m\,sat} = 200$ mA/V, and the gate delay $\tau = 1.5$ ps. The expected IV characteristics are shown in Fig. 3 for different gate voltages. The linear IV dependence at high currents in Fig. 3a corresponds to the space-charge limited current above threshold, while the exponential characteristics in Fig. 3b describe the charge-injection regime. In order to estimate the voltage swing ΔV_G on the gate required to change the output current from a value I_{off} to I_{on} at a constant V_A we integrate the relation $dV_G = g_m^{-1} dI$ using (17). We obtain

$$\Delta V_G = \frac{nkT}{q} \ln\left(\frac{I_{on}}{I_{off}}\right) + \frac{L_2^2 (I_{on} - I_{off})}{2\varepsilon v_s S}. \tag{23}$$

Using this expression with the above structure parameters and taking $I_{on}/I_{off} = 10$ with $I_{on} = 20$ mA we find $\Delta V_G = 220$ mV at room temperature. Our results for this example are summarized in the first row of Table 1. The other two rows of the table describe examples in which we optimized the power-delay product W rather than the transconductance g_m. Clearly, the W characteristic improves with decreasing total area S of the

Table 1. Gate delay τ, power-delay product W, and transconductance g_m for exemplary TET structures

d [μm]	S [μm^2]	ΔV_G [V]	I_{on} [mA]	τ [ps]	W [J]	g_m [mA/V]
0.3	100	0.2	20	2	4×10^{-15}	200
0.3	15	0.2	3	2	6×10^{-16}	30
0.6	30	0.2	1.5	4	6×10^{-16}	15

device. However, in scaling down the area we are limited by the minimum current necessary for charging the interdevice wiring capacitances without a significant additional delay. Our estimates show that a current $I_{on}\sim3\,\mu A$ will charge these capacitances to $\Delta V_G\sim0.2\,V$ in less than 1 ps.

3. Discussion and Conclusions

Fabrication of the proposed thermionic emission transistor requires advanced material and processing technologies. From the material point of view, the use of MBE appears inescapable because of the stringent requirements of thin and sharply defined modulation doped layers. Fortunately, this technology has now matured to the point where high quality layers can be grown on standard silicon substrate wafers [11]. Requirements to the processing tolerance become progressively more demanding with the expected speed of operation and start pushing the limits of the state-of-art technology as we get into the picosecond range. In this range we need an i layer of thickness $L\sim0.3\,\mu m$ which implies $\lambda\lesssim500\,Å$. The period $d=2\pi\lambda$ of stripes on the surface must, therefore, be of order 3000 Å and each stripe approximately 1500 Å. This may be achievable by using x-ray or electron beam lithography. However, more promising is the use of the so-called interference photolithography (IPL) [12] which is capable of even higher resolution. In the IPL method the photoresist is exposed to an interference pattern of a split laser beam. The period of the resulting pattern of stripes equals half the laser wavelength. The latter can be further reduced by placing the wafer surface in an immersion liquid of high refractive index. Using ultraviolet lasers the IPL technology has been successfully used for a long time for the fabrication of diffraction gratings with the period as short as 0.1 μm. It should be emphasized that IPL is only capable of producing periodic patterns on the surface. Application of this method to TET is possible because of the special simplicity of its surface terminal structure which is a grid of alternating electrodes resembling a diffraction grating. We devised several schemes for making separate contacts to the anode and the gate electrodes. The one which appears most promising to us at this time involves a silicon polygate, grooved according to the IPL pattern by reactive ion etching and insulated by oxide on all sides, as shown in Fig. 1. The anode terminal can then be provided by a silicide Schottky-drain structure. The requirement of two surface terminals seems to be essential for the use of IPL in device fabrication. For example, IPL can probably be used for the fabrication of vertical junction FET's and we shall discuss this possibility elsewhere.

Our conclusions may be summarized as follows. We have proposed a new type of device called interchangeably potential-effect, charge-injection, or thermionic emission transistor. The principle of operation of this device is based on gate-controlled charge injection in unipolar semiconductor structures which is a generalization of the situation which takes place in the subthreshold regime of FET. However in the proposed device the current flows perpendicularly to the semiconductor surface and is controlled by planar potential barriers. Because of this the charge injection regime and associated with it exponential dependence of the output current on gate voltage is extended in current compared to the subthreshold regime in a short-channel FET. This allows us to achieve higher values of the transconductance and therefore faster speed and lower power-delay product and noise. The characteristic response time is limited by the time of flight of electrons across the structure. Because the built-in barrier is removed from the semiconductor surface the height and the shape of this barrier is mainly determined in the process of crystal growth by MBE with modulated doping. Therefore, high accuracy of the threshold control is expected. The gate voltage required to switch the output current by 1 decade at room temperature is of the order of 0.2 V.

Acknowledgements. We wish to thank J. R. Brews, D. Kahng, and G. E. Smith for a critical reading of the manuscript and helpful discussions.

Appendix A

Charge Injection Mode in the Operation of FET

In analyzing the operation of IGFET one usually distinguishes two characteristic regimes, namely the weak and the strong inversion [1]. The weak inversion regime is characterized by an exponential dependence of the density σ of the mobile charge in the channel on voltage applied to the gate and by the fact that σ is a small quantity. In the strong inversion regime σ is large and depends linearly on the gate voltage. Transition between the two regimes occurs in the range of about two orders of magnitude for σ and is usually defined by considering the band bending Ψ_s. It is taken conventionally that the threshold point corresponds to $\Psi_s=2\Psi_B$ where Ψ_B is the bulk Fermi level potential referred to the midgap. Such a definition is not quite general to be applied to other FET devices like, e.g., MESFET's

which also possess two analogous modes of operation. It appears reasonable to give a more general definition based on the physics of the transition process which is screening of the applied field by the mobile charge in the channel.
Consider first an MOS capacitor. In weak inversion the field due to the mobile charge of density σ in the channel gives negligible contribution to the band bending near the surface. In this case the dependence of σ on the gate voltage V_G is of the form

$$\sigma = \sigma_0 e^{\beta V_G/n}, \tag{A.1}$$

where the ideality factor n is given by

$$n = 1 + \varepsilon t/\varepsilon_{ox} w \tag{A.2}$$

with t and ε_{ox} being, respectively, the thickness and the permitivity of the oxide layer. As seen from (A.2), the quantity n depends on V_G through the depletion layer thickness w. Similarly, σ_0 in (A.1) depends on V_G through an effective thickness of the channel. Both these dependences can be neglected when one considers the transition to the strong inversion regime because the dominant effect, the screening, pins σ_0 and w to a constant value [1]. The screening effect consists in the extra field σ/ε_{ox} applied between the inversion layer and the gate electrode. Because of this field the gate voltage necessary to achieve the same band bending is shifted by the amount $\sigma t/\varepsilon_{ox}$ and instead of (A.1) we have

$$\sigma = \sigma_0 e^{\beta(V_G - \sigma t/\varepsilon_{ox})/n}. \tag{A.3}$$

Equation (A.3) describes the transition from the exponential dependence (A.1) to a linear dependence corresponding to strong inversion. It is convenient to cast (A.3) into a form similar to (A.1) but with an effective $\tilde{n}$ which itself depends on σ. Differentiating (A.3) we find

$$\tilde{n} \equiv \left[\frac{d(\ln\sigma)}{d(\beta V_G)}\right]^{-1} = \left(1 + \frac{\beta t\sigma}{n\varepsilon_{ox}}\right) n. \tag{A.4}$$

From (A.4) it is seen that $\tilde{n} = 2n$ at a characteristic value $\sigma = \sigma_c$ given by

$$\sigma_c = \frac{n\varepsilon_{ox} kT}{qt}. \tag{A.5}$$

For a short-channel IGFET where the electron velocity in the channel is saturated, equation (A.5) gives

$$I_c = \sigma_c v_s Z, \tag{A.6}$$

at which the ideality factor for the transconductance characteristic is degraded by a factor of 2. This current can be considered the threshold value separating the two regimes of operation of IGFET. One can readily apply the above approach to other kinds of field-effect transistors as is convenient to do when comparing their characteristics from the point of view of the maximum achievable charge injection current.

Appendix B

Calculation of the Input Capacitance

The input capacitance C_{in} is defined by the relation

$$\delta Q_G = C_{in} \delta V_G, \tag{B.1}$$

where δQ_G is the variation of the charge on the gate electrodes due to a change δV_G of the gate voltage. To calculate δQ_G it is convenient to use the neutrality condition

$$\delta Q_G + \delta Q_A + \delta Q_C = 0, \tag{B.2}$$

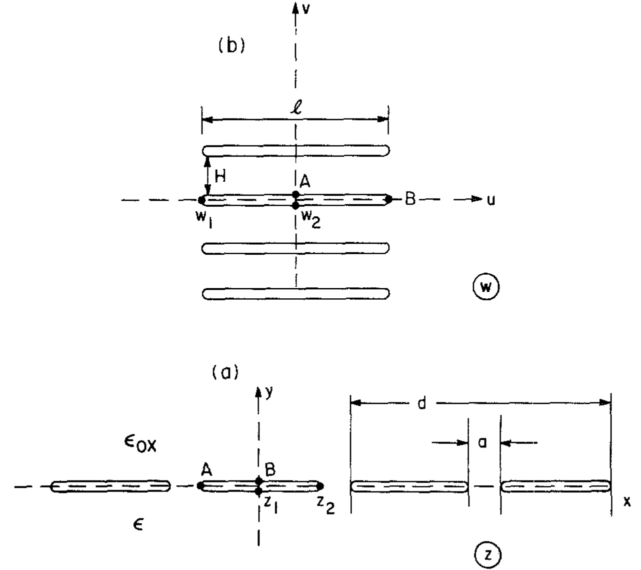

Fig. 4a and b. Domains related by the conformal mapping, (B.4). (a) Domain corresponding to our model of surface electrodes. (b) Domain corresponding to an electrode geometry with known potential distribution

where δQ_A and δQ_C are, respectively, the charge variations on the anode and the cathode (substrate) electrodes. In (B.2) we have neglected the mobile charge in the i-layer and this point will be further discussed below.
It is straightforward to find δQ_C from (3), since near the substrate the multipolar terms in (3) are exponentially small. Taking the first term in (3) and using Gauss's law we find[2]

$$\delta Q_C = -\frac{\varepsilon S}{2L} \delta V_G. \tag{B.3}$$

Equation (B.3) can be interpreted as a simple parallel-plate condenser expression with the plate area equal to the area of the gate.
Evaluation of δQ_A is more involved because one has to consider the near zone. For this calculation it is permissible to ignore the substrate conductor and consider the set of anode and gate electrodes as shown in Fig. 4a. Indeed, the gate to anode field is entirely multipolar and does not penetrate beyond λ. Therefore, the charges induced by this field in the substrate are negligible. To calculate δQ_A we now perform a conformal mapping of the region shown in Fig. 4a (the z-plane) onto that in Fig. 4b (the w-plane). This mapping is effected by the following function (which can be derived by using the symmetry principle of conformal mapping):

$$w(z) = \frac{H}{\pi} \ln[(\cos z + \sqrt{\cos^2 z - \sin^2\alpha})/\sin\alpha], \tag{B.4}$$

where $\alpha = \pi a/d$ and a is the gap between the anode and gate electrodes. In the w-plane we have a system of parallel-plate capacitors with plates of length l separated by the distance H. The mapping (B.4) implies the following relation between the geometric dimensions in the planes z and w:

$$\sin\alpha = \frac{1}{\cosh(\pi l/2H)}. \tag{B.5}$$

[2] For mathematical simplicity in performing the conformal mapping, cf. (B.4), we are taking $S_G = S_A \sim S/2$. In doing so we are overestimating the gate-anode capacitance

For a vanishingly small gap $\alpha \to 0$, we see from (B.5) that $l/H \to \infty$ (logarithmically). For sufficiently small gap, $a \ll d$, we can, therefore, neglect the edge effects when calculating the field in the w-plane. Hence in each of the capacitors the complex potential $F(w)$ can be chosen in the form

$$F(w) = \pm \frac{i(V_A - V_G)}{H} w \tag{B.6}$$

with + or − sign depending on whether V_A or V_G is the potential on the upper plate. The corresponding potential in the z-plane is given by

$$F(z) \equiv \Psi(z) - iA(z) = F[w(z)], \tag{B.7}$$

where Ψ and $-A$ are the real and the imaginary parts of F. As is well known [13], the flux of the electric field through an equipotential surface between points z_1 and z_2 is given by the difference $A(z_2) - A(z_1)$. Separating the imaginary part of (B.7) by using (B.4) and (B.6) we find

$$A(z) = \pm \frac{V_G - V_A}{\pi} \ln \left| \frac{\cos z + \sqrt{\cos^2 z - \sin^2 \alpha}}{\sin \alpha} \right|. \tag{B.8}$$

Let us choose for the equipotential $z_1 z_2$ one half of the lower surface of an anode strip between its midpoint and its edge. With such choice the image points $w(z_1)$ and $w(z_2)$ lie in the same capacitor and one can keep only the + sign in (B.8). By Gauss' law the surface charge on this portion of the electrode is given by

$$\begin{aligned} Q_A(z_1 z_2) &= \varepsilon[A(z_1) - A(z_2)] \\ &= -\frac{\varepsilon(V_G - V_A)}{\pi} Z \ln\left(\operatorname{ctg}\frac{\pi a}{2d}\right), \end{aligned} \tag{B.9}$$

where Z is the length of electrodes in the direction perpendicular to the plane of the figure. By symmetry we have $Q_A = 4Q_A(z_1 z_2)$. A further refinement is to include the difference in dielectric permitivities of the materials above and below the surface electrodes. Assuming $\varepsilon = \varepsilon_{ox}$ in the upper half-plane (Fig. 4a) we obtain the variation of the total charge on one anode electrode in the form

$$Q_A = -\delta V_G \frac{2\varepsilon Z}{\pi}\left(1 + \frac{\varepsilon_{ox}}{\varepsilon}\right) \ln\left(\operatorname{ctg}\frac{\pi a}{2d}\right). \tag{B.10}$$

Collecting eqs. (B.1–3) and (B.10) and expressing Z in terms of S and d in (B.10) we find

$$C_{in} = \frac{\varepsilon S}{2}\left[\frac{1}{L} + \frac{2}{\pi d}\left(1 + \frac{\varepsilon_{ox}}{\varepsilon}\right) \ln\left(\operatorname{ctg}\frac{\pi a}{2d}\right)\right]. \tag{B.11}$$

We now return to the effect of the mobile charge, Q_m, in the i layer which was neglected in (B.2). As discussed in Sect. 1, this charge is due to accumulation of carriers drifting with saturated velocity on the downhill slope of TB and is proportional to the current, $Q_m = L_2 I/v_s$, whence

$$\delta Q_m = \frac{L_2}{v_s} \delta I = \frac{L_2}{v_s} g_m \delta V_G. \tag{B.12}$$

Inclusion of this charge clearly does not change the input capacitance given by (B.11) because the screening effect reduces the charge δQ_C exactly by the amount δQ_m. In the high-current limit, i.e., when $g_m \to g_{m\,\mathrm{sat}}$, one has $\delta Q_C = 0$ and instead of charging the substrate one charges only the i layer. Its capacitance, $C_m \equiv \partial Q_m / \partial V_G$, grows with the transconductance but the time delay associated with charging the i layer remains constant, $\tau_m = L_2/v_s$, which represents the limiting speed of charging the gate-cathode capacitance. Thus, in the high-current limit we can write the total capacitance in the form

$$C_{in} = \frac{L_2}{v_s} g_m + \frac{\varepsilon S}{\pi d}\left(1 + \frac{\varepsilon_{ox}}{\varepsilon}\right) \ln\left(\operatorname{ctg}\frac{\pi a}{2d}\right). \tag{B.13}$$

References

1. J.R.Brews: "Physics of the MOS Transistor", In Supplement to *Advances in Applied Solid State Science*, Vol. 1, ed. by D.Kahng (Academic Press, New York 1981)
2. A.Y.Cho: J. Vac. Sci. Technol. **16**, 275 (1979)
3. J.M.Shannon: Appl. Phys. Lett. **35**, 63 (1979)
4. C.L.Allyn, A.C.Gossard, W.Wiegmann: Appl. Phys. Lett. **36**, 373 (1980)
5. R.J.Malik, K.Board, C.E.C.Wood, L.F.Eastman, T.R.AuCoin, R.L.Ross: Electron. Lett. **16**, 837 (1980)
6. R.F.Kazarinov, S.Luryi: Appl. Phys. Lett. **38**, 810 (1981)
7. A.Chandra, L.F.Eastman: Electron. Lett. **15**, 91 (1979)
8. D.J.Coleman, Jr., S.M.Sze: Bell Syst. Tech. J. **50**, 1695 (1971)
9. C.O.Bozler, G.D.Alley: IEEE Trans. ED-**27**, 1128 (1980)
10. S.M.Sze: *Physics of Semiconductor Devices* (Wiley-Interscience, New York 1969)
11. J.C.Bean: "Growth of doped layers by silicon molecular beam epitaxy", In *Doping Processes in Silicon*, ed. by F.F.Y.Wang (North-Holland, Amsterdam 1981) Chap. 4; "Silicon MBE", J. Vac. Sci. Technol. **18**, 769 (1981)
12. L.F.Johnson, G.W.Kammlott, K.A.Ingersoll: Appl. Opt. **17**, 1165 (1978)
13. L.D.Landau, E.M.Lifshitz: *Electrodynamics of Continuous Media* (Pergamon Press, London 1960)